Ανασκόπηση Στοιχείων Πιθανοτήτων

Αθανάσιος Ροντογιάννης

Αν. Καθηγητής ΣΗΜΜΥ-ΕΜΠ

Συμβολισμοί (Notation)

- Διανύσματα/πίνακες και ανάστροφοί τους:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_l \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{l1} & \cdots & a_{lm} \end{bmatrix}, \quad \boldsymbol{x}^T = [x_1 \quad x_2 \quad \cdots \quad x_l], \quad B = A^T \Leftrightarrow b_{ij} = a_{ji}, \quad A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_m], \quad A^T = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_m^T \end{bmatrix}$$

- Ένας διαγώνιος πίνακας με στοιχεία $a_1, a_2, \dots, a_l$ στη διαγώνιό του θα συμβολίζεται ως $A = \mathrm{diag}\{a_1, a_2, \dots, a_l\}$
- $I = \mathrm{diag}\{1,1,\dots,1\}$ είναι ο μοναδιαίος πίνακας
- Το ίχνος (άθροισμα διαγώνιων στοιχείων) ενός *τετραγωνικού* πίνακα A συμβολίζεται ως trace$\{A\}$ και η ορίζουσά του ως $|A|$.
- Ακολουθίες αριθμών (διανυσμάτων) συμβολίζονται ως $x_n (\boldsymbol{x}_n)$ ανάλογα με την περίπτωση
- Συμβολισμός συναρτήσεων: $f, f(x), f(\cdot)$
- Τα στοιχεία των διανυσμάτων/πινάκων και οι βαθμωτές ποσότητες που θα χρησιμοποιήσουμε θεωρούνται γενικά πραγματικοί αριθμοί.
- *Εσωτερικό γινόμενο* διανυσμάτων $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^l$:

$$\boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{l} x_i y_i \equiv \boldsymbol{y}^T \boldsymbol{x}$$

Χρήσιμες σχέσεις

Αν A, B, C πίνακες με κατάλληλες διαστάσεις και ιδιότητες, θα είναι:

- $(AB)^T = B^T A^T$
- $(AB)^{-1} = B^{-1} A^{-1}$
- $(A^T)^{-1} = (A^{-1})^T$
- $\mathrm{trace}\{AB\} = \mathrm{trace}\{BA\}$
- $\mathrm{trace}\{ABC\} = \mathrm{trace}\{CAB\} = \mathrm{trace}\{BCA\}$
- Αν $A = \boldsymbol{a}\boldsymbol{b}^T$ (εξωτερικό γινόμενο των διανυσμάτων $\boldsymbol{a}, \boldsymbol{b}$) εύκολα προκύπτει:

$$\mathrm{trace}(A) = \mathrm{trace}(\boldsymbol{b}^T \boldsymbol{a}) = \boldsymbol{b}^T \boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{b}$$

- $|AB| = |A||B|$
- $|A^{-1}| = 1/|A|$

Ανάδελτα συνάρτησης - Ιδιότητες

Έστω μια συνάρτηση $f(\boldsymbol{x})$ μιας διανυσματικής ποσότητας $\boldsymbol{x}$. Το ανάδελτα ή παράγωγος της f ως προς $\boldsymbol{x}$ ορίζεται ως εξής:

$$\nabla_{\boldsymbol{x}} f \equiv \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_l} \end{bmatrix}$$

Μπορούν να δειχτούν τα παρακάτω:

- $\frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$
- $\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = (A + A^T)\boldsymbol{x}$, που γίνεται $2A\boldsymbol{x}$ αν ο A είναι συμμετρικός.
- $\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = A^T$

Διακριτές τυχαίες μεταβλητές

Μια διακριτή τυχαία μεταβλητή x μπορεί να πάρει οποιαδήποτε τιμή από ένα πεπερασμένο ή άπειρα αριθμήσιμο σύνολο $\mathcal{X}$, που ονομάζεται δειγματικός χώρος. Η πιθανότητα του γεγονότος "$\mathrm{x} = x \in \mathcal{X}$“ συμβολίζεται ως:

$$P(\mathrm{x} = x) \text{ ή απλά } P(x)$$

Η συνάρτηση P ονομάζεται *συνάρτηση μάζας πιθανότητας* και ισχύει:

$$\sum_{x \in \mathcal{X}} P(x) = 1$$

<u>Από κοινού και δεσμευμένες πιθανότητες</u>:

- Κανόνας αθροίσματος: $P(x) = \sum_{y \in \mathcal{Y}} P(x, y)$
- Κανόνας γινομένου: $P(x, y) = P(x|y)P(y)$
- Θεώρημα Bayes: $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$

Συνεχείς τυχαίες μεταβλητές

Μια συνεχής τυχαία μεταβλητή x παίρνει τιμές στον άξονα των πραγματικών αριθμών, $\mathbb{R}$. Η *συνάρτηση κατανομής πιθανότητας* της x ορίζεται ως:

$$F_{\mathrm{x}}(x) \triangleq P(\mathrm{x} \leq x)$$

και η *συνάρτηση πυκνότητας πιθανότητας (ΣΠΠ)* ως:

$$p_{\mathrm{x}}(x) \triangleq \frac{dF_{\mathrm{x}}(x)}{dx}$$

Από τα παραπάνω προκύπτουν τα εξής (ο δείκτης x παραλείπεται):

$$F(x) = \int_{-\infty}^{x} p(x)dx\,, \qquad P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x)dx \qquad \text{και} \qquad \int_{-\infty}^{\infty} p(x)dx = 1$$

Επίσης ισχύουν για την συνάρτηση πυκνότητας πιθανότητας όλοι κανόνες που ισχύουν για τις πιθανότητες διακριτών τυχαίων μεταβλητών. Για παράδειγμα:

$$p(x|y) = \frac{p(x,y)}{p(y)}, \qquad p(x) = \int_{-\infty}^{\infty} p(x,y)dy.$$

Μέση τιμή και μεταβλητότητα

- Μέση τιμή: $\mathbb{E}[\mathrm{x}] = \int_{-\infty}^{\infty} xp(x)dx$
- Μεταβλητότητα: $\sigma_{\mathrm{x}}^2 = \int_{-\infty}^{\infty} (x - \mathbb{E}[\mathrm{x}])^2 \, p(x)dx$
- Γενικότερα: $\mathbb{E}[f(\mathrm{x})] = \int_{-\infty}^{\infty} f(x)p(x)dx$
- Συμμεταβλητότητα (covariance) τυχαίων μεταβλητών x, y:

$$\mathrm{cov}(\mathrm{x}, \mathrm{y}) = \mathbb{E}[(\mathrm{x} - \mathbb{E}[\mathrm{x}])(\mathrm{y} - \mathbb{E}[\mathrm{y}])]$$

- Συσχέτιση (correlation) τυχαίων μεταβλητών x, y:

$$r_{x,y} = \mathbb{E}[xy] = \mathrm{cov}(\mathrm{x}, \mathrm{y}) + \mathbb{E}[x]\mathbb{E}[y]$$

- Ένα τυχαίο διάνυσμα $\mathbf{x} = [\mathrm{x}_1, \dots, \mathrm{x}_l]^T$ είναι ένα διάνυσμα με στοιχεία τυχαίες μεταβλητές. Ορίζεται η $p(\mathbf{x})$ ως η από κοινού ΣΠΠ:

$$p(\mathbf{x}) = p(\mathrm{x}_1, \dots, \mathrm{x}_l)$$

Πίνακες συμμεταβλητότητας και συσχέτισης

- Πίνακας συμμεταβλητότητας ενός τυχαίου διανύσματος $\mathbf{x} \in \mathbb{R}^l$:

$$\mathrm{Cov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] \qquad \text{ή}$$

$$\mathrm{Cov}(\mathbf{x}) = \begin{bmatrix} \mathrm{cov}(\mathrm{x}_1, \mathrm{x}_1) & \cdots & \mathrm{cov}(\mathrm{x}_1, \mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathrm{x}_l, \mathrm{x}_1) & \cdots & \mathrm{cov}(\mathrm{x}_l, \mathrm{x}_l) \end{bmatrix}$$

- Πίνακας συσχέτισης ενός τυχαίου διανύσματος $\mathbf{x} \in \mathbb{R}^{\boldsymbol{l}}$:

$$R_x = \mathbb{E}[\mathbf{x}\mathbf{x}^{\boldsymbol{T}}] \qquad \text{ή}$$

$$R_x = \begin{bmatrix} \mathbb{E}(\mathrm{x}_1, \mathrm{x}_1) & \cdots & \mathbb{E}(\mathrm{x}_1, \mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(\mathrm{x}_l, \mathrm{x}_1) & \cdots & \mathbb{E}(\mathrm{x}_l, \mathrm{x}_l) \end{bmatrix} = \mathrm{Cov}(\mathbf{x}) + \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}^T]$$

- Πολύ σημαντικό: Οι πίνακες συμμεταβλητότητας και συσχέτισης είναι θετικά ημιορισμένοι.

