TEK5020/9020 Mønstergjenkjenning
Høsten 2023

Forelesning 6 – Parametriske metoder (3)

Idar Dyrdal (idar.dyrdal@its.uio.no)

UiO : Institutt for teknologisystemer

18. september 2023

Innhold i kurset

- Introduksjon til mønstergjenkjenning
- Beslutningsteori (desisjonsteori)
- Parametriske metoder (forts.)
- Ikke-parametriske metoder
- Lineære og generaliserte diskriminantfunksjoner
- Evaluering av klassifikatorer
- Ikke-ledet læring
- Klyngeanalyse.

Bayesisk estimering – oversikt

Tettheten i punkt $\boldsymbol{x}$ er her gitt ved

$$p(\boldsymbol{x}|\mathscr{X}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathscr{X})\mathrm{d}\boldsymbol{\theta}$$

der $p(\boldsymbol{\theta}|\mathscr{X})$ er *a posteriori parameterfordeling* gitt ved

$$p(\boldsymbol{\theta}|\mathscr{X}) = \frac{p(\mathscr{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathscr{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

Her er $p(\mathscr{X}|\boldsymbol{\theta})$ likelihoodfunksjonen som kan beregnes fra treningssettet ved

$$p(\mathscr{X}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{x_k}|\boldsymbol{\theta}),$$

der $p(\boldsymbol{\theta})$ (*a priori parameterfordeling*) og $p(\boldsymbol{x}|\boldsymbol{\theta})$ er kjente funksjoner av $\boldsymbol{\theta}$.

Bayesisk estimering av forventning i multivariat normalfordeling

Antar her følgende:

$$p(\mathbf{x}|\boldsymbol{\mu}) = N(\boldsymbol{\mu}, \Sigma), \qquad \boldsymbol{\mu} = \text{ukjent}, \Sigma = \text{kjent},$$

$$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu}_0, \Sigma_0), \qquad \boldsymbol{\mu}_0 = \text{kjent}, \Sigma_0 = \text{kjent},$$

$$\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \quad \text{treningssett med n sampler fra én og samme klasse.}$$

Dette gir likelihoodfunksjonen

$$p(\mathscr{X}|\boldsymbol{\mu}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{nd/2}|\Sigma|^{n/2}} \prod_{k=1}^{n} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_k - \boldsymbol{\mu})\right].$$

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

A posteriori parameterfordeling blir da

$$p(\boldsymbol{\mu}|\mathscr{X}) = \frac{p(\mathscr{X}|\boldsymbol{\mu})p(\boldsymbol{\mu})}{\int p(\mathscr{X}|\boldsymbol{\mu})p(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}}$$

$$= \alpha \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\boldsymbol{x}_k - \boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})\right] \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^t \Sigma_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right]$$

$$= \alpha e^{-\frac{1}{2}\left[n\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^t\Sigma_0^{-1}\boldsymbol{\mu} + \sum_{k=1}^{n}\boldsymbol{x}_{\boldsymbol{k}}^t\Sigma^{-1}\boldsymbol{x}_k - 2\boldsymbol{\mu}^t\Sigma^{-1}\sum_{k=1}^{n}\boldsymbol{x}_k - 2\boldsymbol{\mu}_0^t\Sigma_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}_0^t\Sigma_0^{-1}\boldsymbol{\mu}_0\right]},$$

der nevneren inngår i skalaren α.

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

Videre regning gir

$$p(\boldsymbol{\mu}|\mathscr{X}) =$$

$$= \alpha e^{-\frac{1}{2}\left\{ n\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^t\Sigma_0^{-1}\boldsymbol{\mu} + \underbrace{\sum_{k=1}^{n}\boldsymbol{x}_k^t\Sigma^{-1}\boldsymbol{x}_k}_{\text{strykes}} - 2\boldsymbol{\mu}^t\Sigma^{-1}\sum_{k=1}^{n}\boldsymbol{x}_k - 2\boldsymbol{\mu}^t\Sigma_0^{-1}\boldsymbol{\mu}_0 + \underbrace{\boldsymbol{\mu}_0^t\Sigma_0^{-1}\boldsymbol{\mu}_0}_{\text{strykes}} \right\}}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left\{ \boldsymbol{\mu}^t(n\Sigma^{-1} + \Sigma_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^t(\Sigma^{-1}\sum_{k=1}^{n}\boldsymbol{x}_k + \Sigma_0^{-1}\boldsymbol{\mu}_0) \right\} \right]$$

$$= \alpha' \exp\left[-\frac{1}{2}\{\boldsymbol{\mu}^t\Sigma_n^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^t\Sigma_n^{-1}\boldsymbol{\mu}_n\} \right].$$

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

Her er innført størrelsene $\boldsymbol{\mu}_n$ og Σ_n som er løsninger av likningssystemet

$$
\begin{aligned}
\Sigma_n^{-1} &= n\Sigma^{-1} + \Sigma_0^{-1} \\
\Sigma_n^{-1}\boldsymbol{\mu}_n &= n\Sigma^{-1}\boldsymbol{m}_n + \Sigma_0^{-1}\boldsymbol{\mu}_0, \\
\text{der } \boldsymbol{m}_n &= \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k.
\end{aligned}
$$

