MCMC

Η Monte Carlo μεθοδολογία για την δημιουργία αριθμητικών προσεγγίσεων διαφόρων τιμών της εκ των υστέρων κατανομής, όπως του μέσου και της τυπικής απόκλισης, στηρίζεται στους Ασθενείς Νόμους των Μεγάλων Αριθμών: σε IID δείγμα, οι Monte Carlo εκτιμητές είναι συνεπείς, που σημαίνει ότι είναι πολύ κοντά στις πραγματικές τιμές με υψηλή πιθανότητα, καθώς ο αριθμός των επαναλήψεων m τείνει στο άπειρο. Πριν δούμε πως οι Metropolis et al. προσπάθησαν να επιτύχουν το ίδιο αποτέλεσμα για ένα μη IID Monte Carlo δείγμα ας δούμε το πρακτικό επακόλουθο της εναλλαγής των δεδομένων από IID σε Μαρκοβιανά.

Παράδειγμα

- Πίσω στο παράδειγμα με το ποσοστό ποδηλάτων στο Berkeley. Τρέχοντας τον IID αλγόριθμο απόρριψης για m επαναλήψεις, δημιουργούμε την επονομαζόμενη Monte Carlo βάση δεδομένων:

Iteration	θ	$I(\theta \leq 0.15)$
1	$\theta_1^* = 0.244$	$I_1^* = 0$
2	$\theta_2^* = 0.137$	$I_2^* = 1$
$\vdots$	$\vdots$	$\vdots$
$m = 31,200$	$\theta_m^* = 0.320$	$I_m^* = 0$
Mean	0.2183 (0.00025)	0.0556 (0.0013)
SD	0.04528	—

Παράδειγμα

- Τρέχοντας τον αλγόριθμο του Metropolis για το ίδιο παράδειγμα, δημιουργούμε την επονομαζόμενη MCMC βάση δεδομένων, η οποία έχει περίπου την ίδια δομή με την προηγούμενη με την διαφορά ότι οι γραμμές χωρίζονται σε 3 φάσεις:
  - Iteration (επανάληψη) 0 είναι η αρχική τιμή.
  - Iterations (1-b) αφορούν την burn-in περίοδο, επαναλήψεις μέχρις ότου επιτευχθεί η στασιμότητα (δεν τις λαμβάνουμε υπόψιν).
  - Iterations (b+1)-(b+m), οι υπό παρακολούθηση επαναλήψεις (monitoring run), από τις οποίες θα λάβουμε εκτιμήσεις για τις ποσότητες της εκ των υστέρων που ενδιαφερόμαστε.

Παράδειγμα

Iteration	Phase	θ	$I(\theta \leq 0.15)$
0	Initialization	$\theta_0^* = 0.200$	—
1	Burn-in	$\theta_1^* = 0.244$	—
⋮	⋮	⋮	⋮
$b = 500$	Burn-in	$\theta_b^* = 0.098$	—
$(b+1) = 501$	Monitoring	$\theta_{b+1}^* = 0.275$	$I_{b+1}^* = 0$
⋮	⋮	⋮	⋮
$(b+m) = 31,700$	Monitoring	$\theta_{b+m}^* = 0.120$	$I_{b+m}^* = 1$
Mean	(Monitoring	0.2177 (0.009)	0.0538 (0.004)
SD	Phase	0.04615	—

Παράδειγμα

- Στις χρονοσειρές σημαντικό ρόλο παίζει η **αυτοσυσχέτιση**: η αυτοσυσχέτιση ρ_κ μίας στάσιμης χρονοσειράς θ_t^* για υστέρηση κ (lag k) ισούται με γ_κ / γ_0 όπου $\gamma_\kappa = C(\theta_t^*, \theta_{t-k}^*)$ είναι η συνδιακύμανση της σειράς με τον εαυτό της πριν από κ επαναλήψεις (δηλαδή πρόκειται για μία μέτρηση του βαθμού εξάρτησης της σειράς από το παρελθόν της).

Παράδειγμα

- Για τον IID αλγόριθμο απόρριψης η χρονοσειρά που δημιουργούμε έχει μηδενική αυτοσυσχέτιση σε κάθε υστέρηση. Αντιθέτως η χρονοσειρά που δημιουργείται από τον αλγόριθμο του Metropolis έχει μή μηδενική αυτοσυσχέτιση, συνήθως θετική, με άλλα λόγια κάθε φορά που προσομοιώνουμε κάποια νέα τιμή από την Μαρκοβιανή αλυσίδα παίρνουμε κάποια νέα πληροφορία για την εκ των υστέρων κατανομή η οποία συνδυάζεται με την παλιά πληροφορία που είχαμε.
- Βασιζόμενοι όμως στο Εργοδικό Θεώρημα καταλήγουμε στο αποτέλεσμα πως όλα τα περιγραφικά μέτρα της εκ των υστέρων κατανομής είναι ακόμα και σε αυτή την περίπτωση συνεπείς εκτιμητές αρκεί η αλυσίδα να έχει συγκλίνει στην στάσιμη κατανομή.
- Άρα οι δυο μεθοδολογίες δίνουν ισοδύναμα αποτελέσματα και διαφέρουν μόνο στην αποτελεσματικότητά τους, μιας και λόγω της θετικής αυτοσυσχέτισης με τον αλγόριθμο του Metropolis μαθαίνουμε πληροφορίες με πιο αργό ρυθμό (μεγαλύτερα τυπικά σφάλματα).

