Introduction aux processus stochastiques

Travail Matlab : Distribution normale multivariée

Consignes de style pour le travail

Vous pouvez remettre un document rédigé dans le style d'un bref rapport de laboratoire.

Simulation de variables suivant une distribution normale multivariée.

But de l'exercice

On considère la génération de vecteurs de variables aléatoires de dimension n, selon une *distribution normale multivariée* de moyenne $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ et de matrice de covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ définie positive spécifiées :

$$\begin{aligned} \mathbb{E}\{\boldsymbol{x}\} = \bar{\boldsymbol{x}} \quad &\Leftrightarrow \quad \mathbb{E}\{x_i\} = \bar{x}_i \\ \mathbb{E}\{(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{x}-\bar{\boldsymbol{x}})^T\} = \boldsymbol{\Sigma} \quad &\Leftrightarrow \quad \mathbb{E}\{(x_i-\bar{x}_i)(x_j-\bar{x}_j)\} = \Sigma_{ij} \end{aligned}$$

La distribution normale multivariée $\mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma})$ généralise la distribution normale $\mathcal{N}(\bar{x}, \sigma^2)$:

	support	paramètres	densité en v (pdf : probability distribution function)
$n=1$	$\mathbb{R}$	$\bar{x}$, σ^2	$p_x(v) = (2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp\{-\frac{1}{2}\left(\frac{v-\bar{x}}{\sigma}\right)^2\}$
$n>1$	$\mathbb{R}^n$	$\bar{\boldsymbol{x}}$, $\boldsymbol{\Sigma}$	$p_{\boldsymbol{x}}(\boldsymbol{v}) = (2\pi)^{-n/2}(\det\boldsymbol{\Sigma})^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{v}-\bar{\boldsymbol{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{v}-\bar{\boldsymbol{x}})\}$

Pour rappel, on interprète comme suit la notion de densité. Dans le cas $n = 1$, on a pour un intervalle $I = [v, v + dv[$ de longueur infinitésimale dv : Prob$(x \in I) = p_x(v)dv$.
Spécifier Prob$(x = v)$ n'aurait pas de sens dans la mesure où cette probabilité est nulle.
Dans le cas $n > 1$, on a pour un hyper-rectangle $R = [v_1, v_1+dv_1[\times[v_2, v_2+dv_2[\times\ldots\times[v_n, v_n+dv_n[$ de volume infinitésimal $dv_1dv_2\ldots dv_n$: Prob$(\boldsymbol{x} \in R) = p_{\boldsymbol{x}}(\boldsymbol{v})dv_1dv_2\ldots dv_n$.

La généralisation au cas $n > 1$ de la notion de *fonction de distribution cumulée* (cdf : cumulative distribution function) va par contre soulever plus de questions.

Pour rappel, dans le cas $n = 1$, la cdf évaluée en v est

$$F(v) = \text{Prob}(x \le v) = \int_{-\infty}^{v} p_x(v)dv = \Phi(\frac{v-\bar{x}}{\sigma})$$

où $\Phi(y)$ est la cdf d'une distribution de moyenne 0 et de variance 1, qui peut être évaluée numériquement (via `normcdf` ou indirectement via `erf`, dans MATLAB). La *fonction de distribution cumulée inverse* (inverse cdf : inverse cumulative distribution function) évaluée en $p \in [0, 1]$ est $F^{-1}(p) = y$, où y est tel que $F(y) = p$. Pour une distribution normale $\mathcal{N}(\bar{x}, \sigma^2)$ on a $F^{-1}(p) = \bar{x} + \sigma\Phi^{-1}(p)$, où Φ^{-1} est la cdf inverse de $\mathcal{N}(0, 1)$ (évaluée via `norminv` ou indirectement via `erfinv` dans MATLAB).
La cdf permet d'évaluer Prob$(x \in B)$ pour une large classe de sous-ensembles $B \subset \mathbb{R}$ (à savoir ceux qui peuvent se décomposer en union dénombrable d'intervalles I de $\mathbb{R}$ ou de leur complémentaire $\mathbb{R} \setminus I$, classe qui suffit en pratique).