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

A posteriori parameterfordeling kan da skrives som

$$\begin{aligned} p(\boldsymbol{\mu}|\mathscr{X}) &= \alpha'' \exp\left[-\frac{1}{2}\left\{\boldsymbol{\mu}^t\Sigma_n^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^t\Sigma_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{\mu}_n^t\Sigma_n^{-1}\boldsymbol{\mu}_n\right\}\right] \\ &= \alpha'' \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^t\Sigma_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right] \\ &= \underline{\underline{N(\boldsymbol{\mu}_n, \Sigma_n)}}, \end{aligned}$$

dvs. en multivariat normalfordeling med $\boldsymbol{\mu}_n$ og Σ_n gitt av likningssystemet

$$\begin{aligned} \Sigma_n^{-1} &= n\Sigma^{-1} + \Sigma_0^{-1} && (I) \\ \Sigma_n^{-1}\boldsymbol{\mu}_n &= n\Sigma^{-1}\boldsymbol{m}_n + \Sigma_0^{-1}\boldsymbol{\mu}_0. && (II) \end{aligned}$$

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

Fra likning *(I)* får man

$$\Sigma_n = [(\frac{1}{n}\Sigma)^{-1} + \Sigma_0^{-1}]^{-1} = \frac{1}{n}\Sigma(\frac{1}{n}\Sigma + \Sigma_0)^{-1}\Sigma_0 = \frac{1}{n}\Sigma_0(\frac{1}{n}\Sigma + \Sigma_0)^{-1}\Sigma$$

der relasjonene

$$(A^{-1} + B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A,$$

er benyttet. Her er A og B er ikke-singulære, kvadratiske matriser. Ved å løse for Σ_n fra *(I)* og sette inn i *(II)*, får man

$$\begin{aligned}
\boldsymbol{\mu}_n &= n\Sigma_n\Sigma^{-1}\boldsymbol{m}_n + \Sigma_n\Sigma_0^{-1}\boldsymbol{\mu}_0 \\
&= \Sigma_0(\frac{1}{n}\Sigma + \Sigma_0)^{-1}\Sigma\Sigma^{-1}\boldsymbol{m}_n + \frac{1}{n}\Sigma(\frac{1}{n}\Sigma + \Sigma_0)^{-1}\Sigma_0\Sigma_0^{-1}\boldsymbol{\mu}_0 \\
&= \Sigma_0(\frac{1}{n}\Sigma + \Sigma_0)^{-1}\boldsymbol{m}_n + \frac{1}{n}\Sigma(\frac{1}{n}\Sigma + \Sigma_0)^{-1}\boldsymbol{\mu}_0 \\
&= C_1\boldsymbol{m}_n + C_2\boldsymbol{\mu}_0 \\
&= \underline{\underline{\text{veiet middel av } \boldsymbol{m}_n \text{ og } \boldsymbol{\mu}_0.}}
\end{aligned}$$

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

Her er

$$\begin{aligned} C_1 + C_2 &= \Sigma_0\left(\frac{1}{n}\Sigma + \Sigma_0\right)^{-1} + \frac{1}{n}\Sigma\left(\frac{1}{n}\Sigma + \Sigma_0\right)^{-1} \\ &= \left(\Sigma_0 + \frac{1}{n}\Sigma\right)\left(\frac{1}{n}\Sigma + \Sigma_0\right)^{-1} \\ &= I. \end{aligned}$$

I grensen $n \to \infty$ (det asymptotiske tilfellet) får man da

$$\begin{aligned} \boldsymbol{\mu}_n &\to \boldsymbol{m}_n \\ \Sigma_n &\to \frac{1}{n}\Sigma \to 0, \end{aligned}$$

dvs. Bayes løsning $\to$ maksimum likelihood løsningen når n vokser.

Bayesisk estimering av forventning i multivariat normalfordeling (forts.)

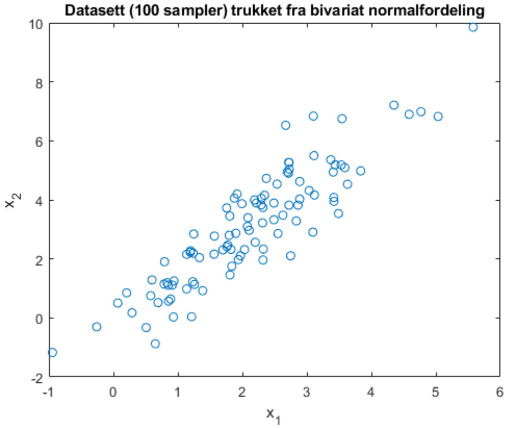
Tettheten i punktet $\boldsymbol{x}$ blir da

$$
\begin{aligned}
p(\boldsymbol{x}|\mathscr{X}) &= \int p(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\mathscr{X})\mathrm{d}\boldsymbol{\mu} \\
&= \int N(\boldsymbol{\mu},\Sigma) \times N(\boldsymbol{\mu}_n,\Sigma_n) \\
&= \underline{\underline{N(\boldsymbol{\mu}_n,\Sigma+\Sigma_n)}} \quad \text{(kan vises).}
\end{aligned}
$$

Den kjente kovariansmatrisen Σ har altså fått et tillegg Σ_n som representerer usikkerheten i estimatet av forventningen.

Asymptotisk blir $p(\boldsymbol{x}|\mathscr{X}) = N(\boldsymbol{m}_n,\Sigma)$.

