1 Klassifikation von Mischverteilungen

Computational Neuroscience II

Andreas Wendemuth

Lecture 17, June 27, 2005.

1.1 Mischverteilungen

Die Normalverteilungsannahme stellt eine starke Einschränkung dar. Anstatt eine einfache Normalverteilung für jede Klasse anzunehmen, können wir auch eine **Mischverteilung** durch Linearkombination von C_k vielen Normalverteilungen für die k-te Klasse ansetzen (wobei $k \in [1, K]$):

$$P(\mathbf{x}|\Omega_k) = \sum_{\nu=1}^{C_k} c_{k,\nu} \cdot N(\mathbf{x}|\boldsymbol{\mu}_{k,\nu}, \boldsymbol{\Sigma}_{\mathbf{k},\nu})$$

$$\text{mit} \quad N(\mathbf{x}|\boldsymbol{\mu}_{k,\nu}, \boldsymbol{\Sigma}_{\mathbf{k},\nu}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{k,\nu})^T \cdot {\boldsymbol{\Sigma}_{\mathbf{k},\nu}}^{-1} \cdot (\mathbf{x}-\boldsymbol{\mu}_{k,\nu})\right\}}{\sqrt{(2\pi)^D \cdot |\boldsymbol{\Sigma}_{\mathbf{k},\nu}|}} \tag{1.1}$$

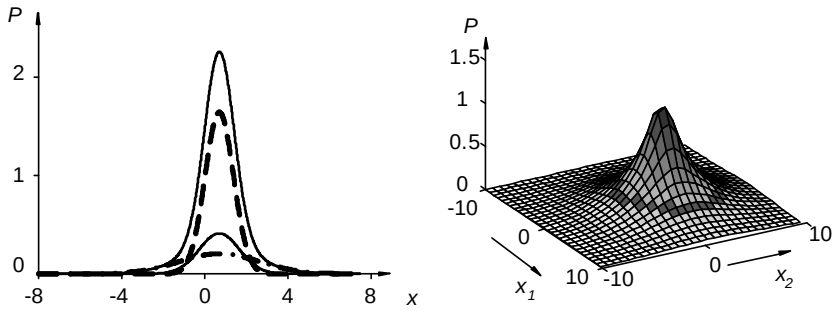

Abbildung 1.1: *1D- und 2D-Richterverteilung*

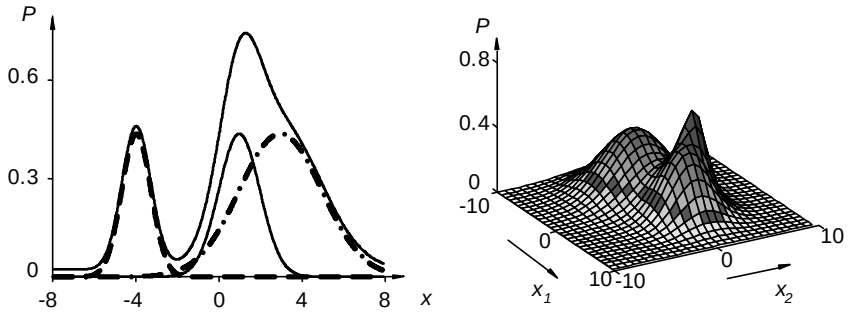

Abbildung 1.2: *1D- und 2D-Mischverteilung*

Dabei muss die Normierungsbedingung (**??**) gelten, aus der mit Gl. (1.1) unmittelbar folgt:

$$\sum_{\nu=1}^{C_k} c_{k,\nu} = 1 \qquad \forall\, k \in [1, K] \tag{1.2}$$

Diese Modellierung von Mischverteilungen ist zwar mathematisch schwerer zu behandeln, jedoch ermöglicht sie für große C_k im Prinzip beliebige Verteilungen zu approximieren. Abb. 1.2 zeigt eine 1-dimensionale und eine 2-dimensionale Mischverteilung. Sind bei diesen Verteilungen die Mittelwerte μ geich, so spricht man von einer **Richterverteilung** (Abb. 1.1). Richterverteilungen unterscheiden sich von den Normalverteilungen insofern, dass sie nicht so schnell auf 0 abfallen wie einfache Normalverteilungen. Die Verteilungen realer Sachverhalte lassen sich so besser modellieren, da oft auch in größerer Entfernung vom Klassenmittelpunkt eine gewisse Wahrscheinlichkeit vorhanden ist. Die Schätzung von Mischverteilungen wird in Kap. 1.3 behandelt.

1.1.1 Maximum Likelihood Schätzungen bei Mischverteilungen

Setzt man nach Gl. (1.1) eine M-dimensionale, multivariate Normalverteilung $P(\mathbf{x}|\Omega_k) = N(\mathbf{x}|\boldsymbol{\mu}_{k,\nu}, \boldsymbol{\Sigma}_{k,\nu})$ an, so erhält man nach Nullsetzen der partiellen Ableitungen nach μ_i^k und $\Sigma_{i,j}^k$ bei gegebener Stichprobe $\mathbf{S}$ analog zum obigen Ergebnis (hier ohne Rechnung):

$$\boldsymbol{\mu}^k = \frac{1}{N_k} \cdot \sum_{i=1}^{N_k} \mathbf{S}_{k,i} \tag{1.3}$$

$$\boldsymbol{\Sigma}^k = \frac{1}{N_k} \sum_{i=1}^{N_k} (\mathbf{S}_{k,i} - \boldsymbol{\mu}^k) \cdot (\mathbf{S}_{k,i} - \boldsymbol{\mu}^k)^T \tag{1.4}$$

Wir zeigen jetzt, dass die Schätzung für Mischverteilungen nicht mehr geschlossen gelöst werden kann. Die Mischverteilung ist angesetzt gemäß Gl. (1.1) und Abb. 1.2. Der verteilungsabhängige Teil der Likelihood sieht dann wie folgt aus:

$$L(\boldsymbol{\Theta}) = L(\{\theta_\nu\}) \tag{1.5}$$

$$= \sum_{k=1}^{N} \{ \sum_{i=1}^{N_k} \log \sum_{\nu=1}^{C_k} c_{k,\nu} N(\omega_{k,i}|\boldsymbol{\mu}_{k,\nu}, \boldsymbol{\Sigma}_{k,\nu}) + \lambda_k (c_{k,\nu} - 1) \} \tag{1.6}$$

Zur Erinnerung an die Notation: Es gibt N Klassen mit jeweils N_k vielen Beobachtungen $\omega_{k,i}$ und für jede Klasse wird eine Mischverteilung aus C_k vielen Normalverteilungen N mit Mittelwerten $\boldsymbol{\mu}_{k,\nu}$, Kovarianzmatrizen $\boldsymbol{\Sigma}_{k,\nu}$ und Gewichten $c_{k,\nu}$ angesetzt.

