Schätzung der Kovarianzmatrix

Aus einem Ensemble von Beobachtungen $\{\mathbf{x}_i\}$ kann die *Kovarianzmatrix* (Zentralmomente) geschätzt werden:

$$\mathbf{C}_{xx} = E\{(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T\} = \mathbf{R}_{xx} - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T$$

Schätzwert (endliche Summe): $\hat{\mathbf{C}}_{xx} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_x)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_x)^T$

und dem Schätzwert: $\hat{\boldsymbol{\mu}}_x = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$

$\hat{\mathbf{C}}_{xx}$ wird also aus der Summe von Matrizen vom Rang 1 berechnet:

$(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_x)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_x)^T$

da in dem dyadischen Produkt nur Vielfache des Zeilenvektors

$(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_x)^T$ bzw. Spaltenvektors $(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_x)$ vorkommen, wegen:

$$\mathbf{x}\mathbf{y}^T = \begin{bmatrix} x_1\cdot\mathbf{y}^T \\ x_2\cdot\mathbf{y}^T \\ \vdots \\ x_N\cdot\mathbf{y}^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}\cdot y_1 & \mathbf{x}\cdot y_2 & \cdots & \mathbf{x}\cdot y_N \end{bmatrix}$$

Problem der hohen Merkmalsdimensionalität

$\hat{\mathbf{C}}$ ist somit singulär, wenn weniger als $n=N$, mit $N=\dim(\mathbf{x})$, unabhängige Beobachtungen des Ensembles verfügbar sind!!

Dies ist ein Problem, wenn die Anzahl der Merkmale sehr groß ist und nur wenige Stichproben des Ensembles zur Verfügung stehen.

Die Güte der Schätzung wird allerdings erst mit $n>>N$ verbessert.

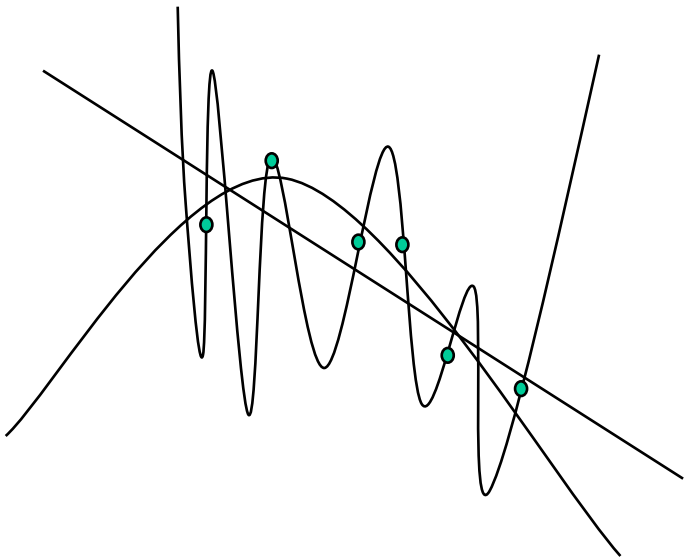
Außerdem wird nicht $\hat{\mathbf{C}}$, sondern $\hat{\mathbf{C}}^{-1}$ benötigt!

Was kann man tun, wenn eine zu geringe Stichprobe zur Verfügung steht? Man kann

- die Anzahl der Merkmale durch eine KLT reduzieren, oder
- Man vereinfacht das Modell und damit die Anzahl der Parameter: man nimmt z.B. Unkorreliertheit der Merkmale an und setzt alle Nebendiagonalelemente zu Null, wodurch die Invertierbarkeit erzwungen wird. Obwohl diese Vorgehensweise eigentlich inkorrekt ist, ergeben sich durch diese Heuristik häufig brauchbare Ergebnisse.