Metropolis Algorithm

- Όπως έχουμε ήδη αναφέρει με την μέθοδο απόρριψης στο χρονικό σημείο t προσομοιώνεις τιμή θ^* από την κατανομή εισήγησης $g(\theta|\mathbf{y})$ και την αποδέχεσαι ή την απορρίπτεις σύμφωνα με την πιθανότητα αποδοχής $\alpha_R(\theta^*|\mathbf{y})$. Αν την δεχτείς μετακινείσαι στο θ^*, αλλιώς προσομοιώνεις νέα τιμή. Με τον τρόπο αυτό δημιουργείς μία IID σειρά προσομοιωμένων τιμών από την εκ των υστέρων κατανομή $p(\theta|\mathbf{y})$.
- Οι Metropolis et al. γενίκευσαν την παραπάνω ιδέα σε περιπτώσεις όπου το IID δείγμα είναι δύσκολο. Επέτρεψαν στην κατανομή εισήγησης στον χρόνο t να εξαρτάται από την τωρινή κατάσταση θ_t της αλυσίδας και εν συνεχεία, για να επιτύχουν την ζητούμενη στάσιμη κατανομή, όταν μία προτεινόμενη τιμή απορρίπτονταν ανάγκαζαν την αλυσίδα το μείνει στην κατάσταση που βρισκόταν για άλλη μια επανάληψη.
- Η αλυσίδα που καταλήγουμε τότε είναι Μαρκοβιανή αφού (α) οι τιμές είναι εξαρτημένες αλλά (β) από όλο το "παρελθόν" της μόνο η πιο πρόσφατη κατάσταση καθορίζει το "μέλλον".

Metropolis – Hastings Algorithm

- Με την παραπάνω μέθοδο υπάρχει μεγάλη ελευθερία στην επιλογή της κατανομής εισήγησης $g(\theta^*|\theta_t,\mathbf{y})$, όπου με θ^* συμβολίζουμε την προτεινόμενη τιμή και με θ_t την τωρινή κατάσταση. Η αρχική ιδέα των Metropolis et al. ήταν η χρησιμοποίηση συμμετρικής κατανομής εισήγησης, δηλ. $g(\theta^*|\theta_t,\mathbf{y}) = g(\theta_t|\ \theta^*,\mathbf{y})$, αλλά ο Hastings το 1970 γενίκευσε την ιδέα αυτή για μη συμμετρικές κατανομές εισήγησης, δημιουργώντας τον αλγόριθμο Metropolis – Hastings. Βασιζόμενος στις ιδέες των Metropolis et al. ο Hastings απέδειξε ότι καταλήγουμε στη σωστή στάσιμη κατανομή αρκεί να χρησιμοποιήσουμε ως πιθανότητα αποδοχής την ακόλουθη:

$$\alpha_{MH}(\theta^*|\theta_t, y) = \min\left\{1, \frac{\frac{p(\theta^*|y)}{g(\theta^*|\theta_t,y)}}{\frac{p(\theta_t|y)}{g(\theta_t|\theta^*,y)}}\right\}. \qquad (1)$$

Metropolis – Hastings Algorithm

Algorithm (Metropolis-Hastings sampling). To construct a **Markov chain** whose **equilibrium distribution** is $p(\theta|y)$, choose a **proposal distribution** $g(\theta^*|\theta_t, y)$, define the **acceptance probability** $\alpha_{MH}(\theta^*|\theta_t, y)$ **and**

Initialize θ_0; $t \leftarrow 0$
Repeat {
  Sample $\theta^* \sim g(\theta|\theta_t, y)$
  Sample $u \sim \text{Uniform}(0, 1)$
  If $u \leq \alpha_{MH}(\theta^*|\theta_t, y)$ then $\theta_{t+1} \leftarrow \theta^*$
    else $\theta_{t+1} \leftarrow \theta_t$
  $t \leftarrow (t + 1)$
}

Metropolis – Hastings Algorithm

- Έχει αρκετό ενδιαφέρον να συγκρίνουμε την συγκεκριμένη πιθανότητα αποδοχής με αυτή από τη μέθοδο απόρριψης. Η ουσιαστική διαφορά τους είναι πως η κατανομή εισήγησης τώρα δεν είναι σταθερή αλλά αλλάζει κάθε φορά.
- Παρατηρήστε πως η σχέση (1) είναι μια γενίκευση της πιθανότητας αποδοχής της μεθόδου απόρριψης: η νέα πιθανότητα αποδοχής μπορούμε να πούμε πως είναι το πηλίκο 2 πιθανοτήτων αποδοχής της μεθόδου απόρριψης, μίας που έχει να κάνει με το που είσαι τώρα και μίας που έχει να κάνει με το που σκέφτεσαι να πας (είναι επιπλέον ισοδύναμο να δουλεύεις με την g ή την G μιας και στην περίπτωση αυτή η σταθερά κανονικοποίησης θα απαλειφθεί στο πηλίκο).

Metropolis – Hastings Algorithm

- Αξιοσημείωτο είναι το γεγονός ότι για οποιαδήποτε κατανομή εισήγησης η στάσιμη κατανομή θα είναι η εκ των υστέρων p.

