Kovarianzmatrix

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Die Kovarianzmatrix eines Zufallsvektor besteht aus den Varianzen der einzelnen Einträge auf ihrer Hauptdiagonale und den Kovarianzen zwischen verschiedenen Einträgen außerhalb der Diagonale. Zusammen mit dem Mittelwert Vektor repräsentiert sie die Wahrscheinlichkeitsverteilung des Zufallsvektor durch ihre Statistiken zweiter Ordnung, die die Streuung der Wahrscheinlichkeitsverteilung entlang jeder Richtung festlegen; für einen Gaußschen Zufallsvektor bestimmen die beiden die Wahrscheinlichkeitsverteilung vollständig, und für einen Datenpunkt, der gleichmäßig zufällig aus einem Datensatz ausgewählt wird, ist die Kovarianzmatrix die sample covariance matrix dieses Datensatz. Aus dieser Darstellung ergeben sich zwei Anwendungen. Die Inverse der Kovarianzmatrix, die Präzisionsmatrix, liefert das kleinste Risiko der Vorhersage eines Eintrags aus den anderen unter dem squared error loss: dieses Risiko ist eins geteilt durch einen Diagonaleintrag der Präzisionsmatrix. Für einen Gaußschen Zufallsvektor liefert die Präzisionsmatrix auch die Modellparameter der linearen Hypothese, die dieses Risiko erreicht. Whitening, eine Merkmalstransformation, die aus der Kovarianzmatrix gebildet wird, standardisiert die Wahrscheinlichkeitsverteilung auf Einheits-Varianzen und Null-Kovarianzen, was Gradientenabstieg beschleunigen kann.

Definition

Ein warmer Morgen in Krems an der Donau (Österreich) geht normalerweise mit einem warmen Nachmittag einher, und die Morgenstunden variieren weniger als die Nachmittagsstunden. Solche informellen Behauptungen können mithilfe eines Wahrscheinlichkeitsmodell zur Generierung von Daten in präzise mathematische Aussagen umgewandelt werden. Hierbei entspricht jede tägliche Wetterbeobachtung der Beobachtung eines Zufallsvektor $\mathbf{x} = (x_1, x_2)^T$, der aus der Lufttemperatur x_1 um 06:00 Uhr und der Lufttemperatur x_2 um 15:00 Uhr (UTC) des Tages besteht.

Die Kovarianzmatrix $\mathbf{C}^{(\mathbf{x})}$ eines Zufallsvektor $\mathbf{x}$ charakterisiert die Streuung der Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{x})}$. Insbesondere messen die Einträge von $\mathbf{C}^{(\mathbf{x})}$, wie stark jeder Eintrag von $\mathbf{x}$ variiert und wie stark Paare von Einträgen gemeinsam variieren. Verschiedene Wahrscheinlichkeitsverteilungen können denselben Mittelwert Vektor und dieselbe Kovarianzmatrix aufweisen. Für einen

Gaußschen Zufallsvektor (siehe multivariate Normalverteilung) bestimmen die beiden jedoch die Wahrscheinlichkeitsverteilung vollständig (Anderson, 2003, Thm. 2.3.1; Billingsley, 1986, Sect. 29).

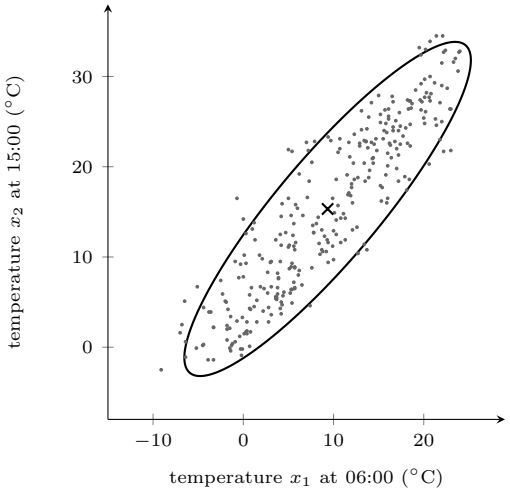
Formal ist die Kovarianzmatrix eines Zufallsvektor $\mathbf{x} \in \mathbb{R}^d$ als der Erwartungswert des äußeren Produkts des zentrierten Vektor $\mathbf{x} - \mathbb{E}\{\mathbf{x}\}$ mit sich selbst definiert (Bishop, 2006, Sect. 1.2.2, Eq. (1.42)):

$$\mathbf{C}^{(\mathbf{x})} := \mathbb{E}\Big\{ \big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)\big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)^\top \Big\}.$$

Dieser Erwartungswert ist wohldefiniert, wann immer jeder Eintrag von $\mathbf{x}$ eine endliche Varianz aufweist (Billingsley, 1986, Sect. 21, Eqs. (21.16)–(21.17)). Der Eintrag $\big(\mathbf{C}^{(\mathbf{x})}\big)_{j,j'}$ in Zeile j und Spalte j' ist die Kovarianz von x_j und $x_{j'}$. Insbesondere sind die Diagonaleinträge $\big(\mathbf{C}^{(\mathbf{x})}\big)_{j,j}$ die Varianzen der einzelnen Einträge von $\mathbf{x}$.

