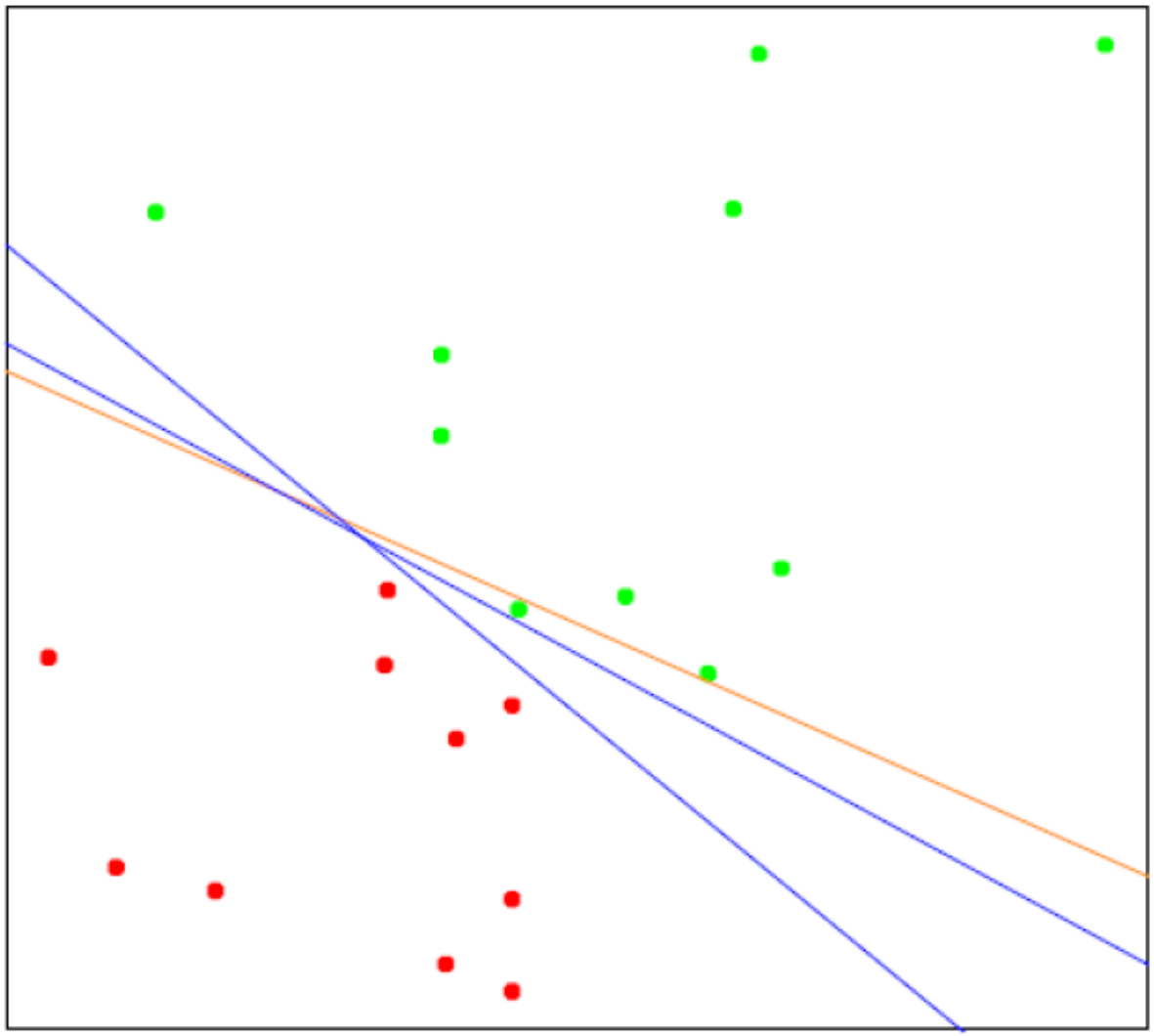
# Klassifikationsmetoder: Support Vectors, LDA, QDA

Peter N. Bakker

08-06-2023

# Separerende hyperplaner

Vi ønsker at adskille to grupper af data med en hyperplan på formen $L = \{x \in R^p : x^T\beta + \beta_0 = 0\}$.