Κανονική (Gaussian) κατανομή

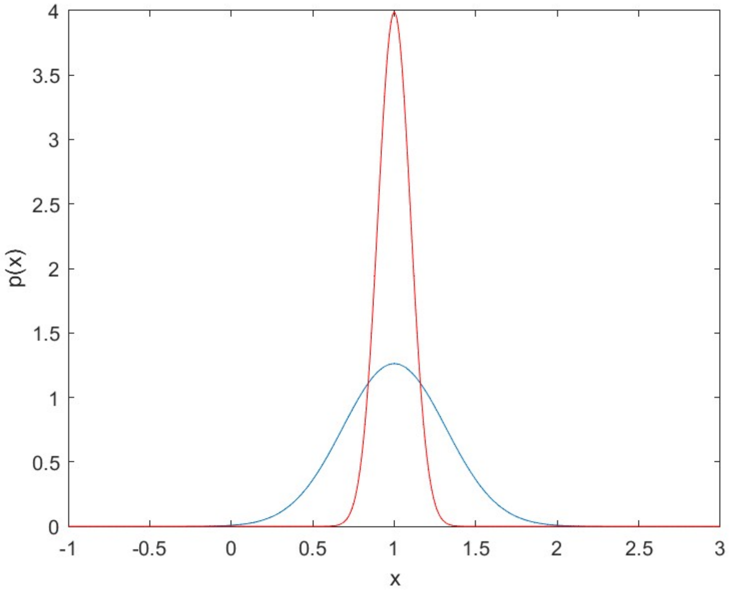
Θα λέμε ότι η τυχαία μεταβλητή x ακολουθεί την κανονική ή Gaussian κατανομή με παραμέτρους μ και σ^2 και γράφουμε $\mathrm{x} \sim \mathcal{N}(\mu, \sigma^2)$ ή $\mathcal{N}(x|\mu, \sigma^2)$ αν

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Μπορεί να δειχτεί ότι:

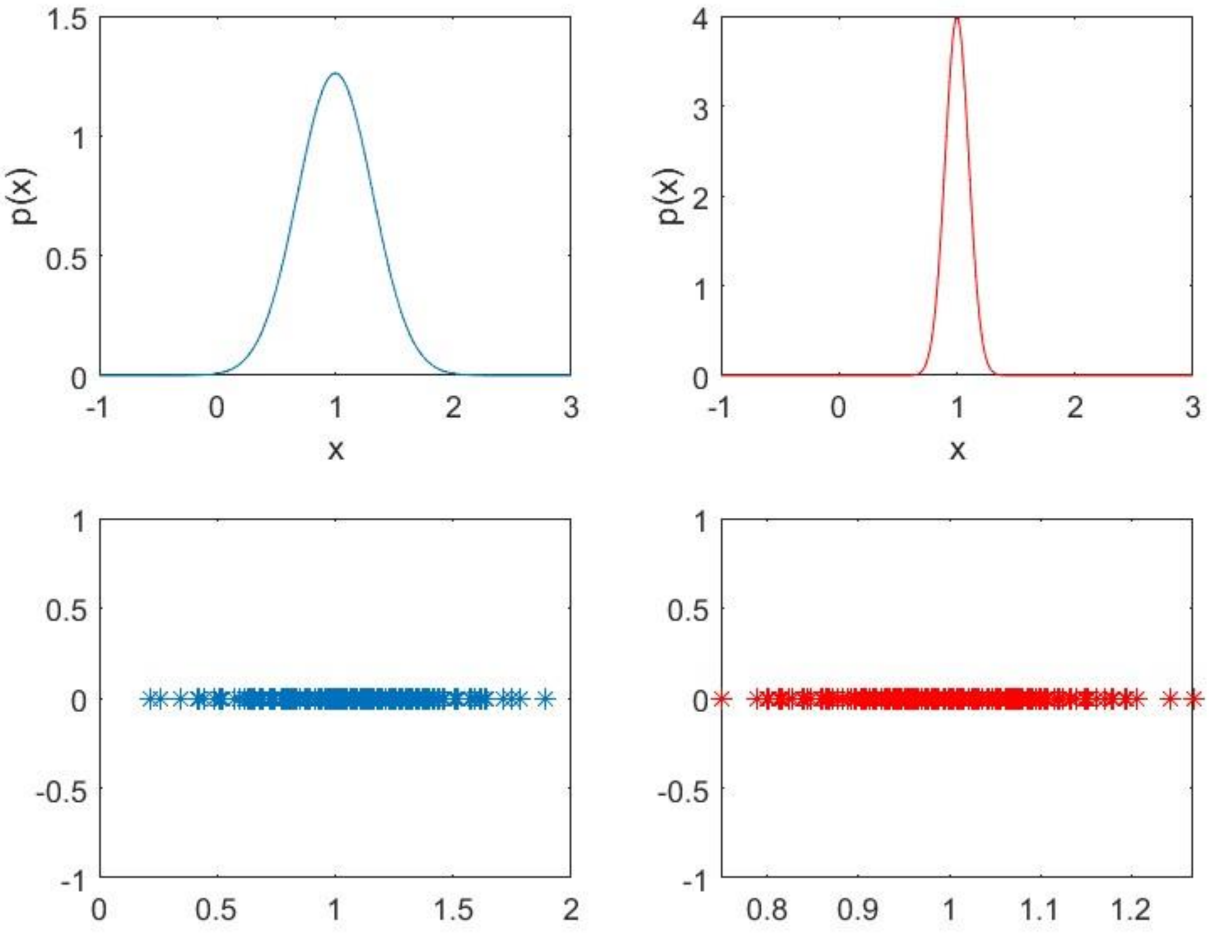
$$\mathbb{E}[\mathrm{x}] = \mu \quad \text{και} \quad \sigma_{\mathrm{x}}^2 = \sigma^2$$

