Lad ω være en amerikaner. Definer

$$\begin{aligned} F_i(\omega) &= \text{ værdien af faktoren } i\text{'te faktor for } \omega \\ D(\omega) &= \text{ dødsårsag for } \omega \\ L(\omega) &= \text{ dødsalder for } \omega \end{aligned}$$

I det kommende undertrykkes afhængigheden af ω.

Faktorsandsynligheder

Vi har blandt andet brug for størrelserne

$$P(F_i \in f), \text{ for forskellige mængder } f \in \mathcal{F}_i$$
$$P(F_{i_1} \in f_1, F_{i_2} \in f_2, \dots F_{i_n} \in f_n) \text{ for } f_i, i = 1, \dots, n$$

hvor f_i er mængder der giver mening for deres tilhørende faktor. I filerne i mappen Factor_frequencies ligger filer af typen.

i_1'te faktor	$\cdots$	i_n'te faktor	freq
f_1^0	$\cdots$	f_n^0	$P(F_{i_1} \in f_1^0, F_{i_2} \in f_2^0, \dots F_{i_n} \in f_n^0)$
f_1^0	$\cdots$	f_n^1	$P(F_{i_1} \in f_1^0, F_{i_2} \in f_2^0, \dots F_{i_n} \in f_n^1)$
$\vdots$	$\cdots$	$\vdots$	$\vdots$
f_1^0	$\cdots$	$f_n^{n_k}$	$P(F_{i_1} \in f_1^0, F_{i_2} \in f_2^0, \dots F_{i_n} \in f_n^{n_k})$
$\vdots$	$\cdots$	$\vdots$	$\vdots$
$f_1^{n_1}$	$\cdots$	$f_n^{n_k}$	$P(F_{i_1} \in f_1^{n_1}, F_{i_2} \in f_2^{n_2}, \dots F_{i_n} \in f_n^{n_k})$

Alternativt kan man skrive $F = (F_{i_1}, \dots, F_{i_n})$

Incidents

Dernæst har vi sandsynlighederne for at dø af en dødsårsag i løbet af et år.

$$p_d(l) = P(D = d, L \leq l \mid L > l - 1), d \in \mathcal{D}, l \in \mathbb{R}_+$$

hvor $\mathcal{D}$ er en mængde af alle dødsårsager i programmet. For beregning, kender vi $p_d(l)$ som stykvis konstant funktion på mængderne

$$l_1, l_2, \dots, l_{22} = [0, 1), [1, 5), [5, 10), \dots, [95, 100), [100, \infty)$$

Vi er interesserede i vektoren

$$p_d(l_1) \quad p_d(l_2) \quad \cdots \quad p_d(l_{22})$$

hvor vi, med den lidt misbrugte notation $p_d(l_i)$, mener $p_d(l)$ for et $l \in l_i$. Disse estimeres med

$$p_d(l_i) \leftarrow \frac{\text{antal amerikanere døde af } d \text{ i aldersgruppen } l_i \text{ i år } Y_1}{\text{antal amerikanere i aldersgruppen } l_i \text{ i år } Y_2} =: \frac{a_{di}}{a_i}$$

(Lige nu har vi $Y_1 = 2014, Y_2 = 2013$). Filerne af formen *ICDcode.txt* indikerer et d med deres titel og indholdet er

$$a_{d1} \quad a_{d2} \quad \cdots \quad a_{d22}$$

Og filen *population.txt* indeholder

$$a_1 \quad a_2 \quad \cdots \quad a_{22}$$

Risk ratios

Lad $F_1, \ldots, F_k$ være nogle faktorer. Risk ratios i dette program fortolkes som

$$\mathrm{RR}_d(f) := \frac{P\big(D = d, L \leq l \mid L > l-1, (F_1, \ldots, F_k) \in f\big)}{P\big(D = d, L \leq l \mid L > l-1, (F_1, \ldots, F_k) \in f_0\big)}$$

Egentlig skulle der et f_0 på i notationen for $\mathrm{RR}_d(f)$, for at indikere at det er riskratio med hensyn til baselinen f_0. Det er en antagelse, at riskratioen ikke afhænger af l. Risk ratioerne er kendt for en mængde af faktorinddelinger $f \in \mathcal{F}$. Vi kræver mere eller mindre at $\mathcal{F}$ kan skrives på formen

$$\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \mathcal{F}_k$$

hvor

$$\mathcal{F}_i = \{f_1^0, f_1^1, \ldots, f_1^{n_i}\}$$

og så er de kodet ved hjælp af

$$\begin{array}{rcrl}
i_1\text{'te faktor} & \cdots & i_k\text{'te faktor} & \text{RR} \\
f_1^0 & \cdots & f_k^0 & RR_d((f_1^0, \ldots, f_k^0)) \\
f_1^0 & \cdots & f_k^1 & RR_d((f_1^0, \ldots, f_k^1)) \\
\vdots & \cdots & \vdots & \vdots \\
f_1^0 & \cdots & f_k^{n_k} & RR_d((f_1^0, \ldots, f_k^{n_k})) \\
\vdots & \cdots & \vdots & \vdots \\
f_1^{n_1} & \cdots & f_k^{n_k} & RR_d((f_1^{n_1}, \ldots, f_k^{n_k}))
\end{array}$$

0.1 Udregning

I første omgang vil vi gerne udregne

$$P\big(D = d, L \leq l \mid L > l-1, F = f\big), \text{ for } f \in \mathcal{F} \tag{1}$$

hvor F er en vektor af faktorer og $\mathcal{F}$ er den endelige mængder af faktor-sammensætninger. (1) kan senere(i javascript-delen) bruges til at udregne

$$P\big(D = d, L \leq l \mid L > l-1, F = f_{\text{personal}}\big) \tag{2}$$

hvor f_{personal} ikke (nødvendigvis) ligger i $\mathcal{F}$. Det gøres ved

$$(1) = \frac{\mathrm{RR}_d(f)}{\sum_{f\in\mathcal{F}} RR_d(f)\cdot P(F=f)} p_d(l) \tag{3}$$

Der er dog nogle forhindringer før vi bare kan stoppe tallene fra filerne ind i (3).

- Vi ikke har nok information til at kende $P(F = f), f \in \mathcal{F}$ 100%.
- Vi har flere riskratio filer for samme d.
- Hvordan man skal tage højde for alders specifikke riskratios og alders specifikke $P(F = f)$'er

Lad os tage dem i voksende sværhedsgrad

Flere risk ratio filer

Hvis vi der er to riskratiofiler baseret på to vektorer af faktorer F^1, F^2 og to tilhørende krydsmængder, $\mathcal{F}^1$ og $\mathcal{F}^2$, så ville den mest rigtige måde at kombinere dem på være

$$\begin{aligned} &P\big(D = d, L \leq l \mid L > l-1, (F^1, F^2) = (f^1, f^2)\big) \\ &\quad= \frac{g(\mathrm{RR}_d(f^1), \mathrm{RR}_d(f^2))}{\sum_{f^1,f^2\in\mathcal{F}^1\times\mathcal{F}^2} g(\mathrm{RR}_d(f^1), \mathrm{RR}_d(f^2))\cdot P((F^1,F^2) = (f^1,f^2))} p_d(l) \end{aligned} \tag{4}$$

hvor g er en passende interaktionsfunktion. Hvis F^1 og F^2 er uafhængige og $g(x, y) = x \cdot y$, kan man dog skrive det som

$$\begin{aligned} &P\big(D = d, L \leq l \mid L > l-1, (F^1, F^2) = (f^1, f^2)\big) \\ &\quad= \frac{\mathrm{RR}_d(f^1)}{\sum_{f^1\in\mathcal{F}^1} RR_d(f^1)\cdot P(F^1=f^1)} \frac{\mathrm{RR}_d(f^2)}{\sum_{f^2\in\mathcal{F}^2} RR_d(f^2)\cdot P(F^2=f^2)} p_d(l) \end{aligned} \tag{5}$$

Fordelen ved (5) er, at man kan lægge et tallene

$$\frac{\mathrm{RR}_d(f^1)}{\sum_{f^1\in\mathcal{F}^1} RR_d(f^1)\cdot P(F^1=f^1)}, f^1\in\mathcal{F}^1$$

i en fil og tallene

$$\frac{\mathrm{RR}_d(f^2)}{\sum_{f^2\in\mathcal{F}^2} RR_d(f^2)\cdot P(F^2=f^2)}, f^2\in\mathcal{F}^2$$

i en anden fil og tallene

$$p_d(l_i),\ i=1,\dots,22$$

i en tredje fil. Selvom betingelserne for at bruge (5) ikke er helt er opfyldt, kan det måske alligevel være en god ide at bruge den.

Vi kender ikke $P(F=f), f\in\mathcal{F}$

Her tænkes $\mathcal{F}$ som den mængde hvor vi kender $\mathrm{RR}_d(f)$ hvis og kun hvis $f\in\mathcal{F}$.

Der er flere slags udfordringer her.

1. Det hænder at vi kun kender

   $$P(F=f), f\in\mathcal{F}'$$

   hvor $\mathcal{F}'\not\supseteq\mathcal{F}$.