# Optimal Margin Classifier

**Motivation:**

- Optimal Margin Classifier sigter mod at finde den plan med maksimal margin mellem de to klasser.
- Ved at maksimere marginen forsøger Optimal Margin Classifier at opnå bedre generaliseringsevne og robusthed over for støj.

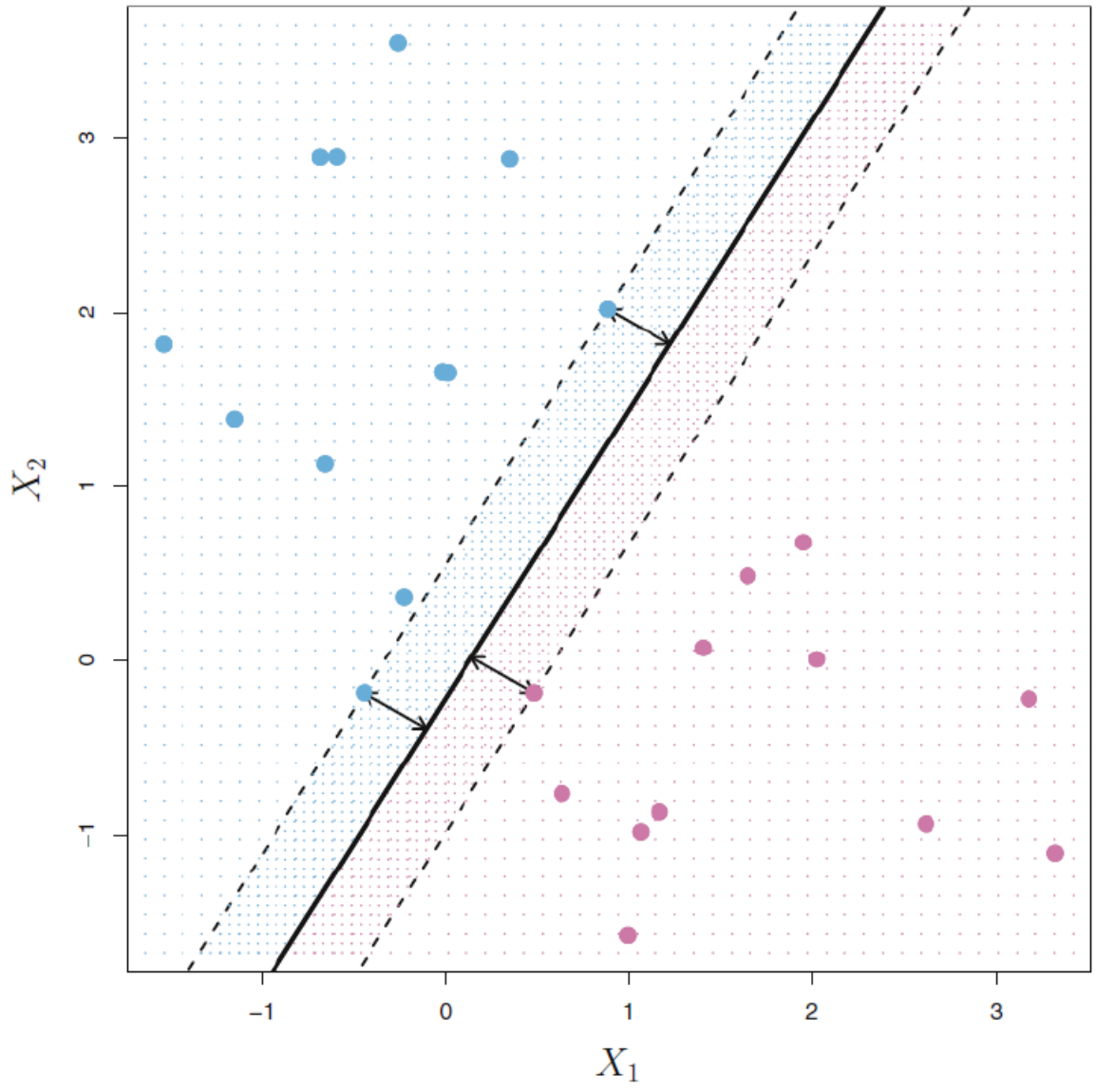
$$\max_{\beta,\beta_0,M,||\beta||=1} M$$
$$s.t. y_i(x_i^T \beta + \beta_0) \geq M$$

- Det er kun observationerne **på kanten af** margninen som har indflydelse på planens retning. Disse kaldes **support vectors**.