Leiten wir jetzt z.B. nach den Gewichten partiell ab, ergibt sich:

$$\frac{\delta}{\delta c_{k',\nu'}} L(\boldsymbol{\Theta}) = \sum_{i=1}^{N_{k'}} \frac{N(S_{k',i}|\boldsymbol{\mu}_{k',\nu}, \boldsymbol{\Sigma}_{k',\nu'})}{\sum_{\nu=1}^{C_k} c_{k,\nu} N(S_{k,i}|\boldsymbol{\mu}_{k,\nu}, \boldsymbol{\Sigma}_{k,\nu})} + \lambda_{\nu'} = 0 \;\; \forall (k', \nu') \tag{1.7}$$

Die Gleichungen entkoppeln für die Klassen, nicht aber für die Komponenten der Mischverteilung: Da die Summe $\sum_{i=1}^{N_{k'}}$ über Brüche verläuft, deren Zähler von i abhängige Normalverteilungen und deren Nenner gewichtete Summen solcher Normalverteilungen enthalten, können aus dieser Gleichung bereits Ausdrücke für die einzelnen $c_{k,\nu}$ nicht mehr gewonnen werden. Anders gesagt: Die Gleichungen können nicht entkoppelt werden. Dasselbe gilt für $\boldsymbol{\mu}_{k,\nu}$ und $\boldsymbol{\Sigma}_{k,\nu}$. Es muss ein EM-Ansatz verfolgt werden. (Siehe Kap. 1.3)

1.2 Überwachtes Lernen bei mehrfach stochastischen Prozessen

1.2.1 Zusammenhang – mehrfach stochastische Prozesse und Mischverteilungen

Man werde sich darüber klar, dass aus mehrfach stochastischen Prozessen Mischverteilungen entstehen und damit das Schätzen der Produktionswahrscheinlichkeiten $P(\mathbf{x}|\Omega_k)$ der einzelnen Klassen auf das Schätzen einer Mischverteilung hinausläuft.

Rein formal lässt sich diese Äquivalenz ebenfalls zeigen. Eine Mischverteilung war definiert durch:

$$P(\mathbf{x}) = \sum_{i=1}^{C_N} c_i P_i(\mathbf{x}) \tag{1.8}$$

Unser doppelt stochastisches Erzeugungsmodell ist definiert durch:

- diskreten Prozess mit Priori-Klassenwahrscheinlichkeiten $P(\Omega_k) = p_k$
- kontinuierlichen Zufallsprozess mit $P(\mathbf{x}|\Omega_k)$

Die Produktionswahrscheinlichkeit unseres doppelt stochastischen Prozesses für einen Vektor $\mathbf{x}$ ist:

$$P(\mathbf{x}) = \sum_{k=1}^{K} P(\mathbf{x}, \Omega_k) = \sum_{k=1}^{K} p_k \cdot P(\mathbf{x}|\Omega_k) \tag{1.9}$$

Setzt man nun $p_k = c_i$, so sieht man die Äquivalenz zwischen Mischverteilung und doppelt stochastischem Prozess. Abb. 1.3 verdeutlicht diesen Sachverhalt noch einmal.

1.2.2 Identifikation von Mischverteilungen

Wie oben bereits erwähnt, sind Mischverteilungen besser zur Beschreibung unserer Modelle geeignet als einfache Normalverteilungen. Durch Mischverteilungen lässt sich jede beliebige Verteilung beliebig genau approximieren. Benutzt man ein orthonormales Funktionensystem (z.B. Normalverteilungen), so kann durch Linearkombination jede Verteilung dargestellt werden (manchmal nur in unendlich vielen Summanden). Ist die zu approximierende Verteilungsfunktion bekannt, sind die Parameter der Normalverteilungen und auch die Wichtungskoeffizienten mit etwas Aufwand zu berechnen.

Unsere Stichprobenvektoren wurden jedoch in einem Zufallsprozess erzeugt, von dem der erste Teil (nämlich die Auswahl der entsprechende Klasse) vollständig unbekannt ist. Wie man sich auch rein logisch überlegen kann, geht die Klasseninformation eines jeden Stichprobenvektors wenigstens zum Teil im Ergebnis verloren. Man kann sich das leicht

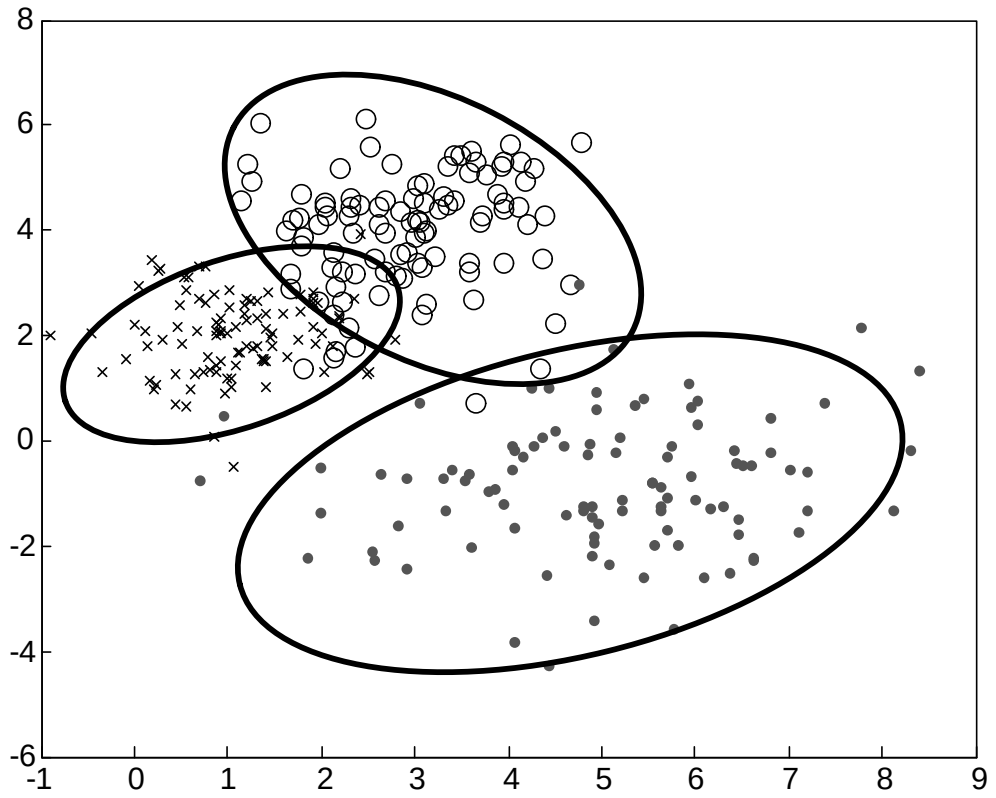

Abbildung 1.3: *Mischverteilung oder doppelt stochastischer Prozess?*

vorstellen, da ein gegebener Stichprobenvektor von jeder Klasse erzeugt sein könnte (vorausgesetzt die Klassenverteilungen werden kontinuierlich und unendlich angesetzt). Wir könnten versuchen, für jede mögliche Linearkombination eine Wahrscheinlichkeit zu berechnen und diese zu maximieren. Unvorteilhafterweise führt das wie oben schon gezeigt meistens auf ein analytisch nicht lösbares Problem.