Zum Problem der geringen Stichprobe

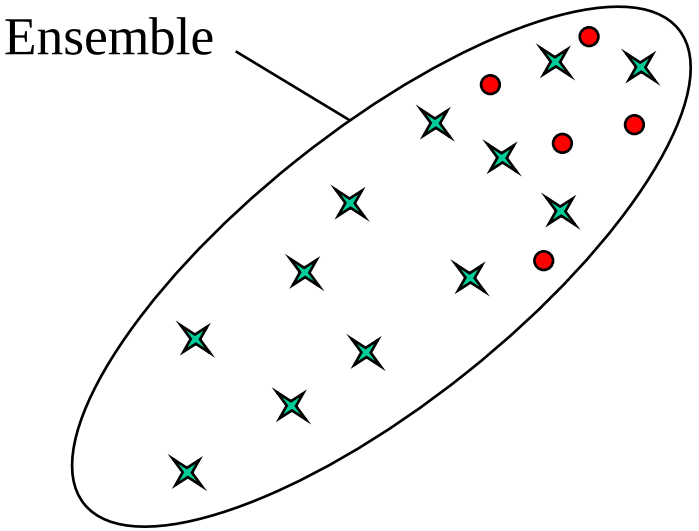
Der resultierende Klasssifikator unter der Zwangsannahme der statistischen Unabhängigkeit ist sicherlich suboptimal. Dies hängt zusammen mit dem Problem der unzureichenden Stichprobe. Man kann es vergleichen mit dem Problem des Kurven-Fitting. Das Bild zeigt 6 Datenpunkte und verschiedene Polynome zum Fitten. Die Datenpunkte wurden erzeugt durch Hinzufügen von mittelwertfreien, unabhängigem Rauschen zu einer Parabel. Deshalb sollte eine Parabel den besten Fit ergeben, wenn wir annehmen, dass weitere Stichproben hinzukommen und die 6 Punkte ergänzen (Generalisierung).



Kurvenapproximation an eine Menge von Punkten

- Die *Gerade* ergibt eine *brauchbare* Näherung.
- Die *Parabel* ergibt eine *bessere* Approximation, aber man kann sich fragen, ob die Stichprobe gut geeignet war, die Parabel festzulegen. Die Parabel für eine größere Stichprobe könnte ganz woanders liegen und im betrachteten Intervall könnte die Gerade die bessere Näherung sein.
- Overfitting: Das Polynom 10. Grades ergibt einen perfekten Fit. Aber man kann nicht erwarten, dass solch eine unterbestimmte Näherung neue Stichproben gut approximiert. Es müßten sehr viel mehr Stichproben zur Verfügung stehen, um eine ähnlich gute Approximation von einem Polynom 10. Grades im Vergleich zu einem Parabelfit zu bekommen, trotz der Tatsache, dass das Letztere ein Sonderfall (n=2) des Ersten ist.

Regel: je kleiner die Stichprobe, desto einfacher sollte auch das Modell gewählt werden



Im allg. gilt: Zuverlässige Inter- und Extrapolation kann nur bei stark überbestimmten Lösungen erwartet werden (hinreichend großer Stichprobenumfang).

Also: Wenn eine exakte statistische Modellierung gegeben wäre, dann ist mit dem MAP-Ansatz unser Problem gelöst. In der Praxis stellt sich jedoch in der Regel das Problem, aus einer endlichen Stichprobe einen guten Klassifikator herzuleiten.

- Stichprobe 1 (repräsentativ)
- Stichprobe 2 (nicht repräsentativ)

Problem der Generalisierungsfähigkeit eines Klassifikators

Wie reagiert ein Klassifikator, welcher auf eine endliche Stichprobe aufbaut, auf neu hinzukommende Experimente (Problem der Inter- und Extrapolation)?

Man unterscheidet deshalb zwischen einer *Trainings*- (*Lern*-) und einer *Testmenge*.