**Problemer:**

- Optimal Margin Classifier fungerer kun, når dataene er lineært separable.
- Hvis dataene ikke er lineært separable, kan der opstå problemer med fejlklassifikation og manglende generaliseringsevne (høj varians).

# Optimal Margin Classifier Illustration



# Support Vector Classifiers (SVC)

**Motivation:**

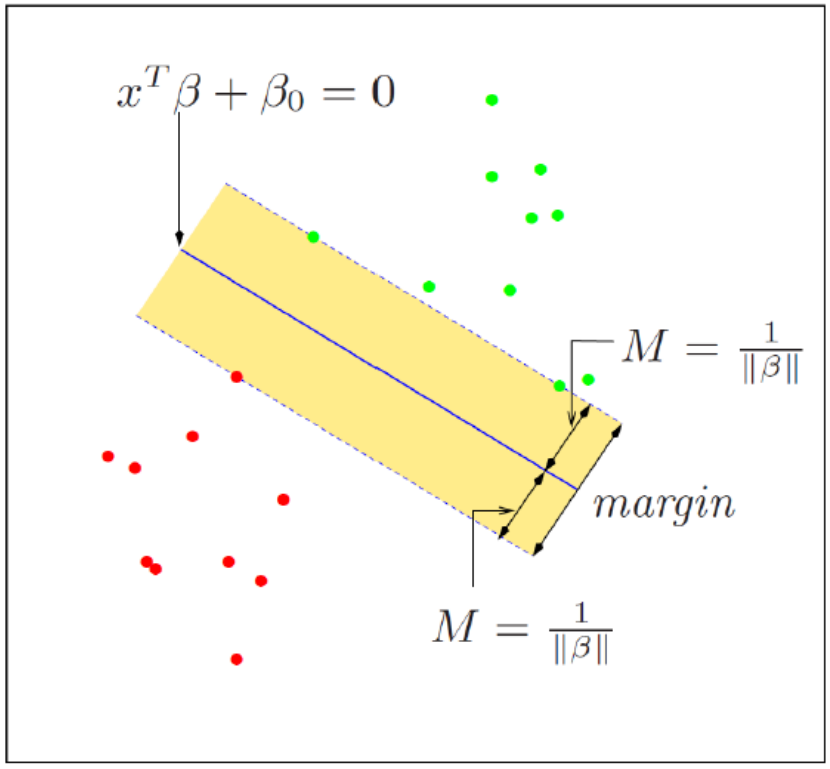
- SVC kan løse problematikken ved at introducere "slack"-variable til alle observationer, og så "straffe" slack-omfanget.

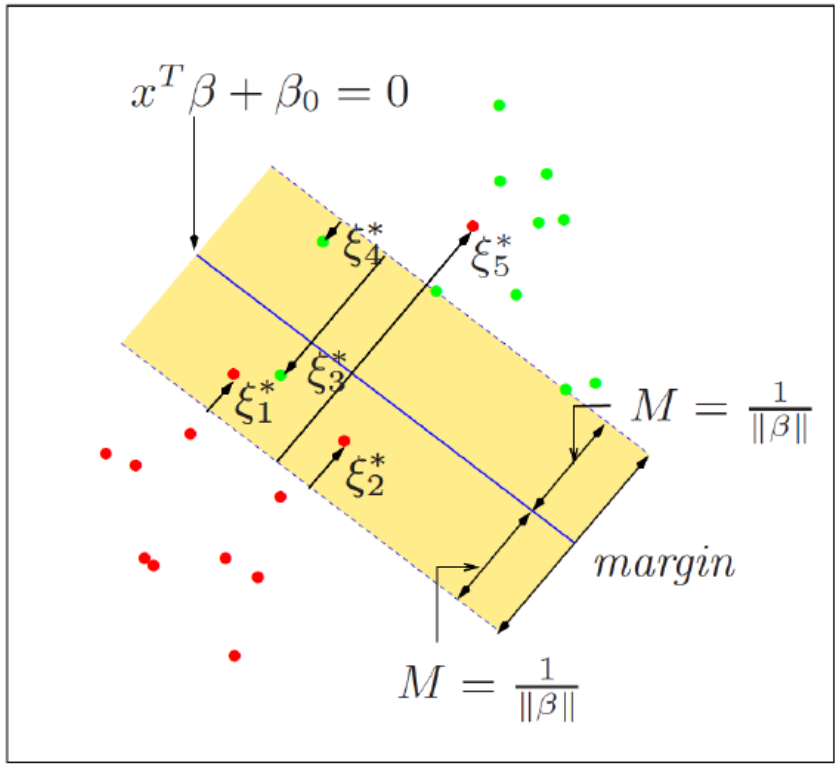
$$\min_{\beta,\beta_0}(\frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i)$$