Wir haben einen zweistufigen Zufallsprozess und eine angenommene Verteilung $P(\mathbf{x}, \mathbf{u}|\theta)$ gegeben, deren Parameter θ wir schätzen wollen. Dabei sei $\mathbf{u}$ der Zufallsprozess der Klassenauswahl, der uns unbekannt ist, während wir den Zufallsprozess $\mathbf{x}$ beobachten können, der die Stichprobenvektoren $\mathbf{S}$ erzeugt. Unser Ziel ist dann die Optimierung der Likelihood:

$$L(\theta) = P(\mathbf{S}|\theta) = \int_u P(\mathbf{S}, u|\theta)\mathrm{d}u \tag{1.10}$$

Das jedoch führt uns auf ein Huhn-Ei-Problem, denn zur Optimierung von θ benötigen

wir den Wert oder die Verteilung von $\mathbf{u}$ und diese hängt wieder von θ ab.

Beispiel 1.1 (Verbundwahrscheinlichkeiten):
Zur Verdeutlichung dieses Sachverhaltes nehmen wir ein Beispiel mit nur zwei Klassen K_1 und K_2. Jede dieser Klassen produziert die Merkmalsvektoren $\mathbf{U}=\mathbf{X}$ bzw. $\mathbf{Y}$. Wir nehmen einmal an, die Prioriwahrscheinlichkeiten $P(K)$, die Verbundwahrscheinlichkeiten $P(U,K)$ und bedingten Wahrscheinlichkeiten $P(U|K)$ sähen wie folgt aus. ($P(U|K) = \frac{P(U,K)}{P(K)}$ liefert die zweite Tabelle.)

$\mathbf{P(U,K)}$	$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{P(K)}$
$\mathbf{K_1}$	1/4	1/12	1/3
$\mathbf{K_2}$	1/3	1/3	2/3
$\mathbf{P(U)}$	7/12	5/12	1

und

$\mathbf{P(U\|K)}$	$\mathbf{X}$	$\mathbf{Y}$
$\mathbf{K_1}$	3/4	1/4
$\mathbf{K_2}$	1/2	1/2

Unsere Stichprobe ist von der Form $\mathbf{S} = \{X, Y, X, X, Y, ...\}$, wobei wir n_x mal X und n_y mal Y erhalten haben. Wir wollen jetzt einzig aus der Stichprobe die Verbund-Wahrscheinlichkeiten schätzen gemäß:

$\mathbf{P}(\mathbf{U}, \mathbf{K}\|\theta)$	X	Y	$P(K)$
$\mathbf{K_1}$	**a**	**b**	$P(K_1)$
K_2	**c**	**d**	$P(K_2)$
$P(U)$	$\mathbf{P}(\mathbf{X})$	$P(Y)$	1

Wir maximieren die Wahrscheinlichkeit

$$\max_\theta[P(\theta|\mathbf{S})] = \max_\theta[\frac{P(\mathbf{S}|\theta) \cdot P(\theta)}{P(\mathbf{S})}] \;\equiv\; \max_\theta P(\mathbf{S}|\theta)$$

Daraus formen wir um:

$$P(\mathbf{S}|\theta) = \prod_{i=1}^{M} P(S_i|\theta) = \prod_{i=1}^{M}[P(S_i, K_1|\theta) + P(S_i, K_2|\theta)]$$

Betrachten wir die Summanden im Produkt genauer:
Für $S_i = X$ würde aus ihnen $(a + c)$, für $S_i = Y$ würde aus ihnen $(b + d)$. Da wir n_x mal X und n_y mal Y in der Stichprobe haben, erhalten wir für das Produkt dann:

$$\prod_{i=1}^{n_x}(a + c) \cdot \prod_{j=1}^{n_y}(b + d) = (a + c)^{n_x} \cdot (b + d)^{n_y}$$

Unsere Likelihood Zielfunktion lautet also:

$$L(a, b, c, d) = \log(P(\mathbf{S}|\theta)) = n_x \log(a + c) + n_y \log(b + d)$$

und ihre Ableitungen:

$$\frac{\delta L}{\delta a} = \frac{n_x}{a + c} + \lambda\,, \quad \frac{\delta L}{\delta b} = \frac{n_y}{b + d} + \lambda\,, \quad \frac{\delta L}{\delta c} = \frac{n_x}{a + c} + \lambda\,, \quad \frac{\delta L}{\delta d} = \frac{n_y}{b + d} + \lambda$$

Wir erhalten durch Nullsetzen:

$$\frac{n_x}{n_y} = \frac{a + c}{b + d}$$

und mit $b + d = 1 - (a + c)$

$$a + c = \frac{n_x}{n_x + n_y} = \frac{n_x}{N} \;\text{ und }\; b + d = \frac{n_y}{N}.$$

Wir sehen, dass wir also nur die Wahrscheinlichkeiten $P(X)$ und $P(Y)$ aufgrund der relativen Häufigkeiten in der Stichprobe per ML schätzen können (was auch ohne Rechnung sofort einsichtig ist). Die Verbundwahrscheinlichkeiten jedoch lassen sich ohne weitere Kenntnisse nicht abschätzen, selbst wenn die Verteilung von K, d.h. $P(K_1)$ und $P(K_2)$, bekannt wären. □

1.3 Schätzung einer Mischverteilung durch den EM-Algorithmus

Wir haben eine Stichprobe $X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$ in D Dimensionen gegeben, von der wir vermuten, sie wurde aus K Klassen erzeugt mit den Priori-Wahrscheinlichkeiten p_k. Wir setzen unsere Verteilung an als Mischverteilung gemäß Gl. (1.1):

$$P(\mathbf{x}) = \sum_{k=1}^{K} p_k P(\mathbf{x}|\theta_k) = \sum_{k=1}^{K} p_k \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)}{\sqrt{(2\pi)^D |\Sigma_k|}} \tag{1.18}$$

Wir benutzen wieder die gewohnte Kurzschreibweise:

$$P(\mathbf{x}|\theta_k, \Omega_k) = P(\mathbf{x}|\theta_k) \tag{1.19}$$

Die Likelihood der Stichprobe ist:

$$L(\theta) = \log P(\mathbf{X}|\theta) = \log(\prod_{i=1}^{N} P(\mathbf{X}_i|\theta_{\mathbf{u}_i})) = \sum_{i=1}^{N} \log P(\mathbf{X}_i|\theta_{\mathbf{u}_i}) \tag{1.20}$$

Die Wahrscheinlichkeit der Stichprobe bei gegebenem Parametersatz θ ist nun abhängig von den unbekannten Klassenzugehörigkeiten der Stichprobenvektoren, die wir $\mathbf{u}$ nennen wollen. Wir haben also als **unbekannte Größe** den Vektor $\mathbf{u} = \{\Omega_1, \Omega_2, \ldots, \Omega_N\}$, der uns angibt, von welcher Klasse der i-te Stichprobenvektor erzeugt wurde.

