ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΚΑΤΕΥΘΥΝΣΗ ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΗΣ ΚΩΝΣΤΑΝΤΙΝΟΥ ΜΑΡΙΑΣ ΖΑΦΕΙΡΟΥΛΑΣ

ΘΕΜΑ

ΑΚΟΛΟΥΘΙΑΚΕΣ ΜΕΘΟΔΟΙ MONTE CARLO ΜΕ ΕΦΑΡΜΟΓΗ ΣΤΗ ΜΠΕΥΖΙΑΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΠΙΒΛΕΠΟΥΣΑ : ΛΟΥΚΙΑ ΜΕΛΙΓΚΟΤΣΙΔΟΥ

ΙΟΥΝΙΟΣ 2013

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Σε αυτή την εργασία θα ασχοληθούμε με τις ακολουθιακές μεθόδους Monte Carlo οι οποίες αποτελούν προηγμένες τεχνικές Monte Carlo. Οι μέθοδοι Monte Carlo είναι ένας πολύ γενικός όρος που αναφέρεται σε στοχαστικές τεχνικές, δηλαδή σε αλγορίθμους που βασίζονται σε τυχαίες δειγματοληψίες για να συγκλίνουν σε μία λύση. Έχουν ιδιαίτερα μεγάλο εύρος εφαρμογών και αποτελούν σημαντικό εργαλείο για την Υπολογιστική Φυσική αλλά και για πολλούς άλλους τομείς.

Οι ακολουθιακές τεχνικές Monte Carlo δημιουργήθηκαν από ερευνητές επιστήμονες για να αντιμετωπίσουν τα μειονεκτήματα που παρουσιάζουν οι απλές μέθοδοι Monte Carlo. Ειδικότερα σε αυτή την εργασία θα ασχοληθούμε με τις μεθόδους Monte Carlo για την επίλυση στατιστικών προβλημάτων με ελλιπή δεδομένα κάτω από το πρίσμα της Μπευζιανής προσέγγισης. Παρακάτω αναφέρουμε εισαγωγικά την χρησιμότητα των τεχνικών Monte Carlo στη Στατιστική.

1.1 Η χρησιμότητα των τεχνικών Monte Carlo

Η Monte Carlo μέθοδος είναι μια αριθμητική τεχνική επίλυσης μαθηματικών προβλημάτων με τη χρήση τυχαίων αριθμών . Η μέθοδος αυτή εφευρέθηκε από επιστήμονες το 1944 περίπου, και ονομάστηκε έτσι από την πόλη του Μονακό, εξαιτίας μιας ρουλέτας, μια απλής γεννήτριας τυχαίων αριθμών. Η μέθοδος **Monte Carlo** είναι μια κατηγορία υπολογιστικών αλγορίθμων που στηρίζονται σε επαναλαμβανόμενες τυχαίες δειγματοληψίες και χρησιμοποιούνται για την προσομοίωση φυσικών και μαθηματικών συστημάτων. Χρησιμοποιούνται κυρίως σε προβλήματα τα οποία είναι πολύ περίπλοκα για να επιλυθούν αναλυτικά ή με τη χρήση κλασικών αριθμητικών μεθόδων.

Η μέθοδος Monte Carlo χρησιμοποιείται ευρέως στα μαθηματικά. Μια κλασική χρήση είναι για τον υπολογισμό ολοκληρωμάτων, ιδιαίτερα πολυδιάστατων, που δεν μπορούν να υπολογιστούν αναλυτικά. Γενικότερα, οι τεχνικές Monte Carlo είναι χρήσιμες για τη μοντελοποίηση φαινομένων με σημαντική αβεβαιότητα όσον αφορά τους διαθέσιμους πόρους, όπως για παράδειγμα ο υπολογισμός των κινδύνων στον τομέα των επιχειρήσεων. Τέτοιες τεχνικές έχουν εφαρμοστεί για την εξερεύνηση και εκμετάλλευση του πετρελαίου, την πραγματική παρατήρηση βλαβών, για τις υπερβάσεις κόστους και χρονοδιαγράμματος κτλ.

Στη στατιστική, η μέθοδος Monte Carlo χρησιμοποιείται συχνά σε προβλήματα που ο υπολογισμός ολοκληρωμάτων της μορφής

$$I = \int_D g(x)dx$$

(όπου το D είναι συχνά μια περιοχή σε ένα χώρο μεγάλης διάστασης και $g(x)$ είναι μια κατανομή στόχος (target)) δεν μπορεί να γίνει αναλυτικά.

Αν μπορούμε να αντλήσουμε, ομοιόμορφα από το D (με έναν υπολογιστή), ανεξάρτητες και ταυτοτικά κατανεμημένες (i.i.d) τυχαίες μεταβλητές $x^1, \dots, x^m$, μια προσέγγιση του I μπορεί να είναι

$$\hat{I}_m = \frac{1}{m}\{g(x^1) + \cdots + g(x^m)\}$$

Ο νόμος των μεγάλων αριθμών αναφέρει ότι ο μέσος όρος πολλών ανεξάρτητων τυχαίων μεταβλητών με κοινή μέση τιμή και πεπερασμένες διακυμάνσεις τείνει να σταθεροποιηθεί στη κοινή τους μέση τιμή που είναι

$\lim_{m\to\infty} \hat{I}_m =$ I με πιθανότητα 1

Ο ρυθμός σύγκλισης τους μπορεί να αξιολογηθεί από το Κ.Ο.Θ

$$\sqrt{m}(\hat{I}_m - \text{I}) \to N(0, \sigma^2) \text{ όπου } \sigma^2 = Var\{g(x)\}\,.$$

Έτσι το σφάλμα της προσέγγισης Monte Carlo είναι της τάξης $O(m^{-1/2})$ ανεξάρτητα από την διάσταση του x. Στην απλούστερη περίπτωση όταν D=[0,1] και $I = \int_0^1 g(x)dx$ μπορούμε να προσεγγίσουμε το I από το

$$\hat{I}_m = \frac{1}{m}\{g(b_1) + \cdots + g(b_m)\} \text{ όπου } b_j = j/m\,.$$

Αυτή η μέθοδος ονομάζεται προσέγγιση Riemann. Όταν η g είναι αρκετά ομαλή η προσέγγιση Riemann μας δίνει σφάλμα της τάξης $O(m^{-1})$ που είναι καλύτερο από το σφάλμα της Monte Carlo μεθοδολογίας. Πιο εξελιγμένες μέθοδοι, όπως αυτές του κανόνα του Simpson και του κανόνα Newton-Cotes δίνουν καλύτερες αριθμητικές προσεγγίσεις . Όμως σημαντικό μειονέκτημα αυτών των μεθόδων είναι ότι δεν μπορούν να εφαρμόστούν καθώς η διάσταση του D αυξάνεται.

Παρά το γεγονός ότι το σφάλμα στην Monte Carlo ολοκλήρωση παραμένει το ίδιο σε προβλήματα μεγάλων διαστάσεων, εμφανίζονται δύο ουσιαστικές δυσκολίες:

α) όταν η περιοχή D είναι μεγάλη σε χώρο μεγάλης διάστασης, η διακύμανση σ^2, (η οποία μετράει πόσο ομοιόμορφα κατανέμεται η συνάρτηση g στην περιοχή D) μπορεί να είναι απαγορευτικά μεγάλη.

β) μπορεί να μην μπορούμε να παράγουμε ομοιόμορφα τυχαία δείγματα σε μια αυθαίρετη περιοχή D.

‘Ετσι για να ξεπεραστούν αυτές οι δυσκολίες οι ερευνητές συχνά χρησιμοποιούν την μέθοδο της δειγματοληψίας σπουδαιότητας (importance sampling) στην οποία κάποιος παράγει τυχαία δείγματα $x^1, \dots, x^m$ από την μη ομοιόμορφη κατανομή π(x) η οποία δίνει μεγαλύτερη πυκνότητα πιθανότητας στις «σημαντικές» περιοχές του D. Μπορούμε τότε να εκτιμήσουμε το I σαν $\bar{\bar{I}}=\frac{1}{m}\sum_{j=1}^{m}\frac{g(x^j)}{\pi(x^j)}$ το οποίο έχει διακύμανση $\sigma_\pi^2 = var_\pi\left\{\frac{g(x)}{\pi(x)}\right\}$. Στην πιο ευνοϊκή περίπτωση μπορούμε να διαλέξουμε ως π(x)$\propto g(x)$ όταν η g είναι μια μη αρνητική συνάρτηση και το I είναι πεπερασμένο και να έχουμε σαν αποτέλεσμα την ακριβή εκτίμηση του ***I***. Αυτή η ευνοϊκή περίπτωση δεν συναντάται σε καμμία εφαρμογή της Monte Carlo μεθοδολογίας.

Η μέθοδος απόρριψης, η μέθοδος δειγματοληψίας σπουδαιότητας (importance sampling) και η μέθοδος Sampling–importance-resampling χρησιμοποιούν δείγματα που παράγονται από μια δοκιμαστική κατανομή (trial distribution) ρ(x) η οποία διαφέρει από την οικογένεια κατανομών π, αλλά θα πρέπει να είναι παρόμοια.

Εξαιτίας των μεγάλων δυνατοτήτων της μεθοδολογίας Monte Carlo, έχουν αναπτυχθεί διάφορες τεχνικές από τους ερευνητές στους αντίστοιχους τομείς τους. Πρόσφατες εξελίξεις στις τεχνικές Monte Carlo είναι η μέθοδος συσσώρευσης (cluster) , η επαύξηση (augmentation), η επέκταση παραμέτρων, η multicanonical δειγματοληψία, η πολυπλεγματική (multigrid) Monte Carlo, η δειγματοληψία umbrella, η προσομοίωση βαφής, η διαδοχική Monte Carlo, κτλ. Οι διάφορες μέθοδοι Monte Carlo είναι πολύ χρήσιμες στις επιστήμες και στα στατιστικά προβλήματα. Όπως για την στατιστική φυσική, μοριακή προσομοίωση, βιοπληροφορική, δυναμικό σύστημα ανάλυσης, Μπευζιανή συμπερασματολογία και για άλλα στατιστικά προβλήματα με ελλιπή δεδομένα.

Ένα κεντρικό θέμα της Στατιστικής είναι η εξαγωγή συμπερασμάτων για μη παρατηρούμενες παραμέτρους από παρατηρούμενα δεδομένα. Έστω θ το διάνυσμα παραμέτρων και y το παρατηρούμενο διάνυσμα με κατανομή $f(y \mid \theta)$, γνωστή πάνω σε μια πεπερασμένης διάστασης παράμετρο θ. Για να βγάλουμε συμπεράσματα για το θ δύο είναι οι πιο δημοφιλείς μέθοδοι : η μέθοδος μέγιστης πιθανοφάνειας (ΕΜΠ) και η μέθοδος Bayes. Στην ΕΜΠ η άγνωστη παράμετρος θ εκτιμάται από την $\hat{\theta}$ η οποία μεγιστοποιεί την $f(y \mid \theta)$. Στις μπευζιανές μεθόδους για να βγάλουμε συμπεράσματα για το θ βασιζόμαστε στην εκ των υστέρων κατανομή του θ που είναι η $p(\theta \mid y)$.

1.2 Συνοπτική περιγραφή της εργασίας

Σε αυτή την εργασία θα ασχοληθούμε με την επίλυση ενός Μπεϋζιανού προβλήματος ελλιπών δεδομένων με χρήση τεχνικών Monte Carlo.

Στο δεύτερο κεφάλαιο αναφερόμαστε στην Μπεϋζιανή στατιστική, περιγράφουμε τα βασικά βήματα της Μπεϋζιανής προσέγγισης και για προβλήματα μιας παραμέτρου και για πολυπαραμετρικά προβλήματα. Επίσης αναφέρουμε τα βήματα του αλγορίθμου Gibbs στα πολυδιάστατα προβλήματα.

Στο τρίτο κεφάλαιο περιγράφουμε βασικές τεχνικές Monte Carlo όπως την μέθοδο απόρριψης, την μέθοδο αντιθετικών μεταβλητών, την μέθοδο μεταβλητών ελέγχου, την στρωματοποιημένη δειγματοληψία και την δειγματοληψία σπουδαιότητας (importance sampling). Ιδιαίτερη προσοχή δίνεται στις μεθόδους που σχετίζονται με την δειγματοληψία σπουδαιότητας και δίνουμε ένα παράδειγμα ενός Μπεϋζιανού προβλήματος με ελλιπή δεδομένα και πως αυτό μπορεί να επιλυθεί με τη χρήση της δειγματοληψίας σπουδαιότητας. Συγκεκριμένα αφορά στη στατιστική συμπερασματολογία για τον πίνακα συνδιακύμανσης μιας διδιάστατης κανονικής κατανομής από ελλιπή δεδομένα (παράδειγμα Murray, 1977). Επίσης περιγράφουμε το βασικό πλαίσιο της διαδοχικής (sequential) δειγματοληψίας σπουδαιότητας.

Στο τέταρτο κεφάλαιο μελετάμε τις εφαρμογές της μεθόδου διαδοχικής δειγματοληψίας σπουδαιότητας. Περιγράφουμε την SAW (Self avoid walk) μέθοδο, την μέθοδο ανάπτυξης (growth method), την διαδοχική υποκατάσταση σε στατιστικά προβλήματα με ελλιπή δεδομένα και πως αυτές οι μέθοδοι συνδέονται μεταξύ τους. Αναφέρουμε ειδικές τεχνικές και περιγράφουμε ένα γενικό πλαίσιο της διαδοχικής Monte Carlo που μπορεί να χρησιμοποιηθεί σαν βάση για καινούργιες τεχνικές.

Στο πέμπτο κεφάλαιο επαναξετάζουμε το παράδειγμα του Murray (1977) και το επιλύουμε με τη χρήση της μεθόδου της διαδοχικής δειγματοληψίας σπουδαιότητας.

Κεφάλαιο 2

Μπεϋζιανή Στατιστική

Τον 18ο αιώνα, ο ιερέας Thomas Bayes προβληματίσθηκε με το πρόβλημα που απασχολεί τον άνθρωπο από την δημιουργία του. Την πιθανότητα της ύπαρξης Θεού. Οι θεωρίες που βασίζονται σε ισοπίθανα ενδεχόμενα ή στην έννοια της σχετικής συχνότητας δεν είναι χρήσιμες στην περίπτωση αυτή. Παρ' όλα αυτά, μιλάμε για μια αβεβαιότητα που απασχόλησε πολλούς ανθρώπους, περιλαμβανομένου και του Bayes. Η πιθανότητα αυτή φαίνεται να είναι, εκ των πραγμάτων, υποκειμενική. Το καλύτερο που μπορεί να κάνει κάποιος στην περίπτωση αυτή, είναι να συνδυάσει τις διαθέσιμες ενδείξεις και τα λογικά επιχειρήματα και να καταλήξει σε μια προσωπική πιθανότητα ύπαρξης του Θεού. Η ιδέα της προσωπικής πιθανότητας έχει από τότε επεκταθεί, διαμορφωθεί και βελτιωθεί από πολλούς Στατιστικούς που χρησιμοποιούν την Μπεϋζιανή προσέγγιση ως βάση για την ανάπτυξη των απόψεών τους.

Η στατιστική κατά Bayes βασίζεται σε μία απλή ιδέα που είναι η εξής: η μόνη ικανοποιητική περιγραφή της αβεβαιότητας μας επιτυγχάνεται μέσω της πιθανότητας. Η αβεβαιότητα υπάρχει καθημερινά στην ζωή μας. Η Μπεϋζιανή προσέγγιση μας δίνει, μέσω του υπολογισμού πιθανοτήτων, ένα ισχυρό εργαλείο να καταλάβουμε, να χειριστούμε και να ελέγξουμε την αβεβαιότητα.

2.1 Συμπερασματολογία κατά Bayes

Γενικότερα μπορούμε να πούμε ότι η στατιστική συμπερασματολογία είναι η επιστήμη η οποία έχει στόχο να εξάγει συμπεράσματα για έναν πληθυσμό μελετώντας ένα δείγμα που προέρχεται από τον πληθυσμό αυτό.

Το πλαίσιο στο οποίο κινείται η συμπερασματολογία κατά Bayes είναι παρόμοιο με αυτό της κλασικής στατιστικής: υπάρχει η παράμετρος θ του πληθυσμού την οποία θέλουμε να εκτιμήσουμε, καθώς και η πιθανότητα $f(x|\theta)$ η οποία καθορίζει την πιθανότητα παρατήρησης διαφορετικών Χ, κάτω από διαφορετικές τιμές της παραμέτρου θ. Όμως η θεμελιώδης διαφορά είναι ότι το θ χρησιμοποιείται σαν τυχαία ποσότητα. Αν και η διαφορά αυτή μπορεί να φανεί όχι και τόσο ουσιαστική, οδηγεί σε μία τελείως διαφορετική προσέγγιση, ως προς την ερμηνεία, από αυτήν την κλασικής στατιστικής.

Στην ουσία, η συμπερασματολογία μας θα βασιστεί στην $f(\theta|x)$ και όχι στην $f(x|\theta)$, δηλαδή στην πιθανότητα της κατανομής της παραμέτρου

δεδομένης της x και όχι της x δεδομένης της παραμέτρου. Σε πολλές περιπτώσεις αυτό οδηγεί σε περισσότερο φυσικά συμπεράσματα σε σχέση με την κλασική στατιστική, για να μπορέσει όμως να επιτευχθεί αυτό θα πρέπει να καθορίσουμε την a-priori κατανομή f(θ) , η οποία αντιπροσωπεύει «τις πεποιθήσεις» μας για την κατανομή του θ προτού αποκτήσουμε οποιαδήποτε πληροφορία για τα δεδομένα μας.

Η ιδέα της a-priori κατανομής της παραμέτρου θ αποτελεί και την «καρδιά» της θεωρίας κατά Bayes, που για τους υποστηρικτές της θεωρίας αποτελεί το μεγαλύτερο πλεονέκτημα αλλά για τους αντιμαχόμενους αποτελεί το σοβαρότερο μειονέκτημα έναντι της κλασικής στατιστικής.

2.1.1 Χαρακτηριστικά της Μπευζιανής προσέγγισης

Τα βασικά χαρακτηριστικά της Μπευζιανής προσέγγισης είναι:

- Η αντιμετώπιση και χρήση της άγνωστης παραμέτρου θ ως τυχαία μεταβλητή.
- Ο καθορισμός της a-priori κατανομής για το θ (η οποία αντιπροσωπεύει τις πεποιθήσεις μας σχετικά με το θ προτού να έχουμε οποιαδήποτε πληροφορία για τα δεδομένα μας)
- Η χρήση του θεωρήματος του Bayes για τον «εκσυγχρονισμό» των a-priori πεποιθήσεων μας σε a-posteriori πιθανότητες και
- Η εξαγωγή της κατάλληλης συμπερασματολογίας.

Για τον σκοπό αυτό υπάρχουν τέσσερα βήματα-κλειδιά για την Μπευζιανή προσέγγιση:

1. Καθορισμός του μοντέλου πιθανοφάνειας f(x |θ).
2. Καθορισμός της a-priori κατανομής f(θ).
3. Υπολογισμός της a-posteriori κατανομής f(θ |x), από το θεώρημα του Bayes.
4. Εξαγωγή συμπερασμάτων από την a-posteriori πληροφορία.

Πιο αναλυτικά τα βασικά βήματα

- **Επιλογή του κατάλληλου μοντέλου πιθανοφάνειας**

Η επιλογή του κατάλληλου μοντέλου πιθανοφάνειας εξαρτάται από τον μηχανισμό με βάση τον οποίο θα αντιμετωπίσουμε το κάθε πρόβλημα και είναι ίδια η περίπτωση αυτή με εκείνη που αντιμετωπίζουμε στην κλασσική στατιστική δηλ. ποιο είναι το μοντέλο των δεδομένων μας. Πολύ συχνά η γνώση της δομής των δεδομένων μας βοηθά στην επιλογή του κατάλληλου μοντέλου.

- **Επιλογή της a-priori κατανομής**

Η επιλογή της a-priori κατανομής αποτελεί βασικό θέμα στην Μπεϋζιανή θεωρία.

Αναφέρουμε κάποια σημεία που είναι απαραίτητα.

1. Από την στιγμή που η a-priori κατανομή αντιπροσωπεύει τις πεποιθήσεις ενός ατόμου για την παράμετρο θ,προτού μελετηθούν τα δεδομένα του, είναι φυσικό ότι η μεταγενέστερη ανάλυση είναι μοναδική για αυτό το άτομο. Δηλαδή η a-priori κατανομή που θέτει κάποιος άλλος, θα οδηγήσει σε διαφορετική μεταγενέστερη (a-posteriori) ανάλυση. Με αυτήν την έννοια η ανάλυση είναι καθαρά υποκειμενική.
2. Στην περίπτωση που η a-priori δεν είναι «εντελώς παράλογη», η επιρροή της γίνεται ολοένα μικρότερη καθώς προστίθενται νέα δεδομένα. Σε αυτήν την περίπτωση ο λάθος προσδιορισμός της a-priori είναι άνευ σημασίας.
3. Πολύ συχνά έχουμε μια «γενική» ιδέα για το ποια θα πρέπει να είναι η a-priori (πιθανότατα να μπορούμε να πούμε ποιος είναι ο μέσος και η διακύμανσή της), χωρίς όμως να μπορούμε να είμαστε πιο συγκεκριμένοι για την μορφή της. Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε μια «βολική» μορφή της a-priori η οποία θα είναι το αποτέλεσμα των πεποιθήσεων μας και ταυτόχρονα θα κάνει τους μαθηματικούς υπολογισμούς σχετικά εύκολους.
4. Μερικές φορές νομίζουμε ότι δεν έχουμε a-priori πληροφορίες σχετικά με την παράμετρο θ. Σε τέτοιες περιπτώσεις είναι αρκετά συνηθισμένο να χρησιμοποιούμε μια a-priori η οποία να ανακλά την άγνοια μας για την παράμετρο. Αυτό είναι συχνά δυνατό να γίνει, αλλά υπάρχουν αρκετές δυσκολίες που σχετίζονται με αυτό.

Υπολογισμός της a-posteriori κατανομής $f(\theta \mid x)$, από το θεώρημα του Bayes

Το Θεώρημα του Bayes

Το θεώρημα του Bayes για τυχαίες μεταβλητές, με πυκνότητες που συμβολίζονται γενικά με f, παίρνει την εξής μορφή:

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{\int f(\theta) f(x \mid \theta)\, d\,\theta}$$

Θα χρησιμοποιήσουμε αυτόν τον συμβολισμό για τις περιπτώσεις που το x είναι είτε συνεχής είτε διακριτή μεταβλητή. Αντίστοιχα, το θ μπορεί να είναι

διακριτό ή συνεχές, αλλά στην διακριτή περίπτωση ο παρανομαστής της παραπάνω σχέσης δηλ. το $\int f(\theta) f(x \mid \theta)\, d\theta$ θα ερμηνευθεί αντίστοιχα όπως το $\sum f(\theta) f(x \mid \theta)$. Θα πρέπει να προσέξουμε ιδιαίτερα το γεγονός ότι από την στιγμή που ολοκληρώσαμε ως προς θ, ο παρανομαστής στο θεώρημα του Bayes είναι συνάρτηση μόνο ως προς x. Συνεπώς, για δεδομένες παρατηρήσεις x, ο παρανομαστής είναι σταθερά και ονομάζεται σταθερά κανονικοποίησης. Με βάσει αυτά ένας εναλλακτικός τρόπος παρουσίασης του θεωρήματος του Bayes είναι ο εξής:

$$f(\theta \mid x) \propto f(\theta) f(x \mid \theta)$$

Δηλαδή ότι η a-posteriori κατανομή (posterior distribution) είναι ανάλογη της a-priori κατανομής (prior distribution) πολλαπλασιαζόμενης με την συνάρτηση πιθανοφάνειας (Likelihood function). Η $f(\theta \mid x)$ είναι συνάρτηση πυκνότητας πιθανότητας και αυτό το εγγυάται η σταθερά κανονικοποίησης. Επίσης η a-priori και η a-posteriori κατανομές είναι έννοιες σχετικές: η σημερινή a-posteriori είναι η αυριανή a-priori.

Επειδή η εφαρμογή του θεωρήματος του Bayes στην πράξη είναι αρκετά δύσκολη όσον αφορά τους μαθηματικούς υπολογισμούς και η δυσκολία αυτή οφείλεται κυρίως στο ολοκλήρωμα το οποίο υπάρχει στον παρανομαστή, με ορισμένες επιλογές των a-priori κατανομών, το ολοκλήρωμα αυτό μπορεί να παραληφθεί, και γενικά χρειάζονται ειδικές τεχνικές για να απλοποιηθούν οι υπολογισμοί .

Συμπερασματολογία

Η Μπευζιανή ανάλυση παρέχει μία περισσότερο πλήρη συμπερασματολογία υπό την έννοια ότι όλη η γνώση σχετικά με την παράμετρο θ η οποία είναι διαθέσιμη λόγω της a-priori κατανομής και των δεδομένων, αντιπροσωπεύεται από την a-posteriori κατανομή . Αυτό σημαίνει ότι η $f(\theta \mid x)$ είναι η συμπερασματολογία.

2.2 Πολυπαραμετρικά προβλήματα

Τα περισσότερα στατιστικά προβλήματα περιλαμβάνουν στατιστικά μοντέλα τα οποία περιέχουν περισσότερες από μία άγνωστες παραμέτρους. Μπορεί να υπάρξει η περίπτωση όπου μονάχα μία από τις παραμέτρους να έχει ενδιαφέρον, αλλά συνήθως θα υπάρχουν και άλλες παράμετροι των οποίων οι τιμές θα είναι άγνωστες.

Η μέθοδος ανάλυσης των πολυπαραμετρικών προβλημάτων στην Μπεϋζιανή στατιστική είναι πολύ πιο άμεση (τουλάχιστον ως προς τις αρχές της), σε σχέση με αυτήν που χρησιμοποιείται στην κλασική στατιστική. Στην περίπτωση των πολυπαραμετρικών προβλημάτων, έχουμε ένα διάνυσμα $\boldsymbol{\theta}=(\theta_1, \ldots \theta_d)$ από παραμέτρους για το οποίο θέλουμε να

εξάγουμε κάποια συμπεράσματα. Καθορίζουμε μία a-priori (πολυμεταβλητή) κατανομή *f(***θ***)* για το διάνυσμα **θ** και συνδυαζόμενο με την πιθανοφάνεια $f(x \mid \boldsymbol{\theta})$ μέσω του θεωρήματος του Bayes, παίρνουμε όπως και προηγουμένως:

$$f(\boldsymbol{\theta} \mid x) = \frac{f(\boldsymbol{\theta}) f(x \mid \boldsymbol{\theta})}{\int f(\boldsymbol{\theta}) f(x \mid \boldsymbol{\theta})\, d\,\boldsymbol{\theta}}$$

Φυσικά, η a-posteriori κατανομή θα είναι και αυτή τώρα μία πολυμεταβλητή κατανομή. Ωστόσο, η απλότητα της Μπεϋζιανής προσέγγισης τώρα, σημαίνει ότι η συμπερασματολογία για οποιαδήποτε υποομάδα παραμέτρων του διανύσματος **θ** μπορεί να υπολογιστεί άμεσα με υπολογισμούς πιθανοτήτων στην μεταβλητή κατανομή. Για παράδειγμα, η περιθώρια a-posteriori κατανομή για το θ_1 μπορεί εύκολα να υπολογιστεί ολοκληρώνοντας ως προς όλες τις άλλες παραμέτρους του διανύσματος **θ**.

$$f(\theta_1 \mid x) = \int_{\theta_2} \dots \int_{\theta_d} f(\theta \mid x) d\theta_2 \dots d\theta_d$$

Για να γενικευτεί το πρόβλημα στις d-διαστάσεις, μία σειρά από πρακτικά προβλήματα δημιουργούνται στα βασικά βήματα της Μπεϋζιανής στατιστικής.

Καθορισμός των a-priori κατανομών

Οι a-priori κατανομές τώρα είναι πολυδιάστατες κατανομές. Αυτό σημαίνει ότι ο καθορισμός των a-priori κατανομών πρέπει να αντιπροσωπεύει τις a-priori πεποιθήσεις μας όχι απλά για κάθε μία παράμετρο ξεχωριστά, αλλά θα πρέπει να αντιπροσωπεύει και τις πεποιθήσεις μας σχετικά με την ανεξαρτησία ανάμεσα σε διαφορετικούς συνδυασμούς παραμέτρων (εάν μία παράμετρος θεωρηθεί μεγάλη, είναι δυνατόν κάποια άλλη παράμετρος να θεωρηθεί μικρή;). Η επιλογή κατάλληλων οικογενειών από a-priori κατανομές και ο συνοψισμός των a-priori πληροφοριών των ειδικών με αυτόν τον τρόπο, είναι αισθητά πιο πολύπλοκο πρόβλημα.

Υπολογισμός

Ακόμα και στα προβλήματα της μίας-διάσταση, είναι σημαντική η χρησιμότητα των συζυγών οικογενειών κατανομών στην απλούστευση της a-posteriori ανάλυσης μέσα από το θεώρημα του Bayes. Με τα πολυδιάστατα προβλήματα, τα ολοκληρώματα γίνονται ακόμα

δυσκολότερα για υπολογισμό. Αυτό κάνει την χρήση των συζυγών a-priori ακόμα πιο απαραίτητη, και φανερώνει την ανάγκη για υπολογιστικές τεχνικές, με σκοπό να βγάλουμε συμπεράσματα για το πότε η χρήση των συζυγών κατανομών είναι κατάλληλη και πότε είναι ακατάλληλη.

Συμπερασματολογία

Ολόκληρη η a-posteriori συμπερασματολογία, περιλαμβάνεται στην a-posteriori κατανομή, η οποία έχει τόσες διαστάσεις όσες και η παράμετρος $\boldsymbol{\theta}$. Η δομή της a-posteriori κατανομής μπορεί να είναι ιδιαίτερα πολύπλοκη, και μπορεί να απαιτεί συγκεκριμένη υποδομή (για παράδειγμα έναν υπολογιστή με δυνατότητες παροχής καλών γραφικών) για να μπορέσει να δοθεί έμφαση στις πιο σημαντικές σχέσεις τις οποίες περιλαμβάνει. Παρόλα τα πρακτικά προβλήματα, είναι πολύ σημαντικό να τονίσουμε το γεγονός για μία ακόμα φορά, ότι για τα πολυπαραμετρικά προβλήματα χρησιμοποιείται η ίδια θεωρία όπως και για τα προβλήματα μίας διάστασης. Το πλαίσιο της Μπεϋζιανής θεωρίας τονίζει ότι όλα τα συμπεράσματα πηγάζουν από τους βασικούς κανόνες των πιθανοτήτων.

2.3 Μεθοδολογία MCMC

Οι μέθοδοι MCMC (Markov Chain Monte Carlo) έχουν βοηθήσει πολύ την πρακτική εφαρμογή της Μπεϋζιανής μεθοδολογίας, αντιμετωπίζοντας διεξοδικά το πρόβλημα του υπολογισμού της posterior κατανομής (ή καλύτερα των χαρακτηριστικών της που μας ενδιαφέρουν). Η βασική ιδέα τους είναι απλή: αντί να υπολογίσουμε αναλυτικά τις ποσότητες που μας ενδιαφέρουν σε πολύπλοκες posterior κατανομές, προσομοιώνουμε ένα δείγμα τιμών από μία κατάλληλη Μαρκοβιανή Αλυσίδα που βρίσκεται σε κατάσταση ισορροπίας. Έτσι μπορούμε να υπολογίσουμε τα χαρακτηριστικά που θέλουμε (μέση τιμή, διασπορά κ.ά.) μέσω των αντίστοιχων τιμών του δείγματος που κατασκευάσαμε.