$$s.t.\ \xi_i \geq 0 \quad \text{og} \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$$

- Ved SVC er det nu muligt at have nogle punkter indenfor marginen. Hvorfor det nu er muligt at kunne seperare, hvis der er overlap. Samt reduceres variansen også.
- Her er det punkterne både **på kanten** og **indenfor** marginen der aggerer support vectors. Og altså de eneste der har betydning for bestemmelse af $\beta$ og $\beta_0$.
- $y_i$ har værdier -1 eller 1. $\hat{G}(x) = sign(x^T\hat{\beta} + \hat{\beta}_0)$ afgør klassificering.

# SVC Illustration



# Support Vector Machines (SVM)

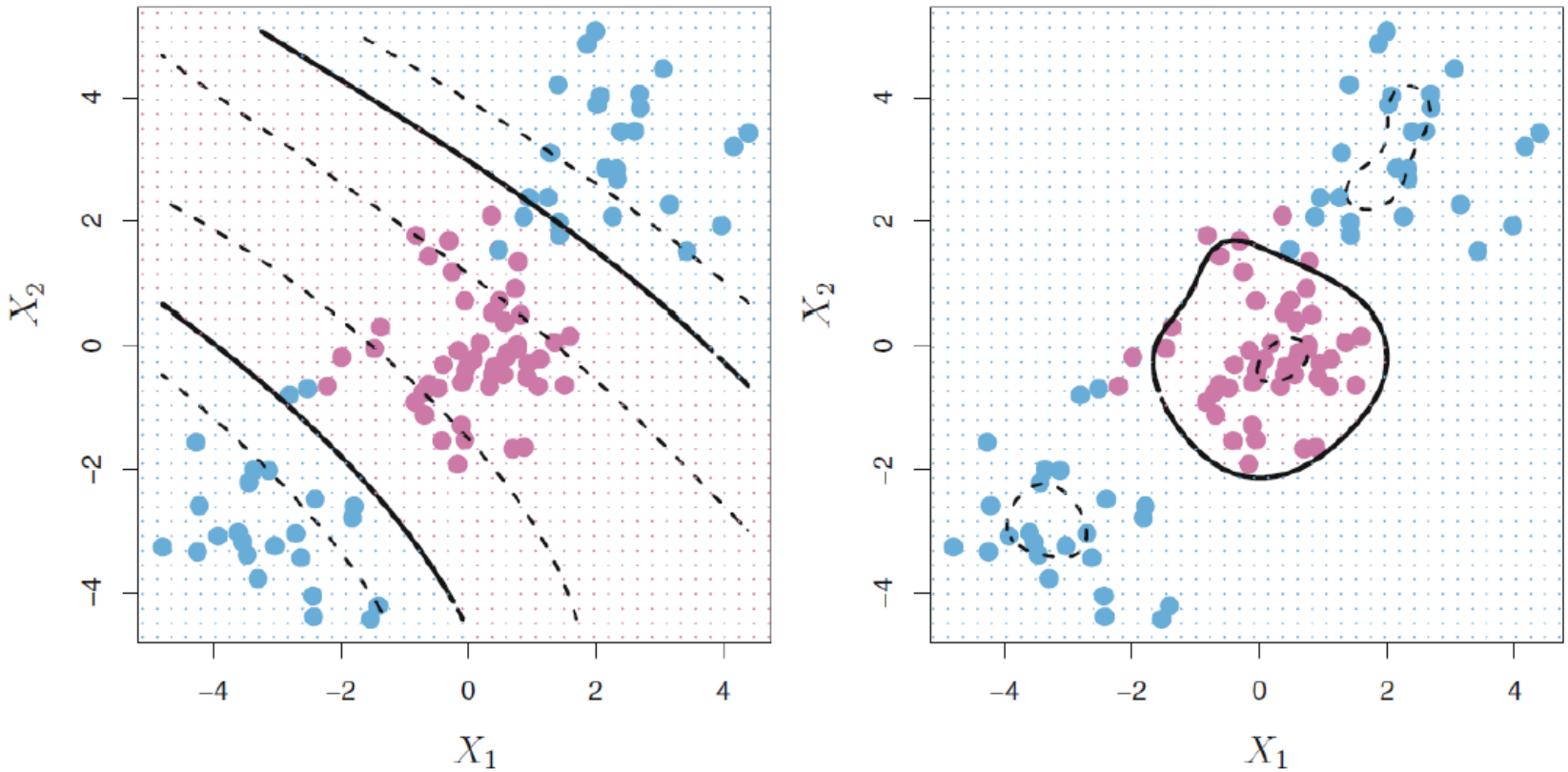
**Motivation:**

- SVM er en generalisering af SVC, der udvider mulighederne for ikke-lineær klassifikation ved hjælp af kernefunktioner.
- SVM introducerer en transformation af dataene til et højdimensionelt rum (gennem kernefunktion), hvor klasserne kan adskilles af SVC (dvs. lineær plan).
- SVM sigter stadig mod at maksimere marginen og udvælge support vektorer, men med flere muligheder for klassifikation.