Γραφική παράσταση δύο Gaussian ΣΠΠ με $\mu = 1$ και $\sigma^2 = 0.1$ (μπλε) και $\sigma^2 = 0.01$ (κόκκινη) αντίστοιχα.



Κανονική (Gaussian) κατανομή

200 δείγματα από καθεμιά από τις παραπάνω κανονικές κατανομές



Πολυμεταβλητή κανονική κατανομή

Η γενίκευση της κανονικής κατανομής για τυχαία διανύσματα $\mathbf{x} \in \mathbb{R}^l$, οδηγεί στη λεγόμενη πολυμεταβλητή κανονική (Gaussian) κατανομή, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, με παραμέτρους $\boldsymbol{\mu}$ και Σ, που ορίζεται ως εξής

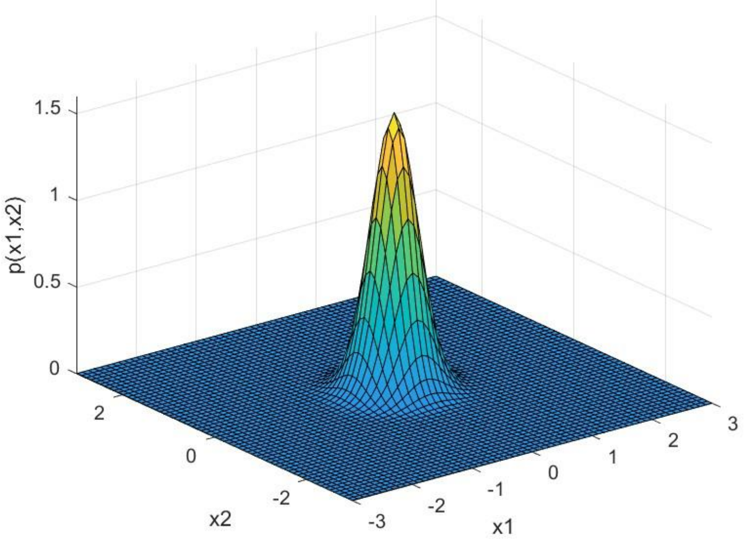
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Μπορεί να δειχτεί ότι:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{και} \quad \mathrm{Cov}(\mathbf{x}) = \Sigma.$$