2.3.1 Δειγματολήπτης Gibbs (Gibbs sampler)

Έστω ότι μας ενδιαφέρει να προσομοιώσουμε δείγμα από την $\pi(\mathrm{x}_1, \dots, \mathrm{x}_k \mid y)$. Ο δειγματολήπτης Gibbs το κατορθώνει προσομοιώνοντας ακολουθιακά και επαναληπτικά από τις δεσμευμένες posterior κατανομές των επιμέρους παραμέτρων.

Αλγόριθμος Gibbs Sampling

Ο αλγόριθμος Gibbs Sampling είναι ιδιαίτερα αποτελεσματικός στα πολυδιάστατα προβλήματα και υλοποιείται ως ακολούθως:

Έστω ένα k-διάστατο διάνυσμα $\mathbf{x} = x_1, \dots, x_\kappa$ με κατανομή π(**x**) , όπου τα x_i δεν είναι απαραίτητα μονοδιάστατα. Υποθέτουμε ακόμα πως μπορούμε να προσομοιώσουμε από τις δεσμευμένες κατανομές :

$$\pi_i(\ x_i\ \mid x_1 \dots\ x_{i-1},\ x_{i+1}, \dots x_\kappa)\ ,\ i=1,\dots,\kappa\ \text{ [full conditionals].}$$

Τότε ο αλγόριθμος που ακολουθεί κατασκευάζει δείγματα από την κατανομή π(**x**) μέσα από διαδοχικές προσομοιώσεις από τις δεσμευμένες κατανομές που περιγράψαμε:

Διαλέγουμε μία τυχαία αρχική τιμή $\boldsymbol{x^0} = (x_1^{\,0}, \dots, x_\kappa^{\,0}\ ) \in X$ και θέτουμε j=1

1. Προσομοιώνουμε , $x_1^{\,j} \sim \pi_1(\ x_1\ \mid x_2^{\,j-1} \dots x_\kappa^{\,j-1})$
   προσομοιώνουμε $x_2^{\,j} \sim \pi_2(\ x_2\ \mid x_1^{\,j}, x_3^{\,j-1} \dots x_\kappa^{\,j-1})$
   . . .
   προσομοιώνουμε $x_\kappa^{\,j} \sim \pi_\kappa(\ x_\kappa\ \mid x_1^{\,j}, \dots x_{\kappa-1}^{\,j}\ )$
2. Θέτουμε *j=j+*1 και πηγαίνουμε στο 1 βήμα

Επαναλαμβάνουμε την διαδικασία μέχρι να επιτευχθεί η σύγκλιση

2.3.2 Αύξηση δεδομένων (Data Augmentation)

Τεχνική που επιτρέπει την χρήση του αλγορίθμου gibbs σε κάποιες κατηγορίες προβλημάτων όπου εξ΄αρχής αυτό δεν ήταν εφικτό. Αυτό γίνεται με την εισαγωγή κάποιων νέων τυχαίων μεταβλητών στο πρόβλημα και έτσι γίνεται εύκολη στο χειρισμό η πιθανοφάνεια.

Χρήσεις αλγορίθμου αύξησης δεδομένων

Σε περιπτώσεις όπου η συνάρτηση πιθανοφάνειας είναι δύσκολη στο χειρισμό (δεν μπορεί να χρησιμοποιηθεί εύκολα για συμπερασματολογία) αλλά με την εισαγωγή νέων μεταβλητών το πρόβλημα λύνεται.

Έτσι ένα μπευζιανό πρόβλημα με ελλιπή δεδομένα (missing data) μπορεί να διαμορφωθεί ως εξής:

Έστω ότι έχουμε την συνάρτηση πιθανότητας με πλήρη δεδομένα (complete data) $f(y \mid \theta)$ από την οποία μπορούμε να πάρουμε την αναλυτική μορφή μιας posterior κατανομής. Το μέρος του y που παρατηρείται το ονομάζουμε y_{obs} και το υπόλοιπο μέρος του y που λείπει

το ονομάζουμε y_{mis}. Θέτουμε y=(y_{obs} , y_{mis}). Η παράμετρος θ εδώ θα είναι Σ (πίνακας συνδιακύμανσης).

●Έστω ότι η prior κατανομή του Σ είναι η $f(\Sigma)$.

●Η από κοινού posterior κατανομή των y_{mis} και Σ είναι :

$f\ (\Sigma, y_{mis}) \propto f(y_{mis}, y_{obs}\ |\ \Sigma) f(\Sigma)$

●και η περιθώρια posterior του Σ είναι :

$f(\Sigma\ | y_{obs}) = f(\Sigma) = \int f(\Sigma, y_{mis})\, dy_{mis}$

Συγκεκριμένα σε προβλήματα με ελλιπή δεδομένα (missing data problems) χρησιμοποιούμε τον αλγόριθμο αύξησης δεδομένων ως εξής:

- Επειδή η πιθανοφάνεια ελλιπών δεδομένων $f(y_{obs}\ |\Sigma)$ είναι δύσκολη στο χειρισμό

- Εισάγουμε νέες τυχαίες μεταβλητές y_{mis} οι οποίες εκφράζουν τα δεδομένα που λείπουν

- Και έχουμε την πιθανοφάνεια πλήρους σετ δεδομένων $f(y_{obs}, y_{mis}\ |\Sigma)$ που είναι εύκολη στο χειρισμό

- Η από κοινού posterior κατανομή των παραμέτρων και των ελλιπών δεδομένων θα είναι

$$f(\Sigma, y_{mis}\ | y_{obs}) \propto f(\Sigma) f(y_{obs}, y_{mis}\ |\Sigma)$$

- Με αλγόριθμο Gibbs :

● Προσομοιώνουμε y_{mis} από $f(y_{mis}\ | y_{obs}, \Sigma)$

● Προσομοιώνουμε Σ από $f(\Sigma\ | y_{obs}, y_{mis})$

Κεφάλαιο 3

Βασικές τεχνικές Monte Carlo

Σε αυτό το κεφάλαιο παρουσιάζουμε τις βασικές τεχνικές Monte Carlo όπως τη μέθοδο της αντιστροφής (inversion), τη δειγματοληψία απόρριψης (rejection sampling), την αντιθετική δειγματοληψία, τη μέθοδο ελέγχου τυχαίας μεταβλητής, την στρωματοποιημένη δειγματοληψία και τη δειγματοληψία σπουδαιότητας (importance sampling). Ιδιαίτερη προσοχή δίνεται στις μεθόδους που σχετίζονται με τη δειγματοληψία σπουδαιότητας (importance sampling) (π.χ. για την επίλυση γραμμικών εξισώσεων, για φυλλογενετικές αναλύσεις και για Μπεϋζιανή συμπερασματολογία με ελλιπή δεδομένα).

3.1 Δημιουργία απλών τυχαίων μεταβλητών

Για να παράγουμε τυχαίες μεταβλητές που ακολουθούν μια γενική κατανομή πιθανότητας μιας συνάρτησης π, θα πρέπει πρώτα να παράγουμε τυχαίες μεταβλητές ομοιόμορφα κατανεμημένες στο [0,1]. Αυτές οι τυχαίες μεταβλητές συχνά ονομάζονται τυχαίοι αριθμοί (random numbers) για ευκολία. Η παραγωγή όμως αυτών των τυχαίων αριθμών είναι ένα έργο που δεν είναι εύκολα εφικτό στον υπολογιστή. Σε ένα πρόγραμμα εντοπισμού σφαλμάτων συχνά πρέπει να επαναλάβουμε τον ίδιο υπολογισμό πολλές φορές. Αυτό απαιτεί από εμάς να αναπαράγουμε την ίδια ακολουθία των τυχαίων αριθμών επανειλημμένα. Για αυτό έχει γίνει αποδεκτή από την επιστημονική κοινότητα η χρήση ψευδοτυχαίων αριθμών.

3.2 Η μέθοδος απόρριψης

Έστω ότι $l(x) = c\pi(x)$, όπου π είναι μια συνάρτηση κατανομής πιθανότητας ή συνάρτηση πυκνότητας και c είναι άγνωστη. Αν μπορούμε να βρούμε μια δειγματοληπτική κατανομή (sampling distribution) $g(x)$ και μια σταθερά M, έτσι ώστε να ισχύει M $g(x) \geq l(x)$ για κάθε x, τότε μπορούμε να εφαρμόσουμε την ακόλουθη διαδικασία.

Μέθοδος απόρριψης (Rejection sampling) [von Neumann (1951)]

α) Παίρνουμε ένα δείγμα x από $g(\quad)$ και υπολογίζουμε την αναλογία

$$r = \frac{l(x)}{\mathrm{M}\, g(x)} (\leq 1)$$

β) Ρίχνουμε ένα κέρμα με πιθανότητα επιτυχίας r

● Αν έρθει κεφαλή δεχόμαστε το x

● Διαφορετικά απορρίπτουμε το x και πηγαίνουμε πίσω στο (α)

Το αποδεκτό δείγμα ακολουθεί την κατανομή στόχο (target) π.

Για να δείξουμε ότι η παραπάνω διαδικασία είναι σωστή θέτουμε I να είναι η δείκτρια συνάρτηση, έτσι ώστε I=1 αν το δείγμα x που προέρχεται από $g(\quad)$ είναι αποδεκτό και I=0 διαφορετικά.

Στη συνέχεια παρατηρούμε ότι

$$P(I=1)=\int P(I=1 \mid X=x)g(x)dx=\int \frac{c\pi(x)}{\mathrm{M}\,g(x)}g(x)dx=\frac{c}{\mathrm{M}}$$

Ως εκ τούτου $\mathrm{p}(x \mid I=1)=\frac{c\pi(x)}{\mathrm{M}\,g(x)}g(x)\,/\,P(I=1)=\pi(\mathrm{x})$

Το κλειδί για μια επιτυχή εφαρμογή του αλγορίθμου είναι να βρεθεί μια καλή δοκιμαστική κατανομή (trial) $g(x)$. Όμως συνήθως είναι πολύ δύσκολο να εφαρμοστεί η απλή μέθοδος απόρριψης για ένα πρόβλημα προσομοίωσης Monte Carlo μεγάλων διαστάσεων.

ΑΛΓΟΡΙΘΜΟΣ ΑΠΟΡΡΙΨΗΣ

Έστω λοιπόν ότι επιθυμούμε την παραγωγή τυχαίων αριθμών από μία κατανομή με συνάρτηση πιθανότητας $\pi_j = P(X = j)$, $j = 0,1,\ldots$. Έστω επίσης ότι μπορούμε εύκολα να παράγουμε τυχαίους αριθμούς από μία κατανομή με συνάρτηση πιθανότητας $g_j = P(Y = j)$, $j = 0,1,\ldots$ για την οποία το μόνο που απαιτούμε είναι το εξής: αν $g_j = 0$ τότε και $\pi_j = 0$ (δηλ. η κατανομή της X είναι απόλυτα συνεχής ως προς την κατανομή της Y).
Σύμφωνα με αυτή τη μέθοδο:
- παράγουμε έναν τυχαίο αριθμό Y από την $\{g_j,\ j=1,2,\ldots\}$ και στη συνέχεια
- δεχόμαστε αυτό το Y (θέτουμε X = Y) με πιθανότητα ανάλογη του πηλίκου π_Y/g_Y .
- εάν δεν γίνει αποδεκτό αυτό το Y, παράγουμε έναν άλλο τυχαίο αριθμό από την $\{g_j, j = 0,1,\ldots\}$ κ.ο.κ.

Πιο συγκεκριμένα, θα πρέπει να υπάρχει μία σταθερά $c < \infty$ για την οποία να ισχύει $\frac{\pi_j}{g_j} \le c$ για κάθε j: $\pi_j > 0$.

Ο αλγόριθμος απόρριψης είναι ο εξής:

ΒΗΜΑ 1 : **Προσομοιώνουμε έναν τυχαίο αριθμό *Y* από κατανομή με σ.π.** $\{g_j, j = 1,2,\ldots\}$.
ΒΗΜΑ 2 : **Προσομοιώνουμε έναν τυχαίο αριθμό** $U \sim U(0,1)$

ΒΗΜΑ 3 : **Εάν** $U < \pi_Y/cg_Y$, **θέτουμε** $X = Y$ **και σταματάμε.**

Εάν όχι, επιστρέφουμε στο 1.

Παρατηρούμε ότι σε κάθε επανάληψη δεχόμαστε ή απορρίπτουμε την τιμή Y ανεξάρτητα από τις προηγούμενες επαναλήψεις.

3.3 Μέθοδοι μείωσης διασποράς

Εδώ περιγράφουμε μερικές τεχνικές που χρησιμοποιούνται συνήθως για την μείωση διασποράς στους υπολογισμούς Monte Carlo.

● **Στρωματοποιημένη δειγματοληψία:**

Είναι μια ισχυρή τεχνική που συχνά χρησιμοποιείται στην έρευνα πληθυσμού και είναι επίσης πολύ χρήσιμη στους υπολογισμούς Monte Carlo. Ο πληθυσμός διαιρείται σε επί μέρους στρώματα, έτσι ώστε να δημιουργηθούν ομογενείς υποομάδες πριν από την έναρξη της δειγματοληψίας. Κάθε παρατήρηση μπορεί να συμπεριληφθεί μόνο σε ένα στρώμα. Με τον τρόπο αυτόν αυξάνεται η αντιπροσωπευτικότητα του δείγματος και μειώνεται το τυχαίο σφάλμα.

Μαθηματικά αυτή η μέθοδος μπορεί να θεωρηθεί ως μια ειδική μέθοδος δειγματοληψίας σπουδαιότητας (importance sampling) με δοκιμαστική κατανομή (trial) $g(x)$. Ας υποθέσουμε ότι μας ενδιαφέρει να εκτιμήσουμε $\int_x f(x)\,dx$. Αν είναι δυνατό, θέλουμε να χωρίσουμε την περιοχή Χ σε κ ξένες υποπεριοχές $D_1, \dots, D_K$ έτσι ώστε σε κάθε υποπεριοχή η συνάρτηση $f(x)$ να είναι σχετικά ομοιογενής (δηλαδή να είναι κοντά σε μια σταθερά). Στη συνέχεια μπορούμε να χρησιμοποιήσουμε m_i τυχαία δείγματα $X^{(i,1)}, \dots, X^{(i,m_i)}$ στην υποπεριοχή D_i και να προσεγγίσουμε το ολοκλήρωμα της υποπεριοχής $\int_{D_i} f(x)\,dx$ με

$$\hat{\mu}_i = \frac{1}{m_i}[f(X^{(i,1)}) + \cdots + f(X^{(i,m_i)})]$$

Το μ μπορεί να προσεγγιστεί από $\hat{\mu} = \hat{\mu}_1 + \cdots + \hat{\mu}_\kappa$ η διακύμανση του οποίου υπολογίζεται εύκολα σαν

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma_1^2}{m_1} + \cdots + \frac{\sigma_m^2}{m_\kappa},$$

όπου η σ_i^2 είναι η μεταβολή της $f(x)$ στην περιοχή D_i.

Αν δεν καταφέρουμε να έχουμε σχετικά ομοιογενείς $f(x)$ σε κάθε περιοχή D_i η στρωματοποιημένη δειγματοληψία στην πραγματικότητα κάνει τον υπολογισμό λιγότερο ακριβή από ότι ένα απλό Monte Carlo. Για αυτό

πρέπει να σκεφτόμαστε προσεκτικά πριν εφαρμόσουμε μια προηγμένη τεχνική.

● Μέθοδος μεταβλητών ελέγχου

Σε αυτή την μέθοδο χρησιμοποιούμε μια μεταβλητή ελέγχου C η οποία συσχετίζεται με το δείγμα X , για να παράγουμε μια καλύτερη εκτίμηση. Ας υποθέσουμε ότι η εκτίμηση του μ=E(x) είναι αυτή που μας ενδιαφέρει και μ_c=E(C) είναι γνωστή. Τότε μπορούμε να κατασκευάσουμε Monte Carlo δείγματα της μορφής X(b)=X+b(C-μ_c) που έχουν την ίδια μέση τιμή με αυτή του X δηλαδή

E(X(b))=E(X+b(C-μ_c))=E(X)+b[E(C)- μ_c]=E(X)+b [μ_c- μ_c]= E(X)

αλλά μια νέα διασπορά:

var{ X(b)}=var{X + b(C − μ_c) } =var(X)+2bcov(X,C) +b^2var(C)

Αν ο υπολογισμός του cov(X,C) και var(C) είναι εύκολος τότε μπορούμε να θέσουμε b= - cov(X,C)/ var(C) που σε αυτή την περίπτωση Var{ X(b)}=

$$= \text{var(X)} + 2[-\text{cov(X, C)}/\ \text{var(C)}]\text{cov(X, C)} + [-\ \text{cov(X, C)}/\ \text{var(C)}]^2\text{var(C)}=$$

$$= \text{var(X)} - \frac{[\text{cov(X,C)}]^2}{\text{var(C)}} \qquad \text{όπου } \rho_{XC} = \frac{\text{cov(X,C)}}{\sqrt{\text{var(X) var(C)}}}$$

άρα έχουμε Var{ X(b)} =(1- ρ_{XC}^2)$var(X) < var(X)$

Μια άλλη περίπτωση είναι όταν γνωρίζουμε μόνο ότι η E(C)=μ. Τότε μπορούμε να σχηματίσουμε X(b)= b X+(1-b)C. Αν C συσχετίζεται με X, μπορούμε πάντα να επιλέγουμε μια κατάλληλη b έτσι ώστε το X(b) να έχει Var{ X(b)} $< var(X)$.

● Μέθοδος αντιθετικών μεταβλητών

Αυτή η μέθοδος δημιουργήθηκε από τους Hammersley και Morton (1956). Η μέθοδος αυτή περιγράφει έναν τρόπο να παράγουμε αρνητικά συσχετισμένα δείγματα. Αν υποθέσουμε U ότι είναι ένας τυχαίος αριθμός που χρησιμοποιείται στην παραγωγή ενός δείγματος X που ακολουθεί μια κατανομή cdf την F[δηλαδή X=$F^{-1}(U)$], [όπου cdf είναι η σωρευτική κατανομή (cumulative distribution function) και είναι πάντοτε μονότονη συνάρτηση με δυνατή περιοχή τιμών από 0 έως 1]. Τότε και η X′=$F^{-1}(1 - U)$ επίσης ακολουθεί κατανομή F.

Γενικότερα αν g είναι μια μονότονη συνάρτηση τότε
$\{g(u_1) - g(u_2)\}\{g(1-u_1) - g(1-u_2)\} \leq 0$ για κάθε $u_1, u_2 \in [0,1]$.

Για δύο ανεξάρτητες ομοιόμορφες τυχαίες μεταβλητές U_1 και U_2 (στην πραγματικότητα αυτό που απαιτείται μόνο είναι αυτές οι δύο τυχαίες μεταβλητές να είναι i.i.d με μια συμμετρική κατανομή στο [0,1]) έχουμε:

$\mathrm{E}[\{g(U_1) - g(U_2)\}\{g(1-U_1) - g(1-U_2)\}] = \mathrm{cov}(\mathrm{X}, \mathrm{X}') \leq 0$

$Cov(X, \mathrm{X}') = E(X\,\mathrm{X}') - E(X)E(\mathrm{X}')$

όπου $\mathrm{X} = g(U)$ και $\mathrm{X}' = g(1-U)$. Έτσι $\mathrm{Var}[(\mathrm{X} + \mathrm{X}')/2] \leq var(X)/2$ όπου φαίνεται ότι παίρνοντας το ζεύγος X και X′ είναι καλύτερα από το να πάρουμε δύο ανεξάρτητα Monte Carlo δείγματα για εκτίμηση του E(X).

Έτσι πιο αναλυτικά

Αν X και X′ ανεξάρτητες τυχαίες μεταβλητές που έχουν μέση τιμή $\theta = \mathrm{E}[x]$ τότε η $\mathrm{Var}[(\mathrm{X} + \mathrm{X}')/2] = 1/4\,[\mathrm{Var}(\mathrm{X}) + \mathrm{Var}(\mathrm{X}') + 2\,\mathrm{cov}(\mathrm{X}, \mathrm{X}')]$.

Αν X και X′ αρνητικά συσχετισμένες θα παίρνουμε μικρότερη διασπορά.

Για να δούμε την αρνητική συσχέτιση υποθέτουμε ότι X είναι συνάρτηση από m τυχαίους αριθμούς με $\mathrm{X} = \mathrm{h}(U_1, U_2, \ldots, U_m)$ όπου $U_1, U_2, \ldots, U_m$ είναι m ανεξάρτητοι τυχαίοι αριθμοί. Αν U είναι τυχαίος αριθμός ο οποίος προσομοιώνεται από ομοιόμορφη κατανομή (0,1) και $1-U$ επίσης είναι τυχαίος αριθμός που προσομοιώνεται από ομοιόμορφη κατανομή (0,1) τότε η τυχαία μεταβλητή $\mathrm{X}' = \mathrm{h}(1-U_1, 1-U_2, \ldots, 1-U_m)$ θα έχει την ίδια κατανομή με τη X. Η U *είναι* αρνητικά συσχετισμένη με την $1-U$. Υποθέτουμε ότι και η X με την X′ είναι αρνητικά συσχετισμένες. Αυτό γίνεται στην περίπτωση όπου η h είναι μια μονότονη (αύξουσα ή φθίνουσα) συνάρτηση . Προσομοιώνουμε τις $U_1, U_2, \ldots, U_m$ για να υπολογίσουμε την X και προσομοιώνουμε τις $1-U_1, 1-U_2, \ldots, 1-U_m$ για να υπολογίσουμε την X′ .

- **Rao –Blackwellization**

Το θεώρημα των Rao-Blackwell:

Η έννοια της επάρκειας συνδέεται με το πρόβλημα της ύπαρξης αμερόληπτων εκτιμητριών ελάχιστης διασποράς. Συγκεκριμένα, αν έχουμε μια αμερόληπτη εκτιμήτρια $\tilde{\theta}$ μιας παραμέτρου θ και μια επαρκή στατιστική συνάρτηση T για την παράμετρο θ, τότε η αναζήτηση μιας "καλύτερης" (με την έννοια της μικρότερης διασποράς) αμερόληπτης εκτιμήτριας $\hat{\theta}$ μπορεί να περιορισθεί μεταξύ συναρτήσεων της T που έχουν την μορφή:

$$\hat{\theta} = \hat{\theta}\,(\mathrm{T}) = \mathrm{E}\,(\tilde{\theta} \mid T)$$

Το αποτέλεσμα αυτό αποδείχθηκε από τον Rao το 1945 και τον Blackwell το 1947 και είναι γνωστό ως θεώρημα Rao-Blackwell.

Η **μέθοδος Rao –Blackwellization** αποτελεί μια βασική αρχή στον υπολογισμό Monte Carlo. Ας υποθέσουμε ότι έχουμε πάρει ανεξάρτητα δείγματα $x^1, \dots, x^m$ από την κατανομή στόχο π(x) και ενδιαφερόμαστε να υπολογίσουμε την $I=E_\pi h(x)$ όπου $h(x)$ στατιστική συνάρτηση. Ένας απλός εκτιμητής είναι :

$$\hat{I} = \frac{1}{m}\{h(x^1) + \cdots + h(x^m)\}$$

Ας υποθέσουμε επιπλέον ότι το x μπορεί να αναλυθεί σε δύο μέρη (x_1, x_2) και θέλουμε να υπολογίσουμε την $\mathrm{E}[h(x) \mid x_2^{(1)}]$. Ένας εναλλακτικός εκτιμητής του I είναι :

$$\tilde{I} = \frac{1}{m}\left\{E[h(x) \mid x_2^{(1)}] + \cdots + E[h(x) \mid x_2^{(m)}]\right\}$$

Και οι δύο εκτιμητές $\hat{I}, \tilde{I}$ είναι αμερόληπτοι γιατί

$$E_\pi h(x) = E_\pi[\mathrm{E}\{h(x) \mid x_2^{(1)}\}]$$

Αν η υπολογιστική προσπάθεια είναι ίδια για τη λήψη των δύο εκτιμήσεων τότε θα προτιμάμε το $\tilde{I}$ γιατί

$$\mathrm{Var}\{h(x)\} = var\{\mathrm{E}[h(x) \mid x_2]\} + \mathrm{E}\{\mathrm{var}[\mathrm{h(x)} \mid x_2]\}$$

Συνεπάγεται ότι $var(\hat{I}) = \frac{\mathrm{Var}\{h(x)\}}{m} \geq \frac{var\{\mathrm{E}[h(x) \mid x_2]\}}{m} = var(\tilde{I})$

Στη στατιστική $\hat{I}$ συχνά αποκαλείται «εκτιμητής ιστόγραμμα » ή εμπειρικός εκτιμητής και $\tilde{I}$ ονομάζεται «εκτιμητής μείγμα».

3.4. Ακριβείς μέθοδοι για μοντέλα με δομή αλυσίδας (chain – structured Models)

Μια σημαντική κατανομή πιθανοτήτων που χρησιμοποιείται σε πολλές εφαρμογές έχει την ακόλουθη μορφή:

$$\pi(\mathbf{x}) \propto \exp\{-\textstyle\sum_{i=1}^{d} h_i(x_{i-1}, x_i)\} \qquad (3.1)$$

όπου το $\mathbf{x}=(x_0, x_1, \dots, x_d)$

Αυτός ο τύπος του μοντέλου έχει μια Μαρκοβιανή δομή επειδή η δεσμευμένη κατανομή $\pi(x_i \mid \boldsymbol{x}_{[-i]})$ όπου $\boldsymbol{x}_{[-i]} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$, εξαρτάται μόνο από τις δύο γειτονικές μεταβλητές x_{i-1} και x_{i+1} που είναι :

$\pi(x_i \mid \boldsymbol{x}_{[-i]}) \propto \exp\{-\mathrm{h}(x_{i-1}, x_i) - \mathrm{h}(x_i, x_{i+1})\}$.

Οι μη παρατηρούμενες μεταβλητές σε ένα μοντέλο χώρου καταστάσεων μπορούν να εκπροσωπούνται από αυτή τη δομή. Η δομή του μοντέλου αυτού συχνά αναφέρεται ως το κρυμμένο (hidden) μοντέλο Markov (HMM). Σε τέτοια μοντέλα μπορεί να χρησιμοποιηθεί η μέθοδος του δυναμικού προγραμματισμού για να βρούμε το ολικό μέγιστο του $\pi(x)$, καθώς και κατάλληλοι αλγόριθμοι για την εύρεση της περιθώριας κατανομής κάθε κατάστασης x_i και για την παραγωγή τυχαίων δειγμάτων από την από κοινού κατανομή $\pi(x)$.

3.4.1 Δυναμικός Προγραμματισμός

Ο Δυναμικός Προγραμματισμός είναι μία υπολογιστική μέθοδος η οποία προκύπτει από τη σύνθεση των επιμέρους αποφάσεων που αλληλοεξαρτώνται. Το αρχικό πρόβλημα διασπάται σε επιμέρους υποπροβλήματα τα οποία συνδέονται με τη βοήθεια κατάλληλων αναδρομικών σχέσεων. Χαρακτηριστικό του Δυναμικού Προγραμματισμού είναι ότι δεν υπάρχει γενικευμένη διατύπωση της μεθόδου που να έχει άμεση λειτουργική ισχύ. Οι αναδρομικές σχέσεις που συνεπάγεται η μέθοδος διαφοροποιούνται ριζικά από πρόβλημα σε πρόβλημα.

Υποθέτουμε ότι κάθε συνιστώσα x_i του διανύσματος $\mathbf{x}$ παίρνει τιμές μόνο στο σύνολο $S=\{s_1, \dots, s_k\}$.

Η μεγιστοποίηση της κατανομής $\pi(\mathbf{x}) \propto \exp\{-\sum_{i=1}^{d} h_i(x_{i-1}, x_i)\}$ είναι ισοδύναμη με την ελαχιστοποίηση του εκθέτη του

$$H(x)=h_1(x_0, x_1) + \cdots + h_d(x_{d-1}, x_d)$$

Μια αναδρομική διαδικασία μπορεί να διεξάγεται ως εξής:

1) Ορίζουμε $m_1(x)=\min_{s_i \epsilon S} h_1(s_i, x)$ για $x=s_1, \dots, s_k$

2) Υπολογίζουμε αναδρομικά την συνάρτηση

$$m_t(x)= \min_{s_i \epsilon S}\{m_{t-1}(s_i) + h_t(s_i, x)\} \text{ για } x=s_1, \dots, s_k$$

3) Η βέλτιστη τιμή $H(x)$ επιτυγχάνεται από την $\min_{s \epsilon S} m_d(s)$

Δεν είναι δύσκολο να παρατηρήσουμε ότι το ελάχιστο της $m_1(x)$ είναι ίσο με το ελάχιστο της $h_1(x_0, x_1)$. Με επαγωγή μπορούμε εύκολα να δείξουμε ότι :

$$\min_{x \epsilon S} m_t(x)=\min_{x_0, \dots, x_t \epsilon S}[h_1(x_0, x_1) + \cdots + h_t(x_{t-1}, x_t)]$$

Έτσι η παραπάνω διαδικασία πράγματι ελαχιστοποιεί την συνάρτηση στόχο H(x).

3.4.2 Ακριβής προσομοίωση (Exact simulation)

Το πρώτο βήμα για να προσομοιώσουμε από την :

$$\pi(\mathbf{x}) \propto \exp\{-\sum_{i=1}^{d} h_i(x_{i-1}, x_i)\}$$

είναι να προσομοιώσουμε x_d από την περιθώριά της κατανομή. Αφού έχουμε παράγει την x_d μπορούμε να εργαστούμε αναδρομικά, δηλαδή παίρνοντας δείγμα x_{d-1} από την $\pi(x_{d-1} \mid x_d)$, x_{d-2} από $\pi(x_{d-2} \mid x_{d-1})$ και να συνεχίσουμε με τον ίδιο τρόπο. Στο βήμα της περιθωριοποίησης βασίζεται η αρχή κατά την οποία το σύνολο μπορεί να αναλυθεί σε αναδρομικά βήματα που είναι :

$$Z \equiv \sum_{\boldsymbol{x}} \exp\{-H(x)\} = \sum_{x_d}[\dots[\sum_{x_1}\{\sum_{x_0} e^{-h_1(x_0, x_1)}\} e^{-h_2(x_1, x_2)}] \dots]$$

Ακριβέστερα, οι ακόλουθες αναδρομές που είναι παρόμοιες με αυτές του δυναμικού προγραμματισμού (DP) μπορούν να πραγματοποιηθούν από έναν υπολογιστή:

1) Ορίζουμε $V_1(x) = \sum_{x_0 e\, S} e^{-h_1(x_0, x)}$

2) Υπολογίζουμε αναδρομικά για t=2,…,d

$$V_t(x) = \sum_{y e S} V_{t-1}(y)\, e^{-h_t(y,x)}$$

Η περιθώρια κατανομή του x_d είναι $\pi(x_d) = V_d(x_d)/Z$ όπου $Z=\sum_{xe\, S} V_d(x)$

Για να προσομοιώσουμε **x** από την π, μπορούμε να κάνουμε τα παρακάτω:

3) Παίρνουμε x_d από το S με πιθανότητα $V_d(x_d)/Z$

4) Για t=d-1,d-2,….,1 παίρνουμε x_t από την κατανομή

$$p_t(x) = \frac{V_t(x) e^{-h_{t+1}(x, x_{t+1})}}{\sum_{y e S} V_t(y)\, e^{-h_{t+1}(y, x_{t+1})}}$$

Το τυχαίο δείγμα $\mathbf{x} = (x_1, \dots, x_d)$ που λαμβάνεται με αυτόν τον τρόπο ακολουθεί την κατανομή π(x).

