Tema: Broer, kæder og kurvefit

At tilnærme en samling datapunkter med en matematisk defineret kurve kaldes at foretage et *kurvefit*, og det er en anvendelse af matematik, som man ofte støder på. Vi kender det fra *lineær regression*, hvor en række datapunkter i planen tilnærmes med en ret linje, som er grafen for en lineær funktion. I det tilfælde vil vi kalde den lineære funktion for den *fittende funktion*. Andre eksempler på fit er *eksponentiel regression* og *potensregression*. Normalt skal man selv foretage et valg blandt de typer funktioner eller familie af funktioner, man ønsker at foretage et fit med i en konkret problemstilling. Én begrundelse for et givet valg kan være, at man teoretisk set forventer, at data følger en bestemt type funktion. I andre tilfælde er man blot interesseret i at finde en tilstrækkelig pæn kurve, som tilnærmer datapunkterne pænt. Et eksempel på sidstnævnte er de såkaldte *splines*, hvor man anvender en kurve, som er stykvist defineret. Vi skal ikke gå nærmere ind på sidstnævnte type, da det er en hel teori for sig. I stedet skal vi kigge på et par eksempler af førstnævnte type, hvor vi tilnærmer data med et andengradspolynomium.

Eksempel 1

Givet tre punkter i planen: $(-2;3{,}5), (1,2)$ og $(3,-4)$. Vi skal se, at der findes netop et andengradspolynomium, hvis graf går igennem disse tre datapunkter. Vi skal med andre ord bestemme koefficienterne a, b og c, så andengradspolynomiet $p(x) = a \cdot x^2 + b \cdot x + c$ opfylder $p(-2) = 3{,}5$, $p(1) = 2$ og $p(3) = -4$. Det giver tre ligninger med tre ubekendte, hvor de ubekendte er a, b og c:

$$(1)\quad \begin{cases} a\cdot(-2)^2 + b\cdot(-2) + c = 3{,}5 \\ a\cdot 1^2 + b\cdot 1 + c = 2 \\ a\cdot 3^2 + b\cdot 3 + c = -4 \end{cases} \Leftrightarrow \begin{cases} 4a - 2b + c = 3{,}5 \\ a + b + c = 2 \\ 9a + 3b + c = -4 \end{cases}$$

En måde at løse de tre ligninger med tre ubekendte på er ved at isolere en af de ubekendte i den ene ligning og indsætte udtrykket i de to andre ligninger. Den anden ligning ser mest simpel ud, så vi vælger at isolere c i denne og indsætte udtrykket for denne i de to øvrige.

$$(2)\quad \begin{cases} 4a - 2b + (2-a-b) = 3{,}5 \\ c = 2-a-b \\ 9a + 3b + (2-a-b) = -4 \end{cases} \Leftrightarrow \begin{cases} 3a - 3b = 1{,}5 \\ c = 2-a-b \\ 8a + 2b = -6 \end{cases}$$

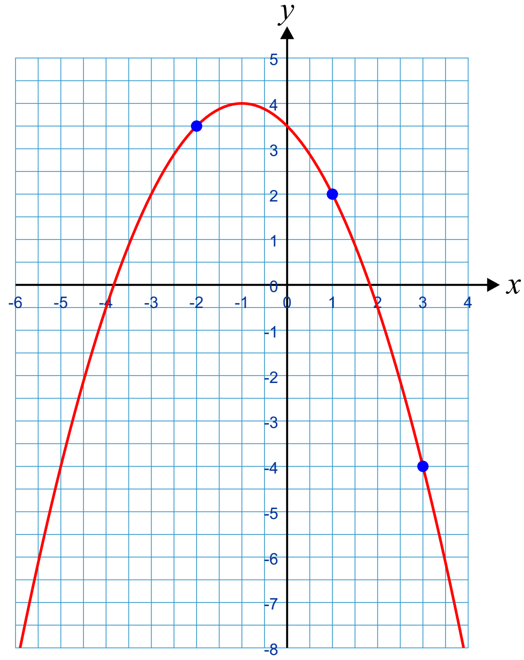
Vi fokuserer nu på første og tredje ligning. De udgør et ligningssystem bestående af to ligninger med to ubekendte. Vi isolerer b i den sidste og indsætter udtrykket for b i den anden ligning:

$$(3)\quad \begin{cases} 3a - 3\cdot(-3-4a) = 1{,}5 \\ b = -3-4a \end{cases} \Leftrightarrow \begin{cases} 15a+9 = 1{,}5 \\ b = -3-4a \end{cases}$$

Den ubekendte a kan herefter bestemmes af den første ligning. Resultatet indsættes i den anden for at bestemme b.