Estimering av forventningsvektor i bivariat normalfordeling (Eksempel)



To egenskaper x_1 og x_2

- Datasett med 100 objekter, trukket fra normalfordeling med

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \text{ og } \Sigma = \begin{bmatrix} 1 & 1,5 \\ 1,5 & 3 \end{bmatrix}.$$

- Antar bivariat normalfordeling med ukjent $\boldsymbol{\mu}$ og kjent Σ.
- Skal finne a posteriori parameterfordeling, $\boldsymbol{\mu}_n$ og Σ_n som funksjon av antall sampler n i treningssettet.

Estimering av forventningsvektor i bivariat normalfordeling (forts.)

La her ukjent tetthetsfunksjon og a priori parameterfordeling være hhv.

$p(\boldsymbol{x}|\boldsymbol{\mu}) = N(\boldsymbol{\mu}, \Sigma)$ der Σ antas kjent og $\boldsymbol{\mu}$ er ukjent, og
$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu}_0, \Sigma_0)$ der $\boldsymbol{\mu}_0$ og Σ_0 er kjente.

A posteriori parameterfordeling blir da

$$p(\boldsymbol{\mu}|\mathscr{X}) = N(\boldsymbol{\mu}_n, \Sigma_n).$$

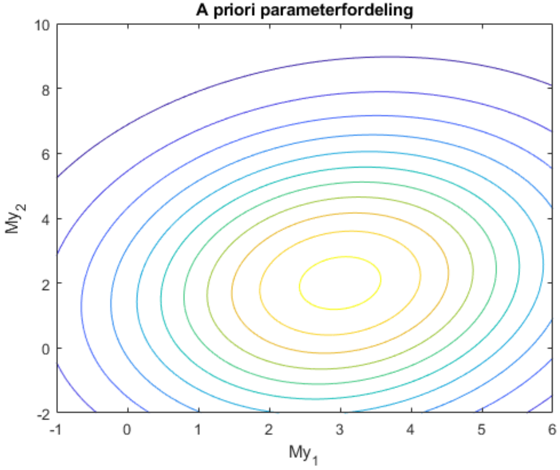
Her er estimatene av forventning og varians

$$\boldsymbol{\mu}_n = n\Sigma_n\Sigma^{-1}\boldsymbol{m}_n + \Sigma_n\Sigma_0^{-1}\boldsymbol{\mu}_0 \qquad \text{og} \qquad \Sigma_n = \left[\left(\frac{1}{n}\Sigma\right)^{-1} + \Sigma_0^{-1}\right]^{-1},$$

der $\boldsymbol{m}_n$ er sampelmiddelet over treningssettet. Skal ut fra dette beregne det Bayesiske tetthetsestimatet

$$p(\boldsymbol{x}|\mathscr{X}) = N(\boldsymbol{\mu}_n, \Sigma + \Sigma_n).$$

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



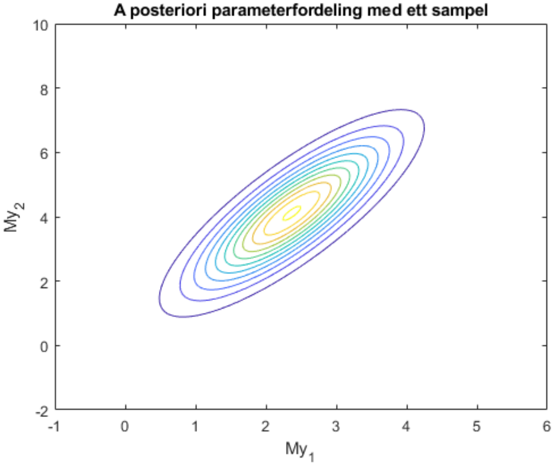
Konturplott av fordelingen

- A priori parameterfordeling $N(\boldsymbol{\mu}_0, \Sigma_0)$ der

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \text{ og } \Sigma_0 = \begin{bmatrix} 5 & 1 \\ 1 & 10 \end{bmatrix}.$$

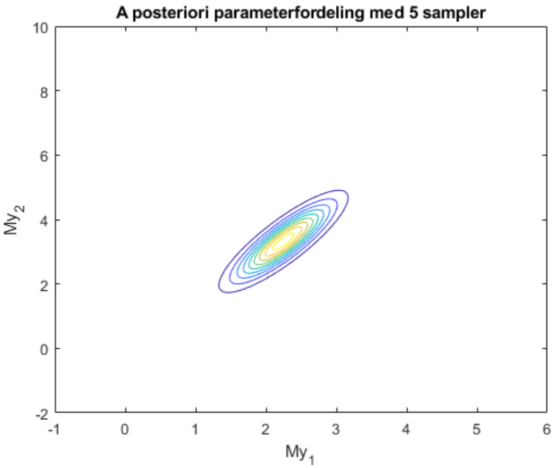
- Her er $\boldsymbol{\mu}_0$ beste gjetning om forventningsvektoren.
- Σ_0 representerer usikkerheten om denne gjetningen.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



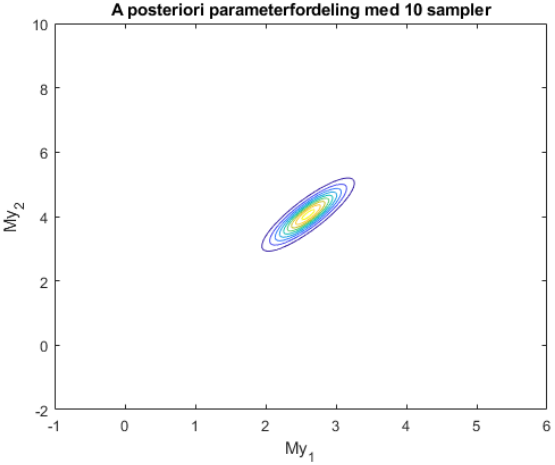
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^1)$. Maksimum for $\boldsymbol{\mu} = [2.37, 4.11]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