2. Vi har $\mathcal{F}=\mathcal{F}_1\times\mathcal{F}_2$ og $F=(F_1,F_2)$, hvor F_1 og F_2 er hver deres faktor. Her hænder det at vi kun kender

   $$P(F_i=f_i), f_i\in\mathcal{F}_i$$

   for $i=1,2$ og altså ingenting om den simultane fordeling af (F_1,F_2)

3. Vi har $\mathcal{F}=\mathcal{F}_1\times\mathcal{F}_2\times F_3$ og $F=(F_1,F_2,F_3)$. Det hænder, at vi kun kender

   $$P((F_1,F_2)=(f_1,f_2)),\ f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2$$
   $$P((F_1,F_3)=(f_1,f_3)),\ f_1\in\mathcal{F}_1, f_3\in\mathcal{F}_3$$

4. Vi har $\mathcal{F} = \mathcal{F}_1$ og $F = F_1$. Det hænder, at vi kender
$$P((F_1, F_2) = (f_1, f_2)),\ f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2$$
men ikke (umiddelbart)
$$P(F_1 = f_1),\ f_1 \in \mathcal{F}_1$$
så vi så om sige har for meget

5. Vi har $\mathcal{F} = \mathcal{F}_1$ og $F = F_1$. Det hænder, at vi slet ikke kender noget til
$$P(F_1 \in f)$$

Problemerne findes også i flere flere dimensioner og med kombinationer.

En løsning af problem 4

Vi laver en funktion som marginaliserer, dvs
$$P(F_1 = f_1) = \sum_{f_2 \in \mathcal{F}_2} P((F_1, F_2) = (f_1, f_2))$$

En løsning af problem 1

Hvis vi laver en funktion, som laver transformationen
$$\begin{aligned} w : \{&P(F = f_1), P(F = f_2), \ldots, P(F = f_n)\} \\ &\mapsto \{P(F = f'_1), P(F = f'_2), \ldots, P(F = f'_k)\} \end{aligned}$$
kan vi løse det første problem. Hvis vi antager
$$\bigcup_{i=1}^{n} f_i \subseteq \bigcup_{j=1}^{k} f'_j \tag{6}$$
kan man lave løsningen
$$\begin{aligned} P(F = f'_j) =& w\big(P(F = f_1), P(F = f_2), \ldots, P(F = f_n)\big) \\ =& \sum_{i=1}^{n} \frac{P(F = f_i) \cdot |f_i \cap f'_j|}{|f_i|} \end{aligned}$$
hvor $|\cdot|$ repræsenterer et mål. Det vil dog nok altid være muligt at bruge Lebesguemålet eller tællemålet og nogle gange kan man måske være nødt til at bruge en mikstur af de to. Det ses foreksempel ved rygning, hvor der er en kategori, der hedder 0 cigaretter. Betingelsen (6) er nødvendig for at $P(F = f'_j), j = 1, \ldots, k$ summer til 1(det er nemlig antaget at $P(F = f_i), i = 1, \ldots, n$ summer til 1).

En løsning af problem 2,3 og 5

For at løse problem 2,3 og 5 kan man bruge tilpasning af marginaler. Man har den ønskede fordeling

$$P(F=f), f \in \mathcal{F}$$

hvor $F=(F_1,\dots,F_k)$ og $\mathcal{F}=\mathcal{F}_1\times\cdots\times\mathcal{F}_k$, og man kender

$$\begin{gathered} P(F^1=f^1), f^1 \in \mathcal{F}^1 \\ \vdots \\ P(F^m=f^m), f^m \in \mathcal{F}^m \end{gathered}$$

hvor $\mathcal{F}^l=\mathcal{F}_{i^l_1}\times\mathcal{F}_{i^l_2}\times\cdots\times\mathcal{F}_{i^l_{r_l}}$. Man starter med en standard uniformfordeling

$$p(f)=\frac{1}{\#\mathcal{F}}, \ f\in\mathcal{F}$$

som estimater for $P(F=f)$. Dernæst *tilpasser man med marginalen* $\mathcal{F}^1$

$$p^{ny}(f)=p(f)\frac{P(F^1=f^1(f))}{\sum_{f':f^1(f')=f^1(f)}p(f')}$$

og derefter med marginalen $\mathcal{F}^2,\mathcal{F}^3$ og så videre (indtil man når til hvad?).

Forskellige credibilities

Vi har mængder, $f\in\mathcal{F}$, for hvilke vi vil finde $P(F=f)$. F kan her skrives $(F_i)_{i\in I}$, hvor I er en mængde i $\{1,\dots,n\}$. Vi kender da 'binnede' fordelinger af $(F_i)_{i\in I_j}$ for $j=1,\dots,k$, hvor I_j også er mængder i $\{1,\dots,n\}$. Credibility-scoren er en funktion $c:\{I_j\}_{j=1,\dots,k}\to\mathbb{R}_+$. De binnede fordelinger, der indgår i konstruktionen $P(F=f)$ er

$$\Big\{(F_i)_{i\in I_j} \mid I_j\cap I\neq\emptyset \wedge \big(\nexists k: I\cap I_j\subseteq I\cap I_k \wedge c(I_k)>c(I_j)\big)\Big\}$$

•

Aldersspecifikke $P(F=f)$'er eller riskratioer

I princippet burde alle ovenstående udregninger laves separat for alle aldersgrupper. Det gøres også, og der er nogle genveje. Definer A til at være den stokastiske variabel der angiver alderen på personen. De nye faktorsandsynlighedsfiler har specificerende kolonner ved

i_1'te faktor	$\cdots$	i_n'te faktor	aldersgr.
f_1^0	$\cdots$	f_n^0	A_1
f_1^0	$\cdots$	f_n^1	A_1
$\vdots$	$\cdots$	$\vdots$	
f_1^0	$\cdots$	$f_n^{n_k}$	A_1
$\vdots$	$\cdots$	$\vdots$	
$f_1^{n_1}$	$\cdots$	$f_n^{n_k}$	A_h

og freq kolonnen indeholder

$$
\begin{array}{l}
\text{freq} \\
P(F_{i_1} \in f_1^0, F_{i_2} \in f_2^0, \dots F_{i_n} \in f_n^0 \mid A \in A_1) \\
P(F_{i_1} \in f_1^0, F_{i_2} \in f_2^0, \dots F_{i_n} \in f_n^1 \mid A \in A_1) \\
\vdots \\
P(F_{i_1} \in f_1^0, F_{i_2} \in f_2^0, \dots F_{i_n} \in f_n^{n_k} \mid A \in A_1) \\
\vdots \\
P(F_{i_1} \in f_1^{n_1}, F_{i_2} \in f_2^{n_2}, \dots F_{i_n} \in f_n^{n_k} \mid A \in A_h)
\end{array}
$$

Det vil sige, at freq-kolonnen skal summe til h. Når man konstruerer $P(F = f)$, bør man konstruere

0.2 Javascript delen

Til hver cause hører noget data på formen

$$d, \big(p_d(l_i)\big)_{i=1,\dots,22} \text{ dataframe, risk ratio data}_d$$

hvor d bare er en string, $\big(p_d(l_i)\big)_{i=1,\dots,22}$ er en data frame udskrevet som liste og risk ratio data'et har formen

$$(\text{Risk ratio datagruppe})_1^d, (\text{Risk ratio datagruppe})_2^d, \dots, (\text{Risk ratio datagruppe})_{n_d}^d$$

En risk ratio datagruppe - indekseret ved (j, d) - består af

$$(\text{norm}^{j,d}(l_i))_{i=1,\dots,22}, (\text{risk ratio dataframe}_i^{j,d})_{i=1,\dots,k_{j,d}}, g_{j,d}$$

hvor $(\mathrm{norm}^j(l_i))_{i=1,\dots,22}$ er en liste af 22 tal som normaliserer for hver af de 22 aldersgrupper, list of risk ratio dataframes er en liste af risk ratio dataframes udskrevet som lister og $g_{j,d}$ er en string, som siger hvilken interaktion der mellem (j,d) dataframesne. En risk ratio dataframe for en faktor $F^{i,j,d}$ har formen

$$(r^{i,j,d}(f))_{f\in\mathcal{F}^{i,j,d}}$$

hvor $\mathcal{F}^{i,j,d}$ er en endelig mængde af faktor levels. Dette r er blot en diskretisering af den underliggende riskfaktor funktion

$$R^{i,j,d}(\psi), \psi \in \Psi^{i,j,d}$$

hvor $\Psi^{i,j,d}$ er en mængde af alle tænkelige værdier af faktoren $F^{i,j,d}$, og derfor kan den være uendelig. Vi vil lave en *polating* funktion, pol, til at evaluere funktionen

$$R^{i,j,d}(\psi) = \mathrm{pol}((r^{i,j,d}(f))_{f\in\mathcal{F}^{i,j,d}}, \psi) \tag{7}$$