3.5 Δειγματοληψία σπουδαιότητας και σταθμισμένα δείγματα (Importance Sampling and Weighted Sample)

3.5.1 Ένα παράδειγμα

Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε την ποσότητα

$$\Theta=\int_x h(x)\pi(x)\,dx = E_\pi[h(X)]$$

όπου το στήριγμα της τυχαίας μεταβλητής X συμβολίζεται ως X και $h(x) \geq 0$. Σε συγκεκριμένες αριθμητικές μεθόδους, κάνουμε διακριτό το χώρο X με κανονικά πλέγματα. Υπολογίζουμε $h(x)\pi(x)$ για κάθε ένα από τα σημεία πλέγματος. Στη συνέχεια χρησιμοποιούμε το άθροισμα Riemann σαν μια προσέγγιση.

Έστω η συνάρτηση στόχος είναι :

$\mathrm{f(x,y)}=0{,}5e^{-90(x-0{,}5)^2-45(y+0{,}1)^4}+e^{-45(x+0{,}4)^2-60(y-0{,}5)^2}$

όπου (x,y) ∈ [-1,1] X [-1,1]. Περισσότερο από τα δύο τρίτα του χρόνου υπολογισμού σπαταλούνται για να υπολογίσουν αυτά τα σημεία πλέγματος στα οποία η συνάρτηση είναι σχεδόν μηδέν. Είναι εύκολο να φανταστούμε ότι η κατάσταση επιδεινώνεται πολύ γρήγορα καθώς η διάσταση του χώρου Xαυξάνεται.

Παίρνουμε m τυχαία δείγματα $(x^1, y^1), \ldots, (x^m, y^m)$ ομοιόμορφα στο [-1,1] X [-1,1] και υλοποιούμε έναν απλό αλγόριθμο (vanilla) Monte Carlo για να εκτιμήσουμε το ολοκλήρωμα $\mu=\iint f(x,y)dxdy$. Επειδή η πυκνότητα για την δειγματοληπτική κατανομή είναι μια σταθερά, ¼ , στην περιοχή, η εκτίμηση του ολοκληρώματος είναι:

$$\hat{\mu} = \frac{4}{m}\{f^1 + \cdots + f^m\} \qquad \text{όπου } f^i = f\left(x^i, y^i\right).$$

Όμως και με αυτόν τον αλγόριθμο σπαταλάται πολύς χρόνος για τον υπολογισμό των τυχαίων δειγμάτων που βρίσκονται στις περιοχές όπου η συνάρτηση είναι σχεδόν μηδέν .

3.5.2 Η βασική ιδέα

Η ιδέα της δειγματοληψίας σπουδαιότητας (Marshall 1956) είναι ότι πρέπει κάποιος να επικεντρωθεί στην περιοχή της «μεγαλύτερης σπουδαιότητας» (importance) έτσι ώστε να εξοικονομούμε υπολογιστικό χρόνο. Η ιδέα της μεροληψίας στις «σημαντικές» περιοχές του δειγματικού χώρου είναι απαραίτητη για τον υπολογισμό Monte Carlo με μοντέλα

μεγάλων διαστάσεων όπως αυτά που έχουμε στη Στατιστική Φυσική, στη Μοριακή Προσομοίωση και στη Μπεϋζιανή Στατιστική. Σε προβλήματα μεγάλων διαστάσεων η περιοχή στην οποία η συνάρτηση στόχος είναι ουσιαστικά διαφορετική από το μηδέν σε σχέση με το σύνολο του χώρου X είναι πολύ μικρή. Απλά συστήματα Monte Carlo (vanilla) (π.χ ομοιόμορφη δειγματοληψία από μια κανονική περιοχή) είναι σίγουρο ότι θα αποτύχουν σε αυτά τα προβλήματα.

Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε την ποσότητα :

$$\mu=E_{\pi}\{h(x)\}=\int h(x)\pi(x)dx$$

Η ακόλουθη διαδικασία είναι μια απλή μορφή του αλγόριθμου δειγματοληψίας σπουδαιότητας (Importance Sampling) :

B1 Προσομοιώνουμε $x^1,\dots.,x^m$ από μια δοκιμαστική κατανομή (trial distribution) g( )

B2 Υπολογίζουμε τα βάρη σπουδαιότητας (Importance Weighted)

$$w^j=\pi(x^j)/g(x^j) \quad \text{για j=1,...,m}$$

B3 Προσεγγίζουμε το μ από

$$\hat{\mu}=\frac{w^1h\,(x^1)+\dots.+w^mh\,(x^m)}{w_1+\dots+w_m}$$

Έτσι για να κάνουμε το σφάλμα εκτίμησης μικρό, πρέπει να διαλέξουμε την g(x) όσο το δυνατόν πιο κοντά στο $h(x)\pi(x)$.

Ένα σημαντικό πλεονέκτημα που μας κάνει να χρησιμοποιούμε την εκτίμηση

$$\hat{\mu}=\frac{w^1h\,(x^1)+\dots.+w^mh\,(x^m)}{w_1+\dots+w_m} \qquad (3.2)$$

αντί της αμερόληπτης εκτίμησης :

$$\tilde{\mu}=\frac{w^1h\,(x^1)+\dots.+w^mh\,(x^m)}{m} \qquad (3.3)$$

είναι γιατί χρειαζόμαστε για την (3.2) μόνο να γνωρίζουμε την αναλογία $\pi(x)/$g(x) σαν μια πολλαπλασιαστική σταθερά, ενώ στην αμερόληπτη εκτίμηση (3.3) πρέπει να γνωρίζουμε ακριβώς την αναλογία. Επίσης αν και η (3.2) εκτίμηση επάγει μια μικρή μεροληψία, όμως έχει συχνά μικρότερο μέσο τετραγωνικό σφάλμα από ότι η αμερόληπτη $\tilde{\mu}$ (3.3) .

΄Ενας άλλος λόγος που χρησιμοποιούμε τη δειγματοληψία σπουδαιότητας είναι όταν θέλουμε να πάρουμε ανεξάρτητα και ισόνομα (i.i.d) δείγματα

από π, αλλά αυτό είναι αδύνατον να γίνει απευθείας. Σ΄ αυτήν την περίπτωση θα μπορούσαμε να παράγουμε τυχαία δείγματα από διαφορετική, αλλά παρόμοια, τετριμμένη κατανομή g( ) και στη συνέχεια να διορθώσουμε την μεροληψία χρησιμοποιώντας την διαδικασία δειγματοληψίας σπουδαιότητας. Όπως και με τη μέθοδο απόρριψης έτσι και για την επιτυχή εφαρμογή της δειγματοληψίας σπουδαιότητας απαιτείται η δειγματοληπτική κατανομή g να είναι κοντά στο π. Το να πάρουμε μια καλή g μπορεί να είναι ένα μεγάλο εγχείρημα (και μερικές φορές αδύνατον) για προβλήματα μεγάλων διαστάσεων.

Κάνοντας εφαρμογή στο παράδειγμα 3.5.1 με συνάρτηση

$$f(x,y)=0{,}5e^{-90(x-0{,}5)^2-45(y+0{,}1)^4}+e^{-45(x+0{,}4)^2-60(y-0{,}5)^2}$$

παίρνουμε μια κατανομή g(x,y) η οποία είναι της μορφής :

$$g(x,y)=0{,}5e^{-90(x-0{,}5)^2-10(y+0{,}1)^4}+e^{-45(x+0{,}4)^2-60(y-0{,}5)^2}$$

όπου (x,y) ∈ [-1,1] X [-1,1]. Αυτή είναι ένα περικομμένο μείγμα της κανονικής κατανομής :

$$0{,}46N[\begin{pmatrix}0{,}5\\-0{,}1\end{pmatrix},\begin{pmatrix}\frac{1}{180}&0\\0&\frac{1}{20}\end{pmatrix}]+0{,}54\ N[\begin{pmatrix}-0{,}4\\0{,}5\end{pmatrix},\begin{pmatrix}\frac{1}{90}&0\\0&\frac{1}{120}\end{pmatrix}]$$

Μπορούμε να πάρουμε δείγμα από αυτή την κατανομή μείγμα με μια διαδικασία δύο σταδίων :

α) στρίβουμε ένα μεροληπτικό κέρμα (με πιθανότητα 0,46 να δείχνει κεφαλές),

β) αν μας δείξει κεφαλή παίρνουμε ένα τυχαίο διάνυσμα από την πρώτη κανονική κατανομή αλλιώς παίρνουμε από τη δεύτερη κανονική κατανομή. Στη συνέχεια η εκτίμηση γίνεται από τον μέσο όρο των λόγων w(x,y)= f(x,y)/ g(x,y) με w=0 όταν το (x,y) πέφτει έξω από την περιοχή *X.*

Για m=2500 η εκτίμηση του μ είναι 0,1259 με ένα τυπικό σφάλμα 0,0005 .

3.5.3 Ο « εμπειρικός κανόνας» για δειγματοληψία σπουδαιότητας

Με τη δειγματοληψία σπουδαιότητας κάνουμε εκτίμηση μ=$E_\pi\{h(x)\}$ από τα πρώτα παραγόμενα ανεξάρτητα δείγματα $x^1,\dots.,x^m$ από μια δοκιμαστική κατανομή g, από την οποία είναι εύκολο να πάρουμε δείγματα και στη συνέχεια διορθώνουμε τη μεροληψία με ενσωμάτωση του βάρους σπουδαιότητας $w\propto\pi(x^j)/g(x^j)$ στην εκτίμηση, χρησιμοποιώντας είτε

$$\hat{\mu} = \frac{w^1 h(x^1) + \cdots + w^m h(x^m)}{w_1 + \cdots + w_m} \quad (3.2) \quad \text{είτε}$$

$$\tilde{\mu} = \frac{w^1 h(x^1) + \cdots + w^m h(x^m)}{m} \quad (3.3) \ .$$

Με σωστή επιλογή της g μπορούμε να μειώσουμε σημαντικά την διακύμανση της εκτίμησης. Άρα η μέθοδος δειγματοληψίας σπουδαιότητας μπορεί να είναι πολύ αποδοτική. Η αποδοτικότητα όμως μιας τέτοιας μεθόδου είναι δύσκολο να μετρηθεί. Ένας χρήσιμος «εμπειρικός κανόνας» για να μετρήσουμε πόσο διαφορετική είναι η δοκιμαστική κατανομή από την κατανομή στόχο είναι να χρησιμοποιήσουμε το αποτελεσματικό μέγεθος δείγματος (effective sample size) (ESS) .

Υποθέτουμε ότι m ανεξάρτητα δείγματα προέρχονται από g(x) . Στη συνέχεια το ESS αυτής της μεθόδου ορίζεται ως :

$$\text{ESS(m)} = \frac{\text{m}}{1 + var_g[w(x)]} .$$

Σε πολλά προβλήματα η κατανομή στόχος π είναι γνωστή μόνο ως μια σταθερά κανονικοποίησης, για αυτό η διακύμανση του κανονικοποιημένου (normalized) βάρους πρέπει να εκτιμάται από τον *συντελεστή της διασποράς* του μη κανονικοποιημένου βάρους

$$cv^2(w) = \frac{\sum_{j=1}^{m} (w^j - \overline{w})^2}{(m-1)\overline{w}^2}$$

Όπου $\overline{w}$ είναι ο μέσος όρος του δείγματος του w^j.

3.5.4 Έννοια σταθμισμένων δειγμάτων

Ας υποθέσουμε ότι ενδιαφερόμαστε για Monte Carlo εκτίμηση του μ=$E_\pi h(x)$ για κάποια αυθαίρετη συνάρτηση $h(x)$. Η αρχή της δειγματοληψίας σπουδαιότητας (Importance Sampling) υποδηλώνει ότι το τυχαίο δείγμα $x^1, \ldots, x^m$ που χρησιμοποιείται για την εκτίμηση του μ δεν χρειάζεται να προέρχεται από π, αλλά μπορεί να προέρχεται από σχεδόν οποιαδήποτε κατανομή με τον όρο ότι ένα κατάλληλο σύνολο βαρών θα πρέπει να συνδέεται με το δείγμα και ότι τα βάρη δεν θα πρέπει να είναι πολύ ασύμμετρα (skewed).

Ορισμός : Ένα σύνολο σταθμισμένων τυχαίων δειγμάτων $\{(x^j , w^j)\}_{j=1}^m$ είναι κατάλληλο σε σχέση με την π, αν για κάθε τετραγωνικά ολοκληρώσιμη συνάρτηση $h()$, ισχύει:

$$\text{E}[\, h(x^j)\, w^j] = \text{c} E_\pi h(x) \ \text{για j=1,...,m}$$

όπου c είναι μια σταθερά κανονικοποίησης κοινή για όλα τα δείγματα m.

Με αυτό το σύνολο των σταθμισμένων δειγμάτων μπορούμε να εκτιμήσουμε το μ ως:

$$\hat{\mu} = \frac{1}{W} \sum_{j=1}^{m} h(x^j)\mathrm{w}(x^j) \qquad (3.4)$$

όπου $W = \sum_{j=1}^{m} w(x^j)$.

Για κάθε τετραγωνικά ολοκληρώσιμη συνάρτηση $h()$,

$$E_g\{h(\mathbf{x})\mathrm{w}\}/E_g(\mathrm{w}) = E_\pi\{h(x)\} \qquad (3.5)$$

Επίσης από αυτή την ισότητα συνεπάγεται ότι

$$\frac{E_g(\mathrm{w} \mid \mathbf{x})}{E_g(\mathrm{w})}\,\mathrm{g}(\mathbf{x}) = \pi(\mathbf{x}) \qquad (3.6)$$

όπου $\mathrm{g}(\mathbf{x})$ είναι η περιθώρια κατανομή του $\mathbf{x}$ της $\mathrm{g}(\mathbf{x}, \mathrm{w})$. Έτσι μια αναγκαία και ικανή συνθήκη για το $\mathbf{x}$ για να είναι σωστά σταθμισμένο από το w σε σχέση με το π είναι η (3.6).

Στο πλαίσιο της δειγματοληψίας σπουδαιότητας το βάρος σπουδαιότητας w είναι μια ντετερμινιστική συνάρτηση του αντίστοιχου δείγματος $\mathbf{x}$ [δηλαδή $\mathrm{w}(\mathbf{x}) = \pi(\mathbf{x})/\ \mathrm{g}(\mathbf{x})$]. Έτσι σε αυτή την περίπτωση η από κοινού κατανομή των $(\mathrm{w}, \mathbf{x})$ είναι εκφυλισμένη.

3.5.5 Περιθωριοποίηση στη δειγματοληψία σπουδαιότητας

Η περιθωριοποίηση στη δειγματοληψία σπουδαιότητας είναι μια τεχνική χρήσιμη για την μείωση της διασποράς του βάρους σπουδαιότητας.

Θεώρημα : Έστω $\mathrm{f}(\boldsymbol{z_1}, \boldsymbol{z_2})$ και $\mathrm{g}(\boldsymbol{z_1}, \boldsymbol{z_2})$ είναι δύο συναρτήσεις πιθανότητας όπου το στήριγμα της f είναι ένα υποσύνολο του στηρίγματος της g. Στη συνέχεια $var_g\{\frac{\mathrm{f}(\boldsymbol{z_1}, \boldsymbol{z_2})}{\mathrm{g}(\boldsymbol{z_1}, \boldsymbol{z_2})}\} \geq var_g\{\frac{\mathrm{f}_1(\boldsymbol{z_1})}{\mathrm{g}_1(\boldsymbol{z_1})}\}$

όπου $\mathrm{f}_1(\boldsymbol{z_1}) = \int \mathrm{f}(\boldsymbol{z_1}, \boldsymbol{z_2})\, \mathrm{d}\boldsymbol{z_2}$ και $\mathrm{g}_1(\boldsymbol{z_1}) = \int \mathrm{g}(\boldsymbol{z_1}, \boldsymbol{z_2})\, \mathrm{d}\boldsymbol{z_2}$ είναι οι περιθώριες κατανομές. Τις διασπορές τις παίρνουμε σε σχέση με τη g.

Απόδειξη:

$$\frac{\mathrm{f}_1(\boldsymbol{z_1})}{\mathrm{g}_1(\boldsymbol{z_1})} = \int \frac{\mathrm{f}(\boldsymbol{z_1}, \boldsymbol{z_2})}{\mathrm{g}_1(\boldsymbol{z_1})\mathrm{g}_{2\,|1}(\boldsymbol{z_2} \mid \boldsymbol{z_1})}\,\mathrm{g}_{2\,|1}(\boldsymbol{z_2} \mid \boldsymbol{z_1})\, \mathrm{d}\boldsymbol{z_2}$$

$$= E_g\{\frac{\mathrm{f}(\boldsymbol{Z_1}, \boldsymbol{Z_2})}{\mathrm{g}(\boldsymbol{Z_1}, \boldsymbol{Z_2})} \mid \boldsymbol{Z_1} = \boldsymbol{z_1}\}$$

Επομένως $var_g\{\frac{f(\boldsymbol{Z_1},\boldsymbol{Z_2})}{g(\boldsymbol{Z_1},\boldsymbol{Z_2})}\} \geq var_g\{ E_g\left[\frac{f(\boldsymbol{Z_1},\boldsymbol{Z_2})}{g(\boldsymbol{Z_1},\boldsymbol{Z_2})} \mid \boldsymbol{Z_1}\right]\} = var_g\{\frac{f_1(\boldsymbol{Z_1})}{g_1(\boldsymbol{Z_1})}\}$

Έτσι η μείωση της διασποράς εκφράζεται ως :

$$var_g\{\frac{f(\boldsymbol{Z_1},\boldsymbol{Z_2})}{g(\boldsymbol{Z_1},\boldsymbol{Z_2})}\} - var_g\left\{\frac{f_1(\boldsymbol{Z_1})}{g_1(\boldsymbol{Z_1})}\right\} = E_g\{ var_g\left[\frac{f(\boldsymbol{Z_1},\boldsymbol{Z_2})}{g(\boldsymbol{Z_1},\boldsymbol{Z_2})} \mid \boldsymbol{Z_1}\right]\}$$

Αυτό το θεώρημα χρησιμοποιήθηκε από τους MacEachern ,Clyde και Liu(1999) σε έναν καινούργιο αλγόριθμο δειγματοληψίας σπουδαιότητας για ένα απαραμετρικό πρόβλημα Μπεϋζιανής συμπερασματολογίας.

3.5.6 Παράδειγμα : Ένα Μπεϋζιανό πρόβλημα με ελλιπή δεδομένα

Στη στατιστική υπάρχει συχνά η περίπτωση να λείπει ένα μέρος των δεδομένων και αυτό κάνει δύσκολο τον υπολογισμό της συνάρτησης πιθανοφάνειας.

Το παρακάτω παράδειγμα κατασκευάστηκε από τον Murray(1977) και αφορά στη στατιστική συμπερασματολογία για τον πίνακα συνδιακύμανσης μιας διδιάστατης κανονικής κατανομής από ελλιπή δεδομένα.

Δίνεται ένας πίνακας με 12 παρατηρήσεις που υποθέτουμε ότι προέρχονται από μια διμεταβλητή κανονική κατανομή με γνωστό διάνυσμα μέσων (0,0) και άγνωστο πίνακα συνδιασποράς. Τα δεδομένα που χρησιμοποίησε ο Murray(1977) στο παράδειγμά του δίνονται στον παρακάτω πίνακα.

1	1	-1	-1	2	2	-2	-2	*	*	*	*
1	-1	1	-1	*	*	*	*	2	2	-2	-2

Στον πίνακα το σύμβολο * υποδεικνύει ότι η τιμή λείπει. Τα πλήρη στοιχεία είναι οι διδιάστατες παρατηρήσεις $y_1 \dots y_{12}$, όπου $y_t = (y_{t,1}, y_{t,2})$ για t=1,…,12. Στα $y_{t,2}$ λείπουν τα στοιχεία για t=5,…,8 ενώ στα $y_{t,1}$ λείπουν για t=9,..12.

Για ευκολία υποθέτουμε μια διδιάστατη κανονική prior για το Σ με μέσο 0 και πίνακα συνδιακύμανσης Σ, δηλαδή $\pi(\Sigma) \propto |\Sigma|^{-\frac{d+1}{2}}$ όπου d είναι η διάσταση και σε αυτό το παράδειγμα είναι d =2. Η εκ των υστέρων κατανομή του Σ με πλήρη δεδομένα είναι η inverse Wishart κατανομή

$$p(\Sigma \mid y_1, \dots y_t) \propto |\Sigma|^{-\frac{t+d+1}{2}} \; exp\left\{-\frac{1}{2} tr[\Sigma^{-1} \cdot S]\right\}$$

όπου S=$(s_{ij})_{2x2}$ είναι το μη διορθωμένο άθροισμα του τετραγωνικού πίνακα δηλαδή

$$(s_{ij} = \sum_{s=1}^{t} y_{s,i} y_{s,j}).$$

Θέτουμε Z=Σ^{-1} και βλέπουμε ότι Z ακολουθεί μια Wishart (t,S) κατανομή

$$p(\mathrm{Z} \mid complete\ data) \propto |Z|^{\frac{t-d-1}{2}} \; exp\left\{-\frac{1}{2} tr[Z \cdot S]\right\}$$

Σ' αυτή την κατανομή, n συχνά αναφέρεται ως ο βαθμός ελευθερίας και S είναι πίνακας κλίμακας. Κάνουμε δειγματοληψία από αυτή την Wishart κατανομή όταν t ≥ d+1 και αυτό μπορεί να γίνει ως εξής: Προσομοιώνουμε t ανεξάρτητα δείγματα $\varepsilon_1 \dots \varepsilon_t$ από μια d-διάστατη κανονική κατανομή, N$(0, S)$, και στη συνέχεια θέτουμε Z=$\sum_{i=1}^{t} \varepsilon_i \, \varepsilon_i^{\,T}$. Υποθέτουμε ότι $S = CC^T$.

Ένας αποτελεσματικός αλγόριθμος που προτείνεται από τους Odell και Feiveson (1966) είναι ο εξής:

B1: Προσομοίωση ανεξάρτητων τυχαίων δειγμάτων $V_j \sim X^2_{(t-j)}$ j=1,...,d

B2: Προσομοίωση ανεξάρτητων τυχαίων μεταβλητών $N_{ij} \sim N(0,1)$ i < j≤d

B3: Κατασκευή ενός συμμετρικού πίνακα B_{dxd} όπου $b_{11}=V_1$, $b_{1j}=N_{1j}\sqrt{V_1}$

$b_{jj}=V_j + \sum_{i=1}^{j-1} N_{ij}^2$, j=2,....,d

$b_{ij}=N_{ij}\sqrt{V_i} + \sum_{k=1}^{j-1} N_{ki} \, N_{kj}$, i < j≤d

B4: Z=CBC^T ακολουθεί την Wishart(t,S) κατανομή

$$p(\mathrm{Z} \mid complete\ data) \propto |Z|^{\frac{t-d-1}{2}} \; exp\left\{-\frac{1}{2} tr[Z \cdot S]\right\}$$

Τώρα επιστρέφουμε στο πάραδειγμα με τα ελλιπή δεδομένα όπου ψάχνουμε να βρούμε την από κοινού posterior κατανομή του Σ και τα ελλιπή δεδομένα $y_{mis}=(y_{5,2}, \dots., y_{8,2}, y_{9,1}, \dots, y_{12,1})$ ως :

$$p(\Sigma, \boldsymbol{y_{mis}} \mid \boldsymbol{y_{obs}}) \propto p(\boldsymbol{y_{mis}}, \boldsymbol{y_{obs}} \mid \Sigma) \; p(\Sigma)$$

$$\propto |\Sigma|^{-\frac{12+3}{2}} \; exp\left\{-\frac{1}{2} tr[\Sigma^{-1} S(\boldsymbol{y_{mis}})]\right\}$$

όπου $S(\boldsymbol{y_{mis}})$ μας δείχνει ότι κάθε τιμή εξαρτάται από την τιμή του $\boldsymbol{y_{mis}}$.

Για να μπορούμε να υπολογίσουμε αυτή την posterior κατανομή εφαρμόζουμε αλγόριθμο δειγματοληψίας σπουδαιότητας (importance sampling) ως εξής :

● Προσομοιώνουμε πίνακα Σ από μια κατανομή $g_o(\Sigma)$

● Προσομοιώνουμε τιμή y_{mis} από την $p(\boldsymbol{y_{mis}} \mid \boldsymbol{y_{obs}}, \Sigma)$

Η κατανομή πρόβλεψης του $\boldsymbol{y_{mis}}$, είναι γινόμενο από κανονικές κατανομές.

Για παράδειγμα $[y_{5,2} \mid \Sigma, \boldsymbol{y_{obs}}] = [y_{5,2} \mid \Sigma, y_{5,1}] \sim N(\mu_*, \sigma_*^2)$

όπου $\mu_* = \rho y_{5,1}\sqrt{\sigma_{11}/\sigma_{22}}$ και $\sigma_*^2 = (1-\rho^2)\,\sigma_{11}$

Το βασικό ερώτημα όμως είναι πως θα προσομοιώσουμε τον Σ. Μια ιδέα είναι να πάρουμε ως δοκιμαστική κατανομή την δεσμευμένη posterior κατανομή από τις 4 πρώτες πλήρεις παρατηρήσεις:

$$g_o(\Sigma) \propto |\Sigma|^{-\frac{7}{2}}\, exp\left\{-\frac{1}{2}tr[\Sigma^{-1}S_4]\right\}$$

Όπου S_4 αναφέρεται στο άθροισμα του τετραγωνικού πίνακα που υπολογίζεται από τις 4 πρώτες παρατηρήσεις. Τώρα μπορούμε να προσομοιώσουμε τον Σ από την $g_o(\Sigma)$ να βρούμε τα y_{mis} όπως είπαμε παραπάνω και τέλος να βρούμε τα βάρη.

Στη συνέχεια παρουσιάζουμε αποτελέσματα από 2 σετ δεδομένων για το παραπάνω πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης διδιάστατης κανονικής κατανομής από ελλιπή δεδομένα.

(1) Αποτελέσματα για δεδομένα Murray

Χρησιμοποιούμε τα δεδομένα του Murray (1977) για n=12 παρατηρήσεις με 8 ελλιπή δεδομένα από μια διδιάστατη κανονική κατανομή με μέσο $\begin{pmatrix}0\\0\end{pmatrix}$ και άγνωστο πίνακα συνδιακύμανσης.

Με την μέθοδο της δειγματοληψίας σπουδαιότητας προσομοιώνουμε τα ελλιπή δεδομένα m=10000 φορές και παίρνουμε:

● τον εκτιμώμενο πίνακα συνδιακύμανσης

$$\hat{\Sigma} = \begin{bmatrix} 2.5764 & 0.0036 \\ 0.0036 & 2.5504 \end{bmatrix}$$

σύμφωνα με τον τύπο $\hat{\Sigma} = \frac{w_1\Sigma^1 + \cdots + w_m\Sigma^m}{w_1 + \cdots + w_m}$

● τον εκτιμώμενο συντελεστή συσχέτισης $\hat{r} = 0.0014$ και

● τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι s_1= 0.4634 και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι s_2=0.4610.

2) Προσομοιωμένα δεδομένα για n=100 παρατηρήσεις με 30 ελλιπή δεδομένα από μια διδιάστατη κανονική κατανομή με μέσο $\begin{pmatrix}0\\0\end{pmatrix}$, πίνακα συνδιακύμανσης $\begin{pmatrix}1 & 0.9899\\0.9899 & 2\end{pmatrix}$ και συντελεστή συσχέτισης r=0,7. **Με την μέθοδο της δειγματοληψίας σπουδαιότητας** προσομοιώνουμε τον πίνακα συνδιακύμανσης m=10000 φορές και παίρνουμε:

● τον εκτιμώμενο πίνακα συνδιακύμανσης

$$\hat{\Sigma}=\begin{bmatrix}1.3119 & 1.3483\\1.3483 & 2.5264\end{bmatrix}$$

σύμφωνα με τον τύπο $\hat{\Sigma}=\frac{w_1\Sigma^1+\cdots.+w_m\Sigma^m}{w_1+\cdots+w_m}$

● τον εκτιμώμενο συντελεστή συσχέτισης $\hat{r}$=0.7406 και

● τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι s_1= 0.1145 και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι s_2=0.1589.

Το πρόγραμμα για την εξαγωγή των παραπάνω αποτελεσμάτων παρατίθεται σε παράρτημα.

3.6 Προηγμένες τεχνικές της δειγματοληψίας σπουδαιότητας

3.6.1 Προσαρμοσμένη (adaptive) δειγματοληψία σπουδαιότητας

Ένας τρόπος για να βελτιώσουμε την δοκιμαστική κατανομή στη δειγματοληψία σπουδαιότητας είναι να κάνουμε τα εξής βήματα :

● Παίρνουμε μια δοκιμαστική συνάρτηση πιθανότητας $g_o(\mathbf{x})$, που είναι t-κατανομή με α βαθμούς ελευθερίας, μέσο μ_0 και διασπορά Σ_0, δηλαδή $g_o(\mathbf{x})= t_a(x,\mu_0,\Sigma_0)$.

● Παίρνουμε σταθμισμένα δείγματα Monte Carlo με τα οποία μπορούμε να εκτιμήσουμε το μέσο μ_1 και τον πίνακα συνδιακυμάσεων Σ_1 της κατανομής στόχου.

● Παίρνουμε μια καινούργια δοκιμαστική συνάρτηση πιθανότητας $g_1(\mathbf{x})$ που είναι t- κατανομή με α βαθμούς ελευθερίας, μέσο μ_1 και διασπορά Σ_1 , δηλαδή την $g_1(\mathbf{x}) = t_a(x, \mu_1, \Sigma_1)$.

Αυτή η διαδικασία επαναλαμβάνεται μέχρι μια ορισμένη διαφορά μεταξύ της δοκιμαστικής κατανομής και της κατανομής στόχου, έτσι ώστε ο συντελεστής διασποράς των βαρών σπουδαιότητας (important weights), να μην βελτιώνεται άλλο.

Ένας άλλος τρόπος για να κάνουμε μια προσαρμογή είναι να πάρουμε μια παραμετρική μορφή για την νέα δοκιμαστική συνάρτηση πιθανότητας $g_1(\mathbf{x})$. Δηλαδή ας υποθέσουμε ότι είναι της μορφής $g(\mathbf{x};\lambda)$. Επιλέγουμε το βέλτιστο λ, δηλαδή αυτό που ελαχιστοποιεί τον εκτιμώμενο συντελεστή διασποράς των βαρών σπουδαιότητας με βάση το τρέχον δείγμα.

Έστω $w(\mathbf{x};\lambda) = \pi(\mathbf{x}) / g_1(\mathbf{x})$

η διασπορά $var_{g_1}(w) = \int \frac{\pi^2(\mathbf{x})}{g_1(\mathbf{x}) g_0(\mathbf{x})} g_0(\mathrm{x})\, \mathrm{d}\mathbf{x} - 1$

Έστω ότι έχουμε $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$ από την δοκιμαστική κατανομή $g_0(\boldsymbol{x})$.

