Einführung in die Hauptkomponentenanalyse

Florian Steinke

16. Juni 2009

1 Vorbereitung: Einige Aspekte der multivariaten Gaußverteilung

Definition 1.1. *Die 1-D Gaußverteilung für $x \in \mathbb{R}$ ist*

$$p(x) \propto \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{1}$$

Notation: $x \sim N(x; \mu, \sigma^2)$.

Eigenschaften:

- Mittelwert $E(x) = \mu$, Varianz $E((x - E(x))^2) = \sigma^2$.

Beipiel 1.2. *Sei $\boldsymbol{x} \in \mathbb{R}^2$ und*

$$\begin{aligned} p(\boldsymbol{x}) &= N(x_1; \mu_1, \sigma_1^2) N(x_2; \mu_2, \sigma_2^2) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \begin{pmatrix} \sigma_1^{-2} & \\ & \sigma_1^{-2} \end{pmatrix} (\boldsymbol{x}-\boldsymbol{\mu})\right) \end{aligned}$$

Beipiel 1.3. *Sei $\boldsymbol{x} \in \mathbb{R}^2$, $\boldsymbol{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$, und*

$$\begin{aligned} p(\boldsymbol{x}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{U}(\boldsymbol{x}-\boldsymbol{\mu}))^T \begin{pmatrix} \sigma_1^{-2} & \\ & \sigma_1^{-2} \end{pmatrix} \boldsymbol{U}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \underbrace{\boldsymbol{U}^T \begin{pmatrix} \sigma_1^{-2} & \\ & \sigma_1^{-2} \end{pmatrix} \boldsymbol{U}}_{\equiv \boldsymbol{\Sigma}^{-1}} (\boldsymbol{x}-\boldsymbol{\mu})\right) \end{aligned}$$

Also $\boldsymbol{\Sigma} = \boldsymbol{U}\begin{pmatrix} \sigma_1^2 & \\ & \sigma_1^2 \end{pmatrix}\boldsymbol{U}^T$.

Definition 1.4. *Die M-D Gaußverteilung für $x \in \mathbb{R}^M$ ist*

$$p(x) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2}$$

Notation: $\boldsymbol{x} \sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Eigenschaften:

- Mittelwert

$$E(\boldsymbol{x}) = \boldsymbol{\mu} \tag{3}$$

  Kovarianz

$$E((\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T) = \boldsymbol{\Sigma}. \tag{4}$$

- $\boldsymbol{\Sigma}$ muss symmetrisch positiv definit (spd) sein.

  Proof. Symmetrisch ist klar nach (4). Zum Beweis, dass $\boldsymbol{\Sigma}$ positiv definit ist, wähle $\boldsymbol{a} \neq 0 \in \mathbb{R}^M$. Sei o.B.d.A. $\boldsymbol{\mu} = 0$. Dann ist nach (4)

  $$\boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a} = E(\boldsymbol{a}^T \boldsymbol{x} \boldsymbol{x}^T \boldsymbol{a}) = E((\boldsymbol{a}^T \boldsymbol{x})^2) \geq 0.$$

  Falls $E((\boldsymbol{a}^T \boldsymbol{x})^2) = 0$, dann ist mit Wahrscheinlichkeit eins $\boldsymbol{x} = 0$ in Richtung $\boldsymbol{a}$, d.h. der Ellipsoid der Gaußverteilung ist **entartet** in Richtung $\boldsymbol{a}$. □

- Jede Gaußverteilung läßt sich wie in Beispiel 1.3 darstellen.

  Proof. Da $\boldsymbol{\Sigma}$ spd, hat es eine Eigenwertdarstellung mit orthogonalen Matrizen, d.h. man kann schreiben $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, wobei $\boldsymbol{U}$ orthogonal, d.h. $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \mathbf{1}$, und $\boldsymbol{D} = \text{diag}(\lambda_1, ..\lambda_M)$ diagonal. Da $\boldsymbol{\Sigma}$ positiv definit, sind die Eigenwerte $\lambda_i > 0$ und $\sigma_i^2 = \lambda_i$. □

  Bemerkung: $\boldsymbol{U}, \boldsymbol{D}$ sind bis auf Permutationen eindeutig bestimmt (falls alle Eigenwerte verschieden). Wir wählen typischerweise die Darstellung $\lambda_1 > \lambda_2 > ... > \lambda_M$.

- Alle 1-D Schnitte einer M-D Gaußverteilung sind 1-D Gaußverteilt, d.h. für $\boldsymbol{x} = \boldsymbol{a} + t\boldsymbol{b}$ ist t immer 1-D Gaußverteilt.