Hier kommt der EM-Algorithmus ins Spiel! Wären unsere Daten komplett, d.h. $\mathbf{u}$ und damit unsere Klassenzugehörigkeiten der Stichprobenvektoren wäre bekannt, so könnten wir den Parametersatz einfach für jede Klasse einzeln (für Normalverteilungen nach Gln. (1.3) und (1.4)) schätzen.

E-Schritt: Bilden der Kullback-Leibler Statistik.

Wir bilden den bedingten Erwartungswert der Likelihood der kompletten Daten $L(\theta)$.

Hier ist es nun essentiell zu erkennen, dass $\mathbf{u}$ ein Vektor mit **diskreten** Werten ist. Der Erwartungswert wird also über den gesamten $\mathbf{U}$-Raum genommen, der

aus allen möglichen Kombinationen der Klassenzugehörigkeiten zu bilden ist. Beispielsweise wären die möglichen Kombinationen für zwei Vektoren und zwei Klassen $\mathbf{u} \in [(1,1),(1,2),(2,1),(2,2)]$.

Zur kurzen Wiederholung:

- Der Erwartungswert einer Zufallsvariablen X mit Verteilung P(x) war definiert als:

$$E(x) = \int_{\mathbb{R}} xP(x)\mathrm{d}x \tag{1.21}$$

  oder im diskreten Fall:

$$E(X) = \sum_i x_i \cdot p_i \tag{1.22}$$

- Der Erwartungswert bezüglich einer Funktion einer Zufallsvariablen (erwarteter Fkt.-Wert):

$$E(f(x)) = \int_R f(x)P(x)\mathrm{d}x \tag{1.23}$$

  bzw. diskret:

$$E(f(X)) = \sum_i f(x_i)p_i \tag{1.24}$$

- Hat man bedingte Wahrscheinlichkeiten, so werden diese einfach mit übernommen:

$$E(f(x)|\Theta) = \int_R f(x)P(x|\Theta)\mathrm{d}x \tag{1.25}$$

  diskret:

$$E(f(X)|\Theta) = \sum_i f(x_i)P(X = x_i|\Theta) \tag{1.26}$$

Der Erwartungswert bei gegebener Stichprobe X und für irgendeinen Parametersatz θ

lautet daher:

$$\begin{aligned}
Q(\hat{\theta},\theta) &= E(\log P(X,U|\hat{\theta})|X,\theta) \\
&= \sum_{U\text{-Raum}} (\sum_{i=1}^{N} \log P(\mathbf{X}_i, u_i|\hat{\theta}_{\mathbf{u}_i}))P(\mathbf{U}|\mathbf{X},\theta) \\
&= \sum_{\text{U-Raum}} (\sum_{i=1}^{N} \log[\hat{p}_{u_i} \cdot P(\mathbf{X}_i|u_i,\hat{\theta}_{u_i})])P(\mathbf{U}|\mathbf{X},\theta) \\
&= \sum_{U\text{-Raum}} (\sum_{i=1}^{N} \log \hat{p}_{u_i} + \log \overbrace{P(\mathbf{X}_i|u_i,\hat{\theta}_{u_i})}^{P(X_i|\hat{\theta}_{u_i})}) \prod_{\nu=1}^{N} P(u_\nu|X_\nu,\theta_{u_\nu}) \\
&= \sum_{U\text{-Raum}} (\sum_{i=1}^{N} \log \hat{p}_{u_i} + \log P(X_i|\hat{\theta}_{u_i})) \cdot \prod_{\nu=1}^{N} \gamma_{\nu,u_\nu} \qquad (1.27)
\end{aligned}$$

Die letzte Abkürzung wird in Gl. (1.30) bestimmt. Die Wahrscheinlichkeit $P(\mathbf{U}|\mathbf{X},\theta)$ ist die Wahrscheinlichkeit für einen Stichprobenvektor $\mathbf{U}$, gegeben eine Stichprobe $\mathbf{X}$ und den Parametersatz θ – und setzt sich multiplikativ aus den Einzelwahrscheinlichkeiten über jedes Stichprobenelement zusammen. Diese Einzelwahrscheinlichkeiten sind jedoch $P(u_k|X_k,\theta_{u_k})$, denn zu jeder Klasse u_k gehört ein Parametersatz θ_{u_k} (das bedeutet übrigens auch, dass $P(u_k|\theta_{u_k}) = 1$ ist). Daher schreiben wir auch kürzer:

$$P(X_i|u_i,\hat{\theta}_{u_i}) = P(X_i|\hat{\theta}_{u_i}) \qquad (1.28)$$

Nun schreiben wir noch die unbekannte Priori-Wahrscheinlichkeit wie folgt um:

$$\begin{aligned}
P(u_\nu|X_\nu,\theta_{u_\nu}) &\overset{\text{Bayes!}}{=} \frac{P(X_\nu|u_\nu,\theta_{u_\nu})\overbrace{P(u_\nu|\theta_{u_\nu})}^{=1}}{P(X_\nu|\theta_{u_\nu})} \\
&= \frac{\overbrace{P(X_\nu|u_\nu,\theta_{u_\nu})}^{\equiv P(X_\nu|\theta_{u_\nu})}}{\frac{P(X_\nu,\theta_{u_\nu})}{\underbrace{P(\theta_{u_\nu})}_{p_{u_\nu}}}} \\
&= \frac{p_{u_\nu} P(X_\nu|\theta_{u_\nu})}{\sum_{\lambda=1}^{K} p_{u_\lambda} P(X_\nu|\theta_{u_\lambda})} \qquad (1.29)
\end{aligned}$$