(4) $$\begin{cases} a = -\frac{1}{2} \\ b = -3-4\cdot(-0{,}5) \end{cases} \quad \Leftrightarrow \quad \begin{cases} a = -\frac{1}{2} \\ b = -1 \end{cases}$$

Disse to værdier indsættes i udtrykket for c i (2): $c = 2-a-b = 2-(-0{,}5)-(-1) = 3{,}5$. Vi har dermed bestemt det andengradspolynomium, hvis graf går igennem de tre datapunkter: $p(x) = -0{,}5x^2 - x + 3{,}5$. Tegner vi grafen for andengradspolynomiet, ser vi da også, at punkterne virkelig ligger på grafen. Der gælder generelt, at hvis man har givet tre punkter, som har forskellige x-værdier og som ikke ligger på linje, så findes der netop én parabel, som passerer igennem punkterne. Vi skal dog undlade at bevise denne påstand her.



□

Det er klart, at et CAS-værktøj lynhurtigt vil kunne udregne koefficienterne i eksempel 1. Typisk kan man definere andengradspolynomiet $p(x) = a\cdot x^2 + b\cdot x + c$ og derefter løse ligningssystemer bestående af de tre ligninger $p(-2) = 3{,}5$, $p(1) = 2$ og $p(3) = -4$ med hensyn til de ubekendte a, b og c. Hvordan det helt præcist gøres, afhænger af ens CAS-værktøj. Men hvad nu, hvis vi har fire datapunkter? Vi ser på det i et eksempel.

Eksempel 2

Lad os sige, at vi udover de tre punkter givet i eksempel 1 også har datapunktet $(-3,1)$. Det vil give anledning til en ekstra ligning, markeret med rødt nedenfor:

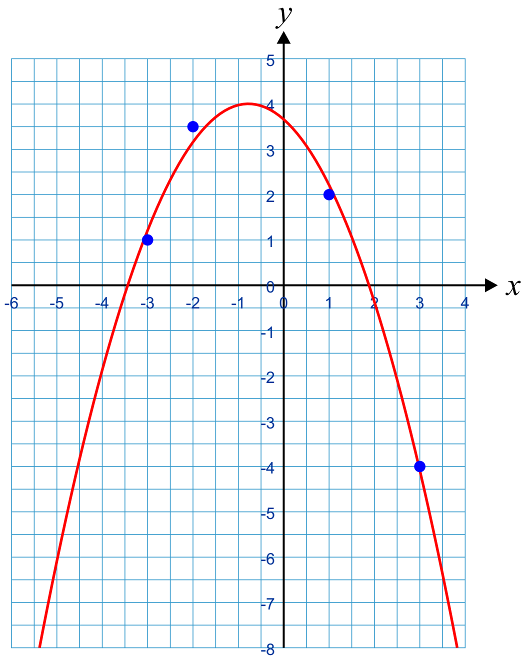
(5) $$\begin{cases} 4a - 2b + c = 3{,}5 \\ a + b + c = 2 \\ 9a + 3b + c = -4 \\ 9a - 3b + c = 1 \end{cases}$$

Fra eksempel 1 ved vi, at de tre første ligninger har en entydig løsning. Den eneste mulighed for at hele ligningssystemet (5) har en løsning er derfor, at løsningen til de første tre ligninger givet ved $a = -0{,}5$, $b = -1$ og $c = 3{,}5$ passer i den sidste ligning. Dette er ikke tilfældet. Hvis vi vil foretage et fit med et andengradspolynomium, må vi med andre ord stille os tilfreds med, at grafen ikke kan passere igennem alle fire datapunkter. Hvordan finder vi da den bedste løsning på problemet. Hvad der er "bedst" kan naturligvis diskuteres. Et meget anvendt valgt er at benytte *mindste kvadraters metode*.