Ο συντελεστής διασποράς του $\pi(\mathbf{x}) / g_1(\mathbf{x})$, δηλαδή των βαρών σπουδαιότητας, μπορεί να προσεγγιστεί ως :

$$\widehat{cv}^2(\lambda) = \bar{H}(\lambda) - 1$$

$$\text{όπου} \quad \bar{H}(\lambda) = \sum_{j=1}^{m} H^{(j)}(\lambda)/m$$

$$\text{και} \quad H^{(j)}(\lambda) = \frac{\pi^2(\mathbf{x}^j)}{g_1(\mathbf{x}^j) g_0(\mathbf{x}^j)} = \frac{\{\pi(\mathbf{x}^j)/g_0(\mathbf{x}^j)\}^2}{g(\mathbf{x}^j,\lambda)/g_0(\mathbf{x}^j)}$$

Εφαρμόζουμε την μέθοδο Newton Raphson για να βρούμε το βέλτιστο λ .

Για $\lambda = (\varepsilon, \mu.\Sigma)$ παίρνουμε την $g(\mathbf{x};\lambda) = \varepsilon g_o(\mathbf{x}) + (1-\varepsilon)\, t_v(\mathbf{x}; \mu, \Sigma)$

που είναι μια βελτιωμένη εκδοχή της μίξης της προηγούμενης δοκιμαστικής κατανομής g_0, με ένα νέο παραμετρικό στοιχείο. Θα πρέπει όμως να είμαστε προσεκτικοί όταν χρησιμοποιούμε αυτές τις προσαρμοστικές μεθόδους γιατί είναι ασταθείς.

3.6.2 Απόρριψη και στάθμιση

Κατά την εφαρμογή της μεθόδου απόρριψης πρέπει να βρούμε μια δοκιμαστική κατανομή $g()$ και μια σταθερή τιμή M, έτσι ώστε να ισχύει $\pi(\mathbf{x}) \leq M\, g(\mathbf{x})$ για όλα τα **x.** Είναι πολύ σημαντικό να βρούμε ένα καλό M. Όμως και αν υποθέσουμε ότι χρησιμοποιήσουμε ένα καλό M δεν είναι σίγουρο ότι η παραπάνω ανισότητα $\pi(\mathbf{x}) \leq M\, g(\mathbf{x})$ θα ισχύει για όλα τα **x**. Τότε μπορούμε να δεχτούμε ότι για αυτά τα **x** ισχύει $\pi(\mathbf{x}) > M\, g(\mathbf{x})$ και να ρυθμίσουμε την μεροληψία δίνοντας σε αυτά τα δείγματα κατάλληλα βάρη. Έτσι μπορούμε να επιτύχουμε πιο γρήγορο υπολογισμό και καλύτερη αποδοτικότητα.

Με τη δειγματοληψία σπουδαιότητας παράγουμε τυχαία δείγματα με πολύ μικρά βάρη σπουδαιότητας .

Αν υποθέσουμε ότι θέλουμε να εκτιμήσουμε την $E_\pi[h(\mathbf{x})]$, με όσο το δυνατόν λιγότερα δείγματα, αλλά χωρίς να χάσουμε πληροφορίες και να μην δημιουργήσουμε μεροληψία, χρησιμοποιούμε την παρακάτω τεχνική που είναι ένας συνδυασμός της απόρριψης και της στάθμισης σπουδαιότητας.

Ας υποθέσουμε ότι έχουμε δείγματα $x^1 \ldots x^m$ από $g(\mathbf{x})$ και έστω $w^j = \pi(\mathbf{x}^j)/g(\mathbf{x}^j)$ τότε τα βήματα που ακολουθούμε για κάθε δεδομένη τιμή $c > 0$ είναι τα εξής:

Έλεγχος απόρριψης [Rejection Control (RC)]

- Για j =1,…,m δεχόμαστε $\mathbf{x}^j$ με πιθανότητα

$$r^j = min\left\{1, \frac{w^j}{c}\right\}$$

- Αν το j-οστό δείγμα $\mathbf{x}^j$ είναι αποδεκτό, το βάρος του γίνεται:

$$w*^j = q_c w^j / r^j \qquad \text{όπου} \quad q_c = \int min\left\{1, \frac{w(\mathbf{x})}{c}\right\} g(\mathbf{x}) d\mathbf{x}$$

Ο αλγόριθμος RC μπορεί να θεωρηθεί σαν μια τεχνική για την προσαρμογή της δοκιμαστικής κατανομής g με την βοήθεια των βαρών σπουδαιότητας. Η καινούργια δοκιμαστική κατανομή $g^*(\mathbf{x})$ που προκύπτει από την προσαρμογή αναμένεται να είναι πιο κοντά στην κατανομή στόχο π(x).

Έτσι έχουμε $g^*(\mathbf{x}) = q_c^{-1}\, min\left\{g(\mathbf{x}), \frac{\pi(\mathbf{x})}{c}\right\}$.

Η σταθερά κανονικοποίησης q_c

$$q_c = \int min\left\{g(\mathbf{x}), \frac{\pi(\mathbf{x})}{c}\right\} d\mathbf{x} = E_g\, min\left\{1, \frac{w(\mathbf{x})}{c}\right\}$$

μπορεί αμερόληπτα να εκτιμηθεί από το δείγμα ως

$$\hat{p}_c = \frac{1}{m}\sum_{j=1}^{m} min\left\{1, \frac{w^j}{c}\right\}.$$

Με την εφαρμογή του ελέγχου απόρριψης συνήθως έχουμε λιγότερα από N δείγματα. Περισσότερα δείγματα μπορούν να εξαχθούν είτε από $g(x)$ είτε από $g^*(x)$.

Θεώρημα: Η μέθοδος ελέγχου απόρριψης πράγματι μειώνει την x^2 απόσταση μεταξύ της κατανομής στόχου και της προσαρμοσμένης δοκιμαστικής κατανομής και είναι:

$$Var\, g^*\left[\frac{\pi(\mathbf{x})}{g^*(\mathbf{x})}\right] \leq Var\; g\left[\frac{\pi(\mathbf{x})}{g(\mathbf{x})}\right]$$

Σύμφωνα με το θεώρημα φαίνεται ότι η $g^*(x)$ είναι καλύτερη από την $g(x)$.

3.6.3 Διαδοχική (Sequential) δειγματοληψία σπουδαιότητας

Είναι δύσκολο να πάρουμε μια καλή δοκιμαστική κατανομή για να κάνουμε δειγματοληψία σπουδαιότητας σε προβλήματα μεγάλων διαστάσεων. Μια από τις πιο χρήσιμες στρατηγικές σ' αυτά τα προβλήματα είναι η δημιουργία δοκιμαστικής κατανομής διαδοχικά. Ας υποθέσουμε ότι το $\mathbf{x}$ μπορούμε να το αναλύσουμε ως $\mathbf{x} = (x_1, \dots, x_d)$ που καθένα από τα x_j μπορεί να είναι πολυδιάστατο. Στη συνέχεια η δοκιμαστική κατανομή μπορεί να κατασκευαστεί ως εξής :

$$g(\mathbf{x}) = g_1(x_1)g_2(x_2 \mid x_1)\dots g_d(x_d \mid x_1, \dots, x_{d-1})$$

και η κατανομή στόχος γράφεται ως:

$$\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \mid x_1)\dots\pi(x_d \mid x_1, \dots, x_{d-1}) \qquad (3.7)$$

υπολογίζουμε το βάρος σπουδαιότητας ως:

$$\mathrm{w}(\mathbf{x}) = \frac{\pi(\mathrm{x})}{\mathrm{g(x)}} = \frac{\pi(x_1)\pi(x_2 \mid x_1)\dots\pi(x_d \mid x_1,\dots,x_{d-1})}{g_1(x_1)g_2(x_2 \mid x_1)\dots g_d(x_d \mid x_1,\dots,x_{d-1})} \qquad (3.8)$$

Αυτή η εξίσωση υποδηλώνει έναν αναδρομικό τρόπο υπολογισμού και παρακολούθησης του βάρους σπουδαιότητας.

Δηλαδή θέτοντας $\mathbf{x_t} = (x_1, \dots, x_d)$ (έτσι $\mathbf{x_d} \equiv \mathbf{x}$)

έχουμε $w_t(\mathbf{x_t}) = w_{t-1}(\mathbf{x_{t-1}}) \frac{\pi(x_t \mid \mathbf{x_{t-1}})}{g_t(x_t \mid \mathbf{x_{t-1}})}$ για t = 1, … , d

Στο τέλος w_d είναι ίσο με $w(\mathbf{x})$ στην (3.8).

Πιο αναλυτικά:

● Προσομοιώνουμε x_1 από $g_1(x_1)$ και το βάρος θα είναι

$$w_1 = \frac{\pi(x_1)}{g_1(x_1)}$$

● Προσομοιώνουμε x_2 από $g_2(x_2 \mid x_1)$

$w_2 = \frac{\pi(x_1)\pi(x_2 \mid x_1)}{g_1(x_1)g_2(x_2 \mid x_1)} = w_1 \frac{\pi(x_2 \mid x_1)}{g_2(x_2 \mid x_1)}$

….

● $w_d = \frac{\pi(x_1)\pi(x_2 \mid x_1)\dots\pi(x_d \mid x_1,\dots,x_{d-1})}{g_1(x_1)g_2(x_2 \mid x_1)\dots g_d(x_d \mid x_1,\dots,x_{d-1})} = \frac{\pi(x)}{g(x)}$

Από τη σχέση $\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \mid x_1)\dots\pi(x_d \mid x_1, \dots, x_{d-1})$ (3.7) και την παραπάνω αναδρομή προκύπτουν τα ακόλουθα :

α) Αν το μερικό βάρος που παίρνουμε από το παραγόμενο διαδοχικό μερικό δείγμα είναι πολύ μικρό τότε μπορούμε να σταματήσουμε να παίρνουμε περισσότερα στοιχεία του **x.**

β) Μπορούμε να επωφεληθούμε από την $\pi(x_t \mid \mathbf{x_{t-1}})$ στο σχεδιασμό της $g_t(x_t \mid \mathbf{x_{t-1}})$. Δηλαδή η περιθώρια κατανομή $\pi(\mathbf{x_t})$ μπορεί να χρησιμοποιηθεί για να μας καθοδηγήσει στην παραγωγή του **x .**

Η παραπάνω ιδέα ακούγεται ενδιαφέρουσα, αλλά το πρόβλημα είναι ότι, τόσο αυτή η ανάλυση της κατανομής π όπως είναι στην σχέση (3.7), όσο και εκείνη του w όπως είναι στην σχέση (3.8), δεν είναι πρακτικές. Ο λόγος είναι ότι προκειμένου να πάρουμε την σχέση (3.7)πρέπει να έχουμε την περιθώρια κατανομή

$$\pi(\mathbf{x_t}) = \int \pi(x_1, \dots, x_d) dx_{t+1} \dots dx_d$$

και επειδή είναι πολύ δύσκολο να γίνει η ολοκλήρωση, υποθέτουμε ότι μπορούμε να πάρουμε μια ακολουθία από « βοηθητικές (auxiliary) κατανομές» $\pi_1(\mathbf{x_1}), \pi_2(\mathbf{x_2}), \dots, \pi_d(\mathbf{x})$ έτσι ώστε η $\pi_t(\mathbf{x_t})$ να είναι μια καλή προσέγγιση της περιθώριας κατανομής $\pi(\mathbf{x_t})$ για t=1,…,d-1 και $\pi_d = \pi$.

Η διαδοχική δειγματοληψία σπουδαιότητας (SIS) είναι μια μέθοδος που μπορεί να ορισθεί επαγωγικά για t=2,…,d ως εξής:

(SIS) Βήμα :

(A) Προσομοιώνουμε $\mathrm{X_t}=\mathrm{x_t}$ από $g_t(x_t \mid x_{t-1})$ και θέτουμε $\mathrm{x_t}=(x_{t-1}, \mathrm{x_t})$

(B) Υπολογίζουμε $u_\mathrm{t}=\frac{\pi_t(\mathbf{x_t})}{\pi_{t-1}(\mathbf{x_{t-1}})g_t(x_t \mid x_{t-1})}$ και θέτουμε $w_\mathrm{t} = w_{\mathrm{t}-1}u_\mathrm{t}$

Όπου u_t είναι το «αυξητικό βάρος»(incremental weight). Το $\mathrm{x_t}$ θα είναι σωστά σταθμισμένο από w_t σε σχέση με το π_t υπό τον όρο ότι x_{t-1} θα είναι σωστά σταθμισμένο από $w_{\mathrm{t}-1}$ σε σχέση με το π_{t-1}. Έτσι το σύνολο του δείγματος $\mathbf{x}$ που πήραμε σταθμίζεται καλά από το τελικό βάρος σπουδαιότητας w_t σε σχέση με την κατανομή στόχο $\pi(\mathbf{x})$. Ένας λόγος που χρησιμοποιούμε την διαδοχική μέθοδο είναι ότι χωρίζει ένα δύσκολο έργο σε διαχειρίσιμα μέρη. Η SIS μέθοδος είναι ιδιαίτερα χρήσιμη στην κατασκευή μιας πιο αποτελεσματικής δοκιμαστικής κατανομής με τη χρήση της ακολουθίας των «βοηθητικών κατανομών» $\pi_1, \pi_2, \dots, \pi_d$.

● Μπορούμε να κατασκευάσουμε δοκιμαστική κατανομή g_t σε σχέση με την κατανομή στόχο π_t.

Για παράδειγμα μπορούμε να επιλέξουμε, αν είναι δυνατόν,

$$g_t(\mathbf{x_t} \mid \mathbf{x_{t-1}}) = \pi_t(\mathbf{x_t} \mid \mathbf{x_{t-1}})$$

Σε αυτή την περίπτωση το αυξητικό βάρος γίνεται

$$u_\mathrm{t}=\frac{\pi_t(\mathbf{x_t})}{\pi_{t-1}(\mathbf{x_{t-1}})}$$

● Όταν παρατηρήσουμε ότι $\mathrm{w_t}$ γίνεται πολύ μικρό, μπορούμε να επιλέξουμε να απορρίψουμε το δείγμα στα μισά της διαδρομής και να επαναλάβουμε την διαδικασία. Σε αυτή την περίπτωση αποφεύγουμε σπατάλη χρόνου στην παραγωγή δειγμάτων που έχουν μικρή επίδραση στην τελική εκτίμηση. Έτσι, καθώς η οριστική απόρριψη δημιουργεί μεροληψίες, η τεχνική του ελέγχου απόρριψης μπορεί να χρησιμοποιηθεί για να διορθώσει αυτές τις μεροληψίες.

Πιο αναλυτικά ο αλγόριθμος SIS με βοηθητικές κατανομές

● Προσομοιώνουμε x_1 από $g_1(x_1)$

$$w_1 = \frac{\pi_1(x_1)}{g_1(x_1)}$$

● Προσομοιώνουμε x_2 από $g_2(x_2 \mid x_1)$

$$w_2 = \frac{\pi_1(x_1)\pi_2(x_1, x_2)}{\pi_1(x_1)g_1(x_1)g_2(x_2 \mid x_1)} = w_1 \frac{\pi_2(x_1, x_2)}{\pi_1(x_1)g_2(x_2 \mid x_1)}$$

…..

$$w_d = w_{d-1} \frac{\pi_d(x_1,\ldots,x_d)}{\pi_{d-1}(x_1,\ldots,x_{d-1})\, g_d(x_d \mid x_1,\ldots,x_{d-1})} =$$

$$= \frac{\pi_d(x_1,\ldots,x_d)}{g_1(x_1) g_2(x_2 \mid x_1) \ldots g_d(x_d \mid x_{d-1})} = \frac{\pi(x)}{g(x)}$$

Επιλέγουμε $g_t(x_t \mid x_{t-1}) = \pi_t(x_t \mid x_{t-1})$

- Προσομοιώνουμε x_1 από $g_1(x_1) = \pi_1(x_1)$

$$w_1 = 1$$

- Προσομοιώνουμε x_2 από $\pi_2(x_2 \mid x_1)$

$$w_2 = \frac{\pi_2(x_1, x_2)}{\pi_1(x_1)\pi_2(x_2 \mid x_1)}$$

$$w_3 = \frac{\pi_2(x_1, x_2)}{\pi_1(x_1)\pi_2(x_2 \mid x_1)} \cdot \frac{\pi_3(x_1, x_2, x_3)}{\pi_2(x_1, x_2)\pi_3(x_3 \mid x_1, x_2)}$$

$$w_d = \frac{\pi_d(x_1,\ldots,x_d)}{\pi_1(x_1)\pi_2(x_2 \mid x_1) \ldots \pi_d(x_d \mid x_1,\ldots,x_{d-1})}$$

3.6.4 Έλεγχος απόρριψης στη διαδοχική δειγματοληψία σπουδαιότητας (rejection control in sequential importance sampling)

Για να βελτιώσουμε την SIS μέθοδο, που παρουσιάζει το πρόβλημα της μεροληψίας, χρησιμοποιούμε αυτή την τεχνική του ελέγχου απόρριψης.

Ας υποθέσουμε ότι έχουμε μια σειρά από σημεία ελέγχου $0 < t_1 < t_2 < t_3 < \ldots < t_k \leq d$ και μια ακολουθία από τιμές κατωφλίου $c_1, \ldots, c_k$ που δίνονται εκ των προτέρων. Μπορούμε να εφαρμόσουμε την ακόλουθη διαδικασία:

- Σε κάθε σημείο ελέγχου t_j αρχίζει η $RC(t_K)$ όπως περιγράψαμε στην ενότητα 3.6.2 με τιμή (κατωφλίου) $c = c_j$. Αν το μερικό δείγμα $(x_1, \ldots, x_{t_j})$ έχει ένα βάρος w_{t_j}, τότε δεχόμαστε αυτό το μερικό βάρος με πιθανότητα

$min\left\{1, \frac{w_{t_j}}{\mathrm{c_j}}\right\}$ και αν γίνει αποδεκτό, αντικαθιστούμε το βάρος του από $w^*_{t_j}=\max\left\{w_{t_j}, \mathrm{c_j}\right\}$.

 Για κάθε επιμέρους δείγμα απόρριψης, κάνουμε επανεκκίνηση ξανά από την αρχή και αφήνουμε το δείγμα να περάσει από όλα τα σημεία ελέγχου $t_1, \ldots . t_j$, με τιμές κατωφλίου $c_1, \ldots, c_j$ αντίστοιχα. Αν απορριφθεί κάποιο ενδιάμεσο σημείο ελέγχου, αρχίζουμε ξανά.

Σημειώνεται ότι, μετά τον πρώτο έλεγχο απόρριψης στο στάδιο t_1, η δειγματοληπτική κατανομή $g_t^*(\mathbf{x_t})$ για $\mathbf{X_t}$, δεν είναι πλέον η ίδια με $g(\mathbf{x}) = g_1(x_1)g_2(x_2 \mid x_1)\ldots g_d(x_d \mid x_1, \ldots, x_{d-1})$ που αναφέρεται στο 3.6.3. Οι Liu, Chen και Wong (1998) έδειξαν ότι για κάθε χρόνο t, το μερικό δείγμα $\mathbf{x_t}$ που προκύπτει από την παραπάνω διαδικασία είναι σωστά σταθμισμένο, σε σχέση με την κατανομή στόχο π_t, από τα τροποποιημένα βάρη τους w^*_t. Για να διατηρήσουμε μια σωστή εκτίμηση της σταθεράς κανονικοποίησης για την κατανομή π, θα πρέπει να εκτιμήσουμε p_c, την πιθανότητα αποδοχής και να ρυθμίσουμε το βάρος $p_c w^*_t$. Η παραπάνω μέθοδος απαιτεί κάθε δείγμα που απορρίπτουμε να ξαναρχίζει από το στάδιο 0, επομένως αυτή η μέθοδος τείνει να γίνει ανέφικτη όταν ο αριθμός των στοιχείων d είναι μεγάλος.

Κεφάλαιο 4

Ακολουθιακές μέθοδοι Monte Carlo

Στο κεφάλαιο 3 περιγράψαμε το βασικό πλαίσιο της διαδοχικής δειγματοληψίας σπουδαιότητας ως εξής : Δημιουργούμε δοκιμαστική κατανομή διαδοχικά και υπολογίζουμε τα βάρη σπουδαιότητας αναδρομικά. Πιο συγκεκριμένα:

Αναλύουμε πρώτα την τυχαία μεταβλητή $\mathbf{x}$ σε ένα αριθμό συνιστωσών $\mathbf{x} = (x_1, \dots, x_d)$ και στη συνέχεια δημιουργούμε την δοκιμαστική κατανομή

$$g(\mathbf{x}) = g_1(x_1)g_2(x_2 \mid x_1)\dots g_d(x_d \mid x_1, \dots, x_{d-1}) \qquad (4.1)$$

όπου οι δεσμευμένες κατανομές g_t (δηλαδή $g_1(x_1)$, $g_2(x_2 \mid x_1)$ κλπ) επιλέγονται έτσι ώστε η από κοινού δειγματοληπτική κατανομή του $\mathbf{x}$ (δηλαδή $g(\mathbf{x})$) να είναι όσο το δυνατόν πιο κοντά στην κατανομή στόχο $\pi(\mathbf{x})$.

Το βάρος σπουδαιότητας του δείγματος που προσομοιώνεται από την $g(\mathbf{x})$ μπορεί να εκτιμηθεί, είτε στο τέλος, όταν όλα τα μέρη του $\mathbf{x}$ είναι τοποθετημένα, είτε αναδρομικά ως :

$$w_t = w_{t-1} \frac{\pi_t(x_1,\dots,x_t)}{\pi_{t-1}(x_1,\dots,x_{t-1})\, g_t(x_t \mid x_1,\dots,x_{t-1})} \qquad \text{για } t=1,\dots,d$$

όπου $\pi_1, \dots, \pi_d$ είναι μια ακολουθία βοηθητικών κατανομών. Μια απαραίτητη προυπόθεση για αυτές τις βοηθητικές κατανομές είναι να ισχύει $\pi_d(x_1, \dots, x_d) = \pi(x_1, \dots, x_d)$. Η βοηθητική κατανομή π_t μπορεί να θεωρηθεί σαν μια προσέγγιση της περιθώριας κατανομής στόχου.

$\pi(x_1, \dots, x_t) = \int \pi(\mathbf{x})\, dx_{t+1} \dots dx_d$ του μερικού δείγματος $\mathbf{x}_t = (x_1, \dots, x_t)$

Αν έχουμε μια καλή ακολουθία από βοηθητικές κατανομές (δηλαδή αυτές οι π_t οι οποίες προσεγγίζουν τις περιθώριες κατανομές της κατανομής στόχου π), θα μπορούμε να κρίνουμε για το αν ένα μερικό δείγμα που παράγεται είναι «καλό». Έτσι μπορούμε να εξετάσουμε αν θα πρέπει να σταματήσουμε τη διαδικασία προσομοίωσης πριν τελειώσει η δημιουργία όλου του δείγματος **x.** Για παράδειγμα αν δούμε $w_2 = 0$ για τα δύο πρώτα στοιχεία που δημιουργούνται τότε, θα πρέπει να σταματήσουμε να προσομοιώνουμε περαιτέρω. Έτσι ένα πολύ μικρό βάρος σπουδαιότητας w_κ στο στάδιο κ , θα μπορούσε να μας οδηγήσει πιθανόν στο να σταματήσουμε νωρίς να προσομοιώνουμε. Όταν σταματήσουμε νωρίς, μειώνεται ο συνολικός αριθμός των δειγμάτων που παράγονται τελικά. Για

να πάρουμε τα δείγματα που χάσαμε επαναλαμβάνουμε την ίδια διαδικασία δειγματοληψίας από την αρχή.

Σε αυτό το κεφάλαιο θα δείξουμε ότι μια πολύ καλή στρατηγική, για να πάρουμε εκείνα τα δείγματα που χάνονται λόγω της πρόωρης παύσης, είναι αυτή που γίνεται με επαναδειγματοληψία από τα τρέχοντα διαθέσιμα «καλά» μερικά δείγματα.

Μια καλή ακολουθία βοηθητικών κατανομών $\{\pi_t\}$ μπορεί επίσης να μας βοηθήσει να διαλέξουμε καλές δοκιμαστικές κατανομές. Μια καλή επιλογή της δοκιμαστικής κατανομής g_t θα είναι $g_t(x_t \mid x_1, \dots, x_{t-1}) = \pi_t(x_t \mid x_1, \dots, x_{t-1})$.

Η SIS μέθοδος έχει εφαρμοστεί σε τουλάχιστον τρεις βασικούς τομείς έρευνας. Η πρώτη εφαρμογή χρονολογείται από το 1950 και είχε ως κίνητρο τα προβλήματα προσομοίωσης μακρομορίων (Hammersley και Morton 1954, Rosenbluth και Rosenbluth 1955). Οι άλλες δύο περιπτώσεις είναι πιο πρόσφατες. Η μια υποκινήθηκε από στατιστικά προβλήματα ελλιπών δεδομένων (Kong et al 1994, Liu και Chen 1995) και η άλλη από μη γραμμικά προβλήματα φιλτραρίσματος σε ραντάρ εντοπισμού (Gordon et al 1993). Στο υπόλοιπο μέρος αυτού του κεφαλαίου θα ασχοληθούμε με αυτές τις ανεξάρτητες εφαρμογές, θα συνοψίσουμε ειδικές τεχνικές σε καθένα από τους τομείς και θα περιγράψουμε ένα γενικό πλαίσιο της διαδοχικής Monte Carlo που μπορεί να χρησιμοποιηθεί σαν βάση για καινούργιες τεχνικές.

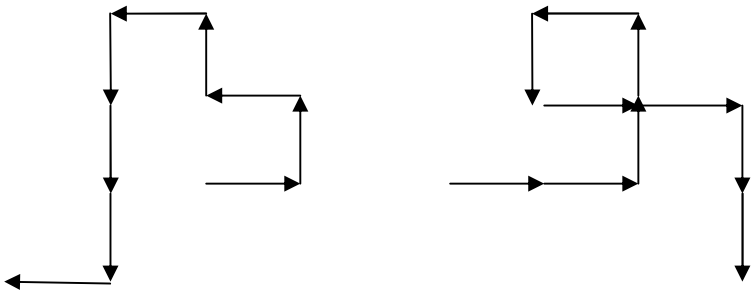
4.1 Πρόωρες εξελίξεις : Αναπτύσσοντας ένα πολυμερές

Η μεθοδολογία της διαδοχικής Monte Carlo μπορεί να αποδοθεί στους Hammersley και Morton (1954), Rosenbluth και Rosenbluth (1955) οι οποίοι εφηύραν μια μέθοδο για να εκτιμήσουν την μέση τετραγωνική επέκταση (the average squared extension) του τυχαίου περίπατου αυτοαποφυγής (self-avoiding random walk) μήκους N σε ένα χώρο πλέγματος (lattice space).

4.1.1 Ένα απλό μοντέλο πολυμερούς: Περίπατος αυτοαποφυγής

Η προσομοίωση αλυσίδας πολυμερών με την βοήθεια των ηλεκτρονικών υπολογιστών είναι ίσως μια από τις πρώτες σοβαρές επιστημονικές προσπάθειες. Αν και αυτές οι προσπάθειες ξεκίνησαν από πολύ παλιά την δεκαετία του 1950, το πρόβλημα της προσομοίωσης μιας μακράς αλυσίδας βιοπολυμερούς εξακολουθεί ακόμα να αποτελεί μια πολύ μεγάλη πρόκληση για την επιστημονική κοινότητα, τόσο λόγω της δύσκολης φύσης του, όσο και της εξαιρετικής του σημασίας στην Βιολογία και στην Χημεία.

Ο τυχαίος περίπατος αυτοαποφυγής (SAW) σε χώρο πλέγματος δύο ή τριών διαστάσεων συχνά χρησιμοποιείται σαν ένα απλό μοντέλο για αλυσίδες πολυμερούς όπως του πολυεστέρα και του πολυαιθυλενίου. Για να κάνουμε μια συνοπτική περιγραφή της βασικής μεθόδου, θα ασχοληθούμε μόνο με το διδιάστατο μοντέλο πλέγματος. Σε αυτό το μοντέλο δημιουργούμε μια αλυσίδα πολυμερούς μήκους N που χαρακτηρίζεται πλήρως από τις θέσεις όλων των μορίων, $\mathbf{x} = (x_1, \dots, x_N)$, όπου x_i είναι ένα σημείο σε διδιάστατο χώρο πλέγματος [δηλαδή $x_i = (\alpha, b)$ όπου α και b είναι ακέραιοι]. Η απόσταση μεταξύ του x_i και x_{i+1} πρέπει να είναι ακριβώς 1 και $x_{i+1} \neq x_\kappa$ για όλα τα $\kappa < i$. Η γραμμή που ενώνει το x_i και το x_{i+1} ονομάζεται (ομοιοπολικός) δεσμός .



Σχήμα 4.1 : Δύο απλοί τυχαίοι περίπατοι σε ένα διδιάστατο χώρο πλέγματος. Ο πρώτος είναι ένας self- avoiding random walk (δηλαδή η αλυσίδα δεν διασταυρώνεται) ενώ ο δεύτερος δεν είναι.

Η κατανομή στόχος που μας ενδιαφέρει στην αλυσίδα πολυμερούς είναι :

$$\pi(\mathbf{x}) = \frac{1}{Z_N}$$

όπου **x** είναι το σύνολο όλων των SAW μήκους N και Z_N είναι η σταθερά κανονικοποίησης, η οποία είναι ακριβώς ο συνολικός αριθμός των διαφορετικών SAW με N στοιχεία.

Ένας απλός τρόπος για να προσομοιώσουμε ένα SAW είναι να ξεκινήσουμε ένα τυχαίο περίπατο από το (0,0) και σε κάθε βήμα i, ο περιπατητής, επειδή δεν του επιτρέπεται να πάει πίσω εκεί από όπου ήρθε δηλαδή στο βήμα i-1 , διαλέγει με ίση πιθανότητα μια από τις τρεις επιτρεπόμενες γειτονικές θέσεις για να πάει. Αν αυτή τη θέση την έχει ήδη επισκεφθεί νωρίτερα ο περιπατητής πρέπει να πάει πίσω στο (0,0) και να ξεκινήσει μια νέα αλυσίδα πάλι. Διαφορετικά , ο περιπατητής συνεχίζει να προχωράει μέχρις ότου το μήκος N να ολοκληρωθεί. Αυτή η διαδικασία προσομοίωσης είναι προφανώς ανεπαρκής.

Το ποσοστό επιτυχίας του SAW (αριθμός επιτυχιών προς τον αριθμό των εκκινήσεων) με δεδομένο μήκος N, μειώνεται εκθετικά σε συνάρτηση με το N.

Για διδιάστατο πλέγμα αυτό το ποσοστό είναι περίπου $\sigma_N = \frac{Z_N}{4x3^{N-1}}$.

Για N=20 το ποσοστό $\sigma_{20} \approx 21.6\%$ και για N=48 το ποσοστό $\sigma_{48} \approx 0.79\%$.