**Support Vector Machines:**

- SVM anvender kernefunktioner til at transformere dataene og skabe ikke-lineære grænseflader i det højdimensionelle rum.
- Lineære (SVC), Polynomium af d'te orden, Radial og "Neuralt netværk" (tanh).

# SVM illustration



# LDA (Lineær Diskriminant Analyse)

- LDA er en lineær klassifikationsmetode, hvor man modellerer den simultane fordeling af $X$ og $G$. Modsat logistisk regression, hvor man modellerer den betingede fordeling af $G$ givet $X$.
- Der antages ofte at den betingede fordeling af $X$ givet $G$ er $(X|G = k) \sim MVN(\mu_k, \Sigma)$, som noteres $f(x|k)$. Vha. bayes kan vi definere for $g$ og $k$ i $0, \ldots, K-1$:

$$P(G = g|X = x) = \frac{f(x|g)\pi_g}{\sum_{\ell=0}^{K-1} f(x|\ell)\pi_\ell}$$

$$f(x \mid g) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu_g)^T \Sigma^{-1}(x-\mu_g)\right)$$

- Ved at indsætte overstående kan vi give vores bud på $\hat{P}(G = g|X = x)$.

# LDA

Udtrykket for $P(G = k \mid X = x)$ fører til, at

$$\log \frac{P(G = k \mid X = x)}{P(G = \tilde{k} \mid X = x)} = \log \frac{f(x \mid k)}{f(x \mid \tilde{k})} + \log \frac{\pi_k}{\pi_{\tilde{k}}}$$