Metoden består i at vælge a, b og c, så *kvadratsummen* af de lodrette afstande fra datapunkterne til parablen bliver mindst mulig. Princippet bag mindste kvadraters metode er illustreret i eksempel 3 nedenfor. Bruger man et CAS-værktøj til at lave et fit med et andengradspolynomium (*polynomiel regression*) vil man få følgende andengradspolynomium som løsning:

$$p(x) = -0{,}56638\,x^2 - 0{,}88277\,x + 3{,}6610$$

Hvor godt polynomiet tilnærmer datapunkterne angives ved den såkaldte *forklaringsgrad* $R^2 = 0{,}99331$. Det er dog altid fornuftigt også at tage en visuel inspektion af grafen, før man fælder nogen dom.



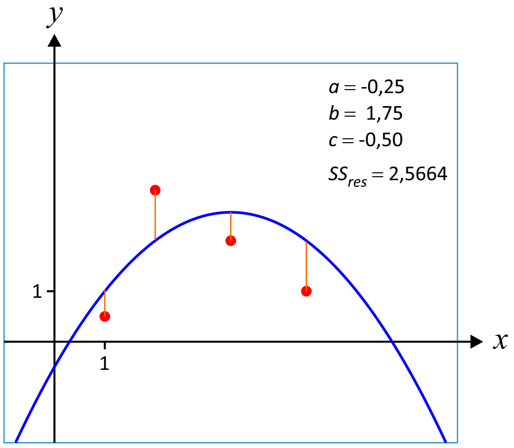
□

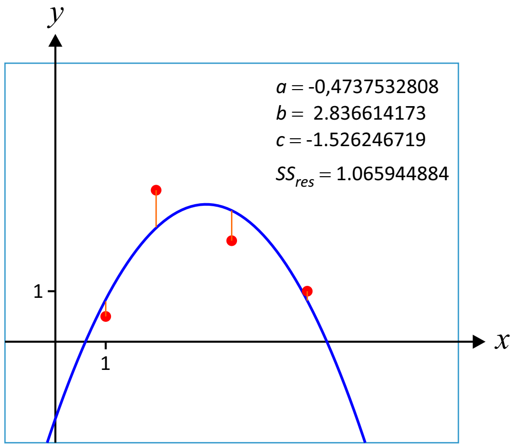
Eksempel 3 (Princippet bag mindste kvadraters metode)

Princippet bag *mindste kvadraters metode* kan nemmest forklares ved et eksempel. Vi tænker os givet de fire datapunkter $(1;0{,}5), (2,3), (3{,}5;2)$ og $(5,1)$. De fire punkter indtegnes i et koordinatsystem. Man kunne overveje at foretage et gæt på et andengradspolynomium, som er en rimelig god tilnærmelse til datapunkterne. Vi vælger andengradspolynomiet $p(x) = -0{,}25x^2 + 1{,}75x - 0{,}50$. Med *residualet* i $x = 1$ menes forskellen mellem dataværdien og funktionsværdien i $x = 1$: $0{,}5 - p(1) = 0{,}5 - 1 = -0{,}5$. Det betyder, at den lodrette afstand fra datapunktet til grafen er 0,5, og det negative fortegn fortæller, at datapunktet ligger *under* grafen (se grafen på næste side). På tilsvarende vis kan residualerne i de øvrige tre x-værdier bestemmes. Kvadreres de og lægger man sammen, fås:

$$\begin{aligned} SS_{res} &= \left(0{,}5 - p(1)\right)^2 + \left(3 - p(2)\right)^2 + \left(2 - p(3{,}5)\right)^2 + \left(1 - p(5)\right)^2 \\ &= \left(0{,}5 - 1\right)^2 + \left(3 - 2\right)^2 + \left(2 - 2{,}5625\right)^2 + \left(1 - 2\right)^2 \\ &= 2{,}5664 \end{aligned} \tag{6}$$

Her er SS en forkortelse af det engelske "Sum of Squares" og indekset *res* er en forkortelse for det engelske "residual". Det er nærliggende at forsøge at vælge koefficienterne a, b og c i andengradspolynomiet på en måde, så den ikke-negative størrelse SS_{res} bliver så lille som muligt. Kan man vælge a, b og c, så den bliver eksakt 0, betyder det for eksempel, at grafen for andengradspolynomiet går eksakt igennem datapunkterne. Dette er normalt ikke muligt, så det gælder om at finde koefficienter, så SS_{res} bliver så lille som overhovedet muligt. Heldigvis findes der en metode til at finde minimum for SS_{res}, men da den involverer funktioner af flere variable, afstår vi fra at forklare mere om det her. Heldigvis har de fleste CAS-værktøj et værktøj til at bestemme det andengradspolynomium, som giver den mindste værdi for SS_{res}. I dette tilfælde er den mindste værdi 1,0659 og den opnås for polynomiet $p(x) = -0{,}4738x^2 + 2{,}8366x - 1{,}5262$. Figuren på næste side viser grafen for gættet samt den optimale løsning.