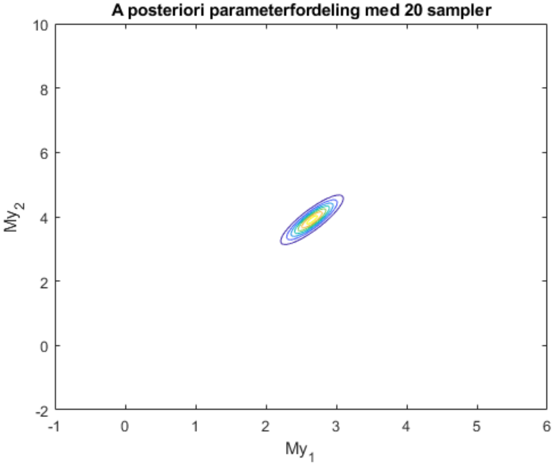
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^5)$. Maksimum for $\boldsymbol{\mu} = [2.25, 3.32]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



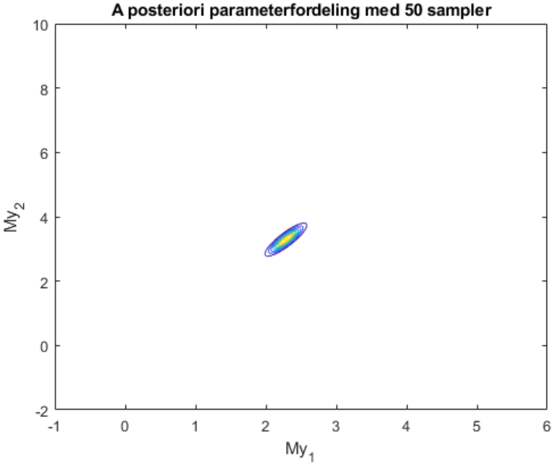
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{10})$. Maksimum for $\boldsymbol{\mu} = [2.60, 4.05]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



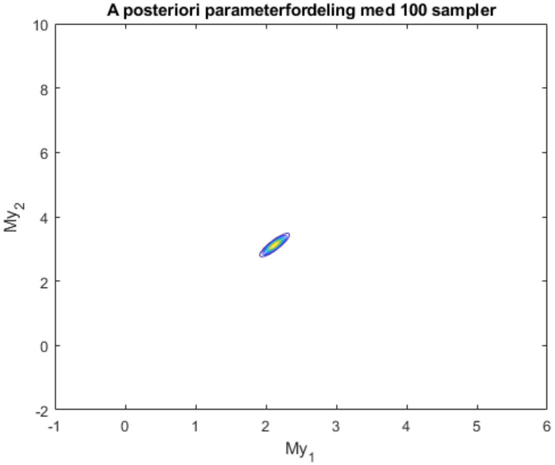
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{20})$. Maksimum for $\boldsymbol{\mu} = [2.65, 3.90]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



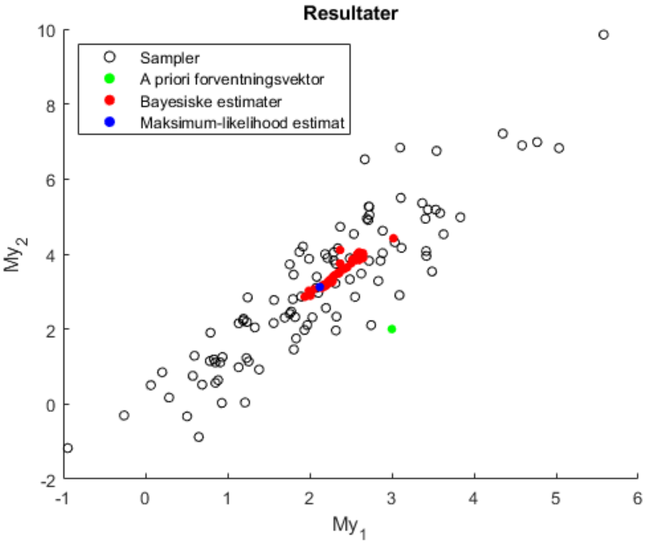
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{50})$. Maksimum for $\boldsymbol{\mu} = [2.28, 3.28]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{100})$. Maksimum for $\boldsymbol{\mu} = [2.12, 3.12]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



Estimering av forventningsvektor i bivariat normalfordeling (forts.)

Estimatet av kovariansmatrisen Σ_n for $n = 100$ blir her

$$\Sigma_n = \left\{ (\frac{1}{n}\Sigma)^{-1} + \Sigma_0^{-1} \right\}^{-1}$$

$$= \left\{ \left(\frac{1}{100} \begin{bmatrix} 1 & 1,5 \\ 1,5 & 3 \end{bmatrix} \right)^{-1} + \begin{bmatrix} 5 & 1 \\ 1 & 10 \end{bmatrix}^{-1} \right\}^{-1}$$

$$= \left\{ \begin{bmatrix} 400 & -200 \\ -200 & 1,3333 \end{bmatrix} + \begin{bmatrix} 0.204082 & -0.020408 \\ -0.020408 & 0.102041 \end{bmatrix} \right\}^{-1}$$

$$= \underline{\underline{\frac{1}{1000} \begin{bmatrix} 9,9629 & 14,934 \\ 14,934 & 29,881 \end{bmatrix}}}$$

Estimering av forventningsvektor i bivariat normalfordeling (forts.)