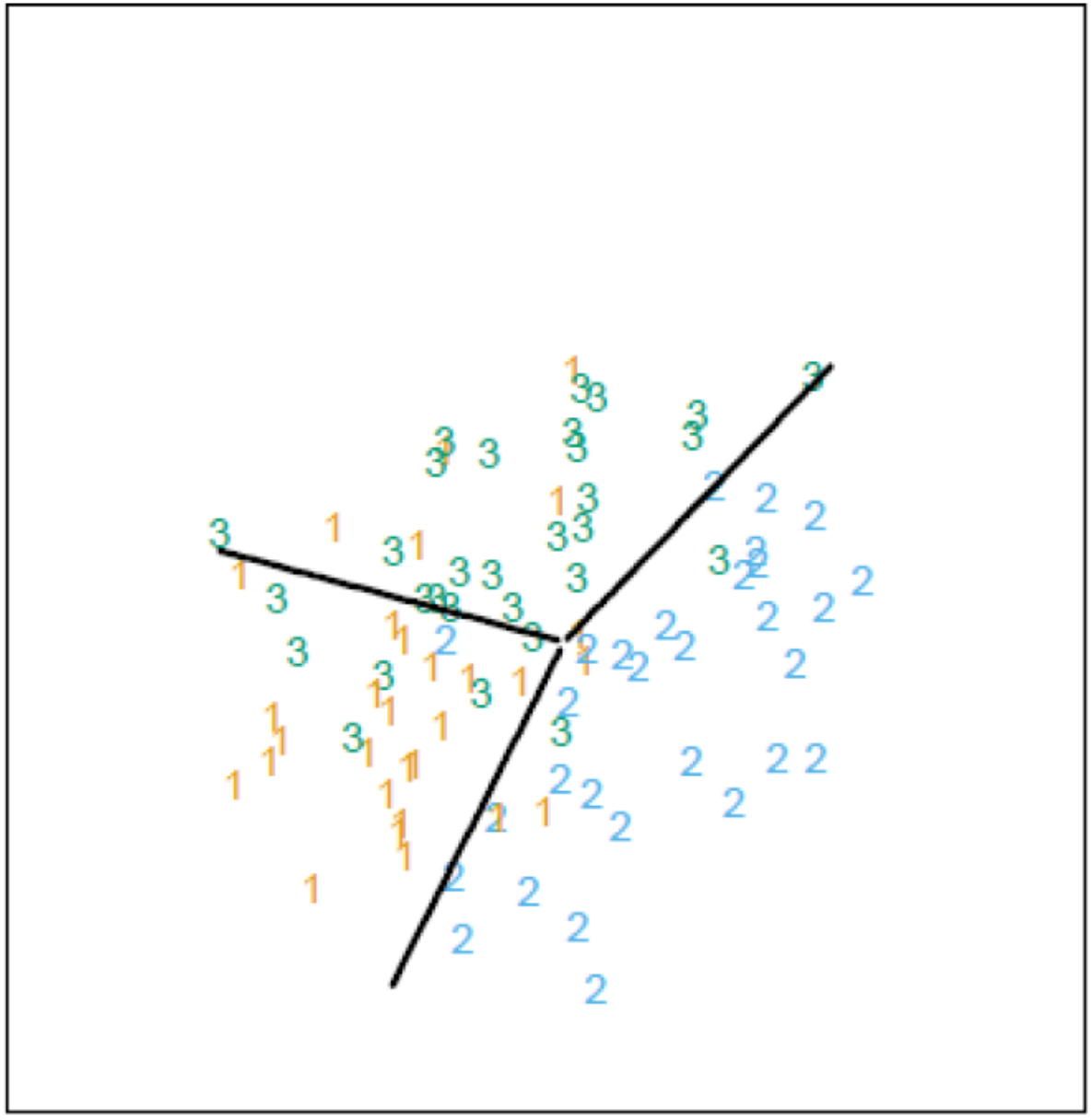
$$= \log \frac{\pi_k}{\pi_{\tilde{k}}} - \frac{1}{2}(\mu_k + \mu_{\tilde{k}})^T \Sigma^{-1}(\mu_k - \mu_{\tilde{k}}) + x^T \Sigma^{-1}(\mu_k - \mu_{\tilde{k}})$$

Hvis udtrykket er:

- $> 0$, er klasse $k$ mere sandsynlig end klasse $\tilde{k}$.
- $= 0$, er klasse $k$ og klasse $\tilde{k}$ lige sandsynlige.
- $< 0$, er klasse $k$ mindre sandsynlig end klasse $\tilde{k}$.

På den måde kan man adskille klasse $k$ fra klasse $\ell$. Minder lidt om SVM.