<u>Απόδειξη</u>

Ο μεταβατικός πυρήνας για τον αλγόριθμο Metropolis-Hastings είναι:

$$P\left(\theta_{t+1} \mid \theta_t\right) = g\left(\theta_{t+1} \mid \theta_t, \mathbf{y}\right)\alpha_{MH}\left(\theta_{t+1} \mid \theta_t, \mathbf{y}\right) + I\left(\theta_{t+1} = \theta_t\right)\left[1 - \int g\left(\theta^* \mid \theta_t, \mathbf{y}\right)\alpha_{MH}\left(\theta^* \mid \theta_t, \mathbf{y}\right)d\theta^*\right], \quad (2)$$

Metropolis – Hastings Algorithm

όπου I(.) είναι η δείκτρια συνάρτηση. Ο πρώτος όρος του δεξιού μέλους της προηγούμενης σχέσης προκύπτει από την αποδοχή του υποψηφίου $\theta^* = \theta^*_{t+1}$, ο δεύτερος όρος από την απόρριψης όλων των πιθανών υποψηφίων θ^*. Χρησιμοποιώντας την σχέση:

$$p\left(\theta_t \mid \mathbf{y}\right) g\left(\theta_{t+1} \mid \theta_t, \mathbf{y}\right) \alpha_{MH}\left(\theta_{t+1} \mid \theta_t, \mathbf{y}\right) = p\left(\theta_{t+1} \mid \mathbf{y}\right) g\left(\theta_t \mid \theta_{t+1}, \mathbf{y}\right) \alpha_{MH}\left(\theta_t \mid \theta_{t+1}, \mathbf{y}\right)$$

η οποία προκύπτει από την (1) καταλήγουμε λόγω της (2) στην σχέση

$$p\left(\theta_t \mid \mathbf{y}\right) P\left(\theta_{t+1} \mid \theta_t, \mathbf{y}\right) = p\left(\theta_{t+1} \mid \mathbf{y}\right) P\left(\theta_t \mid \theta_{t+1}, \mathbf{y}\right)$$

Metropolis – Hastings Algorithm

Ολοκληρώνοντας την προηγούμενη σχέση ως προς θ_t

$$\int p\left(\theta_t \mid \mathbf{y}\right) P\left(\theta_{t+1} \mid \theta_t, \mathbf{y}\right) d\theta_t = p\left(\theta_{t+1} \mid \mathbf{y}\right).$$

Το αριστερό μέρος της παραπάνω σχέσης μας δίνει την περιθώρια κατανομή της θ_{t+1}, υπό την προϋπόθεση ότι η θ_t προέρχεται από την εκ των υστέρων κατανομή. Άρα η παραπάνω σχέση μας λέει πως αν η θ_t προέρχεται από την εκ των υστέρων κατανομή το ίδιο θα συμβαίνει και για την θ_{t+1}. Άρα με το που λάβουμε δείγμα από την στάσιμη κατανομή όλα τα υπόλοιπα δείγματα θα προέρχονται από αυτή και το αποτέλεσμα αυτό ισχύει για οποιαδήποτε g.

Metropolis – Hastings Algorithm

- Η παραπάνω απόδειξη μας αποσαφηνίζει μόνο το γεγονός ότι η στάσιμη κατανομή είναι η εκ των υστέρων, χωρίς να δείχνει αν πράγματι η αλυσίδα συγκλίνει σε στάσιμη κατανομή.
- Όταν η κατανομή εισήγησης είναι συμμετρική τότε η πιθανότητα αποδοχής ισούται με $p(\theta^* \mid \mathbf{y}) \Big/ p(\theta_t \mid \mathbf{y})$, που σημαίνει πως θέλεις να επισκεφτείς σημεία με μεγαλύτερη συχνότητα πιο συχνά.

Παράδειγμα

- Έστω $(Y_i \mid \sigma^2) \overset{IID}{\sim} N(\mu, \sigma^2),\ i = 1,..,n$ (μ γνωστό). Η συνάρτηση πιθανοφάνειας τότε είναι:

$$\begin{aligned} l(\sigma^2|y) &= c\prod_{i=1}^{n}(\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] \\ &= c\,(\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right]. \end{aligned}$$

Αν επιλέξουμε ως εκ των προτέρων κατανομή την:

$$p\left(\sigma^2\right) \propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}$$

Παράδειγμα

δηλαδή $\sigma^2 \sim Inv-\chi^2(\nu_0,\sigma_0^2)$ τότε:

$$p(\sigma^2 \mid y) \propto p(\sigma^2) l(y \mid \sigma^2)$$

$$\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \cdot (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2\sigma^2}u\right\}$$

$$\propto (\sigma^2)^{-\left(\frac{(\nu_0+n)}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2+nu)\right\}, \quad \mu\varepsilon \quad u=\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2$$

Παρατηρούμε ότι

$$\sigma^2 \mid y \sim \text{Inv}-\chi^2\left(\nu_0+n, \frac{\nu_0\sigma_0^2+nu}{\nu_0+n}\right).$$

Παράδειγμα

- Αν και στο συγκεκριμένο παράδειγμα μπορούμε να υπολογίσουμε την εκ των υστέρων κατανομή πλήρως, ας υποθέσουμε ότι θέλουμε να χρησιμοποιήσουμε τον MH algorithm για προσομοίωση τιμών από αυτή. Για να χρησιμοποιήσουμε τον αλγόριθμο θα πρέπει να διαλέξουμε την κατανομή εισήγησης $g(\sigma^2 \mid \sigma_t^2, \mathbf{y}).$

- Όπως αναφέραμε προηγουμένως η επιλογή της g δεν επηρεάζει την σύγκλιση στην εκ των υστέρων κατανομή. Επηρεάζει όμως την ταχύτητα σύγκλισης και πόσο καλά η αλυσίδα αναμιγνύεται (μίξη).