Dans le cas $n > 1$, il existe aussi une cdf évaluée en $\boldsymbol{v}$

$$F(\boldsymbol{v}) = \text{Prob}(x_1 \leq v_1, x_2 \leq v_2, \ldots, x_n \leq v_n) = \int_{-\infty}^{v_1} \int_{-\infty}^{v_2} \ldots \int_{-\infty}^{v_n} p_{\boldsymbol{x}}(\boldsymbol{v}) dx_1 dx_2 \ldots dx_n$$

dont l'évaluation numérique est compliquée. La notion de cdf inverse n'est plus bien définie, car il existe plus d'un vecteur $\boldsymbol{v}$ tel que $F(\boldsymbol{v}) = p$. Pour une analogie, penser à l'infinité de rectangles d'aire S dont le coin inférieur gauche est fixé, alors qu'il n'existe qu'un segment d'extrémité gauche fixée de longueur L.

La notion d'*ellipsoïde de confiance* se révèle plus maniable dans le cas des distributions normales multivariées. En observant l'expression de la densité, on remarque les points $\boldsymbol{v}$ ayant une même densité $p_{\boldsymbol{x}}(\boldsymbol{v})$ doivent satisfaire $(\boldsymbol{v} - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{v} - \bar{\boldsymbol{x}}) = \alpha$ pour un certain $\alpha \geq 0$. Or cette équation s'interprète comme la surface de l'ellipsoïde $\mathcal{E}_\alpha$ couvrant les points $\boldsymbol{v}$ tels que $(\boldsymbol{v} - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{v} - \bar{\boldsymbol{x}}) \leq \alpha$.

En intégrant la densité sur le domaine couvert par l'ellipsoïde, on obtient la probabilité

$$\text{Prob}(\boldsymbol{x} \in \mathcal{E}_\alpha) = \int_{\boldsymbol{v} \,:\, (\boldsymbol{v}-\bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{v}-\bar{\boldsymbol{x}}) \leq \alpha} p_{\boldsymbol{x}}(\boldsymbol{v}) dx_1 \ldots dx_n \triangleq p.$$

Ainsi il exite une relation biunivoque entre certains sous-domaines de $\mathbb{R}^n$ et une probabilité $p \in [0, 1]$, analogue à la relation entre une cdf et son inverse dans le cas unidimensionnel. Cette relation peut être exploitée car l'intégrale est évaluable efficacement.

En effet on peut montrer que la variable aléatoire scalaire $z = (\boldsymbol{v} - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{v} - \bar{\boldsymbol{x}})$ suit une distribution chi-carré à n degrés de liberté, de sorte que

$$\text{Prob}(\boldsymbol{x} \in \mathcal{E}_\alpha) = \text{Prob}(z \leq \alpha) = F_{\chi^2_n}(\alpha)$$

avec $F_{\chi^2_n}(\alpha)$ la cdf de la chi-carré évaluée en α, à évaluer numériquement.

Le but de l'exercice est de pouvoir générer des échantillons $\boldsymbol{x}$ distribués selon $\mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma})$ au départ d'un générateur $\mathcal{N}(0, 1)$, d'afficher le contour de l'ellipsoide de confiance dans le cas $n = 2$ pour certaines valeurs de p, et de vérifier expérimentalement que l'ellipsoide contient bien à peu près la fraction p des points générés.

La section suivante présente un ensemble de résultats dans lequel puiser pour répondre aux questions.

Eléments théoriques utiles

(Les points marqués d'une $\star$ sont à connaître à l'examen.)

$\star$ 1. *Transformations affines de variables normales multivariées.*
Soit $\boldsymbol{x} \sim \mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}_x)$ avec $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}_x \in \mathbb{R}^{n \times n}$. Soient $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ et $\boldsymbol{b} \in \mathbb{R}^m$, fixés.
Alors $\boldsymbol{y} \triangleq \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ suit $\mathcal{N}(\bar{\boldsymbol{y}}, \boldsymbol{\Sigma}_y)$ avec $\bar{\boldsymbol{y}} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b}$ et $\boldsymbol{\Sigma}_y = \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T$.
En effet, l'élément y_i est une v.a. normale en tant que combinaison linéaire de variables normales, et d'autre part,