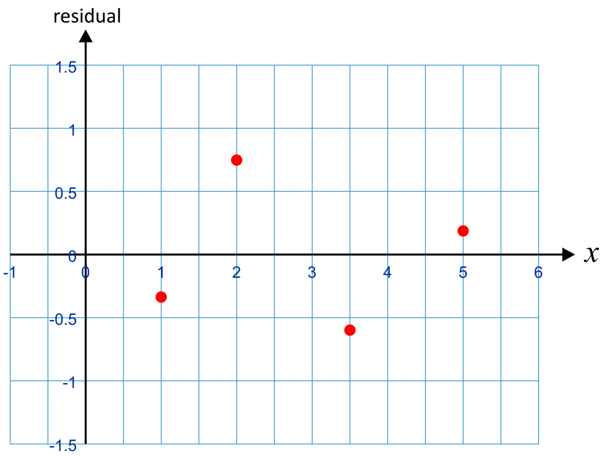
For at kunne udregne den såkaldte forklaringsgrad, skal vi have fat i endnu en størrelse, nemlig *den totale kvadratsum* SS_{tot}. Her får vi først brug for gennemsnittet af y-værdierne for datapunkterne: $\overline{y} = (0{,}5+3+2+1)/4 = 1{,}625$. Den totale kvadratsum fås ved at udregne forskellen mellem de enkelte datapunkters y-værdier og gennemsnittet $\overline{y}$, kvadrere og lægge sammen:

(7) $$SS_{tot} \;=\; (0{,}5-1{,}625)^2+(3-1{,}625)^2+(2-1{,}625)^2+(1-1{,}625)^2 \;=\; 3{,}6875$$

Forklaringsgraden (Eng: *Coefficient of Determination*) betegnet R^2 defineres ved:

(8) $$R^2 \;=\; 1 \;-\; \frac{SS_{res}}{SS_{tot}} \;=\; 1 \;-\; \frac{1{,}0659}{3{,}6875} \;=\; 0{,}7109$$

Hvis forklaringsgraden er tæt på 1 er det normalt et udtryk for et godt fit. I eksempel B2 så vi en forklaringsgrad på 0,99331. Andengradspolynomiet i dette eksempel ses da også tydeligt at tilnærme data bedre end tilfældet er i dette eksempel. En anden visuel måde at afgøre om data følger en bestemt type funktion, såsom et andengradspolynomium her, er at lave et *residualplot*. Her afbildes datapunkternes x-værdier henad 1. aksen og det tilhørende residual opad 2. aksen.



Residualplottet er ikke særlig interessant her, hvor der kun er fire datapunkter. Ofte har man rigtig mange datapunkter, og hvis punkterne i residualplottet så for eksempel ligger konsekvent under grafen i den ene ende og over grafen i den anden, kan det give anledning til at forkaste hypotesen om, at data følger den pågældende graf. Derimod vil det styrke hypotesen, hvis punkterne i residualplottet ser ud til at være spredt tilfældigt.

□

Bemærkning 4

Lad os opsummere til en mere generel situation end det i eksempel B3: Lad der være givet n datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Man ønsker at foretage et fit med en funktion f, som indeholder én eller flere frie *parametre*. Spørgsmålet er, hvilke værdier af parametrene, der giver det "bedste" fit. Lad $\overline{y} = (y_1 + y_2 + \ldots + y_n)/n$ være gennemsnittet af datapunkternes y-koordinater. Da er *kvadratsummen af residualer* givet ved:

(9) $$SS_{res} = \sum_{i=1}^{n} \left(y_i - p(x_i)\right)^2$$

Det er denne kvadratsum, som man ønsker at minimere med hensyn til de parametre, som indgår i funktionen f. *Den totale kvadratsum* er tilsvarende givet ved:

(10) $$SS_{tot} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

Endelig er *forklaringsgraden* R^2 givet ved:

(11) $$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Hvis forklaringsgraden er eksakt 1, betyder det, at vi har at gøre med et perfekt fit: Alle datapunkter ligger på grafen for f. Jo mindre SS_{res} er i forhold til SS_{tot} jo nærmere vil forklaringsgraden være på 1. Det sker typisk, når funktionen er et godt fit til data.