Wir können also wieder durch die Anwendung der Bayes Regel die unbekannte Posteriori-Wahrscheinlichkeit für einen Vektor $\mathbf{u}$ der Klassenzugehörigkeiten durch bekannte Wahrscheinlichkeiten (bekannt durch die parametrischen Klassendichten) ausdrücken. Dieser Ausdruck wird im Folgenden mit einer Abkürzung versehen:

$$\gamma_{k,u_k} := P(u_k|X_k,\theta_{u_k}) = \frac{p_{u_k} P(X_k|\theta_{u_k})}{\sum_{\lambda=1}^{K} p_{u_\lambda} P(X_k|\theta_{u_\lambda})} \qquad (1.30)$$

Nun widmen wir uns wieder der obigen Kullback-Leibler Statistik $Q(\hat{\theta}, \theta)$. Die Summenbildung für den Erwartungswert über den gesamten **U**-Raum muss über alle möglichen Klassenzuordnungen aller Vektoren gebildet werden. Das geschieht durch die folgende Summe:

$$\sum_{U\text{-Raum}} f(u_i) \equiv \underbrace{\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\dots\sum_{k_N=1}^{K}}_{\text{Alle Klassen für alle Stichprobenvektoren}} f(k_1, k_2, \dots, k_n) \tag{1.31}$$

Der obige Kullback-Leibler Abstand wird daher:

$$Q(\hat{\theta}, \theta) = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\dots\sum_{k_N=1}^{K}\left(\sum_{i=1}^{N}\log\hat{p}_{k_i} + \log P(X_i|\hat{\theta}_{k_i})\right)\cdot\prod_{\nu=1}^{N}\gamma_{\nu,k_\nu} \tag{1.32}$$

Das Produkt über alle γ wird nun auseinandergenommen zu $\gamma_{1,k_1}\cdot\gamma_{2,k_2}\cdot\ldots\cdot\gamma_{N,k_N}$ und jeder Faktor wird (als konstanter Faktor für alle anderen Summen) in die entsprechende Summe gezogen. Wer sich wundert, dass der „alte Parametersatz“ θ in der Formel nicht auftritt, der bemerke, dass dieser in den Parametern $\gamma_{i,k}$ steckt.

$$Q(\hat{\theta}, \theta) = \sum_{k_1=1}^{K}\gamma_{1,k_1}\sum_{k_2=1}^{K}\gamma_{2,k_2}\dots\sum_{k_N=1}^{K}\gamma_{N,k_N}\Big(\sum_{i=1}^{N}\log\hat{p}_{k_i} + \log P(X_i|\hat{\theta}_{k_i})\Big) \tag{1.33}$$

Der M-Schritt: Ableitung der Schätzformel für die Parameter.

Wir trennen die innere Summe über alle i in Gl. (1.33) auf und betrachten die entstehenden Terme einzeln:

$$Q(\hat{\theta}, \theta) = \underbrace{\sum_{k_1=1}^{K}\gamma_{1,k_1}\sum_{k_2=1}^{K}\gamma_{2,k_2}\dots\sum_{k_N=1}^{K}\gamma_{N,k_N}\sum_{i=1}^{N}\log\hat{p}_{k_i}+}_{\text{Summe A}} \tag{1.34}$$

$$\underbrace{\sum_{k_1=1}^{K}\gamma_{1,k_1}\sum_{k_2=1}^{K}\gamma_{2,k_2}\dots\sum_{k_N=1}^{K}\gamma_{N,k_N}\sum_{i=1}^{N}\log P(X_i|\hat{\theta}_{k_i})}_{\text{Summe B}} \tag{1.35}$$

Bevor wir nun beide Summenterme betrachten, sollte noch (nebenbei) bemerkt werden, dass:

$$\sum_{k_j=1}^{K}\gamma_{n,k_j} = \frac{\sum_{k_j=1}^{K} p_{k_j}P(X_n|\theta_{k_j})}{\sum_{\lambda=1}^{K} p_{k_\lambda}P(X_n|\theta_{k_\lambda})} = 1 \quad \forall n \tag{1.36}$$

Anschaulich heißt dies nichts anderes, als dass die Summe aller Posteriori- Wahrscheinlichkeiten für die Klassenzuordnungen der Beobachtung X_n zu Eins wird.

Summe A: Wir nehmen die innere Summe auseinander und sortieren die Summen um:

$$\begin{aligned}
&\text{Summe A} = \\
&\sum_{k_1=1}^{K} \gamma_{1,k_1} \sum_{k_2=1}^{K} \gamma_{2,k_2} \ldots \sum_{k_N=1}^{K} \gamma_{N,k_N} (\log \hat{p}_{k_1} + \log \hat{p}_{k_2} + \ldots + \log \hat{p}_{k_N}) \\
= &\sum_{k_1=1}^{K} \gamma_{1,k_1} \log \hat{p}_{k_1} \cdot (\sum_{k_2=1}^{K} \gamma_{2,k_2} \sum_{k_3=1}^{K} \gamma_{3,k_3} \ldots \sum_{k_N=1}^{K} \gamma_{N,k_N}) + \\
&\sum_{k_2=1}^{K} \gamma_{2,k_2} \log \hat{p}_{k_2} \cdot (\sum_{k_1=1}^{K} \gamma_{1,k_1} \sum_{k_3=1}^{K} \gamma_{3,k_3} \ldots \sum_{k_N=1}^{K} \gamma_{N,k_N}) + \\
&\sum_{k_{\ldots}} \quad \ldots \quad (\quad \ldots \quad) + \\
&\sum_{k_N=1}^{K} \gamma_{N,k_N} \log \hat{p}_{k_N} \cdot (\sum_{k_1=1}^{K} \gamma_{1,k_1} \sum_{k_2=1}^{K} \gamma_{2,k_2} \ldots \underbrace{\sum_{k_{N-1}=1}^{K} \gamma_{N-1,k_{N-1}}}_{1})
\end{aligned} \tag{1.37}$$

Die Summen über alle γ werden, wie in obiger Formel exemplarisch gezeigt, alle $= 1$ und damit vereinfacht sich die Formel für Summe A:

$$\text{Summe A} = \sum_{k_1=1}^{K} \gamma_{1,k_1} \log \hat{p}_{k_1} + \sum_{k_2=1}^{K} \gamma_{2,k_2} \log \hat{p}_{k_2} + \ldots + \sum_{k_N=N}^{K} \gamma_{N,k_N} \log \hat{p}_{k_N} \tag{1.38}$$

und durch Indexangleichung: $k_1 = k_2 = \ldots = k_N := \mathbf{k}$

$$\text{Summe A} = \sum_{k=1}^{K} (\sum_{i=1}^{N} \gamma_{i,k}) \log \hat{p}_k \tag{1.39}$$