Παράδειγμα

1. Διαλέξτε μια κατανομή εισήγησης που μοιάζει με μία "υπερκαλυπτόμενη" έκδοση της εκ των υστέρων κατανομή (για αυτό το λόγο συχνά είναι αναγκαίο αρχικά να προβούμε σε μία πιλοτική μελέτη για να αποκτήσουμε μια πρόχειρη εικόνα για το σχήμα της εκ των υστέρων κατανομής).
2. Δημιουργήστε την κατανομή εισήγησης με τέτοιο τρόπο ώστε $\mathrm{E}_{\mathrm{g}}\left[\theta^{*} \mid \theta_{\mathrm{t}}, \mathbf{y}\right]=\theta_{\mathrm{t}}$. Δηλαδή η αναμενόμενη τιμή του πού πρόκειται να μετακινηθείς θ^{*}, δεδομένου ότι έχεις δεχθεί να μετακινηθείς από την τωρινή κατάσταση θ_t ισούται με την τωρινή κατάσταση θ_t. Άρα όταν κινείσαι υπάρχει ένα είδος αριστερής – δεξιάς ισορροπίας στην κατεύθυνση που μετακινείσαι, και άρα επιτυγχάνεις καλύτερη εξερεύνηση στου χώρου.

Παράδειγμα

- Βάση της ιδέας (1) μια ικανοποιητική επιλογή για την κατανομή εισήγησης είναι η

$$g(\sigma^2 \mid \sigma_t^2, \mathbf{y}) = \text{Inv} - \chi^2\left(\nu_*, \sigma_*^2\right).$$

Η κατανομή αυτή έχει μέσο

$$\frac{\nu_*}{\nu_* - 2}\sigma_*^2 \quad \gamma\iota\alpha\ \nu_* > 2.$$

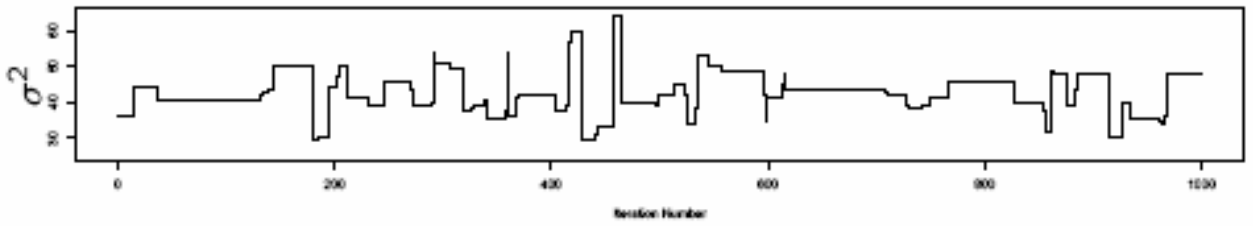
Άρα χρησιμοποιώντας την ιδέα (2) μπορώ να χρησιμοποιήσω $\nu_*>2$ και

$$\sigma_*^2 = \frac{\nu_* - 2}{\nu_*}\sigma_t^2 \quad \delta\eta\lambda\alpha\delta\acute{\eta}:$$

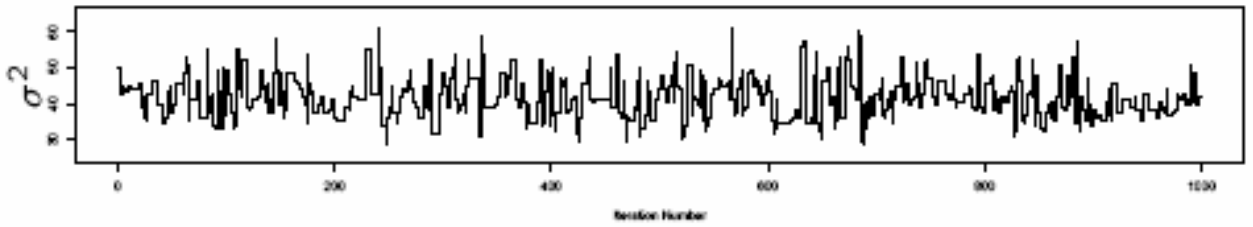
$$g(\sigma^2 \mid \sigma_t^2, \mathbf{y}) = \text{Inv} - \chi^2\left(\nu_*, \frac{\nu_* - 2}{\nu_*}\sigma_t^2\right).$$