□

Øvelse 5

Benyt metoden fra eksempel B1 til at bestemme forskriften for det andengradspolynomium, hvis graf går igennem punkterne $(-2,9), (1;\,1{,}5)$ og $(4,3)$. Udfør først beregningerne manuelt og benyt derefter dit CAS-værktøj til at løse opgaven automatisk. Får du samme resultat?

Øvelse 6

Givet datapunkterne $(-5,8), (-2,-1), (2,1), (3,4)$ og $(4,7)$.

a) Benyt dit CAS-værktøj til at bestemme det bedste fit af formen $p(x) = a\cdot x^2 + b\cdot x + c$ til de fem datapunkter. Angiv a, b og c.

b) Bestem residualerne og lav et residualplot.

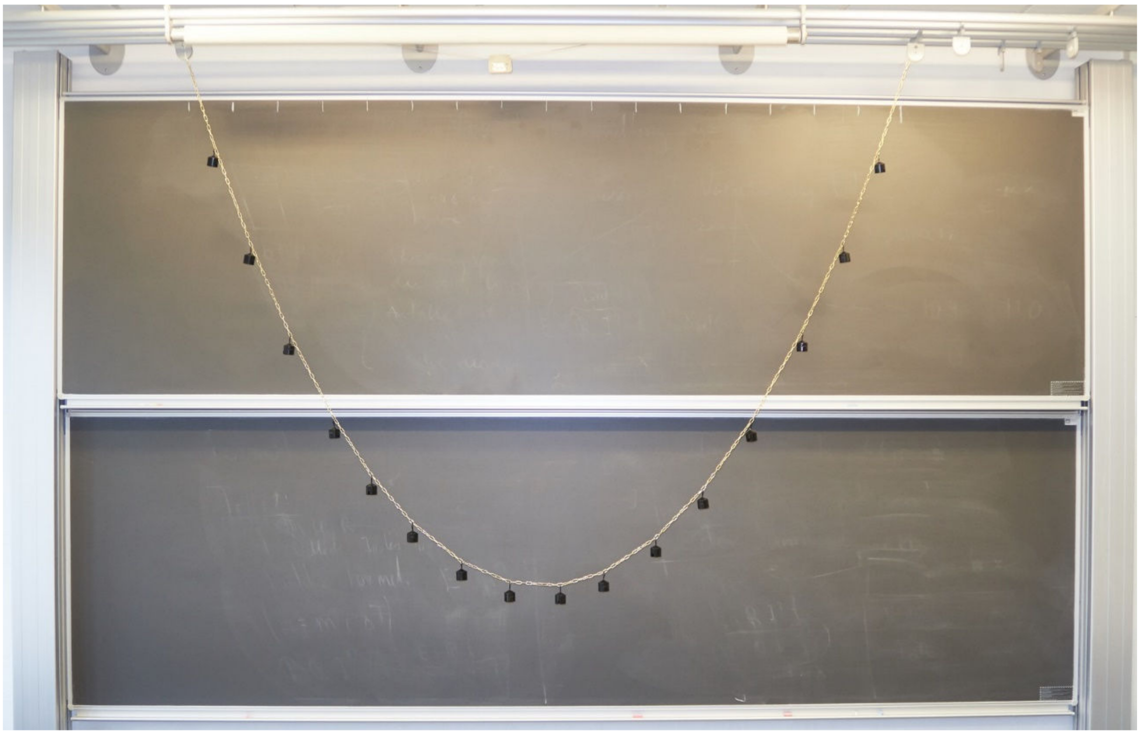
c) Hvilken forklaringsgrad giver CAS-værktøjet?

d) Bestem forklaringsgraden manuelt ved hjælp af formlen i (B11). Stemmer den overens med værdien fra c)?

I de følgende to øvelser skal vi udelukkende bruge et færdigt softwareprogram til at bestemme et kurvefit, uden at dykke ned i, hvordan dette kurvefit bliver beregnet. Et oplagt softwareprogram hertil er *Logger Pro*, da man kan "prikke en kurve ud på et foto".