Lad nu ψ være alle en persons faktorværdier, og lad $\phi^{i,j,d}$ være vektoren af faktorer der er relevante for den (i,j,d)'te risk ratio fil, dvs. en delvektor af ψ. Lad a være en vilkårlig alder og lad $l(a)$ være den af de 22 alderskategorier, som a falder i. Så defineres

$$P_d(a,\psi) = P_d(a) \cdot \prod_{j=1}^{n_d} \frac{g_{j,d}(R^{1,j,d}(\psi^{1,j,d}), \dots, R^{k_{j,d},j,d}(\psi^{k_{j,d},j,d}))}{\mathrm{norm}^{j,d}(l(a))} \tag{8}$$

hvor $P_d(a)$ er 'afdiskretiseringen' af $p_d(l)$. Det kunne måske defineres som

$$P_d(a) = \max\big(1, \mathrm{pol}(\big(p_d(l_i)\big)_{i=1,\dots,22}, a)\big)$$

Størrelsen $P_d(a,\psi)$ tænkes at være det samme som (2), dvs.

$$P_d(a,\psi) = P\big(D = d, L \le a \mid L > a-1, F = \psi\big) \tag{9}$$

Kombinationer af $P_d(a)$'er

Vi definerer nu

$$p(a,\psi) = \sum_{d\in\mathcal{D}} p_d(a,\psi)$$

hvor $\mathcal{D}$ er mængden af alle dødsårsager. Hvis (9) faktisk gælder, er

$$p(a,\psi) = P(L \le a \mid L > a-1, F = \psi\big)$$

Liste over mest interessante kombinationer

Lad nu $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ være en vilkårlig delmængde af dødsårsagerne. Alle de følgende størrelser kan varieres ved at tilføje $d \in \tilde{\mathcal{D}}$ og/eller $L \in [a, b]$ til betingningen

- Forventet levealder

$$E[L \mid \psi] = \sum_{a=0}^{\infty} a \cdot p(a, \psi) \prod_{b=0}^{a-1} (1 - p(b, \psi))$$

- Forventet antal år mistet til sygdom d

$$E[L \mid \psi, D \in \mathcal{D} \setminus \{d\}] - E[L \mid \psi]$$

- Sandsynligheden for at dø af d

$$P(D = d \mid \psi) = \sum_{a=0}^{\infty} p_d(a, \psi) \prod_{b=0}^{a-1} (1 - p(b, \psi))$$

- Ens overlevelseskurve

$$(P(L \geq a \mid \psi))_{a \in \mathbb{N}_0} = \left(\prod_{b=0}^{a-1} (1 - p(b, \psi)) \right)_{a \in \mathbb{N}_0}$$

Forklare døden ved hjælp af ens faktorer

Vi vil nu også lave en optimal ψ-værdi som man kan måle brugerens ψ op imod. Dette er dog en udfordring, fordi en faktor som forårsager en sygdom kan hæmme fremkomsten af en anden sygdom. Det gælder foreksempel rygning, lungekræft og Parkinson's. Vi definerer derfor ψ_0 så $\psi_0^{i,j,d}$ minimerer $R^{i,j,d}(\cdot)$. Det vil sige at i en lungekræftssammenhæng har ψ_0 rygning på 0, mens i en Parkinson's sammenhæng har ψ_0 rygning på 2 (cig/day). Antag af ψ er gjort op af n faktorer. Vi definerer nu for et $J \subseteq \{1, \ldots, n\}$ faktorværdivektoren ψ_{0,F_J}. For en indang i er

$$(\psi_{0,F_J})_i = (\psi_{0,\{F_j\}_{j \in J}})_i = \begin{cases} (\psi)_i & \text{hvis } i \in J \\ \left(\arg\min_{\tilde{\psi}: \tilde{\psi}_j = (\psi)_j \text{ for } j \in J} R(\tilde{\psi}) \right)_i & \text{ellers} \end{cases} \tag{10}$$

Så vi har $\psi_\emptyset = \psi_0$ og $\psi_{0,F_{\{1,\ldots,n\}}} = \psi$. Betragt nu følgende dekomposition

$$\begin{aligned}
P\big(D = d \mid F = \psi\big) &= \sum_{a=0}^{\infty} p_d(a,\psi) \prod_{b=0}^{a-1} (1 - p(b,\psi)) \\
&= \sum_{a=0}^{\infty} (p_d(a,\psi) - p_d(a,\psi_0)) \prod_{b=0}^{a-1} (1 - p(b,\psi)) \\
&\quad + \sum_{a=0}^{\infty} p_d(a,\psi_0) \prod_{b=0}^{a-1} (1 - p(b,\psi))
\end{aligned}$$

Definitionen af ψ_0 gør begge de to led positive. Vi fortolker det første led som sandsynligheden for død på grund af ens faktorværdier, mens det andet led er *death by chance/age/destiny*. Bemærk at andet led ikke er identisk med $P(D = d \mid F = \psi_{0,F})$ fordi parameteren ψ stadig indgår i leddet. Vi er nu interesserede i at dekomponere

$$p_d(a,\psi) - p_d(a,\psi_0)$$

for da kan vi dekompenere $P(D = d \mid F = \psi)$ i flere led. Fra nu af holder vi a og d fast og ser på ovenstående størrelse og definerer $U_* = p_d(a,\psi) - p_d(a,\psi_0)$.

Hvis F bare er en enkelt faktor dvs., $F = (F_1)$, er den fuldt dekomponeret. Men hvis F er en vektor $F = (F_1, F_2, \ldots, F_n)$ ville vi ideelt have positive tal

$$S(F_J) = S(\{F_j\}_{j \in J}), J \subseteq \{1, \ldots, n\}$$

sådan at

$$U_* = \sum_{J \subseteq \{1,\ldots,N\}, J \neq \emptyset} S(F_J) \tag{11}$$

og hver af $S(F_J)$'erne havde en pæn fortolkning; gerne sådan at

$$p_d(a,\psi_{0,F_I}) - p_d(a,\psi_0) = \sum_{J \subseteq I, J \neq \emptyset} S(F_J) \tag{12}$$

for alle $I \subseteq \{1, \ldots, n\}$. Definer for simplicitet $U(F_I) = p_d(a,\psi_{0,F_I}) - p_d(a,\psi_0)$.

Overvejelse

Betragt de simple riskratiointeraktioner

For at undgå komplicerede subskripts definerer vi $S(F_{i_1}, \ldots, F_{i_k}) = S(F_{\{i_1,\ldots,i_k\}})$. Ved at bruge (12) på Tabel 0.1 får man

	f_1^1	f_1^2
f_2^1	1.0	1.1
f_2^2	1.1	2.1

Tabel 0.1: Her er der en positiv interaktion mellem F_1 og F_2

	f_1^1	f_1^2
f_2^1	1.0	2.0
f_2^2	2.0	2.1

Tabel 0.2: Her er der en negativ interaktion mellem F_1 og F_2

$$\begin{aligned} \tilde{S}(F_1) &= 0.1 \\ \tilde{S}(F_2) &= 0.1 \\ \overset{(12)}{\Rightarrow} \tilde{S}(F_1, F_2) &= 0.9 \end{aligned}$$

hvor $\tilde{S} = S/p_d(a, \psi_0)$. (12) virker ikke lige så god på Tabel 0.2 for da er

$$\begin{aligned} S(F_1) &= 1.0 \\ S(F_2) &= 1.0 \\ \overset{(12)}{\Rightarrow} S(F_1, F_2) &= -0.9 \end{aligned}$$

Og den negative værdi besværliggører fortolkningen, så det er måske problematisk at bruge (12).

Løsningsforslag

Man kan godt få lidt mening ud af de negative værdi der kan forekomme under systemet i (12). Systemet medfører at

$$S(F_I) = U(F_I) - \sum_{J \subseteq I, J \neq \emptyset, J \neq I} S(F_J) \tag{13}$$

hvilket ikke ikke leder til en cirkeldefinition, fordi alle led i summen har færre elementer end antallet af elementer i I. For at få mening ud af S'erne foreslås følgende fortolkninger/omregninger, som inddeles i forskellige kompleksitet

(i) 1. ordensfortolkning, $S^{(1)}$. Dette er den simpleste fortolkning for her undgår vi at have interaktioner mellem faktorårsagerne. Det kan opnås ved at normalisere $S(F_i)$'erne

$$S(F_i) = U(F_i)$$
$$S^{(1)}(F_i) = \frac{S(F_i)}{\sum_{j=1}^{n} S(F_j)} \cdot U_*$$

for $i = 1, \ldots, n$. Man kunne sige $S^{(1)}(F_J) = 0$ for $|J| > 1$. Faktoren $\frac{U_*}{\sum_{j=1}^{n} S(F_j)}$ ganges på for at $S^{(1)}(F_i)$'erne fungerer som en fuld dekomposition.