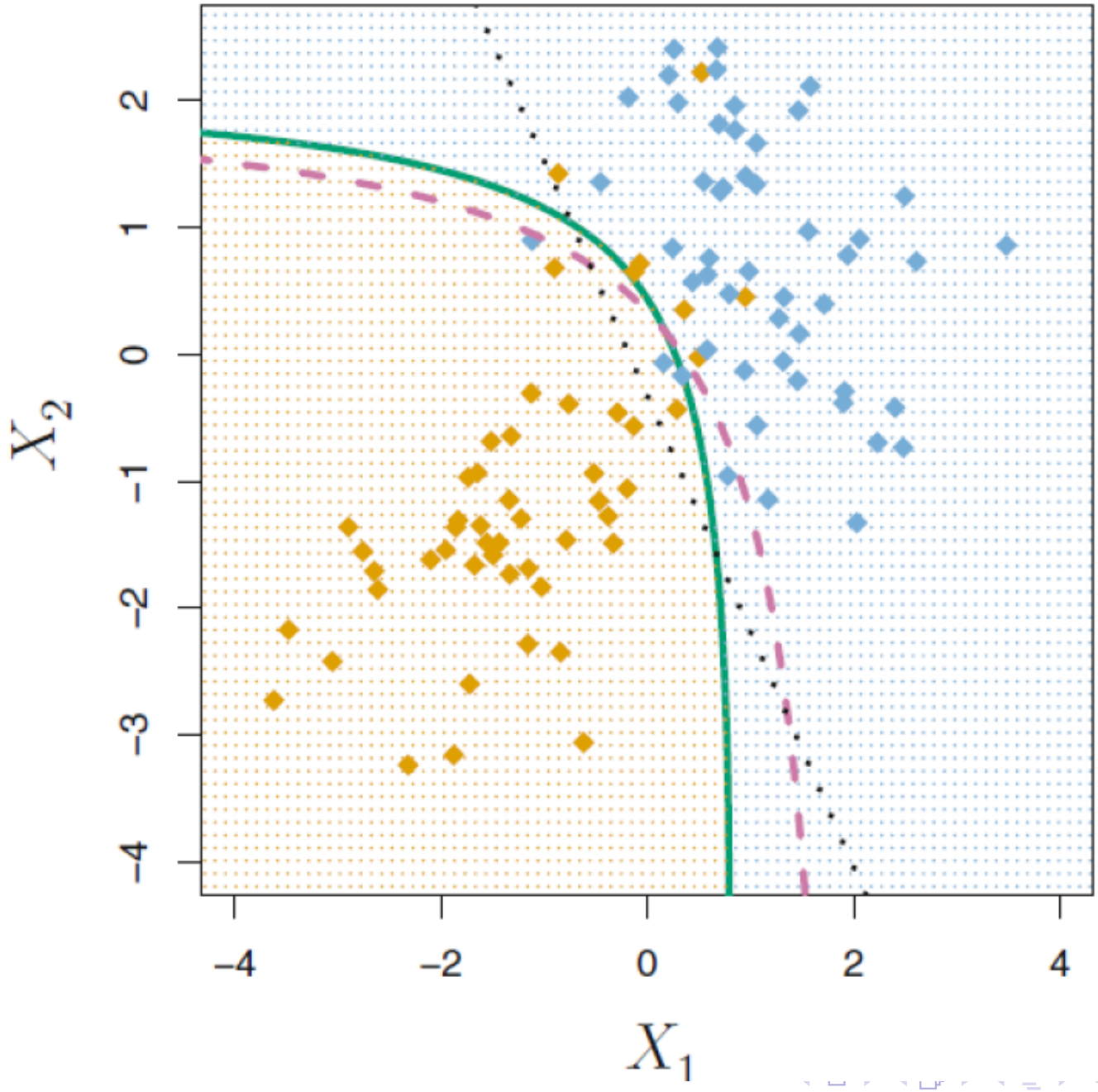
# LDA illustration



# Kvadratisk diskriminansanalyse (QDA)

- Meget det samme som LDA. Der tillades blot forskellig varians i hver klasse.
- Fordelingen bliver nu $(X|G = g) \sim MVN(\mu_g, \Sigma_g)$
- Hvorfor tætheden bliver
  $f(x \mid g) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma_g)}} \exp\left(-\frac{1}{2}(x-\mu_g)^T \Sigma_g^{-1}(x-\mu_g)\right)$
- Estimeringen af sandsynligheden er den samme som LDA. Dog er afgrænsningen mellem de to områder ikke længere lineær.
- Både LDA og QDA gør meget grove antagelser om den simultane fordeling. Dog viser det sig, at modellerne fungerer godt.
- I QDA skal der estimeres flere parametre, fordi der skal estimeres seperate $\Sigma$ (større varians). LDA kan dog give større bias, hvis kovarianserne virkelig er forskellige imellem klasserne.