Βήματα του SAW

- Ξεκινάμε από το (0,0)
- Δεν επιστρέφουμε πίσω δηλαδή από το βήμα i δεν πάμε στο βήμα i-1
- Με ίση πιθανότητα διαλέγουμε μια από τις 3 κατευθύνσεις (γειτονικές θέσεις)
- Κάθε θέση μπορούμε να την επισκεφθούμε μόνο μια φορά αλλιώς ξαναρχίζουμε από το (0,0)

Με κατανομή στόχο του SAW π(**x**)=$\frac{1}{Z_N}$, Z_N σταθερά κανονικοποίησης, N ο αριθμός θέσεων της αλυσίδας, X διάνυσμα μήκους N και σ_N το ποσοστό επιτυχίας αλγορίθμου $\sigma_N = \frac{Z_N}{4x3^{N-1}}$.

4.1.2 Ανάπτυξη ενός πολυμερούς σε τετραγωνικό πλέγμα

Για να ξεπεραστεί το προηγούμενο πρόβλημα της μεθόδου του απλού τυχαίου περίπατου (simple random walk), οι Hammersley και Morton (1954), Rosenbluth και Rosenbluth (1955) εισήγαγαν ουσιαστικά ίδιες μεθόδους, που η μια ονομάζεται «αντιστρόφως περιορισμένη δειγματοληψία (inversely restricted sampling)» και η άλλη «μεροληπτική δειγματοληψία (biased sampling)». Η μέθοδος τους είναι ουσιαστικά μια στρατηγική μονοβηματικής πρόβλεψης (one-step-look-ahead) και είναι μια ειδική διαδοχική δειγματοληψία σπουδαιότητας.

Χωρίς βλάβη της γενικότητας, υποθέτουμε ότι ο προσομοιωμένος SAW ξεκινάει πάντα από τη θέση (0,0) και κάνει την πρώτη επίσκεψη στη θέση (1,0). Έτσι εμείς θέτουμε x_1 =(0,0) και x_2 = (1,0). Υποθέτουμε ότι στο στάδιο t, ο τυχαίος περιπατητής έχει κάνει t-1 βήματα, χωρίς καμία παραβίαση του περιορισμού της αυτοαποφυγής(δηλαδή δεν περνάει δυο φορές από το ίδιο σημείο) και βρίσκεται στη θέση x_t =(i ,j). Στη συνέχεια προκειμένου να πάει στη x_{t+1}, ο περιπατητής πρώτα εξετάζει όλες τις γειτονικές θέσεις του x_t [δηλαδή (i$\pm$1, j) και (i, j$\pm$1)]. Αν όλες οι γειτονικές θέσεις έχουν ήδη επισκεφθεί, ο περίπατος τερματίζεται και έχουμε βάρος 0,

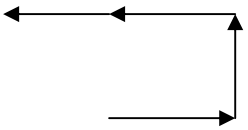
διαφορετικά ο περιπατητής διαλέγει μια από τις ελεύθερες θέσεις (που δεν έχει επισκεφθεί πριν) με ίση πιθανότητα και πηγαίνει στο t+1 βήμα. Μαθηματικά το σύστημα αυτό είναι ισοδύναμο με την δημιουργία της θέσης του x_{t+1} που εξαρτάται από την τρέχουσα κατάσταση του $(x_1, \dots, x_t)$ σύμφωνα με την κατανομή πιθανότητας:

$$P[\, x_{t+1} = (i', j') \mid x_1, \dots, x_t] = \frac{1}{n_t}$$

όπου (i', j') είναι μια από τις μη κατειλημμένες γειτονικές θέσεις του x_t και n_t είναι ο συνολικός αριθμός αυτών των μη κατειλημμένων γειτονικών θέσεων. Αυτό το σύστημα είναι πολύ πιο αποτελεσματικό από την μέθοδο του απλού τυχαίου περίπατου.

Είναι εύκολο να ελέγξουμε ότι αυτοί οι SAW που παράγονται από αυτή τη μέθοδο «ανάπτυξης» δεν είναι ομοιόμορφα κατανεμημένοι. Για παράδειγμα η πιθανότητα δημιουργίας μιας αλυσίδας όπως στο σχήμα 4.2(α) με την μέθοδο ανάπτυξης είναι 1/4 x 1/3 x1/3 x1/2 , ενώ η πιθανότητα δημιουργίας μιας αλυσίδας στο σχήμα 4.2 (b) είναι 1/4x1/3 x1/3 x1/3 . Είναι δύο διαφορετικές καταστάσεις του SAW με N=5 κόμβους και έχουν διαφορετικές πιθανότητες δειγματοληψίας .

(a)



Σχήμα 4.2(α)

- Η αρχή (0,0) έχει πιθανότητα 1/4 για να πάει σε άλλο σημείο
- Το δεύτερο σημείο έχει πιθανότητα 1/3
- Το τρίτο σημείο έχει πιθανότητα 1/3
- Το τέταρτο σημείο έχει πιθανότητα 1/2

P=$\frac{1}{n_t}$ (πιθανότητα να πας από το ένα σημείο στο άλλο)

(b)

Σχήμα 4.2(b)

- Η αρχή (0,0) έχει πιθανότητα 1/4 για να πάει σε άλλο σημείο
- Το δεύτερο σημείο έχει πιθανότητα 1/3
- Το τρίτο σημείο έχει πιθανότητα 1/3
- Το τέταρτο σημείο έχει πιθανότητα 1/3

Για να διορθώσουν την μεροληψία οι Hammersley και Morton, Rosenbluth και Rosenbluth εισήγαγαν αυτό το πιο αποτελεσματικό σύστημα δειγματοληψίας και παρατήρησαν ότι ένας επιτυχημένος SAW που παράγεται με αυτό τον τρόπο θα έχει ένα βάρος που υπολογίζεται ως :

$$w(x) = n_1 \times n_2 \times \dots \times n_{N-1}$$

Επειδή η κατανομή στόχος είναι $\pi(\mathbf{x}) \propto 1$ και η δειγματοληπτική κατανομή του $\mathbf{x}$ είναι $(n_1 \times n_2 \times \dots \times n_{N-1})^{-1}$ η ορθότητα αυτού του βάρους μπορεί να ελεγχθεί από την μέθοδο της δειγματοληψίας σπουδαιότητας.

Η δειγματοληπτική κατανομή $g_t(x_t \mid \boldsymbol{x}_{t-1})$ στην μέθοδο ανάπτυξης είναι ίση με $\frac{1}{n_{t-1}}$ όταν $n_{t-1}>0$ και x_t καταλαμβάνει μια από τις $\boldsymbol{x}_{t-1}$ διαθέσιμες γειτονικές θέσεις και είναι ίση με 0, όταν $n_{t-1} = 0$. Εδώ , n_{t-1} είναι ο αριθμός των διαθέσιμων γειτονικών θέσεων του x_{t-1} . Η ακολουθία των βοηθητικών κατανομών $\pi_t(x_t)$, t=1,…,N-1 είναι ακριβώς η ακολουθία των ομοιόμορφων κατανομών των SAW με t κόμβους (ή $t-1$ βήματα) , που είναι , $\pi_t(x_t) = \frac{1}{Z_t}$ για όλους τους SAW με t κόμβους, όπου Z_t είναι ο συνολικός αριθμός αυτών των SAW. Η περιθώρια κατανομή του μερικού δείγματος $\boldsymbol{x}_{t-1}=(x_1, \dots, x_{t-1})$ κάτω από την ακολουθία βοηθητικών κατανομών π_t είναι :

$$\pi_t(x_{t-1}) = \sum_{\text{όλα τα πιθανά } x_t} \pi_t(x_{t-1}, x_t) = \frac{n_{t-1}}{Z_t}$$

Έτσι η δεσμευμένη κατανομή

$$\pi_t(x_t \mid x_{t-1}) = \frac{1}{n_{\kappa-1}}$$

είναι ακριβώς ίδια με την δειγματοληπτική κατανομή g_κ. Επιλέγουμε η δειγματοληπτική κατανομή g_t να είναι ίση με την ακολουθία των βοηθητικών κατανομών , π_t , t=1,…, .

Για παράδειγμα:

Αν χρησιμοποιήσουμε μια άλλη ακολουθία βοηθητικών κατανομών $\boldsymbol{\pi_t^*}$ όπου είναι η περιθώρια κατανομή του $\boldsymbol{x_t}$ κάτω από την ομοιόμορφη κατανομή(under the uniform distribution) του $(\boldsymbol{x_t}, \boldsymbol{x_{t+1}})$:

$$\pi_t^*(x_t) = \sum_{x_{t+1}} \frac{1}{Z_{t+1}} \propto n_t$$

όπου n_t είναι ο αριθμός των διαθέσιμων γειτονικών θέσεων του x_t, τότε μπορούμε πάλι να θέσουμε g_t να είναι $\pi_t^*(x_t \mid \boldsymbol{x_{t-1}})$. Στη συνέχεια η διαδικασία SIS είναι ισοδύναμη με μια προσέγγιση πρόβλεψης δύο βημάτων (two- step-look-ahead) .

Όπως είπαμε στην SIS το βάρος σπουδαιότητας μπορεί να υπολογιστεί αναδρομικά ως εξής $\boldsymbol{w_{t+1}}=\boldsymbol{w_t} n_t$. Τελικά το συνολικό βάρος της αλυσίδας είναι:

$$w_N = \prod_{t=2}^{N} \frac{1}{g_t(x_t \mid x_1 \ldots x_{t-1})}$$

Επειδή η κατανομή στόχος που έχουμε είναι π(**x**)=$\frac{1}{Z_N}$, το βάρος που μόλις υπολογίσαμε ικανοποιεί την απλή σχέση :

$$w_N = Z_N \, \pi(\mathbf{x}) / \, \rho(\mathbf{x})$$

Συμπερασματικά

Σε αυτή την μέθοδο ανάπτυξης εφαρμόζεται η OSLA στρατηγική δηλαδή το να σταματήσω σήμερα είναι καλύτερο από το να συνεχίσω μέχρι αύριο και να σταματήσω τότε.

Ο αλγόριθμος της μεθόδου ανάπτυξης είναι :

- Ξεκινάμε από το (0,0)
- Πηγαίνουμε στο (1,0)
- Εξετάζουμε αν υπάρχουν ελεύθερα γειτονικά σημεία
- Αν υπάρχουν, με ίση πιθανότητα διαλέγουμε ένα
- Αν όχι, σταματάμε και το βάρος θα είναι w=0

Η διαφορά **της μεθόδου ανάπτυξης** με τον **SAW** είναι ότι ο SAW δεν εξετάζει αν υπάρχουν ελεύθερα γειτονικά σημεία, δηλαδή κάνει το βήμα και αν συναντήσει ένα ήδη κατειλημμένο σημείο, τότε σταματάει ο αλγόριθμος και ξεκινάει πάλι από την αρχή.

 Αν θέλουμε να κάνουμε δειγματοληψία σπουδαιότητας,

- η κατανομή στόχος είναι π(x) ∝ 1

- η δοκιμαστική κατανομή είναι g(x)=$\frac{1}{n_1 n_2 \dots n_{N-1}}$

- το βάρος θα είναι $w(x) = \frac{\pi(x)}{g(x)} = n_1 n_2 \dots n_{N-1}$

 Αν θέλουμε να κάνουμε διαδοχική δειγματοληψία σπουδαιότητας (SIS)

Η δοκιμαστική κατανομή είναι $g_t(x_t \mid x_{t-1}) = \frac{1}{n_{t-1}}$ όταν $n_{t-1} > 0$ και θέλουμε να είναι κοντά στην κατανομή στόχο π(x) .

Θα δείξουμε ότι $\pi_t(x_t \mid x_{t-1})$ είναι ίσο με $\frac{1}{n_{t-1}}$ ως εξής:

Αν η βοηθητική κατανομή είναι $\pi_t(x_t)$ για t=1,…,N-1 $\pi_t(x_t) = \frac{1}{Z_t}$ όπου Z_t είναι ο συνολικός αριθμός αλυσίδων με t κόμβους, η περιθώρια κατανομή του μερικού δείγματος $x_{t-1}=(x_1, \dots, x_{t-1})$ είναι :

$$\pi_t(x_{t-1}) = \sum_{\text{όλα τα πιθανά } x_t} \pi_t(x_{t-1}, x_t) = \frac{n_{t-1}}{Z_t}$$

όπου n_{t-1} ο αριθμός μη απασχολημένων γειτονικών σημείων για x_{t-1} .

Ξέρουμε ότι $\pi_t(x_{t-1}, x_t) = \pi_t(x_{t-1})\pi_t(x_t \mid x_{t-1})$ και δείχνουμε ότι

$$\pi_t(x_t \mid x_{t-1}) = \frac{\frac{1}{Z_t}}{\frac{n_{t-1}}{Z_t}} = \frac{1}{n_{t-1}}$$

Αν πάρουμε μια άλλη βοηθητική κατανομή π_t^* όπου η περιθώρια κατανομή για x_t είναι $\pi_t^*(x_t) = \sum_{x_{t+1}} \frac{1}{Z_{t+1}} \propto n_t$ τότε η $g_t = \pi_t^*(x_t \mid x_{t-1})$.

4.1.3 Περιορισμοί της μεθόδου ανάπτυξης

Ένας περιορισμός της μεθόδου ανάπτυξης εμφανίζεται όταν προσπαθούμε να προσομοιώσουμε πολύ μεγάλα πολυμερή, όπως με N=250. Μέχρι στιγμής το πιο αποτελεσματικό μέσο για να ξεπεραστεί αυτός ο περιορισμός είναι μια προσέγγιση κλαδέματος και εμπλουτισμού (pruning and enrichment). Οι Liu , Chen και Wong (1998) εισήγαγαν μια μέθοδο ελέγχου απόρριψης (Rejection control) μέθοδο που είδαμε στο κεφάλαιο 3, για να βελτιώσουν την διαδοχική δειγματοληψία και την προσομοίωση.

4.2 Διαδοχική υποκατάσταση (Sequential imputation) σε στατιστικά προβλήματα με ελλιπή δεδομένα

Σε ένα πρόβλημα ελλιπών δεδομένων, χωρίζουμε το y σε (y_{obs}, y_{mis}) όπου μόνο τα y_{obs} παρατηρούνται και τα y_{mis} καλούνται ελλιπή δεδομένα. Υποθέτουμε ότι $y \sim f(y \mid \theta)$, όπου $f()$ συνάρτηση πιθανότητας. Η πιθανοφάνεια με τα πλήρη δεδομένα $L(\theta \mid y)$ [η οποία μπορεί να είναι οποιαδήποτε συνάρτηση που είναι ανάλογη με την $f(y \mid \theta)$] και η εκ των υστέρων κατανομή με πλήρη δεδομένα $p(\theta \mid y)$ είναι εύκολο να υπολογιστούν. Το κεντρικό πρόβλημα εδώ είναι ο υπολογισμός της πιθανοφάνειας των παρατηρούμενων δεδομένων $L(\theta \mid y_{obs})$ και της εκ των υστέρων κατανομής των παρατηρούμενων δεδομένων $p(\theta \mid y_{obs})$. Θα πρέπει επίσης να γίνει και ο υπολογισμός των ελλιπών δεδομένων από την $L(\theta \mid y_{obs}, y_{mis})$.

Υποθέτουμε ότι τα y_{obs} και τα y_{mis} μπορούν να αναλυθούν περαιτέρω σε n αντίστοιχα στοιχεία έτσι ώστε :

$$\boldsymbol{y} = (y_1, \dots, y_n) = (y_{obs},\ y_{mis}) = (y_{obs,1},\ y_{mis,1}, \dots, y_{obs,n},\ y_{mis,n})$$

όπου y_t=$(y_{obs,t},\ y_{mis,t})$ για t=1,…,n .

Σε πολλές εφαρμογές τα y_t είναι ανεξάρτητα και ισόνομα κατανεμημένα, δεδομένου του θ, όμως δεν είναι απαραίτητη προυπόθεση .

Πιο συνοπτικά:

Όταν υπάρχουν ελλιπή δεδομένα και θέλουμε να υπολογίσουμε μια πιθανοφάνεια $L(\theta \mid y_{obs})$ και την posterior $p(\theta \mid y_{obs})$ θα πρέπει να υπολογίσουμε πρώτα τα ελλιπή δεδομένα (y_{mis}). Για να το κάνουμε αυτό χρησιμοποιούμε την διαδοχική δειγματοληψία σπουδαιότητας.

4.2.1 Υπολογισμός πιθανοφάνειας (likelihood)

Στο παράδειγμα με ελλιπή δεδομένα, αν θέλουμε να εκτιμήσουμε την παράμετρο θ με την μέθοδο μέγιστης πιθανοφάνειας MLE (Maximum Likelihood Estimation), τότε πρέπει να υπολογίσουμε την συνάρτηση πιθανοφάνειας των παρατηρούμενων δεδομένων. Μπορούμε να βρούμε την συνάρτηση των παρατηρούμενων δεδομένων με περιθωριοποίηση των ελλιπών δεδομένων από την συνάρτηση πιθανοφάνειας με πλήρη δεδομένα:

$$L(\theta \mid y_{obs}) = f(y_{obs} \mid \theta) = \int f(y_{mis}, y_{obs} \mid \theta)\, \mathrm{d}y_{mis} \qquad (4.2)$$

Όμως αυτό το ολοκλήρωμα είναι συχνά δύσκολο να υπολογιστεί αναλυτικά. Υποθέτουμε ότι μπορούμε να προσομοιώσουμε δείγμα $y_{mis}^1, \ldots, y_{mis}^m$ από μια δοκιμαστική κατανομή $g(\,)$.

Στη συνέχεια μπορούμε να εκτιμήσουμε την συνάρτηση πιθανοφάνειας $L(\theta \mid y_{obs})$ ως

$$\hat{L}(\theta \mid y_{obs}) = \frac{1}{m}\left\{\frac{f(y_{obs}, y_{mis}^1 \mid \theta)}{g(y_{mis}^1)} + \cdots + \frac{f(y_{obs}, y_{mis}^m \mid \theta)}{g(y_{mis}^m)}\right\} \qquad (4.3)$$

δηλαδή με τη μέθοδο δειγματοληψίας σπουδαιότητας (Ενότητα 3.5). Αν τα y_{mis}^j προσομοιώνονται, ο τύπος (4.3) μπορεί να εφαρμοστεί για την εκτίμηση του L για πολλά διαφορετικά θ.

Το ερώτημα όμως είναι ποια θα είναι η κατάλληλη δοκιμαστική κατανομή $g(\cdot\,)$.

Αν για παράδειγμα εξετάζαμε μόνο ένα θ , ας πούμε θ_0, τότε θα έπρεπε να προσομοιώσουμε τα ελλιπή δεδομένα y_{mis}^j από την συνάρτηση $f(y_{mis} \mid y_{obs}, \theta_0)$, που σε αυτή την περίπτωση η αναλογία σπουδαιότητας θα είναι:

$$\frac{f(y_{obs}, y_{mis} \mid \theta_0)}{f(y_{mis} \mid y_{obs}, \theta_0)} = f(y_{obs} \mid \theta_0) \equiv L(\theta_0 \mid y_{obs})$$

Αυτό σημαίνει ότι μπορούμε να εκτιμήσουμε την συνάρτηση πιθανοφάνειας L ακριβώς με ένα τυχαίο δείγμα. Όμως σε πολλές εφαρμογές, το να προσομοιώνουμε από τη συνάρτηση $f(y_{mis} \mid y_{obs}, \theta_0)$ είναι ίσως πιο δύσκολο από το να εκτιμήσουμε την $L(\theta_0 \mid y_{obs})$.

Μια απλή επιλογή της δοκιμαστικής κατανομής $g(\,)$ μπορεί να είναι η συνάρτηση $f(\mathbf{y_{mis}} \mid \theta_0)$ (στην οποία δεν χρησιμοποιούμε τις παρατηρούμενες πληροφορίες). Αλλά αυτή η δοκιμαστική κατανομή συχνά διαφέρει πολύ από την ιδανική και αυτό θα έχει σαν αποτέλεσμα να μεταβάλλονται πολύ τα βάρη σπουδαιότητας. Για ευκολία, χρησιμοποιούμε τους εξής συμβολισμούς:

$$\boldsymbol{y_{obs,t}} = (y_{obs,1}, \dots, y_{obs,t})$$

$$\boldsymbol{y_{mis,t}} = (\ y_{mis,1}, \dots,\ y_{mis,t})$$

$$\boldsymbol{y_t} = (\boldsymbol{y_{obs,t}},\ \boldsymbol{y_{mis,t}})$$

Οι Kong et al (1994) πρότειναν την ακόλουθη μέθοδο διαδοχικής υποκατάστασης (sequential imputation) για να προσομοιώνουν Monte Carlo δείγματα $\mathbf{y_{mis}}$ και έφτιαξαν τα κατάλληλα βάρη σπουδαιότητας:

Α. Προσομοιώνουμε $\mathbf{y_{mis}}$ (που είναι ίδια σαν $\mathbf{y_{mis,}}\boldsymbol{_n}$) από

$$g(\boldsymbol{y_{mis}}) = f(y_{mis,1}\ |\ y_{obs,1}, \theta)\ \times\ f\left(y_{mis,2}\ |y_{obs,1}\ , y_{mis,1}, y_{obs,2}, \theta\right)\ \times \dots \times$$
$$f\left(y_{mis,n}\ |y_{obs,n}\ , y_{mis,n-1}, \theta\right)$$

Αυτό το βήμα μπορεί να υλοποιηθεί αναδρομικά αν προσομοιώσουμε $y_{mis,t}$ από την κατανομή πρόβλεψης :

$$f\left(y_{mis,t}\ |y_{obs,t}\ , y_{mis,t-1}\right)$$

Β. Υπολογίζουμε τα βάρη του $\mathbf{y_{mis}}$ αναδρομικά ως :

$$w(\boldsymbol{y_{mis}}) = \mathrm{f}(\mathrm{y_{obs,1}}\ |\theta) f\left(y_{obs,2}|\ \boldsymbol{y}_1\ , \theta\right) \times \dots \times f\left(y_{obs,n}\ \ |\ \ \boldsymbol{y}_{n-1}\ , \theta\right)$$

Παρατηρούμε ότι η δειγματοληπτική κατανομή για τα $\mathbf{y_{mis}}$ ικανοποιεί τη σχέση:

$$\int g(\boldsymbol{y_{mis}})\ w(\boldsymbol{y_{mis}}) d\mathbf{y_{mis}} = f(\boldsymbol{y_{obs}}\ \ |\ \theta)$$

Έτσι μπορούμε να εκτιμήσουμε την τιμή της πιθανοφάνειας $L(\theta\ |\boldsymbol{y_{obs}})$ από :

$$\bar{w} = \frac{1}{m}\{w(y_{mis}^1) + \dots + w(y_{mis}^m)\}$$

Αν ενδιαφερόμαστε για την τιμή της πιθανοφάνειας σε διαφορετικό σημείο θ' το οποίο δεν είναι πολύ μακριά από το θ, μπορούμε να ξαναπάρουμε βάρος (reweight) στα παραπάνω δείγματα Monte Carlo (αντί να δημιουργήσουμε ένα καινούργιο σύνολο) σύμφωνα με το τύπο (4.3) έτσι ώστε να πάρουμε μια καλή εκτίμηση.

Πιο συνοπτικά

Επειδή συχνά είναι δύσκολος ο υπολογισμός του ολοκληρώματος

$$L(\theta\ |y_{obs}) = f(y_{obs}\ |\theta) = \int f(y_{mis}, y_{obs}\ |\theta)\ \mathrm{d}y_{mis}$$

κάνουμε τα εξής βήματα:

 Προσομοιώνουμε τα y_{mis} από μια δοκιμαστική κατανομή g όπου

$$g(y_{mis})= f(y_{mis} \mid y_{obs}, \theta)$$

Μετά εκτιμούμε την $L(\theta \mid y_{obs})$ ως εξής:

$$\hat{L}(\theta \mid y_{obs})=\frac{1}{m}\left\{\frac{f(y_{mis}^{1},y_{obs} \mid \theta)}{g(y_{mis}^{1})}+\cdots+\frac{f(y_{mis}^{m},y_{obs} \mid \theta)}{g(y_{mis}^{m})}\right\}$$

Η κατανομή στόχος είναι $\pi= f(y_{mis}, y_{obs} \mid \theta)$

● Υπολογίζουμε τα βάρη $w(y_{mis})$ για κάποιο θ_0 και έχουμε :

$$w(y_{mis}) = \frac{\pi(y_{mis})}{g(y_{mis})}=\frac{f(y_{mis},y_{obs} \mid \theta_0)}{f(y_{mis} \mid y_{obs}, \theta_0)} = f(y_{obs} \mid \theta_0)$$

Μετά βρίσκουμε την πιθανοφάνεια.

Επειδή όμως μπορεί να είναι δύσκολο να έχουμε $g(y_{mis})= f(y_{mis} \mid y_{obs}, \theta_0)$ για να εκτιμήσουμε την $L(\theta_0 \mid y_{obs})$ κάνουμε τα εξής βήματα:

● Προσομοιώνουμε τα y_{mis} από μια δοκιμαστική κατανομή g όπου

$$g(y_{mis}) = f(y_{mis,1} \mid y_{obs,1}, \theta) \times f\left(y_{mis,2}, \mid y_{obs,1}, y_{mis,1}, y_{obs,2}, \theta\right) \times \ldots \times f\left(y_{mis,n} \mid y_{obs,n}, y_{mis,n-1}, \theta\right)$$

● Υπολογίζουμε τα βάρη

$$w(y_{mis}) = \frac{\pi(y_{mis})}{g(y_{mis})}$$

με κατανομή στόχο $\pi= f(y_{mis}, y_{obs} \mid \theta)$ και

δοκιμαστική κατανομή $g(y_{mis} \mid \theta)$

και έχουμε : $w(y_{mis}) = \frac{f(y_{mis},y_{obs} \mid \theta)}{g(y_{mis} \mid \theta)}$

Το $f(y_{mis}, y_{obs} \mid \theta)= \frac{f(y_{obs} \mid y_{mis},\theta)f(y_{mis} \mid \theta)}{g(y_{mis} \mid \theta)}$

Αν η δοκιμαστική κατανομή είναι $g(y_{mis} \mid \theta)= f(y_{mis} \mid \theta)$

τότε το βάρος θα είναι $w(y_{mis}) = f(y_{obs} \mid y_{mis}, \theta)$

όπου: $f(y_{obs} \mid y_{mis}, \theta) = f\left(y_{obs,1} \mid \theta\right) \times f\left(y_{obs,2} \mid y_{obs,1}, y_{mis,1}, \theta\right) \times \ldots \times f\left(y_{obs,,n} \mid y_{obs,,n-1} y_{mis,n-1}, \theta\right)$

4.2.2 Μπεϋζιανός υπολογισμός

Σε αυτήν την ενότητα θα εξετάσουμε μια Μπεϋζιανή προσέγγιση στο πρόβλημα με ελλιπή δεδομένα.

Υποθέτουμε ότι έχουμε μια εκ των προτέρων κατανομή $f(\theta)$ με άγνωστη παράμετρο θ. Σε αυτό το μπεϋζιανό μοντέλο, η από κοινού κατανομή της παραμέτρου θ και των δεδομένων $(\mathbf{y_{mis}}, \boldsymbol{y_{obs}})$ μπορεί να γραφτεί ως εξής:

$$f(\theta, \mathbf{y_{mis}}, \boldsymbol{y_{obs}}) = f(\boldsymbol{y_{obs}}, \mathbf{y_{mis}} \mid \theta)\, f(\theta)$$

όπου $f(\theta)$ είναι η εκ των προτέρων κατανομή της παραμέτρου θ. Έτσι αν ενδιαφερόμαστε για την εκ των υστέρων κατανομή της παραμέτρου θ, $f(\theta \mid y_{obs})$, θα πρέπει να υπολογίσουμε :

$$f(\theta \mid y_{obs}) = \frac{\int f(\theta, \mathbf{y_{mis}}, \boldsymbol{y_{obs}})\, d\mathbf{y_{mis}}}{\int f(\theta', \mathbf{y_{mis}}, \boldsymbol{y_{obs}}) d\,\theta' d\mathbf{y_{mis}}}$$

Την οποία μπορούμε να ξαναγράψουμε ως:

$$f(\theta \mid y_{obs}) = \int f(\theta \mid \boldsymbol{y_{obs}}, \mathbf{y_{mis}})\; f(\mathbf{y_{mis}} \mid \boldsymbol{y_{obs}})\, d\mathbf{y_{mis}}$$

Δηλαδή, αν ένα τυχαίο δείγμα $y_{mis}^1, \dots, y_{mis}^m$ μπορεί να προσομοιώνεται από την κατανομή $f(\mathbf{y_{mis}} \mid \boldsymbol{y_{obs}})$, μπορούμε να προσεγγίσουμε την παραπάνω εκ των υστέρων κατανομή ως :

$$f(\theta \mid y_{obs}) = \frac{1}{m} \sum_{j=1}^{m} f(\theta \mid \boldsymbol{y_{obs}}, \boldsymbol{y_{mis}^j})$$

(σαν μια μείξη εκ των υστέρων κατανομών με πλήρη δεδομένα).

Δεδομένου ότι η προσομοίωση από την κατανομή $f(\mathbf{y_{mis}} \mid \boldsymbol{y_{obs}})$ είναι τυπικά δύσκολη, μπορούμε να εφαρμόσουμε τη μέθοδο διαδοχικής υποκατάστασης.

Ξεκινάμε με το να προσομοιώσουμε $y_{mis,1}$ από $f(\mathrm{y_{mis,1}} \mid y_{obs,1})$ και να υπολογίσουμε $w_1 = f(y_{obs,1})$.

Σημειώνουμε ότι:

$$f(\mathrm{y_{mis,1}} \mid y_{obs,1}) \propto f(y_{obs,1}, \mathrm{y_{mis,1}} \mid \theta)\, f(\theta)\, d\theta$$

Έτσι για t=2,…,n έχουμε τα ακόλουθα βήματα :

A. Προσομοιώνουμε $y_{mis,\mathrm{t}}$ από την δεσμευμένη κατανομή

$$f\left(y_{mis,\mathrm{t}} \mid y_{obs,\mathrm{t}}, y_{mis,\mathrm{t}-1}\right)$$

B. Υπολογίζουμε τις κατανομές πρόβλεψης $f\left(y_{obs,t} \mid y_{obs,t-1}, y_{mis,t-1}\right)$ και τα βάρη

$$w_t = w_{t-1} f\left(y_{obs,t}, \mid y_{obs,t-1}, y_{mis,t-1}\right) \qquad (4.4)$$

Θέτουμε $w(y_{mis}) = w_n \equiv f(y_{obs,1}) \prod_{t=2}^{n} f\left(y_{obs,1} \mid y_{obs,t-1}, y_{mis,t-1}\right)$

Τα βήματα Α και Β συνήθως γίνονται ταυτόχρονα. Και τα δύο βήματα Α και Β θα πρέπει να υπολογίζονται εύκολα , και αυτό μπορεί να γίνεται στην περίπτωση που οι κατανομές πρόβλεψης με πλήρη δεδομένα $f(y_t \mid \boldsymbol{y}_{t-1})$ είναι απλές.

Υποθέτουμε ότι τα βήματα Α και Β έχουν κάνει m επαναλήψεις ανεξάρτητες, τα αποτελέσματα των οποίων είναι μια υποκατάσταση των ελλιπών δεδομένων, $y_{mis}^1, \ldots, y_{mis}^m$ και τα αντίστοιχα βάρη τους είναι $w^1, \ldots, w^m$. Τότε, μπορούμε να εκτιμήσουμε την εκ των υστέρων κατανομή $f(\theta \mid y_{obs})$ από το σταθμισμένο μείγμα :

$$\frac{1}{W}\sum_{j=1}^{m} w^j \; f(\theta \mid y_{obs}, y^j{}_{mis}) \qquad \text{όπου } W = \sum w^j .$$

Το μειονέκτημα αυτής της μεθόδου είναι ότι όταν το μέγεθος του δείγματος n αυξάνεται, τα βάρη σπουδαιότητας γίνονται πολύ ασύμμετρα. Ο Kong et al (1994) έδειξαν ότι το κανονικοποιημένο βάρος w_n είναι μια ακολουθία martingale (σε σχέση με το n). Έτσι η διασπορά του w_n αυξάνεται στοχαστικά καθώς το n αυξάνεται.