Forventningsestimatet $\boldsymbol{\mu}_n$ for $n = 100$ blir tilsvarende

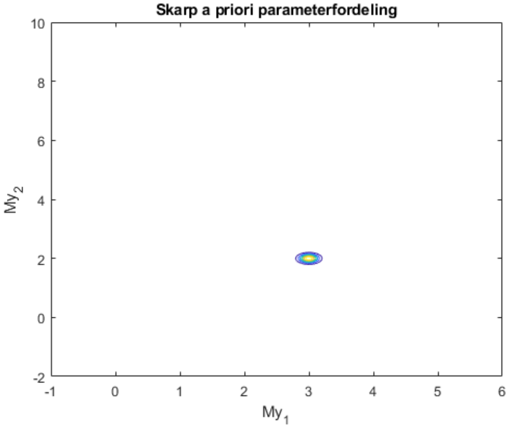
$$\boldsymbol{\mu}_n = n\Sigma_n\Sigma^{-1}\boldsymbol{m}_n + \Sigma_n\Sigma_0^{-1}\boldsymbol{\mu}_0$$

$$= \frac{1}{10}\begin{bmatrix} 9,9629 & 14,934 \\ 14,934 & 29,881 \end{bmatrix}\begin{bmatrix} 1 & 1,5 \\ 1,5 & 3 \end{bmatrix}^{-1}\boldsymbol{m}_n + \frac{1}{10}\begin{bmatrix} 9,9629 & 14,934 \\ 14,934 & 29,881 \end{bmatrix}\begin{bmatrix} 5 & 1 \\ 1 & 10 \end{bmatrix}^{-1}\boldsymbol{\mu}_0$$

$$= \begin{bmatrix} 0,99827 & -0,00132 \\ -0,00244 & 0,99726 \end{bmatrix}\begin{bmatrix} 2,1231 \\ 3,1217 \end{bmatrix} + \frac{1}{1000}\begin{bmatrix} 1,7285 & 1,3206 \\ 2,4380 & 2,7443 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$= \underline{\underline{\begin{bmatrix} 2,1231 \\ 3,1207 \end{bmatrix}}}$$

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



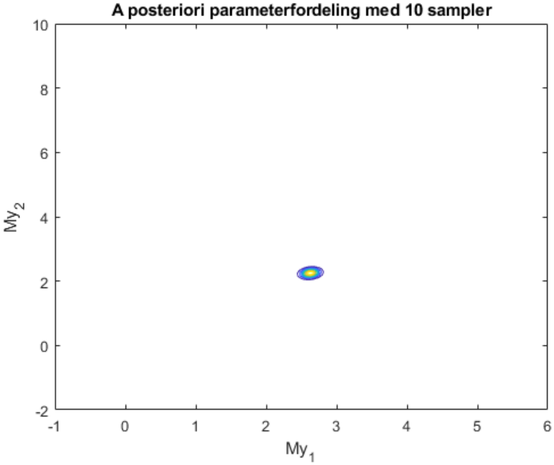
La nå a priori parameterfordeling være gitt ved $N(\boldsymbol{\mu}_0, \Sigma_0)$ med

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

og

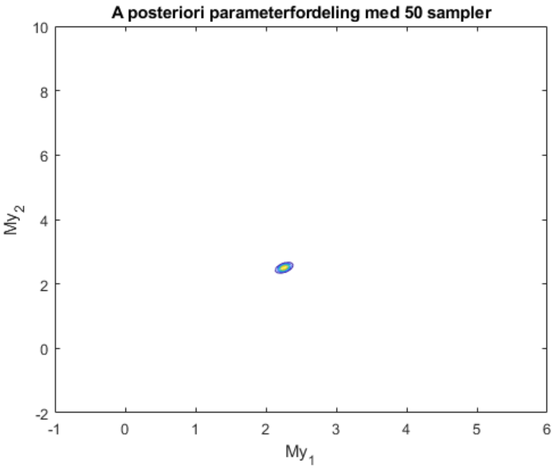
$$\Sigma_0 = \begin{bmatrix} 0,01 & 0 \\ 0 & 0,01 \end{bmatrix}.$$

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



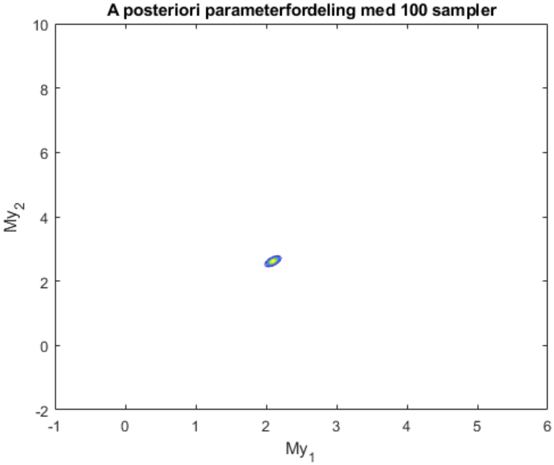
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{10})$. Maksimum for $\boldsymbol{\mu} = [2.63, 2.25]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