Øvelse 7 (Hængebro-model)

Teorien forudsiger, at de kabler, som bærer vejbanen på en hængebro, former en parabel. Vi skal undersøge, om det er rigtigt ved at bruge programmet *Logger Pro*. Anskaf for det første en metalkæde fra et byggemarked. Kæden hænges op i to kroge. Som bekendt er vejbanen båret oppe af lodrethængende kabler, som er forbundet til det bærende kabel. Vi vil simulere denne belastning ved at hænge en række lodder op i forskellige kædeled. Lodderne skal alle have samme vægt, og de skal hænge nogenlunde med *samme vandrette afstand*, som antydet på figuren herunder. Tag et foto af situationen. Vær omhyggelig med at fotografere vandret samt vinkelret ind på det plan, som kæden hænger i.



a) Indsæt fotoet i Logger Pro via menuen *Indsæt* > *Billede* > *Billede med fotoanalyse*…. Ved at vælge med fotoanalyse, får man rådighed over en række værktøjer i højre side af billedet. Dem skal vi bruge i det følgende.

b) Indsæt begyndelsespunktet for koordinatsystemet (Origo) ved at prikke det ud på fotoet med det rette værktøj. Hvor man gør det i billedet er lidt ligegyldigt her.

c) Man skal sætte *skalaen*, dvs. fortælle programmet, hvor meget en bestemt afstand i billedet svarer til i virkelig afstand. Man vælger skala-værktøjet og trækker fx en lodret eller vandret linje ud, hvorefter man i en boks skriver, hvor langt det angivne stykke i virkeligheden er - i samme plan som kæden! Har du ikke umiddelbart mulighed for at vurdere dette, sæt da bare stykket til et eller andet. Det betyder ikke så meget her, eftersom en skaleret parabel stadig er en parabel.

d) Du er nu klar til at afsætte punkter langs kæden ved hjælp af punkt-værktøjet. Afsæt mange punkter langs kæden og gør det så præcist, som du kan. Kommer du til at afsæt et punkt lidt skævt, kan du bruge **Ctrl+Z** for at fortryde det sidst afsatte punkt.

e) Klik på kurvetilpasningsværktøjet. Vælg i det fremkomne vindue AX^2+BX+C og klik derefter på *Prøv tilpasning*. Programmet vil da forsøge at bestemme parametrene A, B og C i andengradspolynomiet, således at man opnår det bedste fit. Resultatet kan findes på en graf bag billedet. Du kan vælge, om billedet eller grafen skal ligge bagerst eller forrest ved at højreklikke på det forreste element og evt. vælge *Flyt bagest*.

f) Punktet *Korrelation* i den lille boks, der indeholder oplysninger om fittet, er kvadratroden af forklaringsgraden. Som bekendt er en forklaringsgrad tæt på 1 – og dermed også en Korrelation tæt på 1 – som oftest et udtryk for, at der er tale om et godt fit. Husk dog, at det især er en visuel inspektion, som bør afgøre, hvor godt andengradspolynomiet tilnærmer datapunkterne. Kan du godtage at de bærende kabler på en hængebro former en parabel?

Øvelse 8 (Kædelinje)

En *kædelinje* (Eng: *Catenary*) er den kurve, som en kæde danner, når den hænger frit ned fra to kroge. Teorien siger, at kurven er graf for følgende funktion:

(12) $$f(x) = a \cdot \cosh\left(\frac{x}{a}\right)$$

hvor $\cosh(x)$ er den funktion, der betegnes "cosinus hyperbolsk". Du skal nu udføre et forsøg ligesom det med hængebroen med henblik på at afgøre, om kæden virkelig hænger i den kurve, teorien forudsiger. Bemærk, at hvis du forsøger at foretage et fit med funktionen i (B12), hvor der kun er parameteren a, så skal du være meget omhyggelig med at udvælge origo. Dette er ikke ønskeligt og vil give væsentlig usikkerhed i forsøget. Bedre er det at definere en ny funktion, som har en graf, som er en parallelforskydning af grafen for f. Dermed behøver man ikke være præcis med at sætte origo nøjagtigt, men kan sætte det et vilkårligt sted på fotoet.

a) Benyt sætning 2.3 til at vise, at hvis man parallelforskyder grafen for f med vektoren (b,c), så får man grafen for følgende funktion:

(13) $$g(x) = a \cdot \cosh\left(\frac{x-b}{a}\right) + c$$

Denne funktion har nu de tre parametre a, b og c.

b) Gennemfør forsøget ligesom i øvelse B7, blot med den forskel, at der ikke skal hænges lodder på kæden. Foretag et fit med funktionen g. Kan dit forsøg og analyse bekræfte, at en frithængende kæde følger en kædelinje?