Παράδειγμα

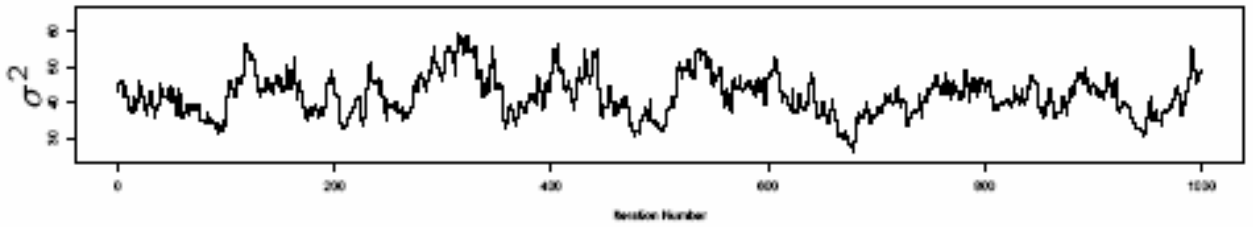
- Διάλεξε μια τιμή για το ν_* έτσι ώστε η αλυσίδα να αναμιγνύεται καλύτερα



→ $\nu_*=2.5$
$\alpha_{MH}=0.07$



→ $\nu_*=20$
$\alpha_{MH}=0.44$



→ $\nu_*=500$
$\alpha_{MH}=0.86$

Παράδειγμα

- Η τυπική απόκλιση (SD) της κατανομής εισήγησης που χρησιμοποιούμε είναι ανάλογη της ποσότητας: $\frac{1}{\sqrt{\nu_* - 4}}$

η οποία είναι φθίνουσα συνάρτηση του ν_*. Όταν το SD είναι μεγάλο (μικρό ν_* όπως στο πρώτο γράφημα) ο αλγόριθμος κάνει μεγάλα άλματα γύρω από τον χώρο του σ^2 (πράγμα καλό), αλλά τα περισσότερα από αυτά απορρίπτονται (πράγμα κακό), και άρα υπάρχουν μεγάλες περίοδοι στις οποίες δεν έχουμε κίνηση. Αντίθετα όταν το SD είναι μικρό (μεγάλο ν* όπως στο τρίτο γράφημα) ο αλγόριθμος δέχεται σχεδόν όλες τις κινήσεις (πράγμα καλό) αλλά αυτές είναι τόσο μικρές έτσι ώστε να χρειάζεται πολύ χρόνο για να εξερευνήσει πλήρως τον χώρο (πράγμα κακό).

Παράδειγμα

- Έχει δειχθεί ότι σε απλά προβλήματα με προσεγγιστικά κανονικές εκ των υστέρων κατανομές η βέλτιστη πιθανότητα αποδοχής είναι περίπου 44%.
- Στο παράδειγμα μας η άγνωστη ποσότητα ήταν μονοδιάστατη, αλλά ο αλγόριθμος δουλεύει χωρίς προβλήματα και με πολυδιάστατο $\boldsymbol{\theta}$.
- Το μεγάλο προτέρημα είναι ότι για να εφαρμόσουμε τον αλγόριθμο αρκεί να γνωρίζουμε την εκ των υστέρων χωρίς την σταθερά κανονικοποίησης.

Single Component M-H

- Όταν το $\boldsymbol{\theta} = (\theta_1,..,\theta_k)$ τότε μπορούμε να εφαρμόζουμε τον αλγόριθμο M-H και να προσομοιώνουμε διανύσματα $\boldsymbol{\theta}$ ή μπορούμε να δημιουργούμε ξεχωριστά τις συνιστώσες βάση διαφορετικών κατανομών εισήγησης. Κάθε δηλαδή επανάληψη του αλγορίθμου αποτελείτε από κ βήματα, στην αρχή δηλαδή της επανάληψης t, ανανεώνεις πρώτα το θ_1, μετά το θ_2 και στο τέλος το $\theta_κ$.
- Ας καλέσουμε $\theta_{t,i}$ την τιμή της i συνιστώσας στο χρόνο t και $\boldsymbol{\theta}_{t,-i}=(\theta_{t+1,1}, \theta_{t+1,2},\ldots, \theta_{t+1,i-1}, \theta_{t,i+1},\ldots, \theta_{t,k}\}$
- Το υποψήφιο σημείο θ_i^* παράγεται από την κατανομή εισήγησης $g_i(\theta_i^* \mid \theta_{t,i}, \boldsymbol{\theta}_{t,-i}, \mathbf{y})$ με πιθανότητα αποδοχής:

$$\alpha_{MH}(\theta_i^* \mid \theta_{t,i}, \boldsymbol{\theta}_{t,-i}, \mathbf{y}) = \min\left[1, \frac{p(\theta_i^* \mid \boldsymbol{\theta}_{t,-i}, \mathbf{y}) g_i(\theta_{t,i} \mid \theta_i^*, \boldsymbol{\theta}_{t,-i}, \mathbf{y})}{p(\theta_{t,i} \mid \boldsymbol{\theta}_{t,-i}, \mathbf{y}) g_i(\theta_i^* \mid \theta_{t,i}, \boldsymbol{\theta}_{t,-i}, \mathbf{y})}\right]$$

Gibbs Sampling

- Ειδική περίπτωση του **Single Component M-H** αποτελεί ο **δειγματολήπτης Gibbs** (**Gibbs Sampling**) όπου η κατανομή εισήγησης $g_i(\theta_i^* \mid \theta_{t,i}, \boldsymbol{\theta}_{t,-i}, \mathbf{y}) = p(\theta_i^* \mid \boldsymbol{\theta}_{t,-i}, \mathbf{y})$ είναι η **πλήρους δέσμευσης εκ των υστέρων κατανομή** (**full conditional posterior distribution**) για το θ_i. Στην περίπτωση αυτή η πιθανότητα αποδοχής ισούται με 1, με άλλα λόγια στο Gibbs Sampling προσομοιώνουμε τιμές από την full conditional posterior distribution (η οποία έχει το πλεονέκτημα ότι στις περισσότερες των περιπτώσεων είναι εύκολα υπολογίσιμη) και δεχόμαστε όλες τις προτεινόμενες κινήσεις.