Η ομοιότητα μεταξύ του Μπεϋζιανού προβλήματος ελλιπών δεδομένων με την προσομοίωση του SAW είναι εμφανής: μαθηματικά , το i-οστό στοιχείο που λείπει $y_{mis,i}$ παίζει τον ίδιο ρόλο σαν την i-οστή θέση του μονομερούς x_i .

Πιο συνοπτικά

Αν θέλουμε να υπολογίσουμε την εκ των υστέρων κατανομή, δηλαδή την $f(\theta \mid y_{obs})$, των ελλιπών δεδομένων ακολουθούμε τα εξής βήματα:

● Προσομοιώνουμε τα y_{mis} από μια δοκιμαστική κατανομή g όπου

$g(y_{mis}) = f(y_{mis,1}\,|y_{obs,1}, \theta) \times f(y_{mis,2}\,|y_{obs,1}\,, y_{mis,1}, \theta) \times .. \times$
$f(y_{mis,n}\,|y_{obs,n}\,, y_{mis,n-1}, \theta)$

● Υπολογίζουμε τις κατανομές πρόβλεψης $f(y_{obs,t}\,|y_{obs,t-1} y_{mis,t-1})$ και τα βάρη είναι

$$w_t = w_{t-1}\, f(y_{obs,t}\,|y_{obs,t-1} y_{mis,t-1})$$

Θέτουμε $w(y_{mis}) = w_n \equiv f(y_{obs,1}) \prod_{t=2}^{n} f(y_{obs,1}\,|y_{obs,t-1} y_{mis,t-1})$

Άρα η εκ των υστέρων κατανομή $f(\theta\,|y_{obs})$ εκτιμάται από την ποσότητα

$$\frac{1}{w}\sum_{j=1}^{m} w^j\; f(\theta\,|y_{obs}, y^j{}_{mis}).$$

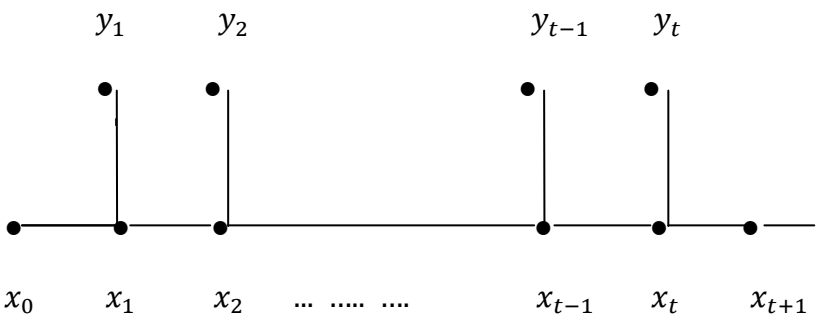
4.3 Μη γραμμικό φιλτράρισμα (Nonlinear Filtering)

Το μοντέλο χώρου καταστάσεων (state space model) είναι ένα πολύ γνωστό δυναμικό σύστημα που αποτελείται από δύο σημαντικά μέρη: 1) την παρατηρούμενη εξίσωση που συχνά γράφεται ως $y_t \sim f_t(\cdot\,|x_t, \varphi)$ και 2) την εξίσωση κατάστασης , η οποία μπορεί να εκφράζεται σαν μια Μαρκοβιανή διαδικασία ως $x_t \sim q_t(\cdot\,|x_{t-1}, \theta)$. Και έχουμε :

(Εξίσωση κατάστασης): $x_t \sim q_t(\cdot\,|x_{t-1}, \theta)$ (4.5)

(Παρατηρούμενη εξίσωση): $y_t \sim f_t(\cdot\,|x_t, \varphi)$

Οι y_t είναι παρατηρούμενες μεταβλητές και οι x_t είναι (οι μη παρατηρούμενες) οι μεταβλητές κατάστασης. Αυτό το μοντέλο σε πολλές εφαρμογές καλείται ως κρυμμένο Μαρκοβιανό μοντέλο (HMM). Ένα τέτοιο μοντέλο μπορεί να αναπαρασταθεί γραφικά όπως στο σχήμα 4.4.



Σχήμα 4.4

Στην πράξη, τα x μπορεί να είναι:

- τα μεταδιδόμενα ψηφιακά σήματα στην ασύρματη επικοινωνία (Liu και Chen 1995),

- η θέση στόχου και η ταχύτητα που ανιχνεύεται από ένα ραντάρ(Gordon et al. 1993, Gordon ,Salmond και Ewing 1995, Avitzour 1995),

- η παρατηρούμενη αστάθεια σε οικονομικές χρονοσειρές (Pitt και Shephard 1999) και πολλά άλλα.

Η κύρια πρόκληση για τους ερευνητές είναι να βρουν αποτελεσματικές μεθόδους για (on-line) εκτίμηση και πρόβλεψη (φιλτράρισμα) των μεταβλητών κατάστασης (δηλαδή x_t) ή άλλων παραμέτρων όταν το σήμα y_t έρχεται διαδοχικά. Για ευκολία υποθέτουμε ότι οι παράμετροι του συστήματος (δηλαδή φ και θ) είναι όλες γνωστές και παραλείπονται από το παραπάνω μοντέλο χώρου καταστάσεων. Ιδέες για την εκτίμηση των παραμέτρων του μοντέλου σε συνδυασμό με τις μεταβλητές κατάστασης μπορούμε να βρούμε στους Liu και Chen (1995), Gordon et al(1995) , Liu και West (2000).

Με γνωστές τις παραμέτρους του συστήματος, η βέλτιστη (on-line) εκτίμηση του x_t είναι η λύση του Bayes (West και Harrison 1989) .

$$\hat{x}_t=\mathrm{E}(x_t \mid y_1, y_2 \dots, y_t)=\frac{\int \dots \int x_t \prod_{s=1}^{t}[f(y_s \mid x_s) q_s(x_s \mid x_{s-1})]dx_1 \dots dx_t}{\int \dots \int \prod_{s=1}^{t}[f(y_s \mid x_s) q_s(x_s \mid x_{s-1})]dx_1 \dots dx_t}$$

Έτσι για κάθε χρόνο t αυτό που μας ενδιαφέρει είναι η «τρέχουσα» εκ των υστέρων κατανομή του x_t, η οποία μπορεί θεωρητικά να προέρχεται αναδρομικά ως :

$$\pi_t(x_t) = P(x_t \mid y_1, y_2 \dots, y_t)$$

$$\propto \int q_t(x_t \mid x_{t-1})\ f_t(y_t \mid x_t)\ \pi_{t-1}(x_{t-1})\ \mathrm{d}x_{t-1} \qquad (4.6)$$

όπου $\pi_{t-1}(x_{t-1}) = P(x_{t-1} \mid y_1, y_2 \dots, y_{t-1})$ είναι η «τρέχουσα» εκ των υστέρων κατανομή του x_{t-1} .

Όταν οι f_t και q_t στην (4.5), είναι κανονικές γραμμικές δεσμευμένες κατανομές, το μοντέλο που προκύπτει ονομάζεται γραμμικό μοντέλο χώρου καταστάσεων, ή δυναμικό γραμμικό μοντέλο, το οποίο έχει ενεργό ρόλο στην έρευνα, στις περιοχές του αυτόματου ελέγχου, των οικονομικών, της μηχανικής και της στατιστικής (Harvey 1990, West και Harrison 1989). Σε αυτή την περίπτωση, ο υπολογισμός της εκ των υστέρων κατανομής μπορεί να γίνει αναλυτικά με επαγωγή (μέσω αναδρομής) (4.7) γιατί όλες οι «τρέχουσες» εκ των υστέρων κατανομές είναι κανονικές. Όταν x_t παίρνει ορισμένες τιμές έστω κ δυνατές τιμές τότε το μοντέλο αναφέρεται ως HMM

έχει πολλές εφαρμογές όπως στην ψηφιακή επικοινωνία, στην αναγνώριση λόγου κτλ.

Βέλτιστες άμεσες εκτιμήσεις του x_t μπορούν να επιτευχθούν μέσω επαγωγικού αλγορίθμου που είναι παρόμοιος με προσομοίωση και μέθοδο δυναμικού προγραμματισμού.

Ο ακριβής υπολογισμός των βέλτιστων εκτιμήσεων του x_t είναι γενικά αδύνατος λόγω των ολοκληρωμάτων και της σταθεράς κανονικοποίησης που γίνονται ανέφικτα καθώς αυξάνεται το t.

Μια πολύ γενική και απλή μέθοδος είναι το φίλτρο Bootstrap για την άμεση εκτίμηση ενός μοντέλου χώρου καταστάσεων, η οποία χρησιμοποιεί την ιδέα της μεθόδου δειγματοληψίας σπουδαιότητας με αναδειγματοληψία (sampling-importance resampling) (SIR) (Rubin 1987).

Έστω ότι τη στιγμή t έχουμε ένα τυχαίο δείγμα $\{x_t{}^1, \dots, x_t{}^m\}$ της μεταβλητής κατάστασης x_t η οποία ακολουθεί προσεγγιστικά την τρέχουσα εκ των υστέρων κατανομή $\pi_t(x_t) = P(x_t \mid y_1, y_2 \dots, y_t)$. Οι Gordon et al (1993) πρότειναν την ακόλουθη διαδικασία προσομοίωσης όταν η y_{t+1} είναι παρατηρούμενη μεταβλητή :

(α) Προσομοιώνουμε x_{t+1}^{*j} από την εξίσωση κατάστασης

$$q_t(x_{t+1} \mid x_t^j) \quad \text{j=1,...,m}$$

(β) Υπολογίζουμε τα βάρη :

$$w^j \propto f_t(y_{t+1} \mid x_{t+1}^{*j}) \qquad \text{j=1,...,m}$$

(γ) Από το δείγμα $\{x_t{}^{*1}, \dots, x_t{}^{*m}\}$ με πιθανότητα ανάλογη του w^j φτιάχνουμε ένα καινούργιο τυχαίο δείγμα $\{x_{t+1}^1, \dots, x_{t+1}^m\}$ για τη στιγμή t+1.

Αν $x_t{}^m$ ακολουθεί την «τρέχουσα» εκ των υστέρων κατανομή $\pi_t(x_t)$ και όταν το m είναι αρκετά μεγάλο τότε το νέο τυχαίο δείγμα $\{x_{t+1}^1, \dots, x_{t+1}^m\}$ ακολουθεί την αναβαθμισμένη εκ των υστέρων κατανομή $\pi_{t+1}(x_{t+1})$ προσεγγιστικά.

Η σύνδεση μεταξύ αυτού του προβλήματος και του στατιστικού προβλήματος ελλιπών δεδομένων είναι προφανής. Η μεταβλητή κατάστασης x_t μπορεί να θεωρηθεί ως ελλιπή δεδομένα. Κάτω από το πλαίσιο της SIS βλέπουμε ότι η τρέχουσα κατανομή στόχος $\pi_t(\cdot)$ ικανοποιεί την αναδρομική σχέση:

$$\pi_t(x_t) \propto f_t(y_t \mid x_t)\, q_t(x_t \mid x_{t-1})\, \pi_{t-1}(x_{t-1}) \qquad (4.7)$$

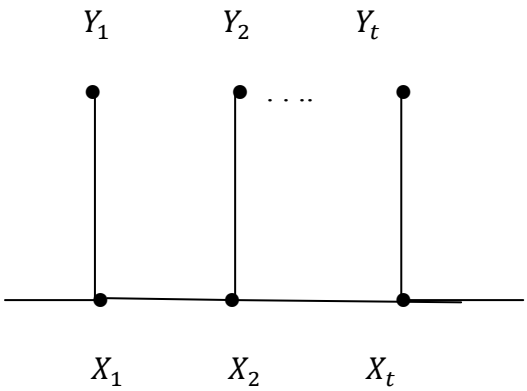
Ένα ιδιαίτερο χαρακτηριστικό του μοντέλου χώρου καταστάσεων είναι ότι η μεταβλητή κατάστασης x_t διαθέτει μια Μαρκοβιανή δομή. Αυτό το χαρακτηριστικό μας δίνει την δυνατότητα να εξετάσουμε μόνο την περιθώρια εκ των υστέρων κατανομή $\pi_t(x_t)$ και όχι την από κοινού κατανομή $\pi_t(\boldsymbol{x_t})$ και είναι σημαντικό ώστε το φίλτρο Bootstrap να είναι εφαρμόσιμο. Πράγματι, αν μετατοπίσουμε το βήμα με δείκτη t για ελλιπή δεδομένα $y_{mis,t}$ προς τα εμπρός κατά μια μονάδα (δηλαδή από $y_{mis,t-1}$ σε $y_{mis,t}$), τότε, η διαδοχική μέθοδος υποκατάστασης με σταδιακή αναδειγματοληψία προκαλεί τον ίδιο αλγόριθμο.

Το πλεονέκτημα της μεθόδου του φίλτρου Bootstrap είναι η ευκολία της ενώ τα μειονεκτήματα της μεθόδου είναι δύο:

(α) Δεν χρησιμοποιεί την τρέχουσα διαθέσιμη πληροφορία, y_{t+1}, στο βήμα δειγματοληψίας.

(β) Η υπερβολική χρήση της αναδειγματοληψίας μπορεί να μειώσει την αποτελεσματικότητά της .

Πιο συνοπτικά :



Τα $Y_1, Y_2 \dots, Y_t$ είναι παρατηρούμενα σήματα και θέλουμε, δεδομένων αυτών των παρατηρήσεων, να εκτιμήσουμε τις μεταβλητές κατάστασης $X_1, X_2 \dots, X_t$, δηλαδή βρίσκουμε την τρέχουσα εκ των υστέρων κατανομή για χρόνο t της x_t $x_t \mid y_1, y_2 \dots, y_t$ που είναι:

$$\pi_t(x_t) = P(x_t \mid y_1, y_2 \dots, y_t)$$

Όταν οι f_t και q_t είναι γραμμικές και κανονικές τότε η ζητούμενη εκ των υστέρων κατανομή υπολογίζεται επαγωγικά από τη σχέση :

$$\pi_t(x_t) \propto \int f_t(y_t \mid x_t) q_t(x_t \mid x_{t-1})\, \pi_{t-1}(x_{t-1})\, \mathrm{d}x_{t-1}$$

όπου $\pi_{t-1}(x_{t-1}) = P(x_{t-1} \mid y_1, y_2 \dots, y_{t-1})$

4.4 Ένα γενικό πλαίσιο

Η σύνδεση των μεθόδων [μέθοδος ανάπτυξης, διαδοχική μέθοδος υποκατάστασης και φίλτρο Bootstrap] που περιγράφτηκαν στις τρεις προηγούμενες ενότητες συνοψίζεται ως εξής:

● Η μέθοδος ανάπτυξης είναι μια ειδική εφαρμογή της γενικής SIS μεθοδολογίας.

● Η διαδοχική μέθοδος υποκατάστασης του Kong et al (1994) είναι η αντίστοιχη μορφή της SIS με ελλιπή δεδομένα.

● Το φίλτρο Bootstrap χρησιμοποιεί μια SIS δομή αλλά έχει εισαχθεί ένα επιπρόσθετο βήμα αναδειγματοληψίας πριν από κάθε βήμα διαδοχικής δειγματοληψίας.

Σε αυτή την ενότητα θα δείξουμε ότι οι τροποποιήσεις της μεθόδου ανάπτυξης που έγιναν από τον Wall και Erpenbeck (1959) και Grassberger (1997) είναι πρακτικά ισοδύναμες με ένα σύστημα SIS με αναδειγματοληψία (Liu και Chen 1995).

Συνοψίζοντας τα κοινά χαρακτηριστικά των προβλημάτων που αντιμετωπίζονται στις προηγούμενες ενότητες θα ορίσουμε κατά αρχάς ένα πιθανοθεωρητικό δυναμικό σύστημα (probabilistic dynamic system (PDS)). Στο πλαίσιο της SIS αυτό το σύστημα εξυπηρετεί σαν μια βοηθητική ακολουθία των κατανομών.

Ορισμός: Μια ακολουθία από κατανομές πιθανοτήτων $\pi_t(\boldsymbol{x_t})$, που αναπροσαρμόζεται με διακριτό χρόνο t=0,1,2,..., θα την ονομάζουμε πιθανοθεωρητικό δυναμικό σύστημα. Η μεταβλητή κατάσταση x_t μπορεί, είτε να αυξάνει την διάστασή της όσο αυξάνει το t (δηλαδή $\boldsymbol{x_{t+1=}}(\boldsymbol{x_t}, x_{t+1})$), είτε να μένει όπως είναι (δηλαδή $\boldsymbol{x_{t+1=}x_t}$)

Εδώ η π( ) πάντα αναφέρεται στην κατανομή στόχο του δυναμικού συστήματος και p( ) είναι ένα γενικό σύμβολο για κατανομές πιθανοτήτων.

Για το πρόβλημα προσομοίωσης SAW (στην ενότητα 4.1) το PDS ορίζεται ως $\pi_t(x_t) = \frac{1}{Z_t}$ στο σύνολο όλων των SAWS με t σημεία, όπου Z_t είναι ο συνολικός αριθμός τέτοιων διαμορφώσεων. Η π_t είναι η ομοιόμορφη κατανομή στο σύνολο όλων των SAW στα t-1 βήματα. Στη βιβλιογραφία της μοριακής προσομοίωσης ενδιαφερόμαστε σε πιο πολύπλοκες δομές (π.χ. δομές πρωτείνης) όπου η π_t λαμβάνεται ως Boltzmann κατανομή της μορφής $\exp\left\{\frac{-U_t}{K_B T}\right\}/Z$. Εδώ το U_t είναι η αλληλεπίδραση της ενέργειας που παράγεται από τα μονομερή της πολυμερούς αλυσίδας (με t μονομερή), T είναι η απόλυτη θερμοκρασία και K_B είναι η σταθερά Boltzmann .

Το παράδειγμα του μοντέλου χώρου κατάστασης (ενότητα 4.3) επίσης ταιριάζει πολύ καλά σε ένα PDS πλαίσιο στο οποίο μπορούμε να ορίσουμε την $\pi_t(x_t)$ σαν την εκ των υστέρων κατανομή $p(x_t \mid y_t, x_0)$. Με αυτό το PDS, η εκτίμηση της x_t μπορεί να είναι ο στατιστικός υπολογισμός (Monte Carlo) του μέσου όρου μιας τυχαίας μεταβλητής που έχει κατανομή π_t. Η παραπάνω μεθοδολογία γενικά ονομάζεται διαδοχική Monte Carlo. Είναι φυσικό να θεωρούμε το PDS σαν μια ακολουθία εκ των υστέρων δεσμευμένων κατανομών με πληροφορία « μέχρι τη στιγμή t», για ένα σύστημα με τυχαίες μεταβλητές $x_1, x_2 \ldots$. Δηλαδή, η τελική κατανομή πιθανότητας «έχει κατασκευαστεί» διαδοχικά με όλο και περισσότερες δομικές λεπτομέρειες (δηλαδή πληροφορίες) να ενσωματώνονται. Όταν ένα κατάλληλο PDS εφαρμόζεται τότε μας βοηθάει στο να επιλέξουμε μια καλή δειγματοληπτική κατανομή $g_t(x_t \mid x_{t-1})$.

Είναι χρήσιμο να θυμηθούμε πως η μέθοδος ανάπτυξης χρησιμοποιήθηκε στη προσομοίωση ενός SAW. Στο βήμα t, ένα μονομερές έχει προστεθεί στην υπάρχουσα μερική αλυσίδα x_{t-1} , σύμφωνα με μια κατανομή που χρησιμοποιεί τις παραγόμενες πληροφορίες από ένα μελλοντικό βήμα (forward-looking). Η παραπάνω διαδικασία ισοδυναμεί με την χρήση της ακολουθίας των κατανομών, { $\pi_t(x_t)$ t=1,2,…}, όπως είναι το PDS, όπου π_t είναι η ομοιόμορφη κατανομή σε όλους τους SAW με t κόμβους και με $g_t(x_t \mid x_{t-1}) = \pi_t(x_t \mid x_{t-1})$. Δεν υπάρχει κάποιος λόγος να μην χρησιμοποιούμε γενικές πληροφορίες από περισσότερα μελλοντικά βήματα. Αλλά καθώς η στρατηγική του ενός μελλοντικού βήματος απαιτεί από εμάς να ελέγξουμε τρεις γειτονικές θέσεις του x_t , μια μέθοδος με κ μελλοντικά βήματα γενικά απαιτεί από εμάς να ελέγξουμε 3^{κ} θέσεις και αυτός ο υπολογισμός γρήγορα γίνεται μη πρακτικός. Η επόμενη ενότητα περιγράφει με λεπτομέρειες πως θα επιλέξουμε την δειγματοληπτική κατανομή.

4.4.1 Η επιλογή της δειγματοληπτικής κατανομής (The choice of the sampling distribution)

Η επιλογή της δειγματοληπτικής κατανομής, g_t, σχετίζεται άμεσα με την αποτελεσματικότητα της SIS. Σε πολλά προβλήματα μια καλή επιλογή της g_t με τη χρήση της ακολουθίας των βοηθητικών κατανομών $\{\pi_t\}$ είναι

$$g_t(x_t \mid x_{t-1}) = \pi_t(x_t \mid x_{t-1})$$

Με αυξητικό βάρος $u_t = \frac{\pi_t(x_{t-1})}{\pi_{t-1}(x_{t-1})}$

Σημαντική παρατήρηση είναι ότι το u_t δεν εξαρτάται από την τιμή του x_t. Είναι πιο επιθυμητό να παίρνουμε x_t από $\pi_t(x_t \mid x_{t-1})$ από ότι από μια αυθαίρετη συνάρτηση .

Ξαναγράφουμε το αυξητικό βάρος ως:

$$u_t = \frac{\pi_t(x_{t-1})}{\pi_{t-1}(x_{t-1})} \frac{\pi_t(x_t \mid x_{t-1})}{g_t(x_t \mid x_{t-1})}$$

Διαισθητικά η δεύτερη αναλογία χρειάζεται για να διορθώσει την ασυμφωνία μεταξύ $g_t(x_t \mid x_{t-1})$ και $\pi_t(x_t \mid x_{t-1})$ όταν αυτές είναι διαφορετικές.

Άλλες επιλογές της g_t μπορούν επίσης να είναι επιθυμητές. Για παράδειγμα μπορούμε επίσης να πάρουμε μια δειγματοληπτική κατανομή με δύο βήματα προς τα εμπρός (two-step-forward):

$$q_t(x_t \mid x_{t-1}) \propto \int \pi_{t+1}(x_{t+1}, x_t) \mathrm{d}x_{t+1}$$

Στην προσομοίωση πολυμερών, αυτό αντιστοιχεί σε στρατηγική πρόβλεψης δύο βημάτων.

Μια άλλη στρατηγική είναι ότι μπορούμε να ομαδοποιήσουμε τα $(x_{bt+1}, \ldots, x_{b(t+1)})$ ως μια μεγάλη κατάσταση x_t' στο σύστημα. Τότε η SIS εφαρμόζεται για να δημιουργήσουμε ένα πλέγμα (block) b των μεταβλητών x για κάθε χρονική στιγμή. Αυτή η μέθοδος έχει επίσης αποδειχθεί χρήσιμη στην προσομοίωση μακρών αλυσίδων πολυμερών.

4.4.2 Σταθερά κανονικοποίησης (Normalizing constant)

Σε πολλά επιστημονικά προβλήματα, ο υπολογισμός της σταθεράς κανονικοποίησης μιας μη κανονικοποιημένης συνάρτησης κατανομής πιθανότητας είναι πολύ σημαντικός. Υπάρχουν πάρα πολλά τέτοια παραδείγματα στην Φυσική και στην Χημεία. Μερικές φορές ακόμα και στα μαθηματικά και στη στατιστική οι επιστήμονες ενδιαφέρονται για τέτοια προβλήματα, ένα από τα οποία είναι για να μετρήσουν το συνολικό αριθμό Z των διακριτών πινάκων που περιέχουν μόνο 0 και 1 και έχουν σταθερό άθροισμα γραμμών $r_1, \ldots, r_m$ και στηλών $c_1, \ldots, c_n$.

Για παράδειγμα υπάρχουν 27 (Z=27) 0-1 πίνακες με άθροισμα γραμμών 2, 2, 2, 3 και άθροισμα στηλών 3, 2, 3, 1. Ένας τέτοιος τυπικός πίνακας 0-1 (από τους 27 που υπάρχουν) είναι

$$\begin{matrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{matrix}$$

$$\begin{matrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{matrix}$$

Αυτό το πρόβλημα μπορεί να τυποποιηθεί, σαν μια εκτίμηση της σταθεράς κανονικοποίησης Z, με ομοιόμορφη κατανομή του χώρου (όλων αυτών των πινάκων) να είναι του τύπου $\frac{1}{Z}$. Προβλήματα όπως η εκτίμηση πιθανοφάνειας και υπολογισμός του παράγοντα Bayes (Bayes factor) ανήκουν σε αυτή την κατηγορία.

Η προσέγγιση SIS μπορεί να είναι ένα αποτελεσματικό μέσο για εκτίμηση της σταθεράς κανονικοποίησης. Υποθέτουμε ότι η κατανομή στόχος είναι π(**x**). Ένα βοηθητικό PDS μπορεί να είναι $\Pi=\{\pi_t(x_t)\,, t=1,\dots,N\}$ (με $\pi_N \equiv \pi$) και κάθε κατανομή πιθανότητας, π_t , είναι γνωστή σαν μια σταθερά κανονικοποίησης. Μπορούμε να γράψουμε την μη κανονικοποιημένη κατανομή, $q_t(\boldsymbol{x_t}) = Z_t\pi_t(\boldsymbol{x}_t)$, στο PDS. Στη στατιστική συχνά το Z_t αντιστοιχεί στον παράγοντα Bayes ή στην συνάρτηση πιθανοφάνειας που έχει εκτιμηθεί με μια δεδομένη τιμή παραμέτρου.

Υποθέτουμε ότι έχουμε εφαρμόσει ένα πλαίσιο SIS με το σύστημα της διαδοχικής δειγματοληψίας $\{g_t(x_t \mid x_{t-1})\ t=1,\dots,N\}$. Στη συνέχεια το αυξητικό βάρος σπουδαιότητας u_t υπολογίζεται ως εξής:

$$u_t = \frac{q_t(x_t)}{q_{t-1}(x_{t-1})g_t(x_t \mid x_{t-1})} = \frac{Z_t\pi_t(\boldsymbol{x}_t),}{Z_{t-1}\pi_{t-1}(\boldsymbol{x}_{t-1})g_t(x_t \mid x_{t-1})}$$

Το τελικό βάρος θα είναι της μορφής:

$$w_N = \prod_{t=1}^{N} u_t = \frac{Z_2 Z_3 \dots Z_{N-1} Z_N}{Z_1 Z_2 \dots Z_{N-2} Z_{N-1}} \frac{\pi_N(x_N)}{g_1(x_1)g_2(x_2 \mid x_1)\dots g_N(x_N \mid x_{N-1})} =$$

$$= \frac{Z_N}{Z_1} \frac{\pi_N(x_N)}{g_1(x_1)g_2(x_2 \mid x_1)\dots g_N(x_N \mid x_{N-1})} \qquad (4.8)$$

Έτσι το μέσο δείγμα του w_N μας δίνει μια αμερόληπτη εκτίμηση του $\mathrm{E}(w_N) = \frac{Z_N}{Z_1}$, δηλαδή την αναλογία δύο σταθερών κανονικοποίησης. Σε πολλές περιπτώσεις, π_1 είναι ένα πολύ απλό σύστημα στο οποίο η Z_1 υπολογίζεται εύκολα, είτε αναλυτικά, είτε αριθμητικά. Στη συνέχεια μπορούμε να πάρουμε μια μάλλον ακριβή εκτίμηση του Z_N. Με βάση αυτήν την διαδικασία , ο Chen (2001) σχεδίασε έναν αλγόριθμο SIS για να προσεγγίσει τον συνολικό αριθμό των 0-1 πινάκων με δεδομένες περιθώριες τιμές. Η σχέση (4.8) είναι γενικά χρήσιμη για την εκτίμηση της σταθεράς κανονικοποίησης με τις ακόλουθες μεθόδους βελτίωσης.

Τα τελευταία χρόνια έχουν αναπτυχθεί πολλές τεχνικές για την βελτίωση της SIS που είναι βασισμένες στις μεθόδους για μη γραμμικά προβλήματα φιλτραρίσματος.

4.4.3 Περικοπή , εμπλουτισμός και αναδειγματοληψία (Pruning , enrichment and resampling)

Όσο το σύστημα μεγαλώνει, η διακύμανση των βαρών σπουδαιότητας w_t αυξάνει. Έτσι, μετά από ένα ορισμένο αριθμό βημάτων, πολλά βάρη γίνονται πολύ μικρά και μερικά άλλα βάρη πολύ μεγάλα.

Μια από τις πρώτες μεθόδους για τη βελτίωση της SIS διαδικασίας στις προσομοιώσεις πολυμερών, ονομάζεται μέθοδος εμπλουτισμού (enrichment) και προτάθηκε από τους Wall και Erpenbeck (1959). Η εργασία τους, σύμφωνα με τους Kremer και Binder (1988), οδήγησε σε μια από τις πρώτες ενδιαφέρουσες επιτυχίες των μεθόδων Monte Carlo. Οι Briefly, Wall και Erpenbeck πρότειναν τον διαχωρισμό αυτών των επιτυχημένων προσομοιωμένων μερικών αλυσίδων SAW, σε r αλυσίδες, που κάθε μια είναι $\frac{1}{r}$ του αρχικού βάρους. Αυτό μετά συνεχίζεται με r ανεξάρτητους τρόπους. Πιο συγκεκριμένα, από τη στιγμή που έχουμε επιτυχημένη προσομοίωση με την SIS μέθοδο, τότε, από μια αλυσίδα στο βήμα t, έστω $\boldsymbol{x}_t$, με βάρος w_t, μπορούμε να κάνουμε r_t αντίγραφα από αυτή και να δώσουμε σε καθένα, βάρος $\frac{w_t}{r_t}$. Μετά όλα τα αντίγραφα θα συνεχίζονται σαν να είχαν δημιουργηθεί ανεξάρτητα από το μηδέν. Ένα ανεπιθύμητο και αναπόφευκτο χαρακτηριστικό της μεθόδου του εμπλουτισμού είναι ότι οι ολοκληρωμένες αλυσίδες που προκύπτουν δεν είναι πλέον στατιστικά ανεξάρτητες.