(ii) Næst kigger vi på interaktioner mellem par af faktorårsagerne, der udregnes som

$$S(F_i, F_j) = U(F_i, F_j) - S(F_i) - S(F_j).$$

Problemet er her, at det er muligt at $S(F_i, F_j) < 0$. Vi ved dog at

$$U(F_1, F_j) > \max\big[S(F_i), S(F_j)\big] \tag{14}$$

og dermed at hvis $S(F_i, F_j) < 0$ så er

$$|S(F_i, F_j)| < S(F_i) \tag{15}$$

Et negativt $S(F_i, F_j)$ betyder at niveauerne af F_i og F_j ikke forstærker hinanden. Tværtimod, de dækker over de samme sygdomstilfælde. Antag fx

$$D = \{\text{drukne i havnen}\}$$
$$F_1 = \{\text{genstande drukket om ugen}\}$$
$$F_2 = \{\text{max genstande drukket på en enkelt dag om ugen}\}$$

Folk der drikker meget har større sandsynlighed for at falde i havnen og drukne fordi de ikke kan komme op. Derfor forventer vi at højere værdier af F_1 og F_2 øger sandsynligheden for D. Vi forventer at man har en endnu højere risiko for at drukne i havnen hvis man gør begge dele, men vi forventer ikke at den er meget højere end hvis man kun gjorde en af tingene. Vi kan derfor godt være i situationen som i Tabel 0.2. Man kan forstå det som at nogle af de dødsfald som ville være sket på grund af F_1, stadig ville være sket på grund af F_2 - og omvendt.

Derfor foreslås en ny slags faktorårsag; et ELLER led, noteret med $T(F_1, F_2)$. For $n = 2$ ville vi definere det således:

$$T(F_1, F_2) = \begin{cases} |S(F_1, F_2)| & \text{hvis } S(F_1, F_2) < 0 \\ 0 & \text{ellers} \end{cases} \tag{16}$$

$$\tilde{S}(F_1, F_2) = \begin{cases} 0 & \text{hvis } S(F_1, F_2) < 0 \\ S(F_1, F_2) & \text{ellers} \end{cases} \tag{17}$$

$$\tilde{S}(F_i) = S(F_i) - T(F_i, F_j), \qquad \text{for } i, j = 1, 2, j \neq i \tag{18}$$

$$C = \frac{U_*}{\sum_{i \neq j}^{n} \tilde{S}(F_i, F_j) + T(F_i, F_j) + \sum_i \tilde{S}(F_i)} \tag{19}$$

$$S^{(2)}(F_i) = C\tilde{S}(F_i) \tag{20}$$

$$S^{(2)}(F_i, F_j) = C\tilde{S}(F_i, F_j) \tag{21}$$

$$T^{(2)}(F_i, F_j) = C\tilde{T}(F_i, F_j). \tag{22}$$

Men der er et problem; hvis $n > 2$ er der mange forskellige par af faktorer og så ved man ikke hvilket j man skal vælge i (18). For at undgå de problemer kan vi udvide ELLER-leddene til at inkludere så mange led at det ikke bliver et problem. Dertil definerer vi for hvert $i = 1, \dots, n$ mængden af inhiberende faktorer af anden orden

$$\begin{aligned} \mathcal{H}_i^2 = \Big\{ & j \in \{1, \dots, n\} : \exists k, i_1, \dots, i_k \in \{1, \dots, n\} \text{ hvor} \\ & S(F_j, F_{i_1}) < 0, S(F_{i_1}, F_{i_2}) < 0, \dots, \\ & S(F_{i_{k-1}}, F_{i_k}), < 0, S(F_{i_k}, i) < 0 \Big\}. \end{aligned}$$

Alle S'er mellem faktorer inden for $\mathcal{H}_i^2$ og faktorer uden for $\mathcal{H}_i^2$, er positive, hvilket vil sige at der er ingen ELLER-led at tage højde for. Definer nu $V(F_J) = U(F_J) - \sum_{j \in J} S(F_j)$. Man derfor vælge faktorårsagsstørrelserne som

$$T(F_{\mathcal{H}_i^2}) = \begin{cases} |V(F_{\mathcal{H}_i^2})| & \text{hvis } V(F_{\mathcal{H}_i^2}) < 0, |\mathcal{H}_i^2| > 1 \\ 0 & \text{ellers} \end{cases} \tag{23}$$

$$\tilde{S}(F_i, F_j) = \begin{cases} 0 & \text{hvis } i \in \mathcal{H}_j^2 \\ S(F_i, F_j) & \text{ellers} \end{cases} \tag{24}$$

$$\tilde{S}(F_{\mathcal{H}_i^2}) = \begin{cases} V(F_{\mathcal{H}_i^2}) & \text{hvis } V(F_{\mathcal{H}_i^2}) > 0, |\mathcal{H}_i^2| > 2 \\ 0 & \text{ellers} \end{cases} \tag{25}$$

Størrelserne $T(F_{\mathcal{H}_i^2})$ fanger nu den evt. negative interaktion der er mellem grupper af faktorer. Ud fra definitionen af $\mathcal{H}_i^2$ er vi dog ikke sikre på at $V(F_{\mathcal{H}_i^2}) < 0$, og i det tilfælde tager vi den med som en positiv interaktion i $\tilde{S}$. Da vi tog højde for den negative interaktion i (18) havde vi uligheden i (15) til at sikre os at $\tilde{S}(F_i)$ blev positiv. Det har vi ikke længere, hvorfor det foreslås at vi gør det følgende:

a. Hvis

$$T(F_{\mathcal{H}_i^2}) \cdot \frac{2}{|\mathcal{H}_i^2|} \leq \min_{j \in \mathcal{H}_i^2} S(F_i) \tag{26}$$

sættes

$$\tilde{S}(F_i) = S(F_i) - T(F_{\mathcal{H}_i^2}) \cdot \frac{2}{|\mathcal{H}_i^2|} \tag{27}$$

Det vil sige at den negative interaktion mellem variablene $F_{\mathcal{H}_i^2}$ ikke er større end at den kan repræsenteres med en bar der kan tages uniformt fra alle faktorer i $F_{\mathcal{H}_i^2}$.

b. Hvis

$$T(F_{\mathcal{H}_i^2}) \cdot \frac{2}{|\mathcal{H}_i^2|} > \min_{j \in \mathcal{H}_i^2} S(F_i) \text{ og} \tag{28}$$

$$2 \cdot T(F_{\mathcal{H}_i^2}) \leq \sum_{j \in \mathcal{H}_i^2} S(F_j) \tag{29}$$

definerer vi flere, nye T-led der kan "tages" uniformt fra deres marginale $S(F_i)$'er. Lad $S^{(k)}$ være den k'te mindste værdi fra mængden $\{S(F_i)\}_{i \in \mathcal{H}_i^2}$ og lad $h^{(k)}$ være det tilsvarende faktorindeks. Dvs. $S(F_{h^{(k)}}) = S^{(k)}$. Så defineres

$$\tilde{T}(F_{\mathcal{H}_i^2}) = S^{(1)} \cdot \frac{|\mathcal{H}_i^2|}{2} \tag{30}$$

Derefter defineres

$$\tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) = \left(S^{(k+1)} - S^{(k)}\right) \frac{|\mathcal{H}_i^2| - k}{2} \tag{31}$$

for $k = 1, \dots, k_0$, hvor $k_0 \equiv k_0(i)$ er det største tal der ville opfylde denne ulighed:

$$\sum_{k=1}^{k_0} \tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) < T(F_{\mathcal{H}_i^2}) \tag{32}$$

Til sidst defineres

$$\tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k_0+1)}\}}) = T(F_{\mathcal{H}_i^2}) - \sum_{k=1}^{k_0} \tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) \tag{33}$$

Ideen er, at vi nu har dekomponeret $T(F_{\mathcal{H}_i^2})$ i flere mindre ELLER-led. På grund (35) ved vi at $l_0 < k$. Det kan ske at $|\mathcal{H}_i^2 \setminus \{h^{(1)}, \dots h^{(l)}\}| = 1$ i hvilket tilfælde det sidste led ikke er et rigtigt ELLER-led, men blot et normalt led.

c. Hvis

$$T(F_{\mathcal{H}_i^2}) \cdot \frac{2}{|\mathcal{H}_i^2|} > \min_{j \in \mathcal{H}_i^2} S(F_i) \text{ og} \tag{34}$$

$$2 \cdot T(F_{\mathcal{H}_i^2}) > \sum_{j \in \mathcal{H}_i^2} S(F_j) \tag{35}$$

kan vi ikke længere repræsentere den fulde negative interaktion med ELLER-led. En oplagt måde at imødekomme det, er at ændre k_0 til at blive udregnet fra (32) til at være det største tal der opfylder

$$\sum_{k=1}^{k_0} \tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) < \frac{\sum_{j \in \mathcal{H}_i^2} S(F_j)}{2} \tag{36}$$

og dermed sætte

$$\tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k_0+1)}\}}) = \frac{\sum_{j \in \mathcal{H}_i^2} S(F_j)}{2} - \sum_{k=1}^{k_0} \tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) \tag{37}$$

Tilfældet b. kan ses som en generalisering af a. og c. er en generalisering af b. Derfor beskriver jeg kun normaliseringen for c.-tilfældet (hvor $\frac{\sum_{j \in \mathcal{H}_i^2} S(F_j)}{2}$ erstattes af $\min\left(\frac{\sum_{j \in \mathcal{H}_i^2} S(F_j)}{2}, T(F_{\mathcal{H}_i^2})\right)$ i (36) og (37)).