Πολυμεταβλητή Gaussian κατανομή με $\boldsymbol{\mu} = \mathbf{0}$ και

$$\Sigma = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

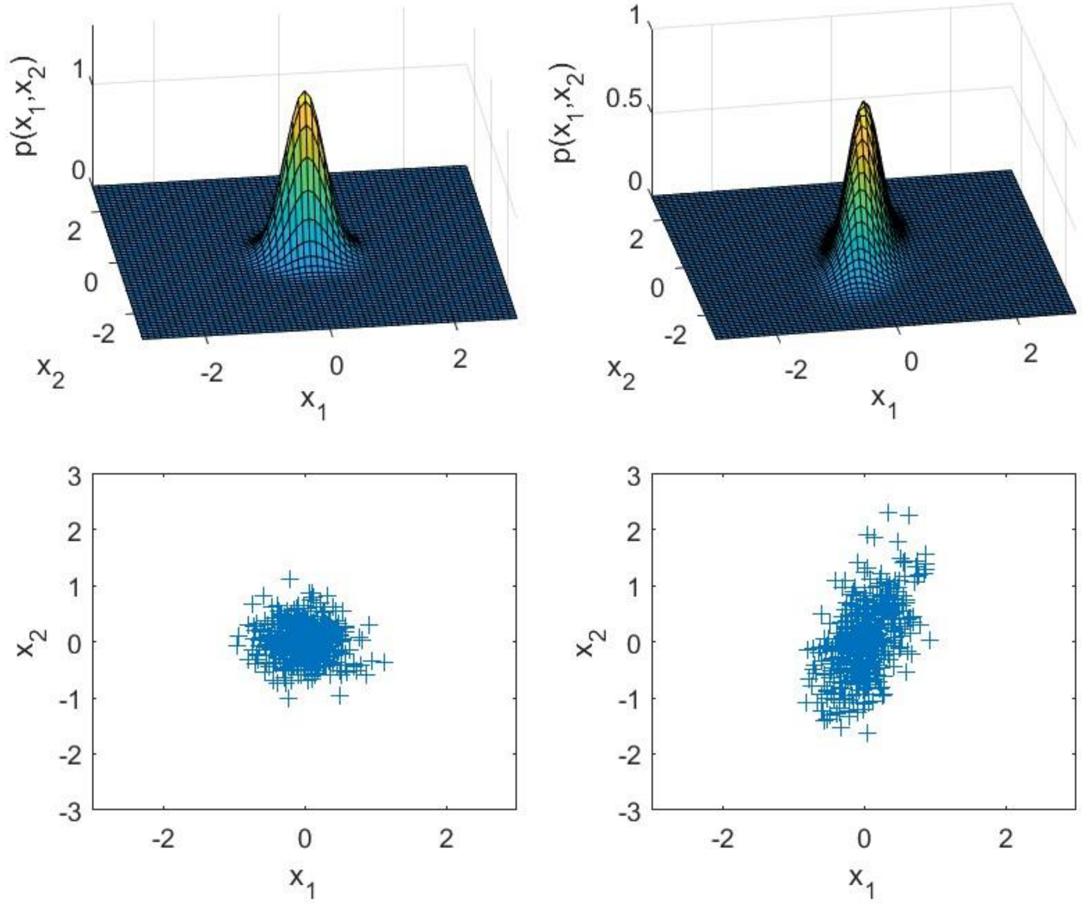


Πολυμεταβλητή κανονική κανανομή

Δύο διμεταβλητές κανονικές κατανομές με $\boldsymbol{\mu} = \mathbf{0}$ και

$$\Sigma_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \qquad \Sigma_2 = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}$$

Στο σχήμα φαίνονται επίσης 500 δείγματα από κάθε κατανομή.