Die Überprüfung der Leistungsfähigkeit nur anhand des Lernsatzes bezeichnet man als *Reklassifikation* (dabei kann man einen idealen Fit erreichen) und die Überprüfung anhand eines unabhängigen Testdatensatzes bezeichnet man als *Generalisierung* (Inter- und Extrapolationsfähigkeit).

Je größer die Anzahl der Parameter der in der Klassifikation verwendeten Schätzfunktion, desto größer muss der Stichprobenumfang der Trainingsmenge sein.

Rekursive Schätzung der statistischen Kenngrößen

Kommen während einer Erkennungsaufgabe fortwährend neue Stichproben hinzu, so ist es vorteilhaft, die statistischen Kenngrößen *rekursiv* zu schätzen. Dies ist mit wesentlich weniger Aufwand verbunden, als von dem erweiterten Stichprobenumfang die Grundgleichungen immer wieder erneut zu lösen (lernende bzw. adaptive Vorgehensweise, *batch* estimate versus *recursive* estimate).

Für die Schätzung des Erwartungswerts gilt:

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k = \frac{1}{n}\left(\sum_{k=1}^{n-1}\mathbf{x}_k + \mathbf{x}_n\right)$$
$$= \left(1-\frac{1}{n}\right)\hat{\boldsymbol{\mu}}_{n-1} + \frac{1}{n}\mathbf{x}_n = \hat{\boldsymbol{\mu}}_{n-1} + \frac{1}{n}\left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1}\right)$$

Die Schätzung wird in jedem Schritt proportional zur Abweichung zwischen der der derzeitigen Schätzung und der derzeitigen Beobachtung verändert.

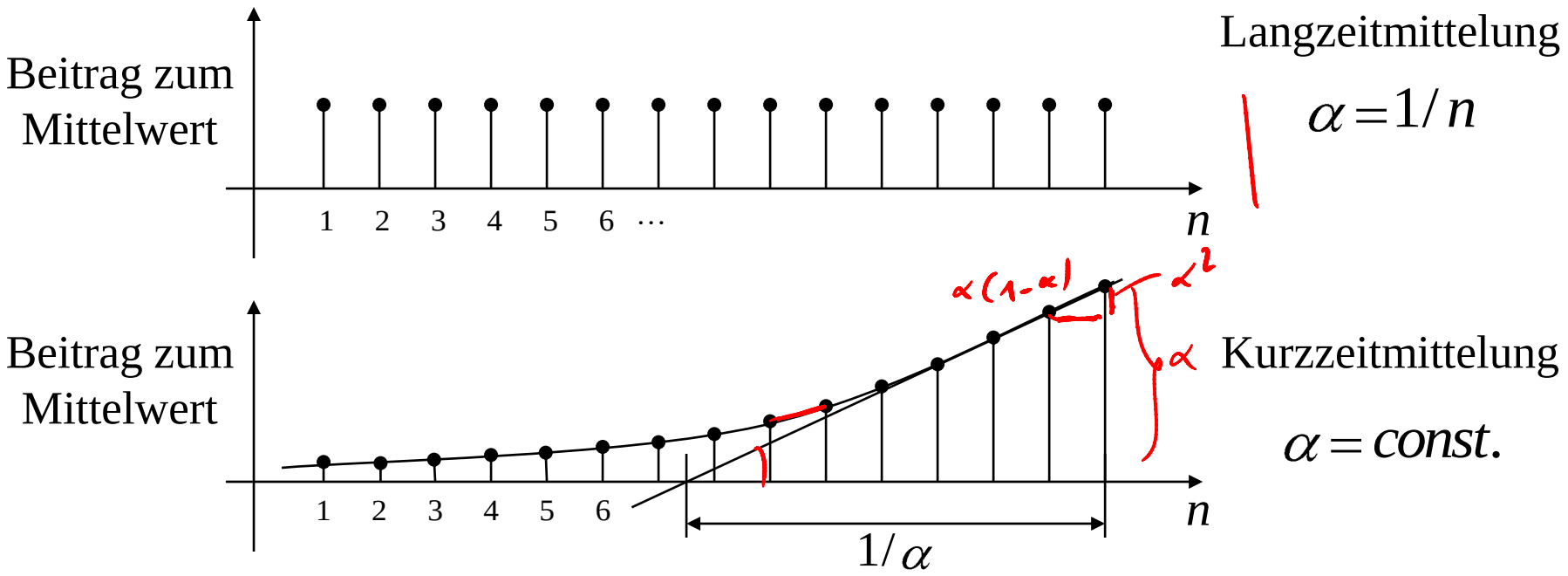
Eine Verallgemeinerung der obigen Rekursion ergibt:

$$\hat{\boldsymbol{\mu}}_n = \hat{\boldsymbol{\mu}}_{n-1} + \alpha\left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1}\right) = \left(1-\alpha\right)\hat{\boldsymbol{\mu}}_{n-1} + \alpha\mathbf{x}_n$$

$$\text{mit:}\quad \alpha = \begin{cases} 1/n & \text{stationär} \\ const. & \text{quasi-stationär} \end{cases}$$

Mit $\alpha=1/n$ werden *stationäre* Verhältnisse angenommen, d.h. alle Beobachtungen haben unabhängig von der Zeit ihres Auftretens das gleiche Gewicht, d.h. die letzten Beobachtungen sind genauso wichtig wie die ersten.

Bei $\alpha=const.$ wird eine *Fluktuation* akzeptiert, d.h. die neueren Beobachtungen haben ein größeres Gewicht als die alten (exponential smoothing). Das Beobachtungsfenster ist näherungsweise gegeben durch $1/\alpha$ mit $\alpha=const.$



Rekursive Schätzung der Kovarianzmatrix

Für die *Korrelationsmatrix* (2. Momente) erhält man die Rekursion:

$$\boxed{\hat{\mathbf{R}}_n = (1-\alpha)\hat{\mathbf{R}}_{n-1} + \alpha \mathbf{x}_n \mathbf{x}_n^T}$$

Für die rekursive Berechnung der *Kovarianzmatrix* wird $\hat{\boldsymbol{\mu}}$ benötigt, was durch eine zweite Rekursion zu ermitteln ist:

$$\begin{aligned}
\hat{\mathbf{C}}_n &= \hat{\mathbf{R}}_n - \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T \\
&= [(1-\alpha)\hat{\mathbf{R}}_{n-1} + \alpha \mathbf{x}_n \mathbf{x}_n^T] - [(1-\alpha)\hat{\boldsymbol{\mu}}_{n-1} + \alpha \mathbf{x}_n][(1-\alpha)\hat{\boldsymbol{\mu}}_{n-1} + \alpha \mathbf{x}_n]^T \\
&= (1-\alpha)\hat{\mathbf{R}}_{n-1} + \alpha \mathbf{x}_n \mathbf{x}_n^T - (1-\alpha)^2 \hat{\boldsymbol{\mu}}_{n-1}\hat{\boldsymbol{\mu}}_{n-1}^T - \alpha(1-\alpha)[\hat{\boldsymbol{\mu}}_{n-1}\mathbf{x}_n^T + \mathbf{x}_n \hat{\boldsymbol{\mu}}_{n-1}^T] - \alpha^2 \mathbf{x}_n \mathbf{x}_n^T \\
&= (1-\alpha)[\hat{\mathbf{R}}_{n-1} - \hat{\boldsymbol{\mu}}_{n-1}\hat{\boldsymbol{\mu}}_{n-1}^T + \alpha(\mathbf{x}_n \mathbf{x}_n^T - \hat{\boldsymbol{\mu}}_{n-1}\mathbf{x}_n^T - \mathbf{x}_n \hat{\boldsymbol{\mu}}_{n-1}^T + \hat{\boldsymbol{\mu}}_{n-1}\hat{\boldsymbol{\mu}}_{n-1}^T)] \\
&= (1-\alpha)[\hat{\mathbf{C}}_{n-1} + \alpha(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})^T]
\end{aligned}$$