Normaliseringen bør være

$$\tilde{S}(F_i) = S(F_i) - \sum_{\substack{k \in \{0,\dots,k_0+1\}: \\ i \in \mathcal{H}_i^2 \setminus \{h^{(1)},\dots,h^{(k)}\}}} \tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots,h^{(k)}\}}) \frac{2}{|\mathcal{H}_i^2| - k} \tag{38}$$

$$C = \frac{U_*}{\begin{pmatrix} \sum_{i \neq j}^{n} \tilde{S}(F_i, F_j) + \sum_i \tilde{S}(F_i) + \\ \sum_{\{H_i^2\}_{i=1,\dots,n}} \Big(\tilde{T}(F_{\mathcal{H}_i^2}) + \tilde{S}(F_{\mathcal{H}_i^2}) + \\ \sum_{k=1}^{k_0(i)} \tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) \Big) \end{pmatrix}} \tag{39}$$

$$S^{(2)}(F_i) = C\tilde{S}(F_i) \tag{40}$$

$$S^{(2)}(F_i, F_j) = C\tilde{S}(F_i, F_j) \tag{41}$$

$$T^{(2)}(F_{\mathcal{H}_i^2}) = C\tilde{T}(F_{\mathcal{H}_i^2}) \tag{42}$$

$$S^{(2)}(F_{\mathcal{H}_i^2}) = C\tilde{T}(F_{\mathcal{H}_i^2}), \quad |\mathcal{H}_i^2| > 2 \tag{43}$$

$$\tag{44}$$

og

$$T^{(2)}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) = C\tilde{T}(F_{\mathcal{H}_i^2 \setminus \{h^{(1)},\dots h^{(k)}\}}) \tag{45}$$

for $k = 1, \dots, k_0(i)$. Se evt. nedenstående eksempel

Eksempel

Betragt følgende Risk-ratio tabel for en sygdom d og en alder A:

	$F_1 = a$		$F_1 = b$	
	$F_2 = a$	$F_2 = b$	$F_2 = a$	$F_2 = b$
$F_3 = a$	1.0	2.0	1.8	1.8
$F_3 = b$	1.5	1.9	1.9	2.1

Antag $\psi = (b, b, b)$. Vi har tydeligvis $\psi_0 = (a, a, a)$. Vi vil nu dekomponere

$$p_d(A, \psi) - p_d(A, \psi_0) \tag{46}$$

som er det samme som at dekomponere

$$\frac{p_d(A, \psi) - p_d(A, \psi_0)}{p_d(A, \psi_0)} = R(\psi) - R(\psi_0) \tag{47}$$

hvor R er ovenstående risk-ratio-tabel. Fremover antager vi derfor uden tab af generalitet at sandsynlighederne S, T, U, V, osv. er riskratios. Fra (13) får vi

$$\begin{aligned}
S(F_1) &= R(\psi_{0,F_1}) - R(\psi_0) = R(b,a,a) - R(a,a,a) = 1.8 - 1.0 = 0.8 \\
S(F_2) &= R(\psi_{0,F_2}) - R(\psi_0) = R(a,b,b) - R(a,a,a) = 1.9 - 1.0 = 0.9 \\
S(F_3) &= R(\psi_{0,F_3}) - R(\psi_0) = R(a,a,b) - R(a,a,a) = 1.5 - 1.0 = 0.5
\end{aligned}$$

Vi har også

$$U_* = R(\psi - R(\psi_0) = R(b,b,b) - R(a,a,a) = 2.1 - 1.0 = 1.1$$

Derfor er

$$C = \frac{U_*}{\sum_{i=1}^{2} S(F_i)} = \frac{1.1}{2.2} = 0.5$$

Dermed får vi

$$\begin{aligned}
S^{(1)}(F_1) &= 0.8 \cdot 0.5 = 0.4 \\
S^{(2)}(F_2) &= 0.9 \cdot 0.5 = 0.45 \\
S^{(3)}(F_3) &= 0.5 \cdot 0.5 = 0.25
\end{aligned}$$

Når vi husker at $R(a,a,a) = 1.0$ får vi sandsynligheden at dø af

$$\begin{aligned}
\text{Ukendt grund: } & \frac{1.0}{1.0 + 0.4 + 0.45 + 0.25} \approx 48\% \\
F_1\text{: } & \frac{0.4}{1.0 + 0.4 + 0.45 + 0.25} \approx 19\% \\
F_2\text{: } & \frac{0.45}{1.0 + 0.4 + 0.45 + 0.25} \approx 21\% \\
F_3\text{: } & \frac{0.25}{1.0 + 0.4 + 0.45 + 0.25} \approx 12\%
\end{aligned}$$

- For at udregne 2. ordens interaktioner starter vi igen med (13)

$$
\begin{aligned}
S(F_1, F_2) &= R(\psi_{0,F_{\{1,2\}}}) - R(\psi_0) - 2.2 \\
&= R(b,b,a) - R(a,a,a) - 2.2 \\
&= 1.8 - 1.0 - 2.2 \\
&= -1.4 \\
S(F_2, F_3) &= R(a,b,b) - R(a,a,a) - 2.2 \\
&= -1.3 \\
S(F_1, F_3) &= R(b,a,b) - R(a,a,a) - 2.2 \\
&= -1.3
\end{aligned}
$$

Altså er

$$\mathcal{H}_1^2 = \mathcal{H}_2^2 = \mathcal{H}_3^2 = \{1,2,3\}$$

Så udregner vi $T(F_{\mathcal{H}_i^2})$

$$
\begin{aligned}
R(b,b,b) - R(a,a,a) - \sum_{i=1}^{3} S(F_i) &= 2.1 - 1.0 - 2.2 \\
&= -1.1
\end{aligned}
$$

Da $-1.1 < 0$ sætter vi $T(F_{\mathcal{H}_i^2}) = |-1.1| = 1.1$ og da $1.1 > \min_{j \in \mathcal{H}_1^2} S(F_j) = S(F_3) = 0.5$ bruger vi formel (30) til at udregne $\tilde{T}(F_{\mathcal{H}_i^2})$

$$
\begin{aligned}
\tilde{T}(F_{\mathcal{H}_i^2}) &= S^{(1)} \cdot \frac{|\mathcal{H}_i^2|}{2} \\
&= 0.5 \cdot \frac{3}{2} \\
&= 0.75
\end{aligned}
$$

Nu tjekker vi om vi er færdige med at udregne $\tilde{T}$-led ved at undersøge kriterium (32), dvs. sammenligne summen af $\tilde{T}$'erne og $T(F_{\mathcal{H}_i^2})$. Vi har $0.75 < 1.1$, så vi udregner endnu et $\tilde{T}$-led vha. (31).

$$
\begin{aligned}
\tilde{T}(F_{\mathcal{H}_i^2 \setminus \{3\}}) &= \left(S^{(2)} - S^{(1)}\right) \frac{|\mathcal{H}_1^2| - 1}{2} \\
&= 0.8 - 0.5 \\
&= 0.3
\end{aligned}
$$

Da $0.3 + 0.75 < 1.1$, fortsætter vi selvom det sidste $\tilde{T}$-led kun indeholder en faktor og derfor ikke er et almindeligt ELLER-led.

$$\begin{aligned}\tilde{T}(F_{\mathcal{H}_i^2\setminus\{3,1\}}) &= \Big(S^{(3)} - S^{(2)}\Big)\frac{|\mathcal{H}_1^2| - 2}{2} \\ &= (0.9 - 0.8)\cdot\frac{1}{2} \\ &= 0.05\end{aligned}$$

Vi har $0.05 + 0.75 + 0.35 \not< T(F_{\mathcal{H}_i^2})$, så vi kan godt stoppe nu og behøver ikke lave nogen korrektion fra c. tilfældet fordi $0.05+0.75+0.35 = T(F_{\mathcal{H}_i^2})$. Korrektion fra c. ville have været at omdefinere $\tilde{T}(F_{\mathcal{H}_i^2\setminus\{3,1\}})$ til

$$\frac{\sum_{j\in\mathcal{H}_i^2} S(F_j)}{2} - \tilde{T}(F_{\mathcal{H}_i^2\setminus\{3\}}) - \tilde{T}(F_{\mathcal{H}_i^2})$$

hvilket også havde givet 0.05. Til sidst, normaliseringerne

$$\begin{aligned}\tilde{S}(F_1) &= S(F_1) - \tilde{T}(F_{\mathcal{H}_1^2})\frac{2}{3} - \tilde{T}(F_{\mathcal{H}_1^2\setminus\{3\}}) \\ &= 0.8 - 0.5 - 0.3 \\ &= 0\end{aligned}$$

og

$$\begin{aligned}\tilde{S}(F_2) &= S(F_2) - \tilde{T}(F_{\mathcal{H}_2^2})\frac{2}{3} - \tilde{T}(F_{\mathcal{H}_2^2\setminus\{3\}}) - 2\tilde{T}(F_{\mathcal{H}_2^2\setminus\{3,1\}} \\ &= 0.9 - 0.5 - 0.3 - 0.1 \\ &= 0.0\end{aligned}$$

og

$$\begin{aligned}\tilde{S}(F_3) &= S(F_3) - \tilde{T}(F_{\mathcal{H}_2^2})\frac{2}{3} \\ &= 0.5 - 0.5 \\ &= 0\end{aligned}$$