  Proof. (für den Fall $\boldsymbol{a} = \boldsymbol{\mu} = 0$). Es ist

  $$\begin{aligned} p(\boldsymbol{x})d\boldsymbol{x} &= p(\boldsymbol{x}_t)\frac{d\boldsymbol{x}}{dt}dt \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{b}t^2\right)dt = N(t; 0, (\boldsymbol{b}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{b})^{-1})dt. \end{aligned}$$

  □

- Für lineare Funktionen $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ gilt: Aus $\boldsymbol{x} \sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ folgt $\boldsymbol{y} \sim N(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
- Mit Rauschen $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}; 0, \boldsymbol{\Sigma}')$ und $\boldsymbol{x} \sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ gilt für $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$, dass $\boldsymbol{y} \sim N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}')$.

Kleiner Ausblick auf weitere wichtige Eigenschaften der Gaußverteilung:

- Konditionierung und Marginalisierung: Sei $\boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_1 \\ \boldsymbol{x}_2\end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}$, $\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}12 \\ \boldsymbol{\Sigma}21 & \boldsymbol{\Sigma}22\end{pmatrix}$ und $\boldsymbol{x} \sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Dann ist

  $$p(\boldsymbol{x}_1) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2)d\boldsymbol{x}_2 = N(\boldsymbol{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \quad (5)$$

  $$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = N(\boldsymbol{x}_1; \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \quad (6)$$

- Die Familie der Gaußverteilungen ist geschlossen unter Multiplikation (Division), Faltung, Marginalisierung und Conditionierung.
- $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ ist die Verteilung, die gegeben den Mittelwert $\boldsymbol{\mu}$ und die Kovarianz $\boldsymbol{\Sigma}$, die maximale Entropie hat! Keine unnötigen Annahmen!
- Die Gaußveteilungen sind Teil der "exponential family" oder der "log-linear models", d.h. man kann schreiben

  $$p(\boldsymbol{x}) \propto \exp\left(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\right).$$

  Zum Beispiel im 1-D Fall $\boldsymbol{\theta} = \begin{pmatrix}\frac{1}{\sigma^2} \\ \frac{\mu}{\sigma^2}\end{pmatrix}$ und $\boldsymbol{\phi}(\boldsymbol{x}) = \begin{pmatrix}\frac{x^2}{2} \\ x\end{pmatrix}$.

  Die Tatsache, dass sich diese Verteilungen linear parameterisieren lassen, ist in vielen Anwendungen in der Statistik/im Machine Learning extrem wichtig und hilfreich.

All diese Eigenschaften machen die Gaußverteilungen zu einer flexiblen, ausdrucksstarken und leicht handhabbaren Klasse von Verteilungen. Sie werden daher sehr gerne im Machine Learning verwendet. (Häretiker könnten sagen, dass derjenige, der die Gaußverteilungen vollständig verstanden hat, auch das Machine Learning komplett beherrscht ;)

2 Grundzüge des Unüberwachten Lernens

Englisch: Unsupervised Learning

Problem 2.1. *Gegeben seien Datenpunkte $\boldsymbol{x}_i \in \mathbb{R}^M$, $i = 1, .., N$ (z.B. digitalisierte Bilder). Wie kann ich diese Daten effizient und intuitiv beschreiben?*

Ein paar einfache Modelle:

- Mittelwert + Noise/Rauschen:

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \tag{7}$$

- mehrere Mittelwerte + Noise/Rauschen

$$\boldsymbol{x}_i = \sum_{j=1}^{r} \delta_{z_j,j} \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_i, \quad z_j \in \{1, .., r\} \tag{8}$$

  Dieses Modell entspricht **Clustering** und wäre genügend Stoff für eine eigene Vorlesung. Einige Stichworte sind k-Means, Bayesian k-Means, spectral clustering, hierarchical clustering, ...

- linear Modelle

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \sum_{j=1}^{r} \boldsymbol{u}_j w_{ji} + \boldsymbol{\epsilon}_i \tag{9}$$

  Dies führt zur heutigen Vorlesung, der **Hauptkomponentenanalyse (Englisch: Principal Component Analysis = PCA)**.

- nicht lineare Modelle. Ein sehr aktives Forschungsgebiet, bekannt als **manifold learning**. Stichworte sind z.B. kernelPCA, locally linear maps (LLE), ISOMAP, bi-directional GP, ...