Summe B: Wir vereinfachen in ähnlicher Weise wie oben durch Summen umsortieren etc. . . :

$$
\begin{aligned}
&\text{Summe B} = \\
&\sum_{k_1=1}^{K} \gamma_{1,k_1} \sum_{k_2=1}^{K} \gamma_{2,k_2} \cdots \sum_{k_N=1}^{K} \gamma_{N,k_N} \times \\
&\qquad \times (\log P(X_1|\hat{\theta}_{k_1}) + \log P(X_2|\hat{\theta}_{k_2}) + \ldots + \log P(X_N|\hat{\theta}_{k_N})) \\
= &\sum_{k_1=1}^{K} \gamma_{1,k_1} \log P(X_1|\hat{\theta}_{k_1}) \cdot (\sum_{k_2=1}^{K} \gamma_{2,k_2} \sum_{k_3=1}^{K} \gamma_{3,k_3} \cdots \sum_{k_N=1}^{K} \gamma_{N,k_N}) + \\
&\sum_{k_2=1}^{K} \gamma_{2,k_2} \log P(X_2|\hat{\theta}_{k_2}) \cdot (\sum_{k_1=1}^{K} \gamma_{1,k_1} \sum_{k_3=1}^{K} \gamma_{3,k_3} \cdots \sum_{k_N=1}^{K} \gamma_{N,k_N}) + \\
&\sum_{k_{\ldots}} \qquad \ldots \qquad (\qquad \ldots \qquad) + \\
&\sum_{k_N=1}^{K} \gamma_{N,k_N} \log P(X_N|\hat{\theta}_{k_N}) \cdot (\sum_{k_1=1}^{K} \gamma_{1,k_1} \sum_{k_2=1}^{K} \gamma_{2,k_2} \cdots \underbrace{\sum_{k_{N-1}=1}^{K} \gamma_{N-1,k_{N-1}}}_{1}) \\
= &\sum_{k_1=1}^{K} \gamma_{1,k_1} \log P(X_1|\hat{\theta}_{k_1}) + \ldots + \sum_{k_N=1}^{K} \gamma_{N,k_N} \log P(X_N|\hat{\theta}_{k_N}) \\
= &\sum_{k=1}^{K} (\sum_{i=1}^{N} \gamma_{i,k} \log P(X_i|\hat{\theta}_k)) \qquad (1.40)
\end{aligned}
$$

Damit lautet unsere Kullback-Leibler Statistik:

$$
Q(\hat{\theta}, \theta) = \underbrace{\sum_{k=1}^{K} (\sum_{i=1}^{N} \gamma_{i,k}) \log \hat{p}_k}_{Q_P} + \underbrace{\sum_{k=1}^{K} (\sum_{i=1}^{N} \gamma_{i,k} \log P(X_i|\hat{\theta}_k))}_{Q_{\mathrm{V}}} \qquad (1.41)
$$

Ableitung der Schätzformeln für die Prioriwahrscheinlichkeiten:

Wir können jetzt Q nach dem Parametersatz $\hat{\theta}$ und den $\hat{p}_k$ getrennt optimieren. Die partiellen Ableitungen lauten:

$$
\begin{aligned}
\frac{d}{d\hat{p}_\nu} Q_p &= \frac{d}{d\hat{p}_\nu} [\sum_{k=1}^{K} (\sum_{i=1}^{N} \gamma_{i,k}) \log \hat{p}_k + \lambda (\sum_{i=1}^{K} \hat{p}_i - 1)] = (\sum_{i=1}^{N} \gamma_{i,\nu}) \frac{1}{\hat{p}_\nu} + \lambda = 0 \\
\underbrace{\sum_{\nu=1}^{K} \hat{p}_\nu}_{=1} \cdot \lambda &= -\sum_{\nu=1}^{K} (\sum_{i=1}^{N} \gamma_{i,\nu}) = -\sum_{i=1}^{N} \underbrace{(\sum_{\nu=1}^{K} \gamma_{i,\nu})}_{=1} = -N \qquad (1.42)
\end{aligned}
$$

und damit schließlich:

$$\hat{p}_k = \frac{1}{N}\sum_{i=1}^{N}\gamma_{i,k} \tag{1.43}$$

Anschaulich ist also der (neue) Schätzwert der Zugehörigkeitswahrscheinlichkeit $\hat{p}_k$ zur Klasse k gleich der mittleren (alten) Posteriori- Wahrscheinlichkeit der Zugehörigkeiten aller Beobachtungen i zur Klasse k. Dieses Ergebnis ist sehr plausibel.

Ableitung der Schätzformeln für die Mittelwerte:

Setzt man für die Klassenverteilungen in Gl. (1.41) jetzt unsere Normalverteilungen ein, d.h.

$$P(\mathbf{X}_i|\hat{\theta}_k) = \frac{1}{\sqrt{(2\pi)^D|\hat{\mathbf{\Sigma}}_k|}}\cdot\exp\left(-\frac{1}{2}(\mathbf{X}_i-\hat{\boldsymbol{\mu}}_k)^T\hat{\mathbf{\Sigma}}_k^{-1}(\mathbf{X}_i-\hat{\boldsymbol{\mu}}_k)\right) \tag{1.44}$$

so erhält man aus dem verteilungsabhängigen Teil der Kullback-Leibler Statistik:

$$Q_V = -\sum_{k=1}^{K}(\sum_{i=1}^{N}\gamma_{i,k}(\log\sqrt{(2\pi)^D|\hat{\mathbf{\Sigma}}_k|}+\frac{1}{2}(\mathbf{X}_i-\hat{\boldsymbol{\mu}}_k)^T\hat{\mathbf{\Sigma}}_k^{-1}(\mathbf{X}_i-\hat{\boldsymbol{\mu}}_k))) \tag{1.45}$$

Um nach dem *Vektor* $\hat{\boldsymbol{\mu}}_k$ und der Matrix $\hat{\mathbf{\Sigma}}^{-1}$ abzuleiten, müssen wir das Matrizenprodukt umschreiben und durch die Einzelkomponenten ausdrücken, wobei $\hat{\Sigma}^{-1} = \sigma$ gesetzt wurde:

$$\begin{aligned} Q_V \;=\; & -\sum_{k=1}^{K}(\sum_{i=1}^{N}\gamma_{i,k}(\log\sqrt{(2\pi)^D|\hat{\mathbf{\Sigma}}_k|} \\ & +\;\frac{1}{2}\sum_{\rho=1}^{D}(\mathbf{X}_i^{(\rho)}-\hat{\boldsymbol{\mu}}_k^{(\rho)}) \overset{\text{Produkt !}}{\cdot} \sum_{\nu=1}^{D}(\mathbf{X}_i^{(\nu)}-\hat{\boldsymbol{\mu}}_k^{(\nu)})\sigma_k^{(\nu,\rho)})) \end{aligned} \tag{1.46}$$

Wir wollen hier und im Folgenden ein wenig genauer die algebraische Formulierung von Ableitungen vektor- und matrizenwertiger Größen darlegen, die nicht jedem Leser bekannt sein mag.