$$\begin{aligned}
\bar{\boldsymbol{y}} &= \mathbb{E}\{\boldsymbol{y}\} = \mathbb{E}\{\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\} = \mathbb{E}\{\boldsymbol{A}\boldsymbol{x}\} + \mathbb{E}\{\boldsymbol{b}\} = \boldsymbol{A}\mathbb{E}\{\boldsymbol{x}\} + \boldsymbol{b} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b} \\
\boldsymbol{\Sigma}_y &= \mathbb{E}\{(\boldsymbol{y} - \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}})^T\} = \mathbb{E}\{\boldsymbol{A}(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T\boldsymbol{A}^T\} = \boldsymbol{A}\mathbb{E}\{(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T\}\boldsymbol{A}^T \\
&= \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T.
\end{aligned}$$

⋆ 2. *Décomposition de Cholesky.*
Soit $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ une matrice définie positive.
Alors $\boldsymbol{S} = \boldsymbol{L}\boldsymbol{L}^T$, où $\boldsymbol{L}$ est triangulaire inférieure.
On a aussi $\boldsymbol{S} = \boldsymbol{R}^T\boldsymbol{R}$, où $\boldsymbol{R} = \boldsymbol{L}^T$ est triangulaire supérieure.
La fonction MATLAB `chol` appelée avec $\boldsymbol{S}$ en argument renvoie $\boldsymbol{R}$.

3. *Centre et axes principaux d'un ellipsoïde.*
Soient $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ une matrice définie positive, $\boldsymbol{c} \in \mathbb{R}^n$, et $\alpha \geq 0$.
Alors l'ellipsoïde $(\boldsymbol{x}-\boldsymbol{c})^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{c}) \leq \alpha$ est centré en $\boldsymbol{c}$ et a pour demi-axes principaux $\sqrt{\alpha\lambda_i}\boldsymbol{u}_i$ (dans le repére centré en $\boldsymbol{c}$). Les vecteurs $\boldsymbol{u}_i \in \mathbb{R}^n$ sont les vecteurs propres orthonormés de $\boldsymbol{S}$ relatifs aux valeurs propres λ_i.

⋆ 4. *Coefficient de corrélation.*
Le coefficient de corrélation $\rho_{ij} \in [0,1]$ entre 2 variables aléatoires x_i, x_j est défini par $\rho_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, où $\boldsymbol{\Sigma}$ désigne la matrice de covariance d'un vecteur dont les éléments i et j sont x_i et x_j.

5. *Approximations de la valeur de α pour les ellipsoïdes de confiance.*
Soit $\boldsymbol{x} \sim \mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}_x)$ avec $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}_x \in \mathbb{R}^{n\times n}$.
Alors la valeur de α telle que Prob$(\boldsymbol{x} \in \mathcal{E}_\alpha) = 0.5$ vaut approximativement n, et la valeur de α telle que Prob$(\boldsymbol{x} \in \mathcal{E}_\alpha) = 0.9$ vaut approximativement $n + 2\sqrt{n}$.

6. *Approximation pour les variables chi-carré à n degrés de liberté.*
(Fisher) Si z suit une χ^2_n, alors $\sqrt{2z}$ suit approximativement une normale $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = \sqrt{2n-1}$ et $\sigma^2 = 1$.

Questions

1. Déterminer comment obtenir des vecteurs $\boldsymbol{x} \in \mathbb{R}^n$ distribués selon une normale multivariée $\mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma}_x)$ à partir de vecteurs $\boldsymbol{z} \in \mathbb{R}^n$ dont les éléments sont des variables aléatoires indépendantes distribuées selon $\mathcal{N}(0,1)$: $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_n)$ où $\mathbf{I}_n$ est la matrice identité.
Indication : Transformations affines et décomposition de Cholesky.
Implémenter la méthode en écrivant une fonction

```
function xx = randn_multi(xbar, sigmax, m)
```

qui renvoie en `xx` une matrice $n \times m$ dont les `m` colonnes sont des échantillons normalement distribués selon les paramètres `xbar` (dimension n) et `sigmax` (matrice symétrique définie positive de dimension $n \times n$).
S'appuyer sur la fonction `randn` de MATLAB : `randn(m,n)` génère une matrice $m \times n$ dont chaque élément est distribué selon $\mathcal{N}(0,1)$.