Rekursive Schätzung der Kovarianzmatrix

Also beide Rekursionen zusammen:

$$\hat{\mathbf{C}}_n = (1-\alpha)[\hat{\mathbf{C}}_{n-1} + \alpha(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})^T]$$
$$\hat{\boldsymbol{\mu}}_n = (1-\alpha)\hat{\boldsymbol{\mu}}_{n-1} + \alpha\mathbf{x}_n$$

Rekursive Schätzung der inversen Korrelationsmatrix

Für die Berechnung des Mahalanobis-Abstandes wird hingegen eine Rekursion für die inverse Kovarianzmatrix benötigt, ohne dass dabei jeweils zusätzlich eine Matrixinversion (O(N^3)) durchzuführen ist!

Mit dem folgenden Satz zur Matrixinversion:

$$(\mathbf{I}+\mathbf{A}\mathbf{B}^T)^{-1} = \mathbf{I}-\mathbf{A}(\mathbf{I}+\mathbf{B}^T\mathbf{A})^{-1}\mathbf{B}^T$$

Erhält man eine Rekursion für die inverse Korrelationsmatrix:

$$\begin{aligned}
\hat{\mathbf{R}}_n^{-1} &= [(1-\alpha)\hat{\mathbf{R}}_{n-1} + \alpha\mathbf{x}_n\mathbf{x}_n^T]^{-1} \\
&= \frac{1}{(1-\alpha)}\hat{\mathbf{R}}_{n-1}^{-1} - \frac{1}{(1-\alpha)^2}\hat{\mathbf{R}}_{n-1}^{-1}\mathbf{x}_n\left(\frac{1}{\alpha}+\frac{1}{(1-\alpha)}\mathbf{x}_n^T\hat{\mathbf{R}}_{n-1}^{-1}\mathbf{x}_n\right)^{-1}\mathbf{x}_n^T\hat{\mathbf{R}}_{n-1}^{-1} \\
&= \frac{1}{(1-\alpha)}\left(\hat{\mathbf{R}}_{n-1}^{-1} - \alpha\frac{\hat{\mathbf{R}}_{n-1}^{-1}\mathbf{x}_n\mathbf{x}_n^T\hat{\mathbf{R}}_{n-1}^{-1}}{1+\alpha(\mathbf{x}_n^T\hat{\mathbf{R}}_{n-1}^{-1}\mathbf{x}_n - 1)}\right)
\end{aligned}$$

Rekursive Schätzung der inversen Kovarianzmatrix

Und für die inverse Kovarianzmatrix:

$$\hat{\mathbf{C}}_n^{-1} = [(1-\alpha)\hat{\mathbf{R}}_{n-1} + \alpha \mathbf{x}_n \mathbf{x}_n^T - \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T]^{-1}$$

$$= \frac{1}{(1-\alpha)}[\hat{\mathbf{C}}_{n-1} + \alpha(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})^T]^{-1}$$

$$= \frac{1}{(1-\alpha)}\left(\hat{\mathbf{C}}_{n-1}^{-1} - \alpha \frac{\hat{\mathbf{C}}_{n-1}^{-1}(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})^T \hat{\mathbf{C}}_{n-1}^{-1}}{1+\alpha(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})^T \hat{\mathbf{C}}_{n-1}^{-1}(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_{n-1})}\right)$$

Rekursives Lernen kann natürlich auch mit der Musterklassifikation kombiniert werden. Das System verbessert sich bei neu hinzukommenden Stichproben. Dies setzt allerdings voraus, dass ein „*Labelling*“ für die Klassen stattfindet (*überwachtes Lernen*), d.h. der menschliche Beobachter trifft eine übergeordnete Entscheidung für die Klassenzugehörigkeit.