Eine Kovarianzmatrix kann auch verwendet werden, um die Streuung eines Datensatz zu messen, wie z.B. des Wetter-Datensatz in Abb. 1. Tatsächlich induziert ein Datensatz $\mathcal{D}$ einen Zufallsvektor $\mathbf{x}$, indem ein Datenpunkt gleichmäßig zufällig aus $\mathcal{D}$ ausgewählt wird. Die resultierende Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{x})}$ ist genau die empirische Verteilung von $\mathcal{D}$, die Wahrscheinlichkeitsverteilung, die jedem ihrer m Datenpunkte die Wahrscheinlichkeit $1/m$ zuweist (Efron and Tibshirani, 1993, Sect. 4.2). Die entsprechende Kovarianzmatrix $\mathbf{C}^{(\mathbf{x})}$ ist die sample covariance matrix von $\mathcal{D}$.

Abb. 1 zeigt den Mittelwert Vektor $\mathbb{E}\{\mathbf{x}\}$ und die Kovarianzmatrix $\mathbf{C}^{(\mathbf{x})}$ für diesen Zufallsvektor. Insbesondere zeigt sie die Niveaumenge $\big\{\mathbf{x} : \big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)^\top \big(\mathbf{C}^{(\mathbf{x})}\big)^{-1} \big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big) = 4\big\}$ (Boyd and Vandenberghe, 2004, Sect. 2.2.2). Diese Niveaumenge ist eine Ellipse mit dem Zentrum $\mathbb{E}\{\mathbf{x}\}$ und Achsen entlang der Eigenvektoren von $\mathbf{C}^{(\mathbf{x})}$. Ihre halbe Länge entlang jeder Achse ist das Doppelte der Quadratwurzel des entsprechenden Eigenwert, sodass die Ellipse die Streuung der Wahrscheinlichkeitsverteilung umreißt.



(a)

$$\mathbf{C}^{(\mathbf{x})} = \begin{pmatrix} 63.01 & 65.40 \\ 65.40 & 85.72 \end{pmatrix}$$

shaded: the Varianzen of x_1 and x_2
dashed: their Kovarianz, positive here

(b)

Abbildung 1: (a) Ein Datensatz, bestehend aus täglichen Temperaturmessungen in Krems an der Donau (Österreich). Ihre Datenpunkte können als Realisierungen eines Zufallsvektors $\mathbf{x} \in \mathbb{R}^2$ interpretiert werden: die Temperatur um 06:00 und um 15:00 Uhr (UTC) an jedem sechsten der 1824 Tage der Jahre 2020 bis 2024, für die beide Messwerte vorliegen (graue Punkte), der Mittelwerts-Vektor $\mathbb{E}\{\mathbf{x}\}$ aller 1824 Datenpunkte (Kreuz), und die Niveaumenge beim Wert vier der quadratischen Form in $\left(\mathbf{C}^{(\mathbf{x})}\right)^{-1}$ (durchgezogene Kurve). Die Kurve ist geneigt, weil x_1 und x_2 eine positive Kovarianz aufweisen. (b) Die Kovarianzmatrix aller 1824 Datenpunkte, in Quadratgrad Celsius: die schattierten Diagonaleinträge sind die Varianzenen von x_1 und x_2, und die gestrichelten Nebendiagonaleinträge sind ihre Kovarianz. Daten generiert von `pythondemos/covmtx.py`

Für jeden Vektor $\mathbf{v} \in \mathbb{R}^d$ ist die Varianz von $\mathbf{v}^\top \mathbf{x}$ gleich $\mathbf{v}^\top \mathbf{C}^{(\mathbf{x})} \mathbf{v} \geq 0$ (Anderson, 2003, Sect. 2.4, Eq. (32)). Eine Kovarianzmatrix ist daher immer positive semi-definite (psd) und, da sie auch eine symmetrische Matrix ist, lässt sie eine spectral decomposition zu (Horn and Johnson, 2013, Thm. 4.1.5)

$$\mathbf{C}^{(\mathbf{x})} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top .$$

Hier sind die Spalten von $\mathbf{V}$ orthonormale Eigenvektoren und $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_d\}$

enthält die nicht-negativen Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$. Die Kovarianzmatrix $\mathbf{C}^{(\mathbf{x})}$ ist genau dann invertible und somit positive definite (pd), wenn alle diese Eigenwerte positiv sind. Äquivalent dazu ist kein Eintrag von $\mathbf{x}$ mit Wahrscheinlichkeit eins eine affine Funktion der anderen Einträge (Anderson, 2003, Sect. 2.4, Eqs. (36)–(41)).