Det vil sige vi har følgende fortolkning af sandsynligheden for at dø af

$$\text{Ukendt grund: } \frac{1.0}{2.1} \approx 48\% \tag{48}$$

$$F_1,\ F_2 \text{ eller } F_3\text{: } \frac{0.75}{2.1} \approx 36\% \tag{49}$$

$$F_1 \text{ eller } F_2\text{: } \frac{0.3}{2.1} \approx 14\% \tag{50}$$

$$F_2\text{: } = \frac{0.05}{2.1} \approx 2\% \tag{51}$$

Forslag til polerende funktion

For hver dødsårsag, d, er der en række riskratiofiler (indekseret med j). I formel (7) diskuterede vi udregningen af en interpolerende funktion

$$R^{i,j,d}(\psi) = \mathrm{pol}((r^{i,j,d}(f))_{f \in \mathcal{F}^{i,j,d}}, \psi).$$

Her er i en (vistnok overflødig) indeksering der angiver aldersgruppen. $\mathcal{F}^{i,j,d}$ er de binnede faktorer i risk ratio-filen og ψ er et niveau af faktorer (som ikke er binnede). For at gøre det efterfølgende mere simpelt bruger vi notationen

$$R(\psi) = \mathrm{pol}(r, \psi)$$

hvor vi altså tænker at r må afhænge af $\mathcal{F}^{i,j,d}$ som vi omdøber til $\mathcal{F}$. For hvert niveau i $\mathcal{F}$ udregner vi et midtpunkt for de kontinuerte variable, $\mathrm{mid} : \mathcal{F} \mapsto \mathcal{S}$. Her skal også tages nogle valg. Hvis $f \in \mathcal{F}$ er på formen

$$\{0\} \times [15, 32] \times [\text{“Yes”}] \times [4, \infty)$$

er det tyer det meget naturligt at midtpunkterne for de tre første faktorer er 0, 16 og “Yes” hhv. Det sidste interval er her lidt tricky da det ikke har noget midtpunkt. Jeg ser to valg

(A) Vælg tallet 4

(B) Kig på det foregående interval, som måske er på formen $[1, 4)$, noter at dets midtpunkt er 1.5 væk fra 4, og sig at midtpunktet for $[4, \infty)$ bør være $4 + 1.5 = 5.5$

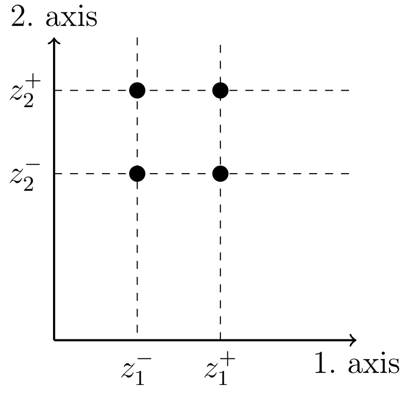
Generelt har vi kun brug for at interpolere og finde midtpunkter, når faktorniveauerne er intervaller. Det antages derfor i det efterfølgende. Når vi har valgt midtpunkterne er det nemmere at bruge standard interpolations metoder til at estimere R. Hvis vi brugte dem direkte ville vi have

$$R(\mathrm{mid}(f)) = r(f) \quad \forall f \in \mathcal{F} \tag{52}$$

Men vi ville ikke vide med sikkerhed om følgende ligning var sand:

$$\int_f R(x)\mathrm{d}x \stackrel{?}{=} r(f) \quad \forall f \in \mathcal{F} \tag{53}$$

Vi ville hellere have at det var omvendt; altså at (53) gjaldt altid, mens vi er lidt ligeglade med (52). For at opnå det foreslås det følgende: Lad $g : (\mathbb{R}^n \times \mathbb{R}^{(n+1)\cdot k}) \mapsto \mathbb{R}$ være en interpolationsmetode der bruger $k = \#\mathcal{F}$ punkter af formen (x_i, y_i), hvor $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$ til at kunne forudsige en



Figur 0.1: A hypercupe in 2 dimensions encapsulated by (z_1^-, z_2^-), (z_1^+, z_2^-), (z_1^-, z_2^+), and (z_1^+, z_2^+).

y-værdi når man kommer med et nyt x-punkt. Vi har allerede x_i-punkterne fra vores udregning af midtpunkterne, men vi mangler y_i-værdierne. Dem vil vi bestemme ud fra (53), dvs. løse ligningerne

$$r(f) = \int_f g(x, \{(x_1, y_1), \ldots, (x_k, y_k)\} \,\mathrm{d}x \qquad \forall f \in \mathcal{F} \tag{54}$$

med hensyn til $y_1, \ldots, y_k$. En specielt attraktiv interpolationsfunktion kunne være den n-dimensionelle lineære interpolation, hvilket vil sige at g inddeler området i et antal begrænsede n dimensionale hyber-rektangler. Hver hyperrektangel er afgrænset af 2^n punkter fra $\{x_1, \ldots, x_k\}$. Vi kan skrive de 2^n punkter

$$(z_1^{\pm}, z_2^{\pm}, \ldots, z_n^{\pm}). \tag{55}$$

Et eksempel ses i Figur 0.1. Lad $y_1, \ldots, y_{2^n}$ være de korresponderende y-værdier for hjørnerne i hyperkuben. Så kan man skrive interpolationsoverfladen som

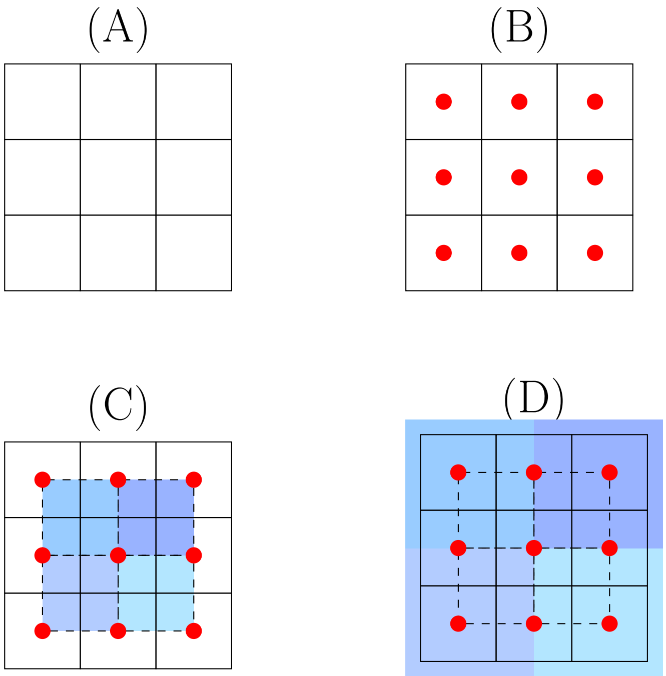
$$f(z) = f((z_1, \ldots, z_n)) = \sum_{J \subseteq \{1,\ldots,n\}} a_J \prod_{j \in J} z_j \tag{56}$$

hvor

$$\begin{pmatrix} a_\emptyset \\ a_{\{1\}} \\ \vdots \\ a_{\{1,\ldots,n\}} \end{pmatrix} = \begin{pmatrix} \vec{v}(z_1^-, z_2^-, \ldots, z_n^-) \\ \vec{v}(z_1^+, z_2^-, \ldots, z_n^-) \\ \vdots \\ \vec{v}(z_1^+, z_2^+, \ldots, z_n^+) \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{2^n} \end{pmatrix} \tag{57}$$

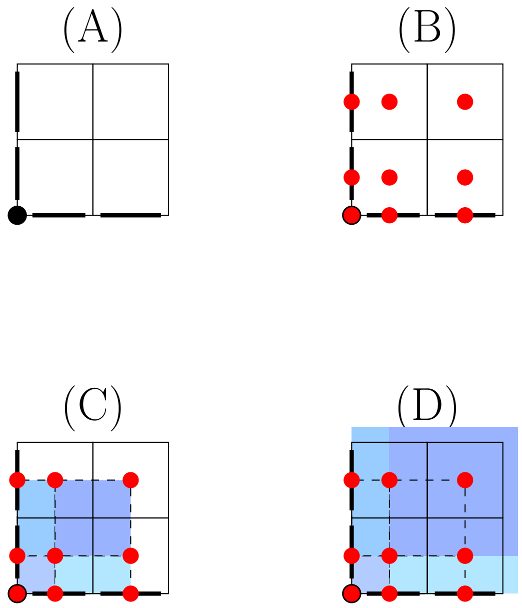
hvor

$$\vec{v}(w_1, \ldots, w_n) = \begin{pmatrix} 1 & w_1 & w_2 & \ldots & w_n & w_1 w_2 & w_1 w_3 & \ldots & \prod_{i=1}^n w_i \end{pmatrix}.$$



Figur 0.2: (A) viser faktorerne i $\mathcal{F}$, når der er tale om et grid af endelige faktorlevels med positivt lebesguemål (i $\mathbb{R}^2$). (B) viser midtpunkterne. (C) viser de 4 interpolationsflader der kan laves af de 9 punkter. (D) viser ekstrapoleringen af de 4 interpolationsområder som potentialt går ud i hele $\mathbb{R}^2$.