Gl. (1.46) mag etwas kompliziert aussehen, jedoch reicht es für die Ableitung aus, nur das Matrizenprodukt (mit *Produkt!* gekennzeichnet) zu betrachten. Die Ableitung dieses inneren Produktes P nach $\hat{\boldsymbol{\mu}}_k^{(n)}$ ist, nach der Produktregel abgeleitet, wobei wir hier zur Anschaulichkeit setzen $\mathbf{V} = (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)$:

$$\begin{aligned} \frac{dP}{d\hat{\boldsymbol{\mu}}_k^{(n)}} \;&:=\; \frac{d}{d\hat{\boldsymbol{\mu}}_k^{(n)}}[\sum_{\rho=1}^{D}\mathbf{V}^{(\rho)}\cdot\sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,\rho)}] \\ &=\; \sum_{\rho=1}^{D}(\underbrace{[\frac{d}{d\hat{\boldsymbol{\mu}}_k^{(n)}}\mathbf{V}^{(\rho)}]\cdot\sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,\rho)}}_{\text{Summe A}} + \underbrace{\mathbf{V}^{(\rho)}\cdot\frac{d}{d\hat{\boldsymbol{\mu}}_k^{(n)}}[\sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,\rho)}]}_{\text{Summe B}}) \end{aligned} \tag{1.47}$$

Wir nehmen die Summe über alle ρ auseinander. Nach der obigen Definition von $\mathbf{V}$ ist dessen Ableitung nach einer Komponente von $\widehat{\boldsymbol{\mu}}_k^{(i)}$:

$$\frac{d\mathbf{V}^{(i)}}{d\hat{\boldsymbol{\mu}}_k^{(n)}} = \frac{d}{d\hat{\boldsymbol{\mu}}_k^{(n)}}(\mathbf{X}_i^{(i)} - \hat{\boldsymbol{\mu}}_k^{(i)}) = \begin{cases} 0 & n \neq i \\ -1 & n = i \end{cases}. \tag{1.48}$$

Summe A wird dann:

$$\sum_{\rho=1}^{D}([\frac{d}{d\hat{\boldsymbol{\mu}}_k^{(n)}}\mathbf{V}^{(\rho)}] \cdot \sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,\rho)}) = -\sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,n)} \tag{1.49}$$

Summe B wird dann:

$$\sum_{\rho=1}^{D}\mathbf{V}^{(\rho)} \cdot [\frac{d}{d\hat{\boldsymbol{\mu}}_k^{(n)}}\sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,\rho)}] = -\sum_{\rho=1}^{D}\mathbf{V}^{(\rho)} \cdot \sigma_k^{(n,\rho)} \tag{1.50}$$

Der gesamte Ausdruck für die Ableitung des Matrizenproduktes reduziert sich dann also auf:

$$\frac{dP}{d\hat{\boldsymbol{\mu}}_k^{(n)}} = -(\sum_{\nu=1}^{D}\mathbf{V}^{(\nu)}\sigma_k^{(\nu,n)} + \sum_{\rho=1}^{D}\mathbf{V}^{(\rho)} \cdot \sigma_k^{(n,\rho)}) \tag{1.51}$$

und nach Indexangleichung $\nu = \rho := i$

$$= -\sum_{i=1}^{D}\mathbf{V}^{(i)}(\sigma_k^{(i,n)} + \sigma_k^{(n,i)}) \tag{1.52}$$

und da σ symmetrisch ist:

$$= -2 \cdot \sum_{i=1}^{D}\mathbf{V}^{(i)}\sigma_k^{(i,n)} = -2(X_i - \hat{\boldsymbol{\mu}}_k)^T \cdot \sigma_k \tag{1.53}$$

Reiht man diese ganzen Vektorelemente wieder zu einem Vektor auf, erlebt man (?) eine Überraschung: Der ganze Aufwand war also recht unnötig! Man sieht, dass sich die Ableitung von $(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)^T \cdot \sigma_k \cdot (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)$ ganz analog zur normalen Differenzialrechnung zu $2 \cdot (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)^T \cdot \sigma_k$ ergibt.

Die Ableitung von Q_v ergibt nun folgenden Nullvektor:

$$\frac{dQ_V}{d\hat{\boldsymbol{\mu}}_k^{(n)}} = -\frac{1}{2} \cdot \sum_{k=1}^{K}\sum_{i=1}^{N}\gamma_{i,k}\frac{dP}{d\hat{\boldsymbol{\mu}}_k^{(n)}} = \sum_{i=1}^{N}\gamma_{i,n}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_n)^T \cdot \sigma_n = \vec{0} \tag{1.54}$$

An dieser Stelle soll erneut daran erinnert werden, dass wir soeben mittels EM- Algorithmus eine Mischverteilung aus N Stichprobenvektoren mit K Klassen in D Dimensionen schätzen wollen. Wir optimieren also die Kullback-Leibler Statistik.

Das σ entfällt durch rechtsseitige Multiplikation mit $\sigma^{-1}(=\Sigma)$.

$$\frac{dQ_V}{d\hat{\boldsymbol{\mu}}_k^{(n)}} = \sum_{i=1}^{N} \gamma_{i,n}(\hat{\boldsymbol{\mu}}_n - \mathbf{X}_i)^T = \vec{0} \iff \sum_{i=1}^{N} \gamma_{i,n}\hat{\boldsymbol{\mu}}_n = \sum_{i=1}^{N} \gamma_{i,n}\mathbf{X}_i \tag{1.55}$$

und schließlich:

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{\sum_{i=1}^{N} \gamma_{i,n}} \cdot \sum_{j=1}^{N} \gamma_{j,n} \cdot \mathbf{X}_j \tag{1.56}$$

Anschaulich ist also der (neue) Schätzwert des Klassenmittelwertes $\hat{\boldsymbol{\mu}}_n$ zur Klasse n gleich dem mittleren (alten) Beobachtungsvektor $\mathbf{X}_j$, gewichtet mit der Posteriori-Wahrscheinlichkeit der Zugehörigkeit *aller* Beobachtungen j zur Klasse n. Der Nenner dient zur Normierung der Posteriori-Wahrscheinlichkeiten.

Dieses Ergebnis ist ebenfalls sehr plausibel. Es geht in eine „normale“ Mittelwertbildung über, wenn alle Beobachtungsvektoren $\mathbf{X}_j$ zu genau einer Klasse gehören würden. In diesem Fall wäre $\gamma_{j,n} = 1$, wenn n die zugehörige Klasse ist, und sonst Null. Wir haben es hier also mit einer erweiterten Mittelwertbildung zu tun, bei der alle Beobachtungen allen Klassen zugehören können.