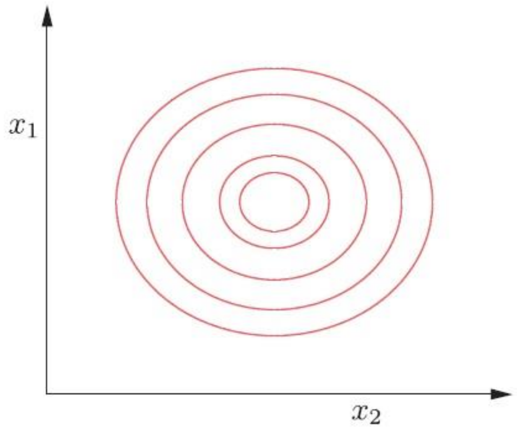


Πολυμεταβλητή κανονική καναναομή

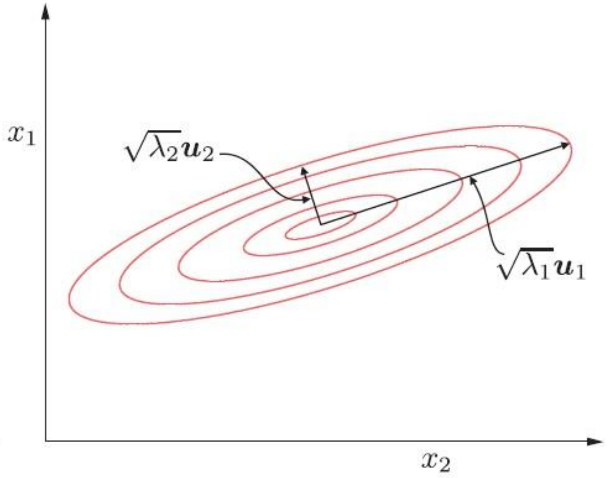
- Ισοσταθμικές καμπύλες της πολυμεταβλητής κανονικής κατανομής: μια ισοσταθμική καμπύλη σχηματίζεται από όλα τα σημεία $\boldsymbol{x}$ τα οποία αντιστοιχούν στην ίδια τιμή της ΣΠΠ, δηλαδή $p(\boldsymbol{x}) = c$,

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) = σταθερο = c$$

- Οι ισοσταθμικές καμπύλες είναι είτε κύκλοι (υπερσφαίρες), είτε ελλείψεις (υπερελλειψοειδή) με κέντρο τη μέση τιμή $\boldsymbol{\mu}$. Τα μήκη των αξόνων καθορίζονται από τις ιδιοτιμές και τα ιδιοδιανύσματα του Σ.



$\Sigma = \sigma^2 I$



$\Sigma \neq \sigma^2 I$

Πολυμεταβλητή κανονική κανανομή

Αν ο πίνακας συμμεταβλητότητας είναι διαγώνιος, δηλαδή

$$\Sigma = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \dots, \sigma_l^2\}$$

τότε οι επιμέρους τυχαίες μεταβλητές θα είναι ασυσχέτιστες, $\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) = 0, \quad i, j = 1,2, \dots, l.$
Επιπλέον όμως μπορεί εύκολα να δειχτεί ότι θα ισχύει:

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right).$$

Με άλλα λόγια,

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} p(x_i)$$

δηλαδή οι τυχαίες μεταβλητές $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_l$ είναι επιπλέον στατιστικά ανεξάρτητες.

Δεσμευμένες κανονικές καναnομές

Έστω ότι το l-διάστατο διάνυσμα $\boldsymbol{x}$ γράφεται ως εξής:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}, \text{ με } \boldsymbol{x}_1 \in \mathbb{R}^k \text{ και } \boldsymbol{x}_2 \in \mathbb{R}^{l-k}$$

και τα $\boldsymbol{\mu}, \Sigma$, διαμερίζονται αντίστοιχα:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \ \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Τότε η δεσμευμένη κατανομή του $\boldsymbol{x}_1$ δοσμένου ότι $\boldsymbol{x}_2 = \boldsymbol{\alpha}$, είναι κανονική $(\boldsymbol{x}_1 | \boldsymbol{x}_2 = \boldsymbol{a}) \sim \mathcal{N}\left(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}\right)$ με:

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}_2) \text{ και } \widehat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Μέθοδος μεγίστης πιθανοφάνειας

- Σε πολλές εφαρμογές μας ενδιαφέρει να εκτιμήσουμε την κατανομή από την οποία προέρχεται ένα σύνολο παρατηρήσεων ή μετρήσεων.

- Πιο συγκεκριμένα, έστω ότι μας δίνεται ένα σύνολο N παρατηρήσεων $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ που έχουν προέλθει από μια κατανομή πιθανότητας. Υποθέτουμε ότι η από κοινού ΣΠΠ αυτών των μετρήσεων είναι γνωστού παραμετρικού συναρτησιακού τύπου και συμβολίζεται με $p(\mathcal{X}; \boldsymbol{\theta})$. Η παράμετρος $\boldsymbol{\theta} \in \mathbb{R}^K$ είναι άγνωστη και ο στόχος μας είναι να εκτιμήσουμε την τιμής της.