Det gør at hvert a i (56) kan skrives som en linearkombination af y_i'er, og dermed kan de k ligninger i (54) skrives som en linearkombinationer af $y_1, \ldots, y_k$. Med andre ord kan (54) (formodentlig) løses entydigt. Hvordan hyperkuberne skal lægges er illustreret i 2 dimensioner i Figur 0.2.

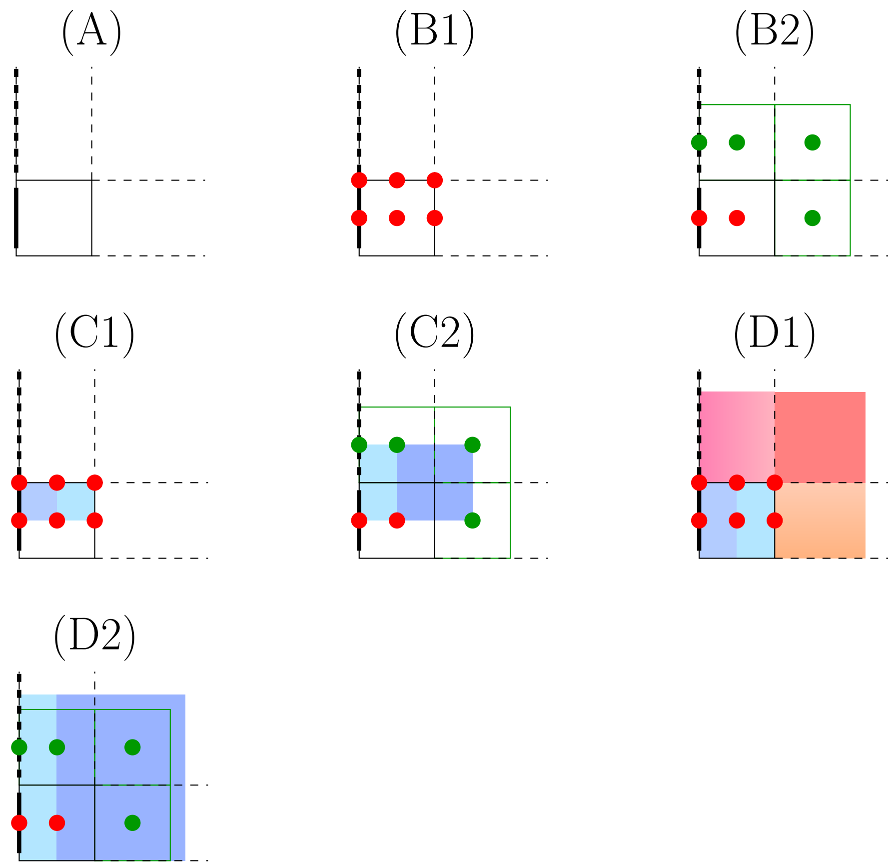


Figur 0.3: (A) viser faktorerne i $\mathcal{F}$, når der er tale om et endeligt grid af faktorlevels hvor der dog er punkter og linjer med positivt mål (markeret med sort prik og fed linje, hhv.). (B) viser midtpunkterne. (C) viser de 4 interpolationsflader der kan laves af de 9 punkter. (D) viser ekstrapoleringen af de 4 interpolationsområder som potentialt går ud i hele $\mathbb{R}^2$. Læg mærke til at y-værdien i nedre venstre hjørne er kendt, og at der derfor ikke er grund til at opstille integralligningen for den.

Løsning af Integralsystem

Vi løser nu (54). Vi ser fra (56) at vi bør integrere

$$\begin{aligned}
&\int_{[z_1^-,z_1^+]\times\cdots\times[z_n^-,z_n^+]} z_1\cdot z_2\cdots z_m \,\mathrm{d}(z_1,\ldots,z_n)\\
&=\int_{[z_2^-,z_2^+]\times\cdots\times[z_n^-,z_n^+]} \frac{1}{2}\Big(\left(z_1^+\right)^2-\left(z_1^-\right)^2\Big) z_2\cdots z_m \,\mathrm{d}(z_2,\ldots,z_n)\\
&=\cdots\\
&=\frac{1}{2^m}\prod_{i=1}^{m}\Big(\left(z_i^+\right)^2-\left(z_i^-\right)^2\Big)\int_{[z_{m+1}^-,z_{m+1}^+]\times\cdots\times[z_n^-,z_n^+]} 1 \,\mathrm{d}(z_1,\ldots,z_n)\\
&=\frac{1}{2^m}\prod_{i=1}^{m}\Big(\left(z_i^+\right)^2-\left(z_i^-\right)^2\Big)\prod_{i=m+1}^{n}\left(z_i^+-z_i^-\right) \qquad (58)
\end{aligned}$$



Figur 0.4: (A) viser faktorerne i $\mathcal{F}$, når der er tale om et grid af faktorlevels hvor der er linjer med positivt mål og faktorniveauer der breder sig ud mod uendelig (markeret med fed linje og stiplet linje, hhv.). (B1) viser en metode for midtpunkterne, mens (B2) viser en anden. I (B2) er der også sat markeringer af et pseudo-område for (B2). (D1) og (D2) viser, hvad vi ser som den mest naturlige måde at udvide interpolationsfladerne. I (D1) angiver det store røde felt at interpolationsværdien bliver sat til konstanten $r(f)$. De to gradientfelter (hhv. orange og lyserød) indikerer at interpolationsværdien bliver udregnet ved at projicere ned på kanten af den endelige hyperkube.

under antagelsen at $z_i^+ > z_i^-$. I det mere generelle tilfælde, hvor $z_i^+ \geq z_i^-$ antager vi at når $z_i^+ = z_i^-$ integrerer vi med hensyn til dirac-målet koncentreret i z_i^+. I så fald bliver (58) i stedet

$$H\big([(z_1^-, z_1^+), \dots, (z_n^-, z_n^+)], \{1, \dots, m\}\big) :=$$
$$\prod_{i=1, z_i^- \neq z_i^+}^{m} \Big(\frac{1}{2}(z_i^+)^2 - \frac{1}{2}(z_i^-)^2\Big) \prod_{i=1, z_i^- = z_i^+} z_i^+ \prod_{i=m+1, z_i^- \neq z_i^+}^{n} \big(z_i^+ - z_i^-\big) \tag{59}$$

Nu udvider vi notationen lidt. Lad

$$\mathrm{mid}(f) = (\mathrm{mid}_1(f_1), \dots, \mathrm{mid}_n(f_n)) \tag{60}$$

være midtpunktet for faktorniveau $f = (f_1, \dots, f_n) \in \mathcal{F}$. Vi antager at $\mathcal{F}$ er et grid, dvs.

$$\mathcal{F} = \Big\{ [a_1^-(j_1), a_1^+(j_1)] \times \dots \times [a_n^-(j_n), a_n^+(j_n)] : \\ j_1 \in \{1, \dots, k_1\}, \dots j_n \in \{1, \dots, k_n\} \Big\} \tag{61}$$

hvor $a_i^\pm(j)$ er passende konstanter i $\mathbb{R} \cup \{-\infty, \infty\}$, der opfylder

$$a_i^+(j) = a_i^-(j+1), \quad j = 1, \dots, k_i - 1$$

Per definition af midpoint ved vi

$$a_i^-(j) \leq \mathrm{mid}_i\big([a_i^-(j), a_i^+(j)]\big) \leq a_i^+(j) \tag{62}$$

Vi forkorter notationen yderligere med

$$z(i,j) = \mathrm{mid}_i\big([a_i^-(j), a_i^+(j)]\big).$$

og dermed kan et midtpunkt skrives som

$$\big(z(1, j_1), z(2, j_2), \dots, z(n, j_n)\big) \tag{63}$$

for passende $j_1 \dots, j_n$. En interpolationsflade (noteret f i (56)), kan nu skrives mere præcist som

$$h^{j_1, \dots, j_n}(z) = \sum_{J \subseteq \{1, \dots, n\}} a_J(j_1, \dots, j_n) \prod_{j \in J} z_j \tag{64}$$

for $j_i \in \{1, \dots, k_i - 1\}$, hvilket giver $(k_1 - 1) \cdots (k_n - 1)$ forskellige interpolationsflader. Interpolationsfladen i (64) gælder umiddelbart kun for z i mængden

$$[z(1, j_1), z(1, j_1 + 1)] \times \dots \times [z(n, j_n), z(n, j_n + 1)] \tag{65}$$

Dog udvider vi ofte denne mængde, se Figur 0.2, 0.3 og 0.4. For at løse ligningssystemet (54) med (B2)-midpoints sætter vi

$$\tilde{z}(i,j) = \begin{cases} a_i^-(1) & \text{hvis } j = 1, a_i^-(j) > -\infty \\ a_i^+(1) - \big(a_i^+(2) - a_i^-(2)\big) & \text{hvis } j = 1, a_i^-(j) = -\infty \\ a_i^+(k_j) & \text{hvis } j = k_j, a_i^+(k_j) < \infty \\ a_i^+(k_j) + \big(a_i^+(k_j - 1) - a_i^-(k_j - 1)\big) & \text{hvis } j = k_j, a_i^+(j) = \infty \\ z(i,j) & \text{ellers.} \end{cases} \tag{66}$$

og lader (64) være gældende for z i mængden

$$[\tilde{z}(1,j_1), \tilde{z}(1,j_1+1)] \times \cdots \times [\tilde{z}(n,j_n), \tilde{z}(n,j_n+1)].$$