# QDA illustration

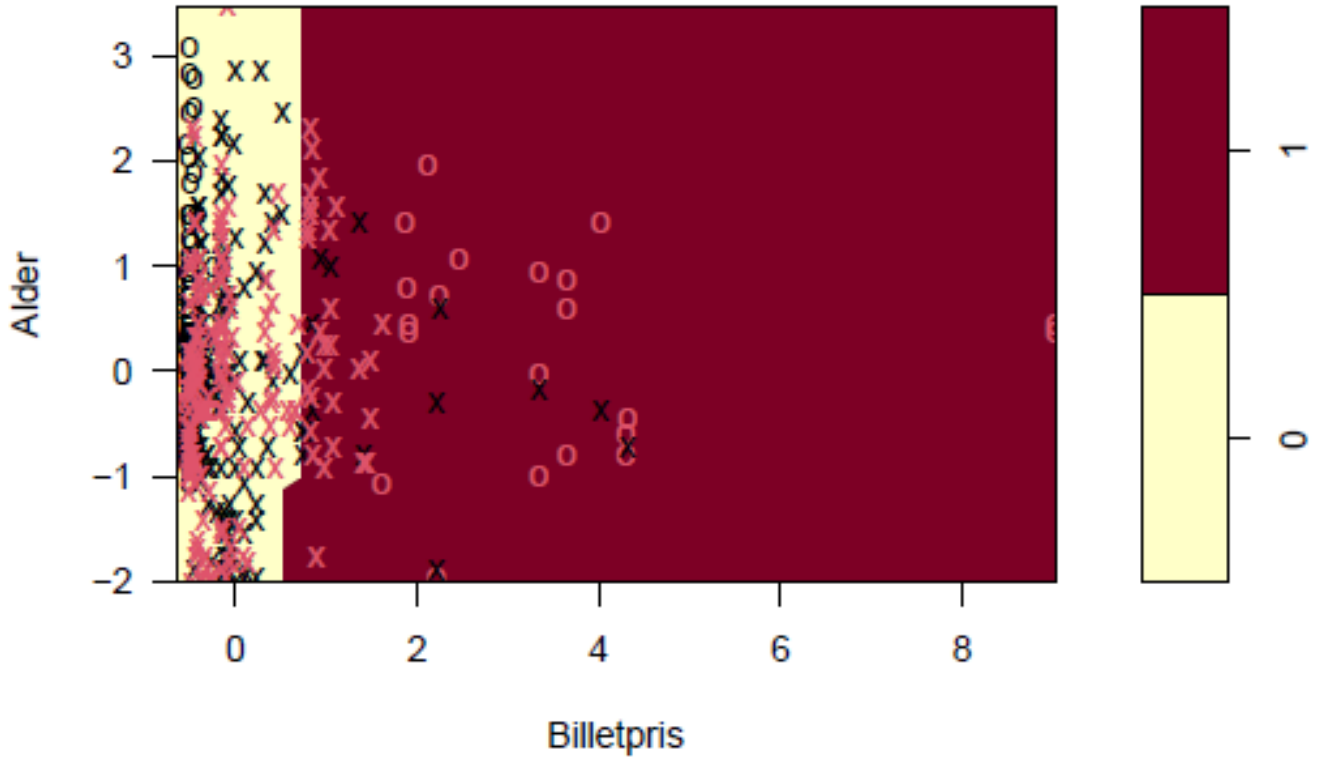


# Dataeksempel titanic (SVC og SVM)

- Vi skal estimere overlevelse vha. alder og billetpris. Jeg splitter 500 i train og 212 i test, og tilbageholder 20% til validering fra train.
- Jeg fitter en SVC (lineær) med optimal cost igennem CV.
- Jeg gør det samme med en radial kerne. Jeg benytter igen CV for at finde optimale værdier af $C$ og $\gamma$.

# Dataeksempel SVC

Fejlrate $= 40.57\%$



# Dataeksempel SVM

Fejlrate = 35.85%. Det virker til at det nu opfanges at børn uanset billetpris har større ssh. for overlevelse.