- Η από κοινού ΣΠΠ $p(\mathcal{X}; \boldsymbol{\theta})$ ονομάζεται συνάρτηση πιθανοφάνειας (likelihood function) του $\boldsymbol{\theta}$ ως προς το δοσμένο σετ παρατηρήσεων $\mathcal{X}$. Σύμφωνα με τη μέθοδο μεγίστης πιθανοφάνειας (maximum likelihood method), η παράμετρος $\boldsymbol{\theta}$ εκτιμάται ως εξής:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} \coloneqq \arg\max_{\boldsymbol{\theta}} p(\mathcal{X}; \boldsymbol{\theta})$$

Μέθοδος μεγίστης πιθανοφάνειας

- Καθώς η λογαριθμική συνάρτηση $\ln(\cdot)$ είναι γνησίως αύξουσα, μπορεί κάνεις να αναζητήσει εναλλακτικά το μέγιστο του λογαρίθμου της συνάρτησης πιθανοφάνειας, δηλαδή:

$$\left.\frac{\vartheta \ln p(\mathcal{X};\boldsymbol{\theta})}{\vartheta \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = \mathbf{0}$$

- Το $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ ονομάζεται εκτιμητής μεγίστης πιθανοφάνειας, είναι συνάρτηση των $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$ και κατά συνέπεια είναι τυχαία μεταβλητή.

- Αν οι παρατηρήσεις $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$ είναι ανεξάρτητες και ακολουθούν την ίδια κατανομή (i.i.d.),

  1. Ο εκτιμητής μεγίστης πιθανοφάνειας είναι ασυμπτωτικά αμερόληπτος. Δηλαδή, αν υποθέσουμε ότι το μοντέλο που επιλέξαμε για την κατανομή των παρατηρήσεων είναι σωστό και η πραγματική παράμετρος του μοντέλου είναι $\boldsymbol{\theta}_o$ θα ισχύει:

$$\lim_{N\to\infty} \mathbb{E}[\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}] = \boldsymbol{\theta}_o$$

  2. Ο εκτιμητής μεγίστης πιθανοφάνειας είναι ασυμπτωτικά συνεπής, δηλαδή για κάθε $\epsilon > 0$

$$\lim_{N\to\infty} \mathrm{Prob}\left\{\left|\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_o\right| > \epsilon\right\} = 0$$

Μέθοδος μεγίστης πιθανοφάνειας

Παράδειγμα. Έστω $x_1, x_2, \ldots, x_N$ βαθμωτές παρατηρήσεις που έχουν προέλθει από μια κανονική κατανομή με γνωστή μεταβλητότητα σ^2 και άγνωστη μέση τιμή μ, δηλαδή,

$$p(x_n; \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right)$$

Υποθέτουμε ότι οι μετρήσεις είναι στατιστικά ανεξάρτητες και θέλουμε να εκτιμήσουμε από αυτές την άγνωστη παράμετρο της κατανομής μ.

Για τις N στατιστικά ανεξάρτητες παρατηρήσεις ο λογάριθμος της από κοινού συνάρτησης πιθανοφάνειας θα είναι:

$$L(\mu) = \ln \prod_{n=1}^{N} p(x_n; \mu) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

Μέθοδος μεγίστης πιθανοφάνειας

Παράδειγμα (συνέχεια). Παραγωγίζοντας ως προς μ παίρνουμε

$$\frac{\vartheta L(\mu)}{\vartheta \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu),$$

το οποίο αν εξισωθεί με το 0 δίνει,

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$: εκτιμητής μεγίστης πιθανοφάνειας του μ

- Παρόμοια διαδικασία μπορεί να εφαρμοστεί και για $\theta = \sigma^2$ με μ γνωστό. Δοκιμάστε!
- Αν οι παρατηρήσεις είναι διανύσματα $\boldsymbol{x}_n \in \mathbb{R}^l$ που ακολουθούν την πολυμεταβλητή κανονική κατανομή με γνωστό Σ και άγνωστο $\boldsymbol{\mu}$, μπορεί να δειχτεί ότι ο εκτιμητής μεγίστης πιθανοφάνειας του $\boldsymbol{\mu}$ είναι:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$$

Βιβλιογραφία

- S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, 2nd Edition, Academic Press, 2020.
- Σ. Θεοδωρίδης, Διαφάνειες του παραπάνω συγγράμματος (στα αγγλικά).