3 Hauptkomponentenanalyse - PCA

Beobachtung 3.1. *Das PCA-Problem, d.h. die Rekonstruktion von $\boldsymbol{u}_j$ und $\boldsymbol{\mu}$ gegeben nur die Datenpunkte $\boldsymbol{x}_i$, ist stark unterbestimmt. Die Anzahl der Unbekannten ist $M + Mr + rN + MN$, die Zahl der Gleichungen nur MN. Dies bedeutet, dass wir starke Annnahmen machen müssen, um eine eindeutige Lösung zu erhalten!*

Annahmen zur Herleitung der PCA:

- $\boldsymbol{\epsilon}_i$ Gaußverteilt:

$$\boldsymbol{\epsilon}_i \sim N(\boldsymbol{\epsilon}_i; 0, \sigma_0^2 \mathbf{1})$$

  Außerdem nehmen wir an, dass $\boldsymbol{\epsilon}_i$ klein ist. D.h. σ_0^2 klein.

- w_{ji} Gaußverteilt,

$$w_{ji} \sim N(w_{ji}; 0, \sigma_j^2)$$

  Dies bedeutet dass w_{ji} unabhängig (**unkorreliert**) von $w_{j'i}$ für $j' \neq j$!

  Annahme zur Notation: $\sigma_1 > \sigma_2 > .. > \sigma_r$.

- Sei $\boldsymbol{U} = [\boldsymbol{u}_1..\boldsymbol{u}_r]$. Wir nehmen an, dass $\boldsymbol{U}$ orthogonal, d.h. $\boldsymbol{u}_i \perp \boldsymbol{u}_j$ für $j \neq i$ und $\|\boldsymbol{u}_j\| = 1$ für $j = 1, .., r$.

Herleitung der PCA:

- Neue, äquivalente Notation:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{w} + \boldsymbol{\epsilon}$$

  Hier ist $\boldsymbol{U} \in \mathbb{R}^{M\times r}$, $\boldsymbol{\mu} \in \mathbb{R}^M$ konstant, und $\boldsymbol{w}$ zufällig verteilt wie $\boldsymbol{w} \sim N(\boldsymbol{w}; 0, \boldsymbol{D})$ mit $\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, .., \sigma_r^2)$. Außerdem ist $\boldsymbol{\epsilon} \in \mathbb{R}^M$ verteilt nach $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}; 0, \sigma_0^2\mathbf{1})$. Damit ist

$$\boldsymbol{x} \sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T + \sigma_0^2\mathbf{1}).$$

- Nun ergänze $\boldsymbol{U}$ zu einer orthonormal Basis $\bar{\boldsymbol{U}} \in \mathbb{R}^{M\times M}$, d.h. $\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T = \mathbf{1}$ und

$$\begin{aligned} \boldsymbol{x} &\sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T + \sigma_0^2\mathbf{1}) \\ &\sim N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T + \sigma_0^2\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T) \\ &\sim N(\boldsymbol{x}; \boldsymbol{\mu}, \underbrace{\bar{\boldsymbol{U}}\bar{\boldsymbol{D}}\bar{\boldsymbol{U}}^T}_{=\boldsymbol{\Sigma}}), \end{aligned}$$

  wobei $\bar{\boldsymbol{D}} = \mathrm{diag}(\sigma_1^2 + \sigma_0^2, .., \sigma_r^2 + \sigma_0^2, \sigma_0^2, .., \sigma_0^2)$.

Idee PCA: Schätze empirische Kovarianz $\hat{\boldsymbol{\Sigma}}$ aus den gegeben Daten und berechne Eigenwert Zerlegung. Die Eigenvektoren entsprechen den gesuchten Basisvektoren!

Kochrezept PCA:

1. Schätze Mittelwert und normalisiere Daten auf Mittelwert 0.

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \boldsymbol{x}_i, \quad \tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \hat{\boldsymbol{\mu}}$$

   (Manchmal, z.B. bei LSA, wird dieser Schritt weggelassen.)

2. Schätze Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_i \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T$$

3. Berechne Eigenwertzerlegung von $\hat{\boldsymbol{\Sigma}}$

$$\hat{\boldsymbol{\Sigma}} = \hat{\bar{\boldsymbol{U}}}\hat{\bar{\boldsymbol{D}}}\hat{\bar{\boldsymbol{U}}}^T$$

   wobei $\hat{\bar{\boldsymbol{D}}} = \mathrm{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_M)$. Permutiere die Eigenwertzerlegung so, dass $\hat{\lambda}_1 > ... > \hat{\lambda}_M$.