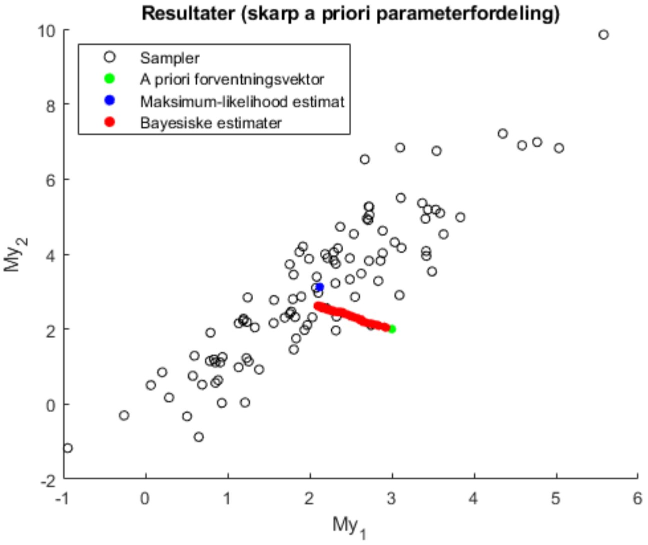
A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{50})$. Maksimum for $\boldsymbol{\mu} = [2.26, 2.50]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



A posteriori parameterfordeling $p(\boldsymbol{\mu}|\mathscr{X}^{100})$. Maksimum for $\boldsymbol{\mu} = [2.10, 2.62]^t$.

Estimering av forventningsvektor til bivariat normalfordeling (forts.)



Estimering av forventningsvektor i bivariat normalfordeling (forts.)

Estimatet av kovariansmatrisen Σ_n for $n = 100$ blir her

$$\Sigma_n = \left\{ (\frac{1}{n}\Sigma)^{-1} + \Sigma_0^{-1} \right\}^{-1}$$

$$= \left\{ \left(\frac{1}{100} \begin{bmatrix} 1 & 1,5 \\ 1,5 & 3 \end{bmatrix} \right)^{-1} + \begin{bmatrix} 0,01 & 0 \\ 0 & 0,01 \end{bmatrix}^{-1} \right\}^{-1}$$

$$= \left\{ \begin{bmatrix} 400 & -200 \\ -200 & 1,3333 \end{bmatrix} + \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix} \right\}^{-1}$$

$$= \underline{\underline{\frac{1}{1000} \begin{bmatrix} 3,0435 & 2,6087 \\ 2,6087 & 6,5217 \end{bmatrix}}}$$

Estimering av forventningsvektor i bivariat normalfordeling (forts.)

Forventningsestimatet $\boldsymbol{\mu}_n$ for $n = 100$ blir tilsvarende

$$\boldsymbol{\mu}_n = n\Sigma_n\Sigma^{-1}\boldsymbol{m}_n + \Sigma_n\Sigma_0^{-1}\boldsymbol{\mu}_0$$

$$= \frac{1}{10}\begin{bmatrix} 3,0435 & 2,6087 \\ 2,6087 & 6,5217 \end{bmatrix}\begin{bmatrix} 1 & 1,5 \\ 1,5 & 3 \end{bmatrix}^{-1}\boldsymbol{m}_n + \frac{1}{1000}\begin{bmatrix} 3,0435 & 2,6087 \\ 2,6087 & 6,5217 \end{bmatrix}\begin{bmatrix} 0,01 & 0 \\ 0 & 0,01 \end{bmatrix}^{-1}\boldsymbol{\mu}_0$$

$$= \begin{bmatrix} 0,6957 & -0,2609 \\ -0,2609 & 0,3478 \end{bmatrix}\begin{bmatrix} 2,1231 \\ 3,1217 \end{bmatrix} + \begin{bmatrix} 0,30435 & 0,26087 \\ 0,26087 & 0,65217 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$= \underline{\underline{\begin{bmatrix} 2,0974 \\ 2,6189 \end{bmatrix}}}$$

Suffisiente observatorer

Vi har sett at Bayesisk estimering gir kompliserte beregninger selv i enkle tilfeller.

Eksistensen av såkalte *suffisiente observatorer*, f.eks.

$$\boldsymbol{m}_n = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k$$

for den multivariate normalfordelingen $N(\boldsymbol{\mu}_n, \Sigma_n)$, gir mulighet til forenklinger.

Observatoren inneholder den informasjon fra treningssettet $\mathscr{X}$ som er *tilstrekkelig* for estimering av parametervektoren $\boldsymbol{\theta}$.

Suffisiente observatorer (forts.)

Definisjon:

$\boldsymbol{s}$ er suffisient for $\boldsymbol{\theta}$ hvis $p(\boldsymbol{\theta}|\boldsymbol{s})$ er uavhengig av $\boldsymbol{\theta}$, dvs. hvis

$$p(\mathscr{X}|\boldsymbol{s},\boldsymbol{\theta}) = p(\mathscr{X}|\boldsymbol{s}).$$

Innsatt i Bayes formel gir dette (med $\boldsymbol{\theta}$ antatt stokastisk)

$$p(\boldsymbol{\theta}|\boldsymbol{s},\mathscr{X}) = \frac{p(\mathscr{X}|\boldsymbol{s},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{s})}{p(\mathscr{X}|\boldsymbol{s})} = p(\boldsymbol{\theta}|\boldsymbol{s}).$$

Suffisiente observatorer (forts.)