Da die Kovarianzmatrix die Streuung der Wahrscheinlichkeitsverteilung repräsentiert, bestimmt sie auch die fundamentale Grenze und die optimale Methode zur Vorhersage eines Eintrags von $\mathbf{x}$ aus den anderen. Die Inverse einer invertiblen Kovarianzmatrix, die als precision matrix bezeichnet wird (Bishop, 2006, Sect. 2.3.1), löst diese Aufgabe durch ihren Diagonaleintrag $\left(\left(\mathbf{C}^{(\mathbf{x})}\right)^{-1}\right)_{j,j}$. Die verbleibenden Merkmale bilden den Vektor $\mathbf{x}_{-j} := \left(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d\right)^\top \in \mathbb{R}^{d-1}$. Für einen Gaußschen Zufallsvektor $\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{C}^{(\mathbf{x})}\right)$ (siehe multivariate Normalverteilung) ist die conditional probability distribution von x_j, gegeben $\mathbf{x}_{-j}$, Gaußsch mit einer Varianz, die nicht vom Wert von $\mathbf{x}_{-j}$ abhängt. Diese bedingte Varianz ist $1/\left(\left(\mathbf{C}^{(\mathbf{x})}\right)^{-1}\right)_{j,j}$ (Bishop, 2006, Sect. 2.3.1, Eq. (2.73)).

Diese bedingte Varianz wird zu einer Grenze für jede machine learning (ML)-Methode, sobald die Datenpunkte als Realisierungen von unabhängig und identisch verteilt (u.i.v.)en Zufallsvektoren $\mathbf{x}^{(r)} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{C}^{(\mathbf{x})}\right)$ approximiert werden können. Die bedingte Varianz ist auch das kleinste Risiko, das jede Hypothese unter dem squared error loss erreichen kann,

$$\min_{h:\mathbb{R}^{d-1}\to\mathbb{R}} \mathbb{E}\left\{\left(x_j - h\left(\mathbf{x}_{-j}\right)\right)^2\right\} = 1/\left(\left(\mathbf{C}^{(\mathbf{x})}\right)^{-1}\right)_{j,j}. \tag{1}$$

Dieses kleinste Risiko ist das Bayes risk des regression-Problems der Vorhersage von x_j aus $\mathbf{x}_{-j}$. Es wird durch den Bayes estimator $\mathbb{E}\{x_j \mid \mathbf{x}_{-j}\}$ erreicht, der im Gaußschen Fall linear von $\mathbf{x}_{-j}$ abhängt (Bertsekas and Tsitsiklis, 2008, Sect. 8.3; Bishop, 2006, Sect. 2.3.1). Für die Wetterdaten aus Abb. 1 hat die beste lineare Hypothese, die die Temperatur eines Tages um 15:00 Uhr aus ihrer Temperatur um 06:00 Uhr vorhersagt, ein Risiko von 17.8 Quadratgrad Celsius, gegenüber der Varianz von 85.7 der Temperatur um 15:00 Uhr, was das Risiko der besten konstanten Vorhersage ist. Ihre beiden Modellparameter werden aus $\mathbf{C}^{(\mathbf{x})}$ und $\mathbb{E}\{\mathbf{x}\}$ berechnet: die Steigung ist das Verhältnis der Kovarianz zur Varianz der Temperatur um 06:00 Uhr, und der Achsenabschnitt wird durch die Mittelwerte der Temperatur um 06:00 Uhr und um 15:00 Uhr festgelegt (Bertsekas and Tsitsiklis, 2008, Sect. 8.4) (siehe lineare Regression). Unter einem Gaußschen Wahrscheinlichkeitsmodell ist dieses Risiko auch das Bayes risk. Durch die Bayes risk-Identität (1) liefert die Diagonale von $\left(\mathbf{C}^{(\mathbf{x})}\right)^{-1}$ eine Baseline.

Die spectral decomposition liefert auch eine Merkmalstransformation, die die dargestellte Wahrscheinlichkeitsverteilung standardisiert: sie beseitigt Skalenunterschiede und lineare Korrelationen zwischen den Einträgen von $\mathbf{x}$. Wenn