Επειδή η επιλογή του r_t μπορεί να είναι ένα δύσκολο έργο γενικά, ο Grassberger (1997) εισήγαγε μια μέθοδο για το διαχωρισμό και τη περικοπή (pruning) της αλυσίδας. Αντί της κοπής κάθε αλυσίδας $\boldsymbol{x}_t$ αδιακρίτως, πρότεινε την χρήση ενός βήματος που εξαρτάται από την ανώτερη τιμή αποκοπής (cutoff) C_t και από μια χαμηλότερη τιμή αποκοπής c_t. Στη συνέχεια αν το βάρος w_t είναι μεγαλύτερο από την C_t, η αλυσίδα $\boldsymbol{x}_t$ διασπάται σε r αντίγραφα, που το καθένα έχει βάρος $\frac{w_t}{r}$, ενώ αν $w_t \leq c_t$, ρίχνουμε ένα αμερόληπτο κέρμα για να αποφασίσουμε αν θα κρατήσουμε το βάρος. Αν το δεχτούμε, το βάρος τους (των αντιγράφων) w_t, διπλασιάζεται σε $2w_t$. Ονόμασε τον αλγόριθμο «PERM» (δηλαδή Prune-Enriched Rosenbluth method). Ωστόσο η επιλογή της αποκοπής των τιμών C_t και c_t μπορεί επίσης να είναι δύσκολη.

Ο Αλγόριθμος PERM

C_t πάνω τιμή και c_t κάτω τιμή (κατώφλια)

- Αν $w_t > C_t$ τότε x_t διασπάται σε r αντίγραφα με βάρος $\frac{w_t}{r}$ το καθένα.

● Αν $w_t \leq c_t$ τότε ρίχνουμε ένα δίκαιο νόμισμα για να αποφασίσουμε αν θα κρατήσουμε το $\boldsymbol{x}_t$. Αν ναι, διπλασιάζεται το βάρος του σε $2w_t$.

Μια άλλη στρατηγική για τη βελτίωση της SIS είναι η μέθοδος της αναδειγματοληψίας που έχει αναπτυχθεί στην στατιστική και τη μηχανική κοινότητα (από τους Liu και Chen (1995), Gordon et al(1993)). Οι Liu και Chen (1995) εισήγαγαν μια παρακολουθούμενη διαδικασία αναδειγματοληψίας για την γενική μέθοδο SIS. Η μέθοδός τους παράγει ίδια αποτελέσματα εμπλουτισμού και περικοπής στα SIS δείγματα με αυτά του Grassberger (1997) και είναι αναμφισβήτητα πιο ευέλικτα από αυτά της μεθόδου PERM. Για την αντιμετώπιση των μοντέλων χώρου κατάστασης, οι Gordon et al (1993) επίσης χρησιμοποίησαν μια στρατηγική αναδειγματοληψίας. Στην πραγματικότητα ένα βήμα αναδειγματοληψίας επιβάλλεται σε κάθε βήμα t του φίλτρου Bootstrap. Η επιτυχία αυτής της μεθόδου βασίζεται στη Μαρκοβιανή δομή των μεταβλητών κατάστασης $x_1, x_2 \ldots$, δηλαδή δεδομένης μιας πραγματοποίησης του x_t, η επόμενη μεταβλητή x_{t+1}, είναι στατιστικά ανεξάρτητη από όλες τις προηγούμενες καταστάσεις x_{t-1}. Έτσι κάνοντας αναδειγματοληψία από το σύνολο $\{\boldsymbol{x_{t-1}^j}, j = 1, \ldots, m\}$, είναι ισοδύναμο με αναδειγματοληψία από $\{x_{t-1}^j, j = 1, \ldots, m\}$, που είναι το σύνολο των τρεχουσών καταστάσεων. Η συχνή αναδειγματοληψία μπορεί γρήγορα να αποδυναμώσει την ποικιλία των μερικών δειγμάτων που δημιουργήθηκαν προηγουμένως.

4.4.4 Περισσότερα για την αναδειγματοληψία

Υποθέτουμε ότι στο βήμα t έχουμε ένα σύνολο από m μερικά δείγματα μεγέθους t, $S_t = \{{x_t}^j, j = 1, \ldots, m\}$, που είναι κατάλληλα σταθμισμένα από το σύνολο των βαρών $W_t = \{w_t^j, j = 1, \ldots, m\}$ σε σχέση με το PDS π_t. Καθένα από αυτά τα μερικά δείγματα , ${x_t}^j$, μπορεί να ονομάζεται ροή (stream). Αντί να μεταφέρουμε τα βάρη w_t^j, καθώς το σύστημα εξελίσσεται, μπορούμε να εισάγουμε το επόμενο βήμα αναδειγματοληψίας μεταξύ SIS αναδρομών (Liu και Chen (1998). Μια τέτοια διαδικασία αναφέρεται ως SIS με αναδειγματοληψία.

Τα επόμενα δύο συστήματα είναι μόνο δύο από τις πολλές πιθανές επιλογές για να επιτύχουμε τον στόχο της αναδειγματοληψίας.

●Απλή τυχαία αναδειγματοληψία

1.) Προσομοιώνουμε ένα καινούργιο σύνολο ροών (δηλαδή μερικά δείγματα) που συμβολίζεται σαν $S_t^{'}$, από S_t σύμφωνα με τα βάρη w_t^j

2) Αναθέτουμε ίσα βάρη W_t/m στις ροές στο σύνολο $S_t^{'}$ όπου

$W_t = w_t{}^1 + \cdots + w_t^m$

●Υπόλοιπο αναδειγματοληψίας

1) Διατηρούμε $k_j = [mw_t^{*j}]$ αντίγραφα των $x_t{}^j$, όπου $w_t^{*j} = w_t^j/W_t$ και $j = 1, \dots, m$. Θέτουμε $m_r = m - k_1 - \cdots - k_m$

2) Παίρνουμε m_r i.i.d δείγματα από το σύνολο S_t με πιθανότητες ανάλογες του $mw_t^{*j} - k_j$

$j = 1, \dots, m.$

3) Επαναφέρουμε όλα τα βάρη σε W_t/m .

Μπορούμε επίσης να χρησιμοποιήσουμε το μέσο όρο των τελικών βαρών για να εκτιμήσουμε την αναλογία Z_n/Z_1 δύο σταθερών κανονικοποίησης ακόμα και μετά από μερικά βήματα αναδειγματοληψίας που θα έχουν πραγματοποιηθεί. Αυτός είναι ο κύριος λόγος που εμείς θέτουμε τα βάρη όλων των ροών στο τρέχοντα μέσο όρο τους μετά από κάθε βήμα αναδειγματοληψίας. Το υπόλοιπο αναδειγματοληψίας υπερτερεί της απλής τυχαίας δειγματοληψίας γιατί έχει μικρότερη Monte Carlo διακύμανση, καλύτερο υπολογιστικό χρόνο και δεν παρουσιάζει μειονεκτήματα.

Μπορούμε να δούμε ότι το σύστημα υπολοίπου αναδειγματοληψίας είναι παρόμοιο με τη μέθοδο εμπλουτισμού και περικοπής (PERM) του Grassberger (Wall και Erpenbeck 1959, Grassberger 1997). Και ο αλγόριθμος PERM, αλλά και το σύστημα αναδειγματοληψίας, δημιουργούν ένα κατάλληλο σύνολο από δείγματα σπουδαιότητας τα οποία μπορούν να χρησιμοποιηθούν στην εκτίμηση όλων των ποσοτήτων που μας ενδιαφέρουν σε σχέση με τη π_t (κατανομή στόχος). Ο αλγόριθμος PERM θα είναι κατάλληλος για όσο διάστημα οι τιμές αποκοπής C_t και c_t προβλέπονται εκ των προτέρων. Η αναδειγματοληψία δεν προκαλεί πρόβλημα γιατί η απόφαση για αναδειγματοληψία δεν εξαρτάται από τις διαμορφώσεις. Ένα σημαντικό πλεονέκτημα της μεθόδου αναδειγματοληψίας σε σχέση με τον PERM είναι ότι η επιλογή των τιμών αποκοπής για τον διαχωρισμό των ροών επιτυγχάνεται αυτόματα από την σύγκριση των διασταυρούμενων ροών των βαρών. Πιο συγκεκριμένα στο υπόλοιπο αναδειγματοληψίας, χωρίζουμε μια ροή σε κ περισσότερα αντίγραφα ακριβώς επειδή το βάρος τους είναι κ φορές καλύτερο από το μέσο βάρος. Αυτό το χαρακτηριστικό κάνει την μέθοδο αναδειγματοληψίας πολύ εφαρμόσιμη και ισχυρή γενικά. Η αναδειγματοληψία έχει μεγάλο πλεονέκτημα στις εφαρμογές που αφορούν την πληθυσμιακή γενετική και την βιολογία.

Αντί να χρησιμοποιήσουμε το w_t στην αναδειγματοληψία μπορούμε να εφαρμόσουμε την παρακάτω πιο γενική στρατηγική αναδειγματοληψίας.

● Για $j' = 1, \dots, \tilde{m}$:

- Προσομοιώνουμε $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}^{j'}$ ανεξάρτητα από το τρέχον δείγμα $\{\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{j}}, j = 1, \dots, m\}$ σύμφωνα με το διάνυσμα πιθανότητας ($\alpha^1 \dots . \alpha^m$) . Υποθέτουμε ότι έχουμε $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}^{j'} = \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{j}}$

- Ένα καινούργιο βάρος $\tilde{w}_t^{j'} = w_t^j \,/\, \alpha^j$ έχει εκχωρηθεί(is assigned) σε αυτό το δείγμα.

● Επιστρέφουμε το νέο σύνολο $\tilde{S}_t = \{\tilde{\boldsymbol{x}}_{\boldsymbol{t}}^{j'}, j' = 1, \dots, \tilde{m}\}$ και $\tilde{W}_t = \{\tilde{w}_t^{j'}, j' = 1, \dots, \tilde{m}\}$

Το καινούργιο σύνολο που σχηματίζεται είναι επίσης κατάλληλα σταθμισμένο από $\tilde{W}_t$ (προσεγγιστικά) σε σχέση με τη κατανομή στόχο π (Rubin 1987). Επειδή ο ρόλος της αναδειγματοληψίας είναι να περικόπτουμε (prune away) τα «κακά» δείγματα και να χωρίζουμε τα καλά, θα πρέπει να επιλέξουμε α^j σαν μια μονότονη συνάρτηση του w_t^j. Μια γενική επιλογή είναι :

$$\alpha^j = [w_t^j]^\alpha$$

όπου $0 < \alpha \leq 1$ μπορεί να ποικίλει ανάλογα με τον συντελεστή μεταβολής του βάρους w_t. Ο καθηγητής W.H. Wong πρότεινε την επιλογή $\alpha = 1/2$ που φαίνεται να δουλεύει καλά σε πολλά παραδείγματα.

Το πρόγραμμα αναδειγματοληψίας μπορεί να είναι είτε δυναμικό είτε ντετερμινιστικό. Όταν το πρόγραμμα είναι δυναμικό μπορούμε να εισάγουμε μια μικρή μεροληψία, η οποία μέσα από εφαρμογές φαίνεται ότι δεν παράγει κάποια αρνητικά αποτελέσματα. Με ένα ντετερμινιστικό πρόγραμμα μπορούμε να κάνουμε αναδειγματοληψία σε χρόνο $t_0, 2t_0, \dots$, όπου t_0 δίνεται εκ των προτέρων και μπορεί να ρυθμιστεί έτσι ώστε να ταιριάζει με το συγκεκριμένο πρόβλημα που μας ενδιαφέρει. Σε ένα δυναμικό πρόγραμμα μια ακολουθία κατωφλίων $c_1, c_2, \dots$. δίνεται εκ των προτέρων. Ελέγχουμε τον συντελεστή διασποράς των βαρών ${cv_t}^2$ και επικαλούμαστε το βήμα αναδειγματοληψίας όταν έχουμε ${cv_t}^2 > c_t$. Μια τυπική ακολουθία κατωφλίων c_t μπορεί να είναι $c_t = \alpha + bt^\alpha$.

Συνοψίζουμε το σύστημα αναδειγματοληψίας στον παρακάτω αλγόριθμο:

1. Ελέγχουμε την κατανομή βάρους, εκτελώντας μια από τις παρακάτω δύο μεθόδους σε χρόνο t:

● Δυναμική μέθοδος : πηγαίνουμε στο βήμα 2 αν ο συντελεστής του βάρους είναι μέτριος (modest) {δηλαδή $cv_t{}^2 < c_t$} διαφορετικά πηγαίνουμε στο βήμα 3.

● Ντετερμινιστική μέθοδος: πηγαίνουμε στο βήμα 2 αν t≠ kt_0 για ακέραιο αριθμό k , διαφορετικά πηγαίνουμε στο βήμα 3.

2. Παρεμβάλουμε ένα βήμα SIS. Θέτουμε t=t+1 και πηγαίνουμε στο βήμα 1

3. Παρεμβάλουμε ένα βήμα αναδειγματοληψίας. Πηγαίνουμε στο βήμα 2

Δεν είναι άμεσα σαφές για ποιο λόγο πρέπει να χρησιμοποιούμε αναδειγματοληψία σε κάποιο στάδιο t. Για αυτό θα αναφέρουμε διαισθητικά μερικούς από αυτούς τους λόγους.

1) Αν τα βάρη w_t^j είναι σταθερά (ή σχεδόν σταθερά) για όλα τα t (μια τέτοια περίπτωση συμβαίνει όταν μπορούμε να πάρουμε από την π_t κατευθείαν), η αναδειγματοληψία μειώνει μόνο τον αριθμό των διακριτών ροών και εισάγει επιπλέον Monte Carlo διακύμανση. Αυτό υποδηλώνει ότι δεν πρέπει να κάνουμε αναδειγματοληψία όταν ο συντελεστής διασποράς [όπως ορίζεται στην (3.5)], $cv_t{}^2$, για w_t^j είναι μικρός. Όπως υποστήριξαν οι Kong at al.(1994), το αποτελεσματικό μέγεθος δείγματος είναι αντιστρόφως ανάλογο με το 1+ $cv_t{}^2$. 2) Μπορούμε να δείξουμε ότι όσο το σύστημα εξελίσσεται, ο συντελεστής διασποράς $cv_t{}^2$ αυξάνεται στοχαστικά (Kong at al.(1994)). Όταν τα βάρη γίνουν πολύ ασύμμετρα σε χρόνο t , δηλαδή μεταφέρουν πολλές ροές με πολύ μικρά βάρη, τότε τα απορρίπτουμε. Η αναδειγματοληψία μπορεί να παρέχει αλλαγές για τις καλές (δηλαδή σημαντικές) ροές για να τις ενισχύσει και ως εκ τούτου να ανανεώσει τον δειγματολήπτη. Το βήμα της αναδειγματοληψίας τείνει να προκαλέσει μια καλύτερη ομάδα από «προγόνους» έτσι ώστε να παράγει καλύτερους «απογόνους» (δηλαδή Monte Carlo δείγματα από μελλόντικές καταστάσεις) , αν και δεν βελτιώνει τα συμπεράσματα της τρέχουσας κατάστασης $\boldsymbol{x_t}$.

4.4.5 Μερικός έλεγχος απόρριψης

Καθώς το t αυξάνεται, η κατανομή του βάρους w_t τυπικά γίνεται όλο και πιο ασύμμετρη, που δείχνει ότι η x^2 απόσταση μεταξύ της δειγματοληπτικής κατανομής του $\boldsymbol{x_t}$ και της τρέχουσας κατανομής π_t αυξάνεται. (Αυτή η απόσταση είναι ισοδύναμη με τον συντελεστή διακύμανσης του w_t). Αυτό έχει ως συνέπεια πολλές ροές που μεταφέρονται από την SIS να επηρεάζουν ελάχιστα την τελική εκτίμηση. Για αυτό θα πρέπει αυτές τις ροές να τις περικόπτουμε σε ένα

προγενέστερο στάδιο. Στις δύο προηγούμενες ενότητες είδαμε τις ιδέες του αλγόριθμου PERM και της αναδειγματοληψίας. Στην ενότητα 3.6.4 είδαμε πως η μέθοδος ελέγχου απόρριψης μπορεί να χρησιμοποιηθεί για να πετύχει περικοπή χωρίς να δημιουργήσει μεροληψία ή συσχετισμούς. Ωστόσο, η εφαρμογή ενός πλήρους ελέγχου απόρριψης απαιτεί να προσομοιώσουμε τις χαμένες ροές με επανεκκίνηση από το στάδιο 1 [δηλαδή να προσομοιώνουμε από την δοκιμαστική κατανομή g_1 και να προχωράμε προς τα εμπρός] και να περάσουμε από όλα τα ενδιάμεσα βήματα απόρριψης (δηλαδή σημεία ελέγχου). Αυτή η διαδικασία μπορεί να είναι εξαιρετικά χρονοβόρα και δεν είναι πολύ πρακτική για ένα μεγάλο πιθανοθεωρητικό σύστημα.

Αντί να ασχοληθούμε με τον πλήρη έλεγχο απόρριψης μπορούμε να επιλέξουμε την παρακάτω πιο πρακτική μέθοδο, δηλαδή το μερικό έλεγχο απόρριψης :

1. Σε κάθε σημείο ελέγχου t_κ αρχίζουμε $\mathrm{RC}(t_\kappa)$ όπως περιγράφεται στην ενότητα 3.6.4 με τιμή κατωφλίου $\mathrm{c} = c_\kappa$. Αν η ροή $\boldsymbol{x}_{t_\kappa}{}^{\boldsymbol{j}}$ με βάρος $w_{t_\kappa}{}^{\boldsymbol{j}}$ περνάει αυτό το σημείο ελέγχου, προχωρούμε με τον ίδιο τρόπο όπως σε ένα πλαίσιο SIS με το παλιό βάρος να αντικαθίσταται από

   $$w_{t_\kappa} *^{\boldsymbol{j}} = \max\{w_{t_\kappa}{}^{j},\ c_\kappa\}$$

2. Όταν απορρίπτεται, πηγαίνουμε πίσω στο σημείο ελέγχου $t_{\kappa-1}$ για να πάρουμε μια ροή $\boldsymbol{x}_{t_{\kappa-1}}{}^{\boldsymbol{j}}$ από το σύνολο $S_{t_{\kappa-1}}$, με πιθανότητα ανάλογη προς το $w_{t_{\kappa-1}}{}^{\boldsymbol{j}}$. Επαναφέρουμε το βάρος του σαν $\widetilde{w}_{t_{\kappa-1}}$ και προχωρούμε με την διαδικασία SIS . Αν η καινούργια ροή που σχηματίζεται με αυτό τον τρόπο περνάει το σημείο ελέγχου t_κ , τότε το βάρος του ορίζεται ως

   $$w_{t_\kappa} *^{\boldsymbol{j}} = \max\{w_{t_\kappa}{}^{j},\ c_\kappa\}$$

3. Επαναφέρουμε όλα τα βάρη σαν $w_{t_\kappa}{}^{\boldsymbol{j}} = \hat{p}_c w_{t_\kappa} *^{\boldsymbol{j}}$

Η μέθοδος του μερικού ελέγχου απόρριψης συνδυάζει τις τρεις βασικές ιδέες που χρησιμοποιούνται στον σχεδιασμό αλγορίθμων της διαδοχικής μεθόδου Monte Carlo : δηλαδή στάθμιση, απόρριψη και αναδειγματοληψία.

4.4.6 Περιθωριοποίηση, πρόβλεψη (Marginalization , look- ahead)

Μια γενική μέθοδος για τη βελτίωση της απόδοσης ενός SIS αλγορίθμου είναι η περιθωριοποίηση. Αυτή είναι μια αναλυτική μέθοδος που όταν

επιτευχθεί, μπορεί να βελτιώσει σχεδόν όλες τις μεθόδους Monte Carlo. Η γενική της αρχή περιγράφτηκε από τους Hammersley, Handscomb (1964) και Rubinstein (1981) σαν μια τεχνική μείωσης διάστασης. Στην ενότητα 3.5.5 είδαμε τη χρήση της περιθωριοποίησης για τη βελτίωση μιας στατιστικής μεθόδου δειγματοληψίας σπουδαιότητας .

Μια άλλη χρήσιμη στρατηγική στην κατασκευή της δειγματοληπτικής κατανομής είναι η πρόβλεψη για μερικά βήματα. Αυτή η μέθοδος έχει επιτυχία για προσομοιώσεις πολυμερών και για μη γραμμικά μοντέλα χώρου καταστάσεων. Πιο συγκεκριμένα, κατασκευάζουμε την δειγματοληπτική κατανομή όσο πιο κοντά γίνεται στον μελλοντικό στόχο. Μια τέτοια κατασκευή είναι να θέσουμε $g_t(x_t \quad | x_{t-1}) = \pi_{t+s}(x_t \quad | x_{t+1})$.

Σημειώνουμε ότι $\pi_{t+s}(x_t \quad | x_{t-1}) = \int \pi_{t+s}(x_{t+s}, \dots, x_t \quad | x_{t-1}) d\, x_{t+1} \dots dx_{t+s}$.

Αν θεωρούμε κάθε $\pi_t(\boldsymbol{x})$ σαν μια των εκ των υστέρων κατανομή σύμφωνα με την πληροφορία στον χρόνο t, τότε η παραπάνω προσέγγιση προσπαθεί να χρησιμοποιήσει όσο το δυνατόν περισσότερες μελλοντικές πληροφορίες. Για παράδειγμα αν είχαμε επιλέξει s έτσι ώστε t+s=n, τότε η προκύπτουσα δειγματοληπτική κατανομή g_t από την πρόβλεψη (forward-looking) είναι στην πραγματικότητα η ιδανική δεσμευμένη κατανομή για x_t κάτω από την κατανομή στόχο (έτσι ώστε η g_t να είναι ίση με π_n).

Κεφάλαιο 5

Εφαρμογή της ακολουθιακής μεθόδου SIS

5.1 Το πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης από ελλιπή δεδομένα

Εδώ θα επανεξετάσουμε το πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης, Σ , μιας διδιάστατης κανονικής κατανομής από ελλιπή δεδομένα , το οποίο περιγράφτηκε στην ενότητα 3.5.6. Η προηγούμενη προσέγγιση βασίστηκε στη δειγματοληψία σπουδαιότητας στην οποία ως δοκιμαστική κατανομή επιλέχθηκε η εκ των υστέρων κατανομή του Σ βασισμένη στις πλήρεις παρατηρήσεις. Εδώ θα εξετάσουμε μια προσέγγιση SIS.

Υπενθυμίζουμε ότι ο πίνακας συνδιακύμανσης αποδίδεται από την εκ των προτέρων μη πληροφοριακή κατανομή του Jeffrey $\pi(\Sigma) \propto |\Sigma|^{-\frac{d+1}{2}}$. Η εκ των υστέρων κατανομή του Σ δοθέντων των πλήρων δεδομένων είναι :

$$p(\mathrm{Z} \mid complete\ data) \propto |Z|^{\frac{3}{2}}\ exp\left\{-\frac{1}{2}tr[Z \cdot S]\right\}$$

όπου $\mathrm{S}=(s_{ij})_{2x2}$ είναι το μη διορθωμένο άθροισμα του τετραγωνικού πίνακα και $\mathrm{Z}=\Sigma^{-1}$.

Έστω ότι τα πλήρη στοιχεία είναι $y_1 \dots y_{12}$, όπου $y_t = (\ y_{t,1}, y_{t,2})$ για t=1,…,12. Επιπλέον, στα $y_{t,2}$ λείπουν τα στοιχεία για t=5,…,8 ενώ στα $y_{t,1}$ λείπουν για t=9,..12. Η κατανομή πρόβλεψης του y_{t+1} δεδομένου του $\boldsymbol{y}_t = (y_1 \dots y_t)$ είναι:

$$y_{t+1} \mid \boldsymbol{y}_t \sim t_2\ (\ 0\ , \frac{S_{\mathrm{t}}}{\mathrm{t}-1}, \mathrm{t}-1)$$

όπου $S_t = (s_{ij})$, $s_{ij} = \sum_{s=1}^{t} y_{s,i} y_{s,j}$ και t_2 είναι η διμεταβλητή t- κατανομή. Συνάγεται ότι, για $\boldsymbol{y}_t = (y_1 \dots y_t)$, η περιθώρια κατανομή και η δεσμευμένη κατανομή είναι :

$$y_{t+1,1} \mid \boldsymbol{y}_t \sim t_1\ (\ 0\ , \frac{s_{11}}{\mathrm{t}-1}, \mathrm{t}-1)$$

$$y_{t+1,2} \mid \boldsymbol{y}_t, y_{t+1,1} \sim t_1\ [\ \frac{s_{12}}{s_{11}} y_{t+1,1}\ , \frac{|S_{\mathrm{t}}|}{\mathrm{t}s_{11}}(1 + \frac{y^2{}_{t+1,1}}{s_{11}}),\ \mathrm{t}\ ]$$

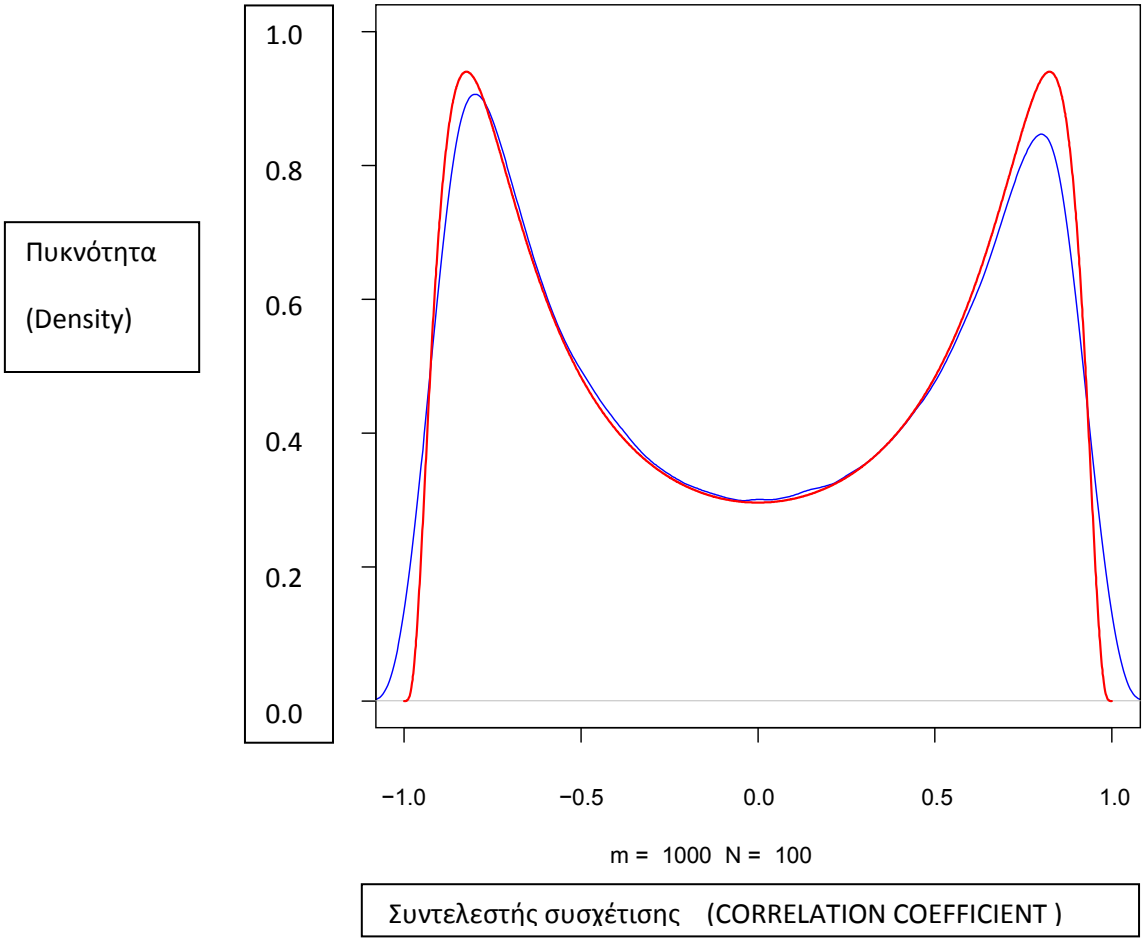
αντίστοιχα. Παρόμοια αποτελέσματα μπορούμε να πάρουμε από δεσμευμένες κατανομές $[y_{t+1,2} \mid \boldsymbol{y}_t]$ και $[y_{t+1,1} \mid \boldsymbol{y}_t, y_{t+1,2}]$. Το σχήμα 5.1 μας δίνει μια προσέγγιση της εκ των υστέρων κατανομής του συντελεστή συσχέτισης ρ του δείγματος για δεδομένα του Murray.

Η προσέγγιση είναι ένα σταθμισμένο μείγμα m=1000 εκ των υστέρων κατανομών από πλήρη δεδομένα.

Η εκ των υστέρων κατανομή έχει τη μορφή :

$$p(\rho \mid \boldsymbol{y}) \propto (1-\rho^2)^{\frac{\nu-2}{2}} \int_0^\infty \omega^{-1} \left(\omega + \frac{1}{\omega} - 2\rho r\right)^{-(\nu+1)} d\omega \quad (5.1)$$

όπου $\boldsymbol{y}$ είναι τα πλήρη δεδομένα μετά την υποκατάσταση του μέρους που λείπει (being imputed), r είναι ο συντελεστής συσχέτισης του δείγματος και ν=n-2 =10 .



Σχήμα 5.1 : Μια προσέγγιση της εκ των υστέρων κατανομής του συντελεστή συσχέτισης ρ του δείγματος για δεδομένα του Murray.

Παρατηρούμε από το σχήμα ότι ο συντελεστής συσχέτισης κατανέμεται συμμετρικά ως προς έναν νοητό άξονα που περνάει κάθετα από το 0.0 που βρίσκεται στο μέσο του άξονα του συντελεστή συσχέτισης. Υπάρχει θετική και αρνητική συσχέτιση συμμετρική.

Πιο αναλυτικά το πρόβλημα του Murray με ελλιπή δεδομένα

Ενδιαφερόμαστε για τον πίνακα συνδιακύμανσης Σ συναρτήσει των y_{obs} και των y_{mis}. Για το λόγο αυτό θα χρησιμοποιήσουμε την μέθοδο SIS.

Συγκεκριμένα**:**

Β1 Προσομοιώνουμε μόνο τον πίνακα που περιέχει τα y_{mis} επαγωγικά

ως εξής : $y_{t+1,1} \mid \boldsymbol{y}_t \sim t_1 \left(0, \frac{s_{11}}{t-1}, t-1\right)$

$$y_{t+1,2} \mid \boldsymbol{y}_t, y_{t+1,1} \sim t_1 \left[\frac{s_{12}}{s_{11}} y_{t+1,1}, \frac{|S_t|}{t s_{11}} \left(1 + \frac{y^2_{t+1,1}}{s_{11}}\right), t \right]$$

Β2 Υπολογίζουμε το βάρος $w_{t+1} = w_t \frac{f(y_{t+1,1}) f(y_{t+1,2})}{g(y_{t+1,1}) g(y_{t+1,2})}$

όπου $f(y_{t+1,1}) \sim N(\mu_1, \sigma_1^2)$

$f(y_{t+1,2}) \sim N(\mu_2, \sigma_2^2)$

$g(y_{t+1,1}) \sim t_1\left(0, \frac{s_{11}}{t-1}, t-1\right)$

$g(y_{t+1,2}) \sim t_1\left[\frac{s_{12}}{s_{11}} y_{t+1,1}, \frac{|S_t|}{t s_{11}} \left(1 + \frac{y^2_{t+1,1}}{s_{11}}\right), t \right]$

όπου μ_1, σ_1^2 είναι ο μέσος και η διασπορά αντίστοιχα της πρώτης γραμμής του πίνακα των δεδομένων και μ_2, σ_2^2 είναι ο μέσος και η διασπορά αντίστοιχα της δεύτερης γραμμής του πίνακα των δεδομένων.