Dette gir

$$\boldsymbol{s} \text{ suffisient} \quad \Rightarrow \quad p(\boldsymbol{\theta}|\boldsymbol{s}, \mathscr{X}) = p(\boldsymbol{\theta}|\boldsymbol{s}),$$

(der ekvivalens kan vises) slik at $\boldsymbol{s}$ inneholder all nødvendig informasjon fra $\mathscr{X}$.

Det kan også vises at

$$\boldsymbol{s} \text{ er suffisient for } \boldsymbol{\theta} \text{ hvis og bare hvis } p(\mathscr{X}|\boldsymbol{\theta}) = g(\boldsymbol{s}, \boldsymbol{\theta})h(\mathscr{X}),$$

(det såkalte *faktoriseringsteoremet*).

Suffisiente observatorer er nyttige dersom $\boldsymbol{s}$ og $g(\boldsymbol{s}, \boldsymbol{\theta})$ er enkle og mest mulig av likelihoodfunksjonen kan skilles ut i faktoren $h(\mathscr{X})$.

Suffisiente observatorer (forts.)

Bruk av suffisiente observatorer

1) Maksimum likelihood estimering:

Her skal likelihoodfunksjonen

$$p(\mathscr{X}|\boldsymbol{\theta}) = g(\boldsymbol{s}, \boldsymbol{\theta})h(\mathscr{X})$$

maksimaliseres med hensyn på parametervektoren $\boldsymbol{\theta}$.

Det er tilstrekkelig å finne maksimum av funksjonen

$$g(\boldsymbol{s}, \boldsymbol{\theta}),$$

(en kjent funksjon så snart $\boldsymbol{s}$ er beregnet fra treningssettet) med hensyn på $\boldsymbol{\theta}$, og glemme $h(\mathscr{X})$.

Suffisiente observatorer (forts.)

Bruk av suffisiente observatorer

2) Bayesisk estimering:

Her må man først finne a posteriori parameterfordeling fra uttrykket

$$p(\boldsymbol{\theta}|\mathscr{X}) = \frac{p(\mathscr{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathscr{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

$$= \frac{g(\mathbf{s},\boldsymbol{\theta})h(\mathscr{X})p(\boldsymbol{\theta})}{\int g(\mathbf{s},\boldsymbol{\theta})h(\mathscr{X})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$ (faktoren $h(\mathscr{X})$ kan her strykes oppe og nede)

$$= \frac{g(\mathbf{s},\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int g(\mathbf{s},\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

Suffisiente observatorer (forts.)

Dersom $p(\boldsymbol{\theta})$ er uniform kan denne fordelingen uttrykkes ved

$$p(\boldsymbol{\theta}|\mathscr{X}) = \frac{g(\boldsymbol{s},\boldsymbol{\theta})}{\int g(\boldsymbol{s},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \tilde{g}(\boldsymbol{s},\boldsymbol{\theta}),$$

den såkalte *kjernetettheten* til g. Kjernetettheten er invariant under skalering av g.

Resultatet ovenfor gjelder også i det asymptotiske tilfellet selv om $p(\boldsymbol{\theta})$ *ikke* er uniform.

Dersom $n >> 1$ vil $p(\boldsymbol{\theta}|\mathscr{X}) \approx \tilde{g}(\boldsymbol{s},\boldsymbol{\theta})$ selv om $p(\boldsymbol{\theta})$ ikke er uniform. I dette tilfellet blir

$$p(\boldsymbol{x}|\mathscr{X}) \approx \int p(\boldsymbol{x}|\boldsymbol{\theta})\tilde{g}(\boldsymbol{s},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

der $p(\boldsymbol{x}|\boldsymbol{\theta})$ er antatt kjent (som funksjon av $\boldsymbol{x}$ og $\boldsymbol{\theta}$), mens $\tilde{g}$ kan finnes i lærebøker.

Eksponensialfamilien

Familie av fordelinger med enkle suffisiente observatorer og faktoriseringer (enkle g eller $\tilde{g}$ funksjoner) på formen

$$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \alpha(\boldsymbol{x})\exp[a(\boldsymbol{\theta}) + \boldsymbol{b}(\boldsymbol{\theta})^t\boldsymbol{c}(\boldsymbol{x})],$$

der a er en skalar og $\boldsymbol{b}$ og $\boldsymbol{c}$ er vektorer. Likelihoodfunksjonen kan da skrives som

$$\begin{aligned} p(\mathscr{X}\,|\,\boldsymbol{\theta}) &= \prod_{k=1}^{n}\alpha(\boldsymbol{x}_k)\exp[a(\boldsymbol{\theta}) + \boldsymbol{b}(\boldsymbol{\theta})^t\boldsymbol{c}(\boldsymbol{x}_k)] \\ &= \underbrace{\exp\left[n\left\{a(\boldsymbol{\theta}) + \boldsymbol{b}(\boldsymbol{\theta})^t\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{c}(\boldsymbol{x}_k)\right\}\right]}_{g(\boldsymbol{s},\boldsymbol{\theta})}\underbrace{\prod_{k=1}^{n}\alpha(\boldsymbol{x}_k)}_{h(\mathscr{X})} \\ &= g(\boldsymbol{s},\boldsymbol{\theta})h(\mathscr{X}), \end{aligned}$$

der den suffisiente observatoren er gitt ved $\boldsymbol{s} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{c}(\boldsymbol{x}_k)$.