4. Bestimme $\hat{r}$ aus dem Abfall des Eigenwertspektrum (Zur Erinnerung: $\sigma_0^2 < \sigma_1^2, .., \sigma_{\hat{r}}^2$)

Die ersten $\hat{r}$ Spalten von $\hat{\bar{\boldsymbol{U}}}$ sind dann die gesuchten Basisvektoren $\hat{\boldsymbol{U}} = [\hat{\boldsymbol{u}}_1, .., \hat{\boldsymbol{u}}_{\hat{r}}]$. Die Eigenwerte $\hat{\lambda}_j$ sind Schätzungen der Varianz in der jeweiligen Richtung $\hat{\sigma}_j^2 = \hat{\lambda}_j$. Das Abschneiden des Eigenwertspektrums durch ein geeignete Wahl von $\hat{r}$ begründet dann auch den Namen "Hauptkomponentenanalyse".

Beweisskizze PCA: Die Eigenwertzerlegung ist eindeutig, falls alle Eigenwerte verschieden und richtig sortiert.

Bemerkungen zur PCA:

- Für einen neuen Datenpunkt $\boldsymbol{z} \in \mathbb{R}^M$ erhalten wir die Hauptkomponenten $\boldsymbol{w}$ einfach durch

$$\boldsymbol{w} = \hat{\boldsymbol{U}}^T(\boldsymbol{z} - \hat{\boldsymbol{\mu}}).$$

- Effiziente Berechnung der PCA: Die oben vorgestellte Methode skaliert $O(M^3)$ wegen der Eigenwertzerlegung. Für große M ist dies unpraktisch. Man kann die Eigenwertzerlegung aber mit Hilfe einer **Singulärwertzerlegung** umgehen. Falls $N \ll M$ ist dies viel effizienter, skaliert nämlich mit $O(MN^2)$.

  Sei $\boldsymbol{X} = [\tilde{\boldsymbol{x}}_1..\tilde{\boldsymbol{x}}_N]$. Die Schätzung der Kovarianz kann dann auch geschrieben werden als $\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$. Jetzt besitzt aber jede Matrix eine Singulärwertzerlegung. D.h. für $\boldsymbol{X}$ können wir schreiben

$$\boldsymbol{X} = \boldsymbol{U}'\boldsymbol{S}\boldsymbol{V}',$$

wobei $\boldsymbol{U'} \in \mathbb{R}^{N\times N}$ und $\boldsymbol{V'} \in \mathbb{R}^{M\times M}$ unitär/orthogonal und $\boldsymbol{S} \in \mathbb{R}^{M\times N}$ diagonal ist, d.h. $\boldsymbol{S} = \text{diag}(\sigma_1, .., \sigma_{\min(N,M)})$. Bisher haben wir berechnet

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T = \hat{\boldsymbol{U}}\hat{\boldsymbol{D}}\hat{\boldsymbol{U}}^T,$$

mit der Singulärwertzerlegung erhalten wir

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T = \frac{1}{N}\boldsymbol{U'}\boldsymbol{S}\boldsymbol{V'}\boldsymbol{V'}^T\boldsymbol{S}^T\boldsymbol{V'}^T = \frac{1}{N}\boldsymbol{U'}\underbrace{\boldsymbol{S}\boldsymbol{S}^T}_{=\boldsymbol{D'}}\boldsymbol{V'}^T,$$

wobei $\boldsymbol{D'} = \text{diag}(\sigma_1^2, .., \sigma_{\min(N,M)}^2, 0, ..., 0)$ diagonal ist.

Zusammen bedeutet dies, dass wir bei der PCA aus einer kostengünstig berechenbaren Singulärwertzerlegung dieselben Ergebnisse wie aus einer Eigenwertzerlegung erhalten können.

(Mann kann sogar $M \to \infty$ gehen lassen, siehe kernelPCA!)

- Wir haben die PCA unter der Annahme eines Gaußverteilten linearen Modells hergeleitet. Oft wird PCA aber auch dann verwandt, wenn die **Modellannahmen verletzt** sind oder gar **kein Modell bekannt** ist. Oft liefert dies trotzdem gute Ergebnisse, siehe der experimentelle Teil der Vorlesung.

  Der wichtige Grundgedanke bei der PCA ist, dass die Datenpunkte auf einer Linie/Fläche (einem linearen Unterraum des $\mathbb{R}^M$) liegen und dass wir diesen aus der Kovarianz der Daten schätzen können.