Επαναλαμβάνουμε τα παραπάνω βήματα m φορές και στο τέλος υπολογίζουμε το ζητούμενο $\hat{\Sigma}$ ως εξής :

$$\hat{\Sigma} = \frac{w_1 \Sigma_1 + \cdots + w_m \Sigma_m}{w_1 + \cdots + w_m}$$

Στη συνέχεια παρουσιάζουμε αποτελέσματα από 2 σετ δεδομένων για το παραπάνω πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης διδιάστατης κανονικής κατανομής από ελλιπή δεδομένα.

(1) Αποτελέσματα για δεδομένα Murray

Χρησιμοποιούμε τα δεδομένα του Murray (1977) για n=12 παρατηρήσεις με 8 ελλιπή δεδομένα από μια διδιάστατη κανονική κατανομή με μέσο $\binom{0}{0}$ και άγνωστο πίνακα συνδιακύμανσης.

Με την μέθοδο SIS προσομοιώνουμε τα ελλιπή δεδομένα m=10000 φορές και παίρνουμε:

● τον εκτιμώμενο πίνακα συνδιακύμανσης

$$\hat{\Sigma}= \begin{bmatrix} 1.0434 & 0.0004 \\ 0.0004 & 1.7407 \end{bmatrix}$$

● τον εκτιμώμενο συντελεστή συσχέτισης $\hat{r}=$ 3.2304e-004 και

● τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι $s_1=$ 0.2949 και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι $s_2=$0.3809.

Παρατηρούμε ότι: ο εκτιμώμενος συντελεστής συσχέτισης είναι περίπου μηδέν $\hat{r} \approx 0$ (από τον πίνακα συνδιακύμανσης φαίνεται ότι η συνδιακύμανση είναι 0,0004 δηλαδή περίπου μηδέν).

(2) Προσομοιωμένα δεδομένα για n=100 παρατηρήσεις με 30 ελλιπή δεδομένα από μια διδιάστατη κανονική κατανομή με μέσο $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, πίνακα συνδιακύμανσης $\begin{pmatrix} 1 & 0.9899 \\ 0.9899 & 2 \end{pmatrix}$ και συντελεστή συσχέτισης r=0,7.
Με την μέθοδο SIS προσομοιώνουμε τον πίνακα συνδιακύμανσης m=10000 φορές και παίρνουμε:

● τον εκτιμώμενο πίνακα συνδιακύμανσης

$$\hat{\Sigma}= \begin{bmatrix} 1.0353 & 0.9265 \\ 0.9265 & 1.9592 \end{bmatrix}$$

σύμφωνα με τον τύπο $\hat{\Sigma} = \frac{w_1\Sigma^1+\cdots.+w_m\Sigma^m}{w_1+\cdots+w_m}$

● τον εκτιμώμενο συντελεστή συσχέτισης $\hat{r}=$0.6506 και

● τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι $s_1=$ 0.1017 και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι $s_2=$0.1400.

Σχόλιο: Οι εκτιμήσεις με SIS είναι πιο ακριβείς σε σχέση με απλό IS (δειγματοληψία σπουδαιότητας).

Το πρόγραμμα για την εξαγωγή των παραπάνω αποτελεσμάτων παρατίθεται σε παράρτημα.

5.2 Σύγκριση μεθόδου SIS με το δειγματολήπτη Gibbs

Εδώ θα συγκρίνουμε τις δύο μεθόδους επίλυσης του προβλήματος εκτίμησης του πίνακα συνδιακύμανσης από ελλιπή δεδομένα. Τη μέθοδο SIS την περιγράψαμε στην παραπάνω ενότητα 5.1 και υπενθυμίζουμε και τη μέθοδο Gibbs για να κάνουμε τη σύγκριση.

Πιο αναλυτικά το πρόβλημα με ελλιπή δεδομένα με τη μέθοδο Gibbs

Και σε αυτή τη μέθοδο το μέρος του y που παρατηρείται το ονομάζουμε y_{obs} και το υπόλοιπο μέρος του y που λείπει το ονομάζουμε y_{mis}. Θέτουμε y=(y_{obs} , y_{mis}).

Η πιθανοφάνεια ελλιπών δεδομένων $f(y_{obs} \mid \Sigma) = \int f(y_{obs}, y_{mis} \mid \Sigma)\, dy_{mis}$ είναι δύσκολη στο χειρισμό για αυτό εισάγουμε νέες τυχαίες μεταβλητές y_{mis} οι οποίες εκφράζουν τα δεδομένα που λείπουν και έχουμε την πιθανοφάνεια πλήρους σετ δεδομένων $f(y_{obs}, y_{mis} \mid \Sigma)$ που είναι εύκολη στο χειρισμό. Η από κοινού posterior κατανομή των παραμέτρων και των ελλιπών δεδομένων θα είναι

$$f(\Sigma, y_{mis} \mid y_{obs}) \propto f(\Sigma) f(y_{obs}, y_{mis} \mid \Sigma)$$

- Προσομοιώνουμε y_{mis} από $f(y_{mis} \mid y_{obs}, \Sigma)$
- Προσομοιώνουμε Σ από $f(\Sigma \mid y_{obs}, y_{mis})$

Έτσι τα βήματα της μεθόδου Gibbs είναι:

Αυθαίρετες αρχικές τιμές για ${y^0}_{mis}$ και Σ^0

Β0) Θέτουμε Σ^0=cov(y_{obs})

Β1) Προσομοιώνουμε

Β1) Προσομοιώνουμε

${y^1}_{mis} \sim f({y^1}_{mis} \mid y_{obs}, \Sigma^0)$

Η κατανομή πρόβλεψης του $\boldsymbol{y_{mis}}$, είναι γινόμενο από κανονικές κατανομές.

Για παράδειγμα $[y_{5,2} \mid \Sigma, \boldsymbol{y_{obs}}] = [y_{5,2} \mid \Sigma, y_{5,1}] \sim N(\mu_*, \sigma_*^2)$

όπου $\mu_* = \rho y_{5,1} \sqrt{\sigma_{11}/\sigma_{22}}$ και $\sigma_*^2 = (1-\rho^2)\,\sigma_{11}$

Σ^1=cov(y_{obs}, ${y^1}_{mis}$)

…

${y^n}_{mis} \sim f({y^n}_{mis} \mid y_{obs}, \Sigma^{n-1})$

$\Sigma^n = \text{cov}(y_{obs}, y^n{}_{mis})$

Τέλος παίρνουμε $\hat{\Sigma} = \frac{\Sigma^1 + \cdots + \Sigma^n}{n}$.

Στη συνέχεια παρουσιάζουμε αποτελέσματα από 2 σετ δεδομένων για το παραπάνω πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης διδιάστατης κανονικής κατανομής από ελλιπή δεδομένα.

(1)Αποτελέσματα για δεδομένα Murray

Χρησιμοποιούμε πάλι τα δεδομένα του Murray (1977) για n=12 παρατηρήσεις με 8 ελλιπή δεδομένα από μια διδιάστατη κανονική κατανομή με μέσο $\begin{pmatrix}0\\0\end{pmatrix}$ και άγνωστο πίνακα συνδιακύμανσης.

Με την μέθοδο Gibbs προσομοιώνουμε τις άγνωστες ποσότητες του προβλήματος m=10000 φορές (με burn-in period 1000) και παίρνουμε:

- τον εκτιμώμενο πίνακα συνδιακύμανσης

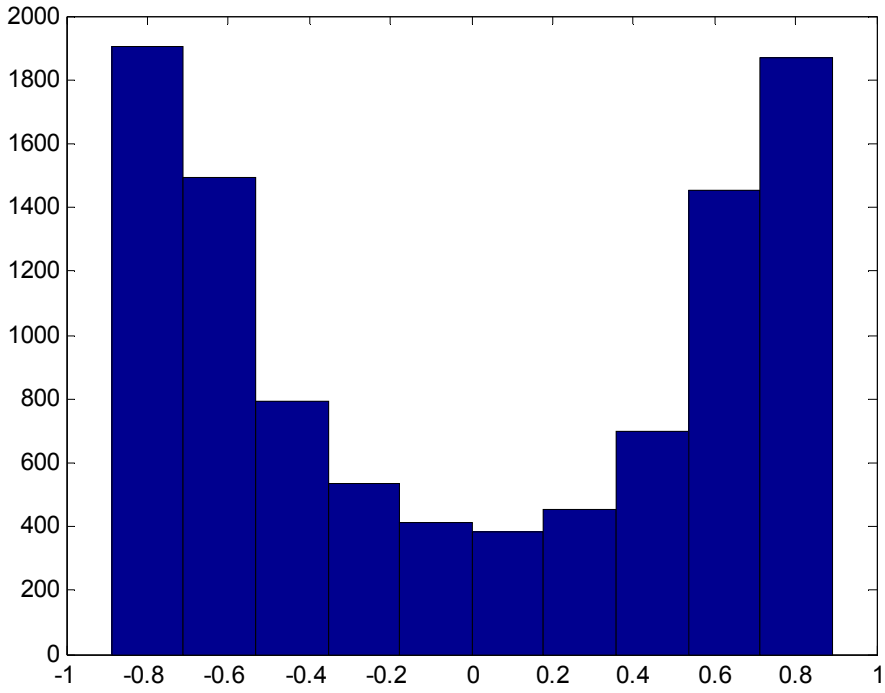
$$\hat{\Sigma} = \begin{bmatrix} 3.0637 & -0.0240 \\ -0.0240 & 3.0221 \end{bmatrix}$$

όπου $\hat{\Sigma} = \frac{\Sigma^1 + \cdots + \Sigma^n}{n}$

- τον εκτιμώμενο συντελεστή συσχέτισης $\hat{r} = -0.0079$

- τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι $s_1 = 0.5053$ και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι $s_2 = 0.5018$.

- και το παρακάτω ιστόγραμμα για το δείγμα MCMC για τον συντελεστή συσχέτισης



Ιστόγραμμα συντελεστή συσχέτισης με συχνότητα

Παρουσιάζεται για m=10000 φορές η συχνότητα εμφάνισης στο δείγμα MCMC των τιμών του συντελεστή συσχέτισης .

Σχόλιο: Αυτό το γράφημα είναι ίδιο με το προηγούμενο 5.1 γράφημα που είχαμε για τη μορφή της posterior του p

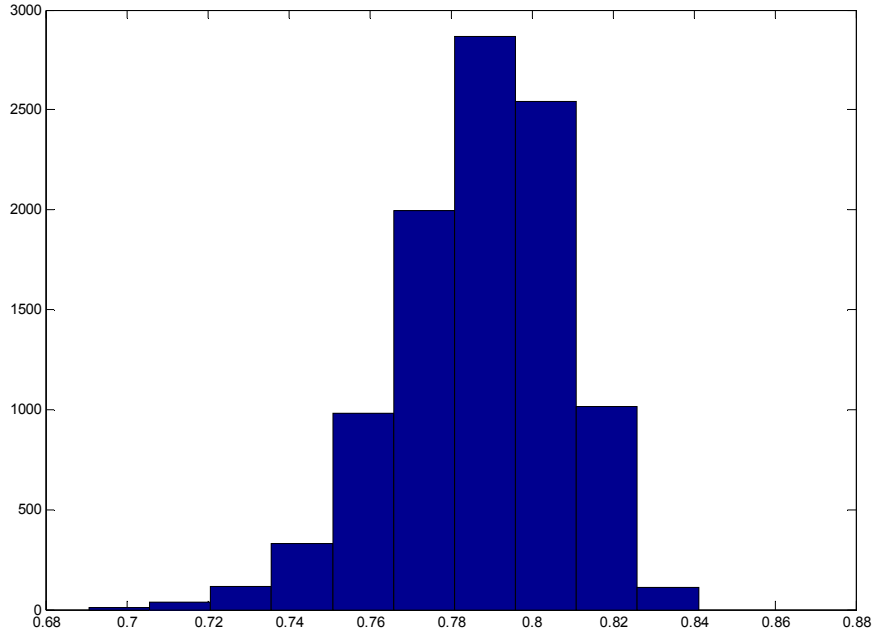
(2) Προσομοιωμένα δεδομένα για n=100 παρατηρήσεις με 30 ελλιπή δεδομένα από μια διδιάστατη κανονική κατανομή με μέσο $\begin{pmatrix}0\\0\end{pmatrix}$, πίνακα συνδιακύμανσης $\begin{pmatrix}1 & 0.9899\\0.9899 & 2\end{pmatrix}$ και συντελεστή συσχέτισης r=0,7. **Με την μέθοδο Gibbs** προσομοιώνουμε τον πίνακα συνδιακύμανσης m=10000 φορές και παίρνουμε:

● τον εκτιμώμενο συνδιακύμανσης

$$\hat{\Sigma}=\begin{bmatrix}1.1990 & 1.3646\\1.3646 & 2.5064\end{bmatrix}$$

● τον εκτιμώμενο συντελεστή συσχέτισης $\hat{r}=0.7872$ και

● τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι $s_1=0.1095$ και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι $s_2=0.1583$.



Ιστόγραμμα συντελεστή συσχέτισης με συχνότητα

Παρουσιάζεται για m=10000 φορές η συχνότητα εμφάνισης στο δείγμα MCMC των τιμών του συντελεστή συσχέτισης .

Το πρόγραμμα για την εξαγωγή των παραπάνω αποτελεσμάτων παρατίθεται σε παράρτημα.

Από τις δύο παραπάνω μεθόδους παρατηρούμε τα εξής:

● Στη μέθοδο Gibbs προσομοιώνουμε τις παρατηρήσεις που λείπουν από την δεσμευμένη posterior κατανομή (κανονική κατανομή) **ενώ** στην SIS από την περιθώρια posterior κατανομή τους (non-standardized student's t-distribution).
● Τρέχοντας τα προγράμματα παρατηρήσαμε ότι ο αλγόριθμος Gibbs χρειάζεται **λιγότερο χρόνο** από τον αλγόριθμο SIS.
● Στη μέθοδο Gibbs δεν χρησιμοποιούμε βάρη και ο πίνακας συνδιακύμανσης προκύπτει ως ο μέσος πίνακας συνδιακυμάνσεων που προκύπτουν από 10000 επαναλήψεις του αλγορίθμου **ενώ** στη μέθοδο SIS προκύπτει ως ο σταθμισμένος μέσος αφού σε κάθε επανάληψη του αλγορίθμου SIS υπολογίζουμε βάρος.
● Στα συγκεκριμένα δεδομένα που τρέξαμε (100 παρατηρήσεις με 30 ελλιπή δεδομένα) παρατηρούμε ότι

με τον **αλγόριθμο Gibbs** προέκυψε

πίνακας συνδιακύμανσης $\hat{\Sigma}=\begin{bmatrix}1.1990 & 1.3646\\ 1.3646 & 2.5064\end{bmatrix}$

και συντελεστής συσχέτισης r= 0.7872

τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι s_1= 0.1095 και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι s_2=0.1583

ενώ με τον **αλγόριθμο SIS** προέκυψε

πίνακας συνδιακύμανσης $\hat{\Sigma}=\begin{bmatrix}1.0353 & 0.9265\\ 0.9265 & 1.9592\end{bmatrix}$

με συντελεστή συσχέτισης r= 0.6506

τα τυπικά σφάλματα: για την διασπορά της πρώτης γραμμής του πίνακα των δεδομένων είναι s_1= 0.1017 και για την διασπορά της δεύτερης γραμμής του πίνακα των δεδομένων είναι s_2=0.1400

Υπενθυμίζουμε ότι στην αρχή τα δεδομένα τα κατασκευάσαμε από διδιάστατη κανονική κατανομή με μέσο $\begin{pmatrix}0\\0\end{pmatrix}$, πίνακα συνδιακύμανσης $\begin{pmatrix}1 & 0.9899\\ 0.9899 & 2\end{pmatrix}$ και συντελεστή συσχέτισης r=0,7. Σημειώνουμε ότι και οι δύο αλγόριθμοι παρουσιάζουν τα ίδια σχεδόν αποτελέσματα με πολύ μικρές αποκλίσεις από τις αρχικές τιμές που είχαμε θέσει.

Επίλογος

Σε αυτή την εργασία περιγράψαμε τις ακολουθιακές μεθόδους Monte Carlo όπως την SAW μέθοδο, την μέθοδο ανάπτυξης, την διαδοχική υποκατάσταση σε στατιστικά προβλήματα με ελλιπή δεδομένα και είδαμε πως αυτές οι μέθοδοι συνδέονται μεταξύ τους. Είδαμε ότι αυτές οι μέθοδοι είναι προηγμένες τεχνικές, οι οποίες βελτιώνουν τις προηγούμενες μεθόδους Monte Carlo γιατί ελαχιστοποιούν τα μειονεκτήματά τους και βοηθούν στην επίλυση πολυδιάστατων προβλημάτων. Δίνουν καλύτερες προσομοιώσεις, μπορούν να χρησιμοποιηθούν σαν βάση για καινούργιες τεχνικές, βοηθούν πάρα πολύ στην επίλυση προβλημάτων με ελλιπή δεδομένα δίνοντας καλύτερες προσεγγίσεις και έχουν μεγάλα πλεονεκτήματα σε πολλούς τομείς της επιστήμης γενικότερα.

Ειδικότερα σε αυτή την εργασία ασχοληθήκαμε με τις μεθόδους Monte Carlo για την επίλυση στατιστικών προβλημάτων με ελλιπή δεδομένα κάτω από το πρίσμα της Μπευζιανής προσέγγισης. Συγκεκριμένα ασχοληθήκαμε με το πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης από ελλιπή δεδομένα και το επιλύσαμε με τρεις διαφορετικές μεθόδους Monte Carlo, 1) με τη μέθοδο της δειγματοληψίας σπουδαιότητας, 2)με τη μέθοδο Gibbs και 3)με τη μέθοδο SIS.

ΠΑΡΑΡΤΗΜΑ

Προγράμματα

Murray δεδομένα

```
y=[1 1 -1 -1 2 2 -2 -2 NaN NaN NaN NaN ; 1 -1 1 -1 NaN NaN NaN
NaN 2 2 -2 -2];
```

Δεδομένα για 100 παρατηρήσεις με 30 να λείπουν

```
sigma1=1; sigma2=2; rho=0.7;
sigma12=rho*sqrt(sigma1*sigma2);
Sigma1=[sigma1 sigma12; sigma12 sigma2];
y=mvnrnd([0;0],Sigma1,100)';
for i=1:15
y(1,i)=NaN;
end
for i=16:30
y(2,i)=NaN;
end
```

discrete

```
function x=discrete(probs)

n=length(probs);

intervals=[0 probs];

cumprobs=cumsum(intervals);

r=rand;

for i=1:n
    x(cumprobs(i)<r&r<cumprobs(i+1))=i;
end
end
```

Non-standardized Student's t-distribution

```
function t=tnstand(X,mean,sigmastar,n)
A=gamma((n+1)/2)/(gamma(n/2)*sqrt(pi*n*sigmastar));
B=1+(1/n)*((X-mean)^2/sigmastar);
t=A*B^(-(n+1)/2);
end
```

Πρόβλημα εκτίμησης του πίνακα συνδιακύμανσης από ελλιπή δεδομένα

1)Με δειγματοληψία σπουδαιότητας

```
M=10000;
sigma_hat=zeros(2,2);
W=zeros(1,M);
weight_hat=0;
p=zeros(1,M);
for u=1:M
[d,t]=size(y);
m=0;
l=0;
Ymis=zeros(d,t);
Yobs=zeros(d,t);
%ανίχνευση των * στα δεδομένα και διαχωρισμός yobs με ymis
for j=1:t
    i=1;
    while i<d+1
        if y(i,j)<10^20
            i=i+1;
        else
            i=d+2;
        end
    end
    if i==d+1
        m=m+1;
        Yobs(:,m)=y(:,j);
    else
        l=l+1;
        Ymis(:,l)=y(:,j);
    end
end
Ynew=Ymis(:,1:l);
Sm=cov(Yobs(:,1:m)');
%κατασκευή των Ymis
s=size(Ynew,2);
r=Sm(1,2)/(sqrt(Sm(1,1)*Sm(2,2)));
c1=r*sqrt(Sm(1,1)/Sm(2,2));
c2=r*sqrt(Sm(2,2)/Sm(1,1));
logfy=0;
for j=1:s
    i=1;
        if Ymis(i,j)<10^20
            meanstar=c2*Ymis(i,j);
            sigmastar=(1-r^2)*Sm(2,2);
            Ynew(i+1,j)=meanstar+sqrt(sigmastar)*randn;

logfy=logfy+log(normpdf(Ynew(i+1,j),meanstar,sigmastar));
        else
            meanstar=c1*Ymis(i+1,j);
            sigmastar=(1-r^2)*Sm(1,1);
            Ynew(i,j)=meanstar+sqrt(sigmastar)*randn;

logfy=logfy+log(normpdf(Ynew(i,j),meanstar,sigmastar));
        end
end
Ynew=[Yobs(:,1:(t-l)) Ynew];
```

```
Sigma=cov(Ynew');
g=det(Sigma)^(-(m+3)/2)*exp(-1/2*trace(inv(Sigma)*Sm));
S=zeros(2,2);
for i=1:t
    S=S+Ynew(:,i)*Ynew(:,i)';
end
f=det(Sigma)^(-(t+3)/2)*exp(-1/2*trace(inv(Sigma)*S));
wu=log(f)-(log(g)+logfy);
wu1=exp(wu);
W(1,u)=wu1;
w=wu1*Sigma;
sigma_hat=sigma_hat+w;
weight_hat=weight_hat+wu1;
p(1,u)=Sigma(1,2)/(sqrt(Sigma(1,1))*sqrt(Sigma(2,2)));
end
disp('o zhtoumenos pinakas einai: ')
SIGMA1=sigma_hat/weight_hat
r0=SIGMA1(1,2)/(sqrt(SIGMA1(1,1))*sqrt(SIGMA1(2,2)))

std1=sqrt(SIGMA1(1,1)/t)
std2=sqrt(SIGMA1(2,2)/t)

Q=W./weight_hat;
rho=zeros(1,10000);
for i=1:100000
    x=discrete(Q);
    rho(1,i)=p(1,x);
end

hist(rho);
```

2)Με δειγματολήπτη Gibbs

```
p=zeros(1,100000);
[d,t]=size(y);
m=0;
l=0;
Ymis=zeros(d,t);
Yobs=zeros(d,t);
sigma_hat=zeros(2,2);
%ανίχνευση των * στα δεδομένα και διαχωρισμός yobs με ymis
for j=1:t
    i=1;
    while i<d+1
        if y(i,j)<10^20
            i=i+1;
        else
            i=d+2;
        end
    end
    if i==d+1
        m=m+1;
        Yobs(:,m)=y(:,j);
    else
        l=l+1;
        Ymis(:,l)=y(:,j);
    end
end
Ynew=Ymis(:,1:l);
s=size(Ynew,2);
```

```
Sigma=cov(Yobs(:,1:m)');
q=0;
M=11000;

while q<=M
r=Sigma(1,2)/(sqrt(Sigma(1,1)*Sigma(2,2)));
c1=r*sqrt(Sigma(1,1)/Sigma(2,2));
c2=r*sqrt(Sigma(2,2)/Sigma(1,1));
for j=1:s
    i=1;
        if Ymis(i,j)<10^20
            meanstar=c2*Ymis(i,j);
            sigmastar=(1-r^2)*Sigma(2,2);
            Ynew(i+1,j)=meanstar+sqrt(sigmastar)*randn;
        else
            meanstar=c1*Ymis(i+1,j);
            sigmastar=(1-r^2)*Sigma(1,1);
            Ynew(i,j)=meanstar+sqrt(sigmastar)*randn;
        end
end
%ο πίνακας y χωρίς να λείπουν στοιχεία (ολοκληρωμένος)
Y=[Ynew Yobs(:,1:(t-l))];
q=q+1;
Sigma=cov(Y');
if q>1000
p(1,q-1000)=Sigma(1,2)/(sqrt(Sigma(1,1))*sqrt(Sigma(2,2)));
sigma_hat=sigma_hat+Sigma;
end
end
SIGMA2=sigma_hat/10000
r=SIGMA2(1,2)/(sqrt(SIGMA2(1,1))*sqrt(SIGMA2(2,2)))

std1=sqrt(SIGMA2(1,1)/t)
std2=sqrt(SIGMA2(2,2)/t)

hist(p(:,1:10000));
```

3)Με μέθοδο SIS

```
[d,t]=size(y);
m=0;
l=0;
Ymis=zeros(d,t);
Yobs=zeros(d,t);
%ανίχνευση των * στα δεδομένα και διαχωρισμός yobs με ymis
for j=1:t
    i=1;
    while i<d+1
        if y(i,j)<10^20
            i=i+1;
        else
            i=d+2;
        end
    end
    if i==d+1
        m=m+1;
        Yobs(:,m)=y(:,j);
    else
        l=l+1;
```

```
        Ymis(:,l)=y(:,j);
    end
end
Ynew=Ymis(:,1:l);
Yobs=Yobs(:,1:m);
l=m;

M=10000;
W=zeros(1,M);
p=zeros(1,M);
weight_hat=0;
sigma_hat=zeros(2,2);
for u=1:M
    m=l;
    Yobs=Yobs(:,1:m);
    w=1;
    while m<t
        S=zeros(2,2);
        for i=1:m
            S=S+Yobs(:,i)*Yobs(:,i)';
        end
        T1=trnd(m-1);
        D=S(1,1)/(m-1);
        Ynew1=sqrt(D)*T1;
        T2=trnd(m);
        A=(det(S)/(m*S(1,1)))*(1+((Ynew1^2)/S(1,1)));
        B=(S(1,2)/S(1,1))*Ynew1;
        Ynew2=B+(sqrt(A)*T2);
        Yobs=[Yobs, [Ynew1 ; Ynew2]];
        meanstar=mean(Yobs');
        sigmastar=var(Yobs');
        [d,m]=size(Yobs);
        f1pdf=normpdf(Ynew1,0,1);
        f2pdf=normpdf(Ynew2,0,sqrt(2));
        t1pdf=tnstand(Ynew1,0,D,m-1);
        t2pdf=tnstand(Ynew2,B,A,m);
        wu=(log(f1pdf)+log(f2pdf))-(log(t1pdf)+log(t2pdf));
        wu1=exp(wu);
        w=w*wu1;
    end
    W(1,u)=w;
    Sigma=cov(Yobs');
    p(1,u)=Sigma(1,2)/(sqrt(Sigma(1,1))*sqrt(Sigma(2,2)));
    wuM=w*Sigma;
    sigma_hat=sigma_hat+wuM;
    weight_hat=weight_hat+w;
end
disp('o zhtoumenos pinakas einai: ')
SIGMA3=sigma_hat/weight_hat
r=SIGMA3(1,2)/(sqrt(SIGMA3(1,1))*sqrt(SIGMA3(2,2)))

std1=sqrt(SIGMA3(1,1)/t)
std2=sqrt(SIGMA3(2,2)/t)

Q=W./weight_hat;
rho=zeros(1,100000);
for i=1:100000
    x=discrete(Q);
    rho(1,i)=p(1,x);
end
hist(rho);
```

Βιβλιογραφία

[1] Avitzour, D. (1995). Stochastic simulation Bayesian approach to multi-target tracking, *IEE Proceedings-Radar Sonar and Navigation*

[2] Chen, Y. (2001). *Sequential Importance Sampling and Its Aplications* Ph.d. , Stanford University

[3] Gordon, N. J., Salmond, D. J. and Ewing, C. (1995). Bayesian state estimation for tracking and guidance using the bootstrap filter *Journal of Guidance Control and Dynamics.*

[4] Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear non-Gaussian Bayesian state estimation., IEE Proceedings on Radar and Signal Processing

[5] Grassberger, P. (1997). Pruned-enriched Rosenbluth method: Simulations of θ polymers of chain length up to 1,000,000, *Physical Review E .*

[6] Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods* Methuen's monographs on applied probability and statistics, Methuen Wlley, London.

[7] Hammersley, J. M. and Morton, K W. (1954). Poor man's Monte Carlo., *Journal* of *the Royal Statistical Society* , *Series B*

[8] Hammersley, J. M. and Morton, K. W. (1956). A new Monte Carlo technique: Antithetic variates, *Proceedings* of *the Cambridge Philosophical Society*.

[9] Harvey, A. C. (1990). *The Econometric Analysis* of *Time Series,* 2nd edn, MIT Press, Cambridge.

[10] Kong, *A.,* Liu, J. S. and Wong, W. H. (1994). SequentiaJ imputations and Bayesian missing data problems, *Journal* of *the American Statistical Associalion.*

[11] Kremer, K. and Binder, K. (1988). Monte CarJo simulation of Jattice models for macromolecules, *Computer Physics Reports*.

[12] Liu, J. and West, M. (2000). Combined parameter and state estimation in simulation-based filtering, *in* A. Doucet, J. F. G. de Freitas and N. J. Gordon (eds), *Sequential Monte Carlo in Practice,* Springer-Verlag, New York.

[13] Liu, J. S. and Chen, R. (1995). *Blind* deconvolution via sequentiaI Imputations, *Journal* of *the Ameτicαn Statisticαl Association.*

[14] Liu, J. S. and Chen, R. (1998). SequentiaI Monte Carlo methods for dynamic systems, *Journal* of *the Ameτicαn Statistical Association*

[15] Liu, J. S., Chen, R. and Wong, W. H. (1998). Rejection control and sequential importance sampling, *Journal* of *the Ameτicαn Statisticαl Association*.

[16] MacEachern, S. N. , CIyde, M. and Liu, J. S. (1999). Sequential Importance sampling for nonparametric Bayes models: The next generation, *Canadian Journal* of *Statistics* .

[17] Marshall, A.(1956). The use of multi-stage sampling schemes in Monte Carlo computations, in M.Meyer (ed.), Symposium on Monte Carlo Methods, Wiley, New York.

[18] Murray, G. (1977). Comment on "Maximum likelihood from incomplete data via the EM algorithm" by A.P. Dempster, N.M. Laird, and D.B. Rubin, *Journαl of the Royal Statisticαl Society, B.*

[19] Odell, P.L. and Feiveson , A.H. (1966). A numerical procedure to generate a sample covariance matrix , Journal of the American Statistical Association

[20] Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters, Journal of the American Statistical Association.

[21] Rosenbluth, M. N. and Rosenbluth, A. W. (1955). Monte CarIo calculation of the average extension of molecular chains, *Journal* of *Chemical Physics*.

[22] Rubin , D.B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest : *the SIR algorithm , Journal of the American Statistical Association*.

[23] Rubinstein, R.Y. (1981). Simulation and the Monte Carlo Method, Wiley series in probability and mathematical statistics,Wiley, New York.

[24] von Neumann, J. (1951). Various techniques used in connection with random digits, *National Bureau* of *Standards Applied Mathematics Series*

[25] Wall, F. T. and Erpenbeck, J. J. (1959). New method for the statistical computation of polymer dimensions, *Journal* of *Chemical Physics*

[26] West, M. and Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models,* Springer-Verlag, New York.

[27] Ross M. Sheldon (2006) Simulation , Fourth Edition, Elsevier Academic Press, University of Southern California

[28] Δελλαπόρτας Π. και Τσιαμυρτζής Π. (2004) Σημειώσεις μαθήματος "*Στατιστική κατά Bayes*" Οικονομικού Πανεπιστημίου Αθηνών Τμήμα Στατιστικής

[29] Liu J.S (2001). Monte Carlo Strategies in scientific computing, springer.