2. Soit $\boldsymbol{x} = [x_1\ x_2]^T$ distribué selon $\mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma})$, avec $\bar{\boldsymbol{x}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 2\,\rho \\ 2\,\rho & 4 \end{bmatrix}$.
Pour *chacune* des 3 valeurs particulières de ρ suivantes, $\rho_1 = 0.05$, $\rho_2 = -0.50$, $\rho_3 = 0.95$, constituer *un* tableau de résultats et *une* figure, reprenant les points suivants qui concernent soit le tableau (T), soit la figure (F).
   (T) La matrice qui résulte de la décomposition de Cholesky intervenant dans la méthode développée en 1.
   (F) 1000 échantillons du vecteur $\boldsymbol{x} = [x_1\ x_2]^T$, affichés dans le plan (x_1, x_2) muni d'un repère orthonormé.

(F) La surface de l'ellipsoïde de confiance à 90 %, i.e. la courbe d'équation $(\boldsymbol{x} - \bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}) = \alpha$ où α est tel que Prob$(\boldsymbol{x} \in \mathcal{E}_\alpha) = 0.90$.
Vous pouvez utiliser une valeur approchée pour α.
Indication : Le cercle unitaire, facile à tracer en coordonnées polaires, a pour équation $||\boldsymbol{x}||_2^2 = \boldsymbol{x}^T \boldsymbol{x} = 1$. *Le cercle peut être déformé en une ellipse : raisonner avec Cholesky.*

(T) La fraction des points générés couverts par l'ellipsoïde. (Cette fraction est-elle proche de 0.9 ?)

(T) La longueur des demi-axes principaux de l'ellipsoïde.

(F) Les axes principaux de l'ellipsoïde.

3. Une formule simplifiée pour α a été donnée dans le cas $p = 0.9$ et $p = 0.5$.
Montrer en partant de Prob$(\boldsymbol{x} \in \mathcal{E}_\alpha) =$ Prob$(z \leq \alpha)$, où z suit une χ_n^2, qu'en utilisant l'approximation normale pour la χ_n^2 (Fisher) on peut établir que

$$\alpha \simeq \frac{(\Phi^{-1}(p))^2 - 1}{2} + n + (\sqrt{2}\Phi^{-1}(p))\sqrt{n}.$$

Expliquer en injectant $p = 0.9$ et $p = 0.5$ comment s'obtiennent les formules approchées.

4. On considère une variante de l'approximation de α dans le cas $p = 0.9$, à savoir $\tilde{\alpha} = 0.32 + n + 1.81\sqrt{n}$. Constituer un tableau, pour les valeurs de n de 2 à 100, qui reporte la valeur *exacte* de α, la valeur de l'approximation $\tilde{\alpha}$, ainsi que la valeur de l'approximation $n + 2\sqrt{n}$. Commenter voire si possible expliquer l'évolution des erreurs d'approximation avec n.
Indication : chi2inv dans la Statistics Toolbox de Matlab. Article "chi-square distribution" dans Wikipedia.

5. On considère $\boldsymbol{x} = [x_1\ x_2\ x_3]^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.2 & -0.4 \\ 0.2 & 1 & 0.6 \\ -0.4 & 0.6 & 1 \end{bmatrix}$.
Calculer la valeur exacte de α pour l'ellipsoïde de confiance à 95%.
Trouver également les valeurs à 2 décimales de a, b pour une approximation de la forme $\tilde{\alpha} = a + n + b\sqrt{n}$.
Générer 5000 échantillons de $\boldsymbol{x}$, et déterminer la fraction des échantillons couverts par l'ellipsoïde exact et par l'ellipsoïde qui utilise $\tilde{\alpha}$.
Constituer une figure avec les points générés dans $\mathbb{R}^3$ et les axes principaux de l'ellipsoïde de confiance, à l'aide de la fonction `plot3`.

Bon travail.