Gibbs Sampling

Algorithm (Single-element Gibbs sampling). To construct a **Markov chain** whose **equilibrium distribution** is $p(\theta|y)$ with $\theta = (\theta_1, \ldots, \theta_k)$,

Initialize $\theta^*_{0,1}, \ldots, \theta^*_{0,k}$; $t \leftarrow 0$
Repeat {
  Sample $\theta^*_{t+1,1} \sim p(\theta_1|y, \theta^*_{t,2}, \theta^*_{t,3}, \theta^*_{t,4}, \ldots, \theta^*_{t,k})$
  Sample $\theta^*_{t+1,2} \sim p(\theta_2|y, \theta^*_{t+1,1}, \theta^*_{t,3}, \theta^*_{t,4}, \ldots, \theta^*_{t,k})$
  Sample $\theta^*_{t+1,3} \sim p(\theta_3|y, \theta^*_{t+1,1}, \theta^*_{t+1,2}, \theta^*_{t,4}, \ldots, \theta^*_{t,k})$
  $\vdots$ $\vdots$ $\vdots$ $\vdots$ $\vdots$ $\vdots$
  Sample $\theta^*_{t+1,k} \sim p(\theta_k|y, \theta^*_{t+1,1}, \theta^*_{t+1,2}, \theta^*_{t+1,3}, \ldots, \theta^*_{t+1,k-1})$
  $t \leftarrow (t+1)$
}

Παράδειγμα

$$\begin{aligned} (\mu, \sigma^2, \nu) &\sim p(\mu, \sigma^2, \nu) \\ (y_i|\mu, \sigma^2, \nu) &\overset{\text{IID}}{\sim} t_\nu(\mu, \sigma^2), \end{aligned}$$

όπου η $t_\nu(\mu, \sigma^2)$ είναι η scaled t-distribution με μέσο μ, παράμετρο κλίμακας σ^2 και παράμετρο μορφής ν. Είναι γνωστό από την Θεωρία Πιθανοτήτων ότι για να προσομοιώσεις τιμές από την συγκεκριμένη t κατανομή μπορείς αρχικά να προσομοιώσεις τιμές από μία αντίστροφη Γάμμα και μετά από μία Κανονική δεδομένων των τιμών που προσομοίωσες από την Γάμμα (Inverse Gamma mixture of Gaussians).

$$\begin{aligned} (\lambda|\nu) &\sim \Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\ (y|\mu, \sigma^2, \lambda) &\sim N\left(\mu, \lambda\sigma^2\right). \end{aligned}$$

Παράδειγμα

Εισάγοντας και τις συζυγείς εκ των προτέρων για τα μ και σ^2 καταλήγουμε στο ακόλουθο **ιεραρχικό μοντέλο** (**hierarchical model**)

$$
\begin{aligned}
\nu &\sim p(\nu) \\
\sigma^2 &\sim \text{SI-}\chi^2\left(\nu_0, \sigma_0^2\right) \\
\left(\mu|\sigma^2\right) &\sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \\
\left(\lambda_i|\nu\right) &\overset{\text{IID}}{\sim} \Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\
\left(y_i|\mu, \sigma^2, \lambda_i\right) &\overset{\text{indep}}{\sim} N\left(\mu, \lambda_i\, \sigma^2\right).
\end{aligned}
$$

Παράδειγμα

Iteration	Phase	μ	σ^2	ν
0	Initializing	μ_0	σ_0^2	ν_0
1	Burn-In	$\mu_1(y, \sigma_0^2, \nu_0)$	$\sigma_1^2(y, \mu_1, \nu_0)$	$\nu_1(y, \mu_1, \sigma_1^2)$
2	Burn-In	$\mu_2(y, \sigma_1^2, \nu_1)$	$\sigma_2^2(y, \mu_2, \nu_1)$	$\nu_1(y, \mu_2, \sigma_2^2)$
.	.	.	.	.
b	Burn-In	μ_b	σ_b^2	ν_b
$(b+1)$	Monitoring	μ_{b+1}	σ_{b+1}^2	ν_{b+1}
$(b+2)$	Monitoring	μ_{b+2}	σ_{b+2}^2	ν_{b+2}
.	.	.	.	.
$(b+m)$	Monitoring	μ_{b+m}	σ_{b+m}^2	ν_{b+m}

Παράδειγμα

- **Ερωτήματα**
  - Υπολογισμός των full Conditionals.
  - Αρχικές τιμές.
  - Πόσο μεγάλα πρέπει να είναι τα b, m;
  - Πως ξέρω ότι έχω σύγκλιση της αλυσίδας;