Eksempel – Univariat normalfordeling

Her er

$$\boldsymbol{\theta} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \; \alpha(x) = \frac{1}{\sqrt{2\pi}}, \; a(\boldsymbol{\theta}) = -\frac{\mu^2}{2\sigma^2} - \ln\sigma, \; \boldsymbol{b}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \text{ og } \; \boldsymbol{c}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix},$$

som gir

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{\mu^2}{2\sigma^2} - \ln\sigma + \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}^t \begin{bmatrix} x \\ x^2 \end{bmatrix} \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} = \underline{\underline{N(\mu, \sigma^2)}}. \end{aligned}$$

Eksempel – Multivariat normalfordeling

Her er

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \Sigma \end{bmatrix}, \ \alpha(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}}, \ a(\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\Sigma|, \ \boldsymbol{b}(\boldsymbol{\theta}) = \begin{bmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix} \text{ og } \boldsymbol{c}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^t \end{bmatrix},$$

som gir

$$\begin{aligned}
p(x|\boldsymbol{\theta}) &= \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\Sigma| + \begin{bmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix}^t \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^t \end{bmatrix} \right\} \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\underbrace{(\Sigma^{-1} : \boldsymbol{x}\boldsymbol{x}^t)}_{\text{Frobeniusprodukt}} \right\} \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x} \right\} (\text{siden } \Sigma^{-1} : \boldsymbol{x}\boldsymbol{x}^t = \boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x}) \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} = \underline{\underline{N(\boldsymbol{\mu}, \Sigma)}}.
\end{aligned}$$

Eksponensialfamilien – noen eksempler på fordelinger

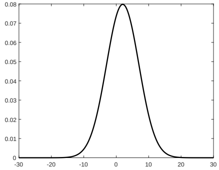
Navn	Fordeling	Parametre	Suffisient observator	$g(s,\theta)$
Univariat normal	$p(x\|\boldsymbol{\theta}) = \sqrt{\frac{\theta_2}{2\pi}} e^{-\frac{1}{2}\theta_2(x-\theta_1)^2}$	$\theta_1 = \mu$, $\theta_2 = \sigma^{-2}, \sigma > 0$	$s_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$, $s_2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2$	$\theta_2^{n/2} e^{-\frac{n}{2}\theta_2(s_2 - 2\theta_1 s_1 + \theta_1^2)}$
Multivariat normal	$p(\mathbf{x}\|\boldsymbol{\theta}) = \frac{\|\Theta_2\|^{\frac{1}{2}}}{(2\pi)^{d/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\theta}_1)^t\Theta_2(\mathbf{x}-\boldsymbol{\theta}_1)}$	$\boldsymbol{\theta}_1 = \boldsymbol{\mu}$, $\Theta_2 = \Sigma^{-1}, \|\Theta_2\| > 0$	$\mathbf{s}_1 = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$, $S_2 = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^t$	$\|\Theta_2\|^{n/2} e^{\frac{n}{2}[Tr(\Theta_2)s_2 - 2\theta_1^t\Theta_2\mathbf{s}_1 + \theta_1^t\Theta_2\boldsymbol{\theta}_1]}$
Eksponensial	$p(x\|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & ellers \end{cases}$	$\theta > 0$	$s = \frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta^n e^{-n\theta s}$
Gamma	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta_2 x} & x \geq 0 \\ 0 & ellers \end{cases}$	$\theta_1 > -1$, $\theta_2 > 0$	$s_1 = \left(\prod_{k=1}^{n} x_k\right)^{1/n}$, $s_2 = \frac{1}{n}\sum_{k=1}^{n} x_k$	$\left[\frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} s_1^{\theta_1} e^{-\theta_2 s_2}\right]^n$
Poisson	$P(x\|\theta) = \frac{\theta^x}{x!} e^{-\theta}$, der $x = 0, 1, 2, \ldots$	$\theta > 0$	$s = \frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta^{ns} e^{-n\theta}$
Multinomial	$P(\mathbf{x}\|\boldsymbol{\theta}) = m! \prod_{i=1}^{d} \frac{\theta_i^{x_i}}{x_i!}$, der $\begin{cases} x_i = 0, 1, \ldots, m \\ \sum_{i=1}^{d} x_i = m \end{cases}$	$\theta_{i, i=1,\ldots,d}$ der $\begin{cases} 0 < \theta_i < 1 \\ \sum_{i=1}^{d} \theta_i = 1 \end{cases}$	$\mathbf{s} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$	$\prod_{i=1}^{d} \theta_i^{ns_i}$

Begrensninger ved parametriske metoder

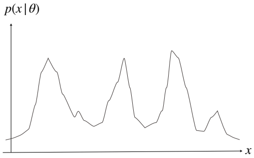
Parametriske metoder har visse begrensninger:

- Tetthetsfunksjonene har som oftest ukjent form slik at gal eller dårlig antakelse om formen gir suboptimalt resultat.
- De fleste kjente (enkle) fordelinger har bare én mode, slik at de passer dårlig til mange virkelige fordelinger med flere moder (multimodale fordelinger).
- Tetthetsfunksjonene kan være kjente, men beregningsmessig kompliserte; de kan f.eks. bestå av en blanding av unimodale komponenter.

Unimodal fordeling



Multimodal fordeling



Innhold i kurset

- Introduksjon til mønstergjenkjenning
- Beslutningsteori
- Parametriske metoder
- Ikke-parametriske metoder
- Lineære og generaliserte diskriminantfunksjoner
- Evaluering av klassifikatorer
- Ikke-ledet læring
- Klyngeanalyse.