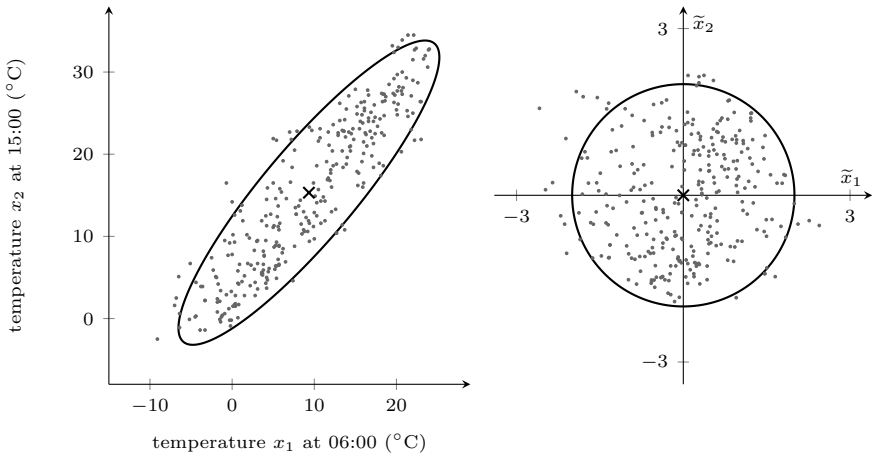
$\mathbf{C}^{(\mathbf{x})}$ invertible ist, ist ihre inverse Quadratwurzel

$$\big(\mathbf{C}^{(\mathbf{x})}\big)^{-1/2} := \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}^\top,$$

wobei $\mathbf{\Lambda}^{-1/2}$ durch Ersetzen jedes Eigenwert durch seine inverse Quadratwurzel erhalten wird. Der transformierte Merkmalsvektor

$$\widetilde{\mathbf{x}} := \big(\mathbf{C}^{(\mathbf{x})}\big)^{-1/2}\big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)$$

hat einen Mittelwert von Null und die Kovarianzmatrix $\mathbf{C}^{(\widetilde{\mathbf{x}})} = \mathbf{I}$. Somit haben ihre Einträge eine Einheits-Varianz und sind wechselseitig unkorreliert, wie in Abb. 2 dargestellt. Diese Transformation ist als Whitening bekannt (Bishop, 2006, Sect. 12.1.3, Eqs. (12.24)–(12.25)). Durch das Beseitigen von Skalenunterschieden und linearen Korrelationen zwischen den Merkmale kann Whitening die Konditionierung von Optimierungsprobleme verbessern und dadurch Gradientenabstieg und verwandte gradientenbasierte Verfahren beschleunigen. Für lineare Regression mit dem squared error loss auf zentrierten Merkmale ist die Hessian des Risiko $2\mathbf{C}^{(\mathbf{x})}$, und ein Gradientenabstieg-Schritt mit einer exakten Liniensuche verringert den Abstand zum kleinsten Risiko einer linearen Hypothese um einen Faktor von höchstens $1 - 1/\kappa\big(\mathbf{C}^{(\mathbf{x})}\big)$, wobei die condition number $\kappa\big(\mathbf{C}^{(\mathbf{x})}\big)$ das Verhältnis des größten zum kleinsten Eigenwert ist (Boyd and Vandenberghe, 2004, Sect. 9.3.1). Nach dem Whitening, $\mathbf{C}^{(\widetilde{\mathbf{x}})} = \mathbf{I}$, ist dieses Verhältnis eins, und ein einziger Schritt erreicht das Minimum. In der Praxis werden der Mittelwert und die Kovarianzmatrix, die für diese Transformation verwendet werden, ausschließlich aus dem Trainingsdatensatz geschätzt.



(a) (b)

Abbildung 2: Die Wetterdaten von Abb. 1 (a) vor und (b) nach der Multiplikation jedes zentrierten Merkmalsvektors mit $\left(\mathbf{C}^{(\mathbf{x})}\right)^{-1/2}$. Die Kovarianzellipse von (a) wird zum Kreis mit Radius zwei in (b), der Niveaumenge beim Wert vier von $\widetilde{\mathbf{x}}^\top\widetilde{\mathbf{x}}$. Die transformierten Merkmalee $\widetilde{x}_1$ und $\widetilde{x}_2$ haben eine Einheits-Varianz und eine Null-Kovarianz. Daten generiert von `pythondemos/covmtx.py`

Siehe auch: Kovarianz, Matrix, Zufallsvariable (ZV), Zufallsvektor, Varianz, sample covariance matrix, multivariate Normalverteilung, spectral decomposition, Eigenwertzerlegung, Eigenwert, lineare Regression, Bayes risk, Baseline, precision matrix.

Querverweise

Literatur

1. Anderson (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., Wiley, Hoboken, NJ.
2. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.
3. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.
4. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.
5. Efron and Tibshirani (1993). An Introduction to the Bootstrap. Chapman & Hall, New York, NY, USA.

6. Boyd and Vandenberghe (2004). Convex Optimization. Cambridge Univ. Press, Cambridge, U.K.

7. Horn and Johnson (2013). Matrix Analysis. 2nd ed., Cambridge Univ. Press, New York, NY, USA.