Υπολογισμός των Full Conditionals

Ας θεωρήσουμε το εξής πιο απλό παράδειγμα:

$$\begin{aligned} \sigma^2 &\sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \\ (\mu|\sigma^2) &\sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \\ (Y_i|\mu, \sigma^2) &\overset{\text{IID}}{\sim} N(\mu, \sigma^2)\,. \end{aligned}$$

$$\begin{aligned} p(\mu|\sigma^2, y) &= \frac{p(\mu, \sigma^2, y)}{p(\sigma^2, y)} \\ &= c\, p(\mu, \sigma^2, y) \\ &= c\, p(\sigma^2)\, p(\mu|\sigma^2)\, p(y|\mu, \sigma^2) \\ &= c \exp\left[-\frac{\kappa_0}{2\sigma^2}(\mu - \mu_0)^2\right] \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \end{aligned}$$

Υπολογισμός των Full Conditionals

Οπότε: $$p(\mu|\sigma^2, y) = c \exp\left[-\frac{\kappa_0 + n}{2\sigma^2}\left(\mu - \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}\right)^2\right],$$

δηλαδή $$(\mu|\sigma^2, y) \sim N\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n}\right).$$

Όμοια

$$\begin{aligned}
p(\sigma^2|\mu, y) &= \frac{p(\sigma^2, \mu, y)}{p(\mu, y)} \\
&= c\, p(\sigma^2, \mu, y) \\
&= c\, p(\sigma^2)\, p(\mu|\sigma^2)\, p(y|\mu, \sigma^2) \\
&= c\, (\sigma^2)^{-(1+\frac{1}{2}\nu_0)} \exp\left(\frac{-\nu_0\,\sigma_0^2}{2\sigma^2}\right) \cdot \\
&\quad (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{\kappa_0}{2\sigma^2}(\mu - \mu_0)^2\right] \cdot \\
&\quad (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right].
\end{aligned}$$

Υπολογισμός των Full Conditionals

Οπότε:

$$p(\sigma^2|\mu,y) = c\ (\sigma^2)^{-(1+\frac{\nu_0+1+n}{2})}\ \exp\left[-\frac{\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2+ns_\mu^2}{2\sigma^2}\right],$$

με $\quad s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2.$

δηλαδή

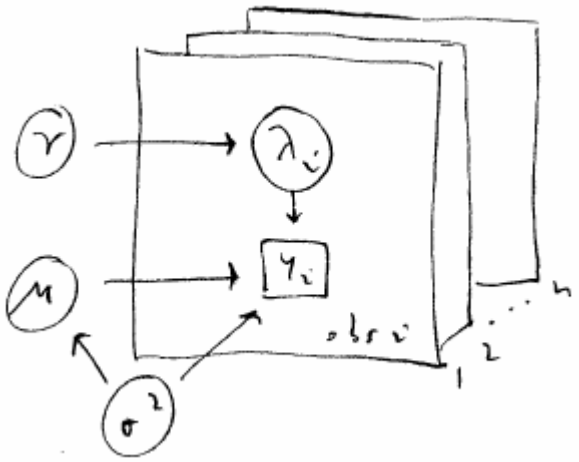
$$(\sigma^2|\mu,y) \sim \text{SI-}\chi^2\left(\nu_0+1+n, \frac{\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2+ns_\mu^2}{\nu_0+1+n}\right).$$

Υπολογισμός των Full Conditionals

- Παρατηρούμε ότι σε συζυγείς περιπτώσεις οι full conditional έχουν συζυγείς μορφές. Άρα είναι εφικτό με την βοήθεια ενός λογισμικού να υπολογιστούν αυτόματα οι full conditionals και να μην χρειαστεί να κάνουμε εμείς κάθε φορά τους υπολογισμούς.
- Ένα τέτοιο λογισμικό είναι το BUGS το οποίο τρέχει κάτω από Unix ή Dos καθώς και το αντίστοιχο λογισμικό κάτω από Windows το **WinBugs**.
- Το Bugs/WinBugs δουλεύουν και για μή συζυγείς εκ των προτέρων κατανομές βασιζόμενο στα παρακάτω:

Directed Acyclic Graphs

1. Βλέποντας τα ιεραρχικά μοντέλα ως **Κατευθυνόμενα Ακυκλικά Γραφήματα - Directed Acyclic Graphs (DAG)**. Η δεσμευμένη ανεξαρτησία των ιεραρχικών μοντέλων (οι ποσότητες στο μοντέλο εξαρτώνται μόνο από αυτές που είναι ένα στάδιο παραπάνω και όχι από τις υπόλοιπες) μας επιτρέπει να δούμε τις ποσότητες ως κόμβους σε ένα γράφημα. Το DAG είναι ένα γράφημα στο οποίο οι ποσότητες απεικονίζονται είτε ως κύκλοι (άγνωστες ποσότητες) είτε ως τετράγωνα (γνωστές ποσότητες) και εν συνεχεία βασισμένοι στην εξάρτηση τους τις συνδέουμε με ένα βέλος. Το γράφημα αυτό είναι ακυκλικό (acyclic) με την έννοια ότι ακολουθούμενος την πορεία του βέλους είναι αδύνατο να γυρίσεις σε κόμβο από τον οποίο πέρασες.