Ableitung der Schätzformeln für die Kovarianzmatrizen:
Sieht man sich die Formel von Q_v einmal in Hinblick auf die Kovarianzen an, so wird man mit erkennen, dass man bei der Optimierung eine Determinante abzuleiten hat: Was ist nun also die Ableitung einer Determinante nach einem der Elemente?

Die Algebra lehrt:

$$|\mathbf{A}| = \sum_{i=1}^{D} \breve{a}_{i,k} \cdot a_{i,k} \qquad , k \text{ beliebig} \rightarrow \text{Entwicklung nach der } k\text{-ten Spalte} \tag{1.57}$$

wobei $\breve{a}_{i,k}$ die Unterdeterminante in A zum Element $a_{i,k}$ ist.

Will man nun die Determinante nach dem Element $a_{m,n}$ ableiten, entwickelt man einfach nach der n-ten Spalte und es ergibt sich:

$$\frac{d}{da_{m,n}}|A| = \frac{d}{da_{m,n}} \sum_{i=1}^{D} \breve{a}_{i,n} \cdot a_{i,n} = \breve{a}_{m,n}, \tag{1.58}$$

da sämtliche Unterdeterminanten in der m-ten Zeile unabhängig von $a_{m,n}$ sind (man erhält die Unterdeterminante durch Streichen der m-ten Zeile und der n-ten Spalte!).

Weiterhin lehrt die Algebra:

$$A^{-1}_{m,n} = \frac{1}{|A|}\breve{a}_{n,m} \tag{1.59}$$

(Cramersche Regel) und weiterhin:

$$|\Sigma| = \frac{1}{|\Sigma^{-1}|} \tag{1.60}$$

daher folgt:

$$\frac{d}{da_{m,n}}|A| = |A| \cdot A^{-1}_{n,m} \tag{1.61}$$

In Gl. (1.46) muss man beim Ableiten nach den Elementen der *inversen Kovarianzmatrix* Σ^{-1} den Ausdruck $\log \frac{1}{|\hat{\Sigma}|}$ ableiten.

Die Ableitung sieht nach obigen Ergebnissen also wie folgt aus:

$$\begin{aligned}
\frac{\delta}{\delta\sigma_k^{(y,x)}} \log|\hat{\Sigma}_k| &= \frac{\delta}{\delta\sigma_k^{(y,x)}} \log\frac{1}{|\hat{\Sigma}_k^{-1}|} = -\frac{\delta}{\delta\sigma_k^{(y,x)}} \log(\sum_{i=1}^{D} \breve{\sigma}_k^{(i,x)} \cdot \sigma_k^{(i,x)}) \\
&= \frac{-1}{|\hat{\Sigma}_k^{-1}|} \cdot \breve{\sigma}_k^{(y,x)} = -\frac{1}{|\hat{\Sigma}_k^{-1}|} \cdot |\hat{\Sigma}_k^{-1}| \cdot \hat{\Sigma}_k^{(x,y)} \\
&= -\Sigma_k^{(x,y)}
\end{aligned} \tag{1.62}$$

Die Ableitung von Gl. (1.46) nach den Elementen der *inversen* Kovarianzmatrix $\sigma_j^{(x,y)}$ ist:

$$\frac{\delta Q_v}{\delta\sigma_j^{(y,x)}} = \sum_{i=1}^{N} \gamma_{i,j}(-\frac{1}{2}\frac{|\hat{\Sigma}_j^{-1}| \cdot \Sigma_j^{(x,y)}}{|\hat{\Sigma}_j^{-1}|} \tag{1.63}$$

$$+ \frac{1}{2}(\mathbf{X}_i^{(x)} - \hat{\boldsymbol{\mu}}_j^{(x)})(\mathbf{X}_i^{(y)} - \hat{\boldsymbol{\mu}}_j^{(y)})) = 0 \tag{1.64}$$

$$\Sigma_j^{(x,y)} = \frac{1}{\sum_{i=1}^{N} \gamma_{i,j}} \sum_{i=1}^{N} \gamma_{i,j}(\mathbf{X}_i^{(x)} - \hat{\boldsymbol{\mu}}_j^{(x)})(\mathbf{X}_i^{(y)} - \hat{\boldsymbol{\mu}}_j^{(y)}) \tag{1.65}$$

$$\hat{\Sigma}_j = \frac{1}{\sum_{i=1}^{N} \gamma_{i,j}} \sum_{i=1}^{N} \gamma_{i,j}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_j)^T$$

Anschaulich ist also der (neue) Schätzwert der Kovarianzmatrix $\hat{\Sigma}_j$ zur Klasse j gleich der mittleren Kovarianzmatrix $(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_j)^T$, gewichtet mit der Posteriori-Wahrscheinlichkeit der Zugehörigkeit *aller* Beobachtungen i zur Klasse j. Der Nenner dient wieder zur Normierung der Posteriori-Wahrscheinlichkeiten. Allerdings werden hier Posteriori-Wahrscheinlichkeiten auf der Grundlage der alten Zugehörigkeiten benutzt, es sind aber bereits die neuen Klassenmittelwerte erforderlich. Dies verlangt einen zweimaligen Durchlauf der Daten, wo zunächst die neuen Klassenmittelwerte und dann die neuen Klassenkovarianzmatrizen geschätzt werden.

Dieses Ergebnis ist ebenfalls sehr plausibel. Es geht wieder in eine „normale“ Kovarianzschätzung über, wenn alle Beobachtungsvektoren $\mathbf{X}_i$ zu genau einer Klasse gehören würden. Wir haben es hier also mit einer erweiterten Kovarianzschätzung zu tun, bei der alle Beobachtungen allen Klassen zugehören können.

Die Schätzformeln beim EM-Algorithmus für einen Iterationsschritt lauten bei Normalverteilungsannahme demnach zusammengefasst:

$$\hat{p}_k = \frac{1}{N} \sum_{j=1}^{N} \gamma_{j,k} \tag{1.66}$$

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{\sum_{i=1}^{N} \gamma_{i,n}} \cdot \sum_{j=1}^{N} \gamma_{j,n} \cdot \mathbf{X}_j \tag{1.67}$$

$$\hat{\Sigma}_n = \frac{1}{\sum_{i=1}^{N} \gamma_{i,n}} \cdot \sum_{j=1}^{N} \gamma_{j,n} \cdot (\mathbf{X}_j - \hat{\boldsymbol{\mu}}_n)(\mathbf{X}_j - \hat{\boldsymbol{\mu}}_n)^T \tag{1.68}$$

1.4 Exercise

Using the data points from lecture 16, Exercise 6/1, perform an estimation of a Gaussian mixture probability density with two components.

For doing so, use the EM algorithm in eqs.(1.66) - (1.68).

How many iterations do you need?