Adaptive Rejection Sampling

2. Εφαρμόζοντας την προσαρμοσμένη μέθοδο απόρριψης (Adaptive Rejection Sampling), για να προσομοιώσεις τιμές από full conditional distributions που δεν έχουν απλή μορφή (μή συζυγείς εκ των προτέρων). Όπως έχουμε δει η μέθοδος απόρριψης είναι μια γενική μέθοδος προσομοίωσης τιμών από την p(θ|**y**) με την βοήθεια ενός φακέλου G(θ|**y**). Ο αλγόριθμος για κανονικοποιημένη G είναι ο ακόλουθος:

```
Repeat {
  Sample a point theta from G ( . | y );
  Sample a Uniform( 0, 1 ) random variable U;
  If U <= p ( theta | y ) / G ( theta | y ) accept theta;
  }
until one theta is accepted.
```

Adaptive Rejection Sampling

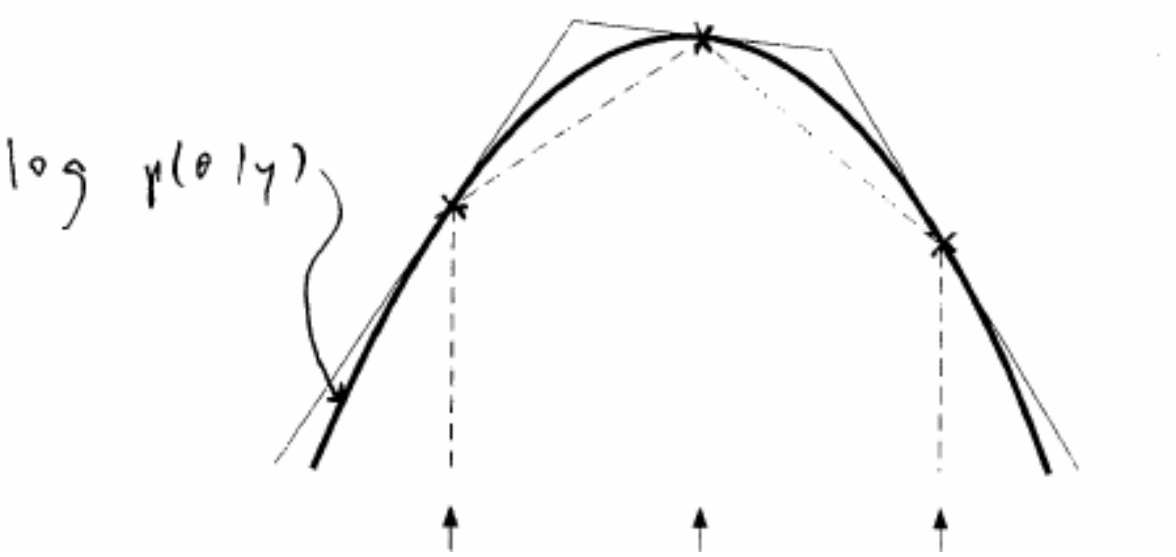
Για την κατασκευή του φακέλου G μπορούμε να εισάγουμε 2 συναρτήσεις a, b (**squeezing functions**) τέτοιες ώστε $b(\theta|\mathbf{y}) \leq p(\theta|\mathbf{y}) \leq a(\theta|\mathbf{y})$ και να αντικαταστήσουμε την γραμμή 4 του προηγουμένου κώδικα (πιθανότητα αποδοχής) με το παρακάτω:

```
If U > a( theta | y ) / G( theta | y ) reject theta;
  else if U <= b( theta | y ) / G( theta | y ) accept theta;
  else if U <= p( theta | y ) / G( theta | y ) accept theta.
```

Η αναπροσαρμοσμένη μέθοδος απόρριψης (ARS) είναι μια αρκετά αποδοτική μέθοδος, που δουλεύει ως βάση του Gibbs Sampling σε περιπτώσεις όπου οι full conditionals είναι log-concave.

Adaptive Rejection Sampling

- Για μονοδιάστατο θ μπορεί εύκολα να κατασκευαστεί η logG(θ|**y**) στην λογαριθμική κλίμακα φέρνοντας εφαπτόμενες στην logp(θ|**y**) από σημεία ενός σύνολου S



Οι εφαπτόμενες δημιουργούν τον φάκελο στην λογαριθμική κλίμακα και οι χορδές την συνάρτηση a.

Adaptive Rejection Sampling

- Ο φάκελος αποτελείτε από γραμμικές συναρτήσεις στην λογαριθμική κλίμακα, οπότε στην αρχική κλίμακα αποτελείτε από εκθετικές κατανομές από τις οποίες μπορούμε εύκολα να προσομοιώσουμε τιμές.
- Το πλεονέκτημα της μεθόδου είναι η αναπροσαρμογή στην κατασκευή του φακέλου που μπορούμε να πραγματοποιήσουμε. Καθώς μαθαίνουμε όλο και περισσότερα για το θ προσομοιώνοντας νέες τιμές, προσθέτουμε νέα σημεία στο σύνολο S και άρα ο φάκελος καλυτερεύει και πλησιάζει όλο και περισσότερο την πραγματική συνάρτηση.