I integralet i (54) skal vi integrere over f. Hvis vi bruger (B2) er det nemmest at formulere integralet ved at justere $a_i^{\pm}$'erne. Det gøres med

$$\tilde{a}_i^-(j) = \begin{cases} a_i^+(1) - \big(a_i^+(2) - a_i^-(2)\big) & \text{hvis } j = 1, a_i^-(j) = -\infty \\ a_i^-(j) & \text{ellers.} \end{cases}$$

og

$$\tilde{a}_i^+(j) = \begin{cases} a_i^+(k_j) + \big(a_i^+(k_j - 1) - a_i^-(k_j - 1)\big) & \text{hvis } j = k_j, a_i^+(j) = \infty \\ a_i^+(j) & \text{ellers.} \end{cases}$$

For at gøre notationen nemmere i kanterne af grid'et, sættes også

$$h^{j_1,\dots,j_n} = h^{(j_1 \wedge k_1 - 1)\vee 1,\dots,(j_n \wedge k_n - 1)\vee 1}, \tag{67}$$

$$a_J(j_1,\dots,j_n) = a_J\big((j_1 \wedge k_1 - 1)\vee 1,\dots,(j_n \wedge k_n - 1)\vee 1\big) \tag{68}$$

for $j_i \in \{0,\dots,k_i\}, i = 1,\dots,n, J \subseteq \{1,\dots,n\}$

Bemærk at et $f \in \mathcal{F}$ er kendetegnet ved $j_1,\dots,j_n$. Da kan vi skrive ligningssystemet (54) som k ligninger, der hver består af en sum af 2^n integraler.

$$\begin{aligned} r(j_1,\dots,j_n) = & \int_{[\tilde{a}_1^-(j_1),z(1,j_1)]\times\cdots\times[\tilde{a}_n^-(j_n),z(n,j_n)]} h^{j_1-1,\dots,j_n-1}(z)\,\mathrm{d}z \\ & + \int_{[z(1,j_1),\tilde{a}_1^+(j_1)]\times[\tilde{a}_2^-(j_2),z(2,j_2)]\times\cdots\times[\tilde{a}_n^-(j_n),z(n,j_n)]} h^{j_1,j_2-1\dots,j_n-1}(z)\,\mathrm{d}z \\ & \vdots \\ & + \int_{[z(1,j_1),\tilde{a}_1^+(j_1),]\times\cdots\times[z(n,j_n),\tilde{a}_n^+(j_n)]} h^{j_1,j_2\dots,j_n}(z)\,\mathrm{d}z \end{aligned} \tag{69}$$

Ved at bruge H fra tidligere kan vi sætte

$$\begin{aligned}&\int_{[\tilde{a}_1^-(j_1),z(1,j_1)]\times\cdots\times[\tilde{a}_n^-(j_n),z(n,j_n)]} h^{j_1-1,\ldots,j_n-1}(z)\ \mathrm{d}z\\ &= \sum_{J\subseteq\{1,\ldots,n\}} \Big(a_J(j_1-1,\ldots,j_n-1)\\ &\qquad \cdot H\big([(\tilde{a}_1^-(j_1),z(1,j_1)),\ldots,(\tilde{a}_n^-(j_n),z(n,j_n)],J\big)\Big)\end{aligned} \tag{70}$$

Definer nu

$$\mathbf{V}(j_1,\ldots,j_n) = \begin{pmatrix} \vec{v}(z(1,j_1),z(2,j_2),\ldots,z(n,j_n)) \\ \vec{v}(z(1,j_1+1),z(2,j_2),\ldots,z(n,j_n)) \\ \vec{v}(z(1,j_1),z(2,j_2+1),\ldots,z(n,j_n)) \\ \vdots \\ \vec{v}(z(1,j_1+1),z(2,j_2+1),\ldots,z(n,j_n+1)) \end{pmatrix}. \tag{71}$$

med

$$\mathbf{V}(j_1,\ldots,j_n) = \mathbf{V}\big((j_1\wedge[k_1-1])\vee 1,\ldots,(j_n\wedge[k_n-1])\vee 1\big) \tag{72}$$

for $j_i \in \{0,\ldots,k_i\}, i=1,\ldots,n$, samt

$$\mathbf{y}(j_1,\ldots,j_n) = \begin{pmatrix} y(j_1,j_2,\ldots,j_n) \\ y(j_1+1,j_2,\ldots,j_n) \\ \vdots \\ y(j_1+1,\ldots,j_n+1) \end{pmatrix} \tag{73}$$

for $j_i \in \{1,\ldots,k_i-1\}, i=1,\ldots,n$. Vi definerer

$$\mathbf{y}(j_1,\ldots,j_n) = \mathbf{y}\big((j_1\wedge[k_1-1])\vee 1,\ldots,(j_n\wedge[k_n-1])\vee 1\big) \tag{74}$$

for generelle $j_i \in \{0,\ldots,k_i\}, i=1,\ldots,n$.

Vi fortolker $y(j_1,\ldots,j_n)$ som interpolationens y-værdi i $(z(1,j_1),\ldots,z(n,j_n))$. Altså er $\mathbf{y}$ vektoren af y-værdier, der indgår i $h^{j_1,\ldots,j_n}$. Indsat i (70) giver det

$$\begin{aligned}\sum_{J\subseteq\{1,\ldots,n\}} \Big(&\mathbf{e}_J^T\mathbf{V}(j_1-1,\ldots,j_n-1)^{-1}\mathbf{y}(j_1-1,\ldots,j_n-1)\\ &\cdot H\big([(\tilde{a}_1^-(j_1),z(1,j_1)),\ldots,(\tilde{a}_n^-(j_n),z(n,j_n)],J\big)\Big)\end{aligned} \tag{75}$$

hvor $\mathbf{e}_J$ er en vektor af 0'er på nær et 1-tal i den indgang der svarer til den koordinat der svarer til J. Hvis vi definerer vektoren af længde 2^n

$$\mathbf{H}\big([(\tilde{a}_1^-(j_1),z(1,j_1)),\ldots,(\tilde{a}_n^-(j_n),z(n,j_n)]\big) \tag{76}$$

$$= \begin{pmatrix} H\big([(\tilde{a}_1^-(j_1),z(1,j_1)),\ldots,(\tilde{a}_n^-(j_n),z(n,j_n)],\emptyset\big) \\ H\big([(\tilde{a}_1^-(j_1),z(1,j_1)),\ldots,(\tilde{a}_n^-(j_n),z(n,j_n)],\{1\}\big) \\ \vdots \\ H\big([(\tilde{a}_1^-(j_1),z(1,j_1)),\ldots,(\tilde{a}_n^-(j_n),z(n,j_n)],\{1,\ldots,n\}\big) \end{pmatrix} \tag{77}$$

kan vi skrive (75) som

$$\begin{aligned}&\Big[\mathbf{H}\big([(\tilde{a}_1^-(j_1), z(1,j_1)), \ldots, (\tilde{a}_n^-(j_n), z(n,j_n)]\big)^T \\ &\cdot\mathbf{V}(j_1-1,\ldots,j_n-1)^{-1}\Big]\mathbf{y}(j_1-1,\ldots,j_n-1)\end{aligned} \tag{78}$$

Vi kan derfor skrive ligningssystemet fra (69) som

$$\begin{aligned} r(j_1,\ldots,j_n) = \Bigg\{ &\Big[\mathbf{H}\big([(\tilde{a}_1^-(j_1), z(1,j_1)), \ldots, (\tilde{a}_n^-(j_n), z(n,j_n)]\big)^T \\ &\cdot\mathbf{V}(j_1-1,\ldots,j_n-1)^{-1}\Big]\mathbf{y}(j_1-1,\ldots,j_n-1) \\ &+\Big[\mathbf{H}\big([(z(1,j_1), \tilde{a}_1^+(j_1)), \ldots, (\tilde{a}_n^-(j_n), z(n,j_n)]\big)^T \\ &\quad\cdot\mathbf{V}(j_1, j_2-1\ldots,j_n-1)^{-1}\Big]\mathbf{y}(j_1, j_2-1\ldots,j_n-1)+ \\ &\vdots \\ &+\Big[\mathbf{H}\big([(z(1,j_1), \tilde{a}_1^+(j_1)), \ldots, (z(n,j_n), \tilde{a}_n^+(j_n))]\big)^T \\ &\quad\cdot\mathbf{V}(j_1,\ldots,j_n)^{-1}\Big]\mathbf{y}(j_1,\ldots,j_n)\Bigg\} \end{aligned} \tag{79}$$

Ud fra (79) kræver det et stykke (trivielt) omrokeringsarbejde at få det på formen

$$\big\{r(j_1,\ldots,j_n)\big\}_{j_i\in\{1,\ldots,k_i\},i=1,\ldots,n} = B\big\{y(j_1,\ldots,j_n)\big\}_{j_i\in\{1,\ldots,k_i\},i=1,\ldots,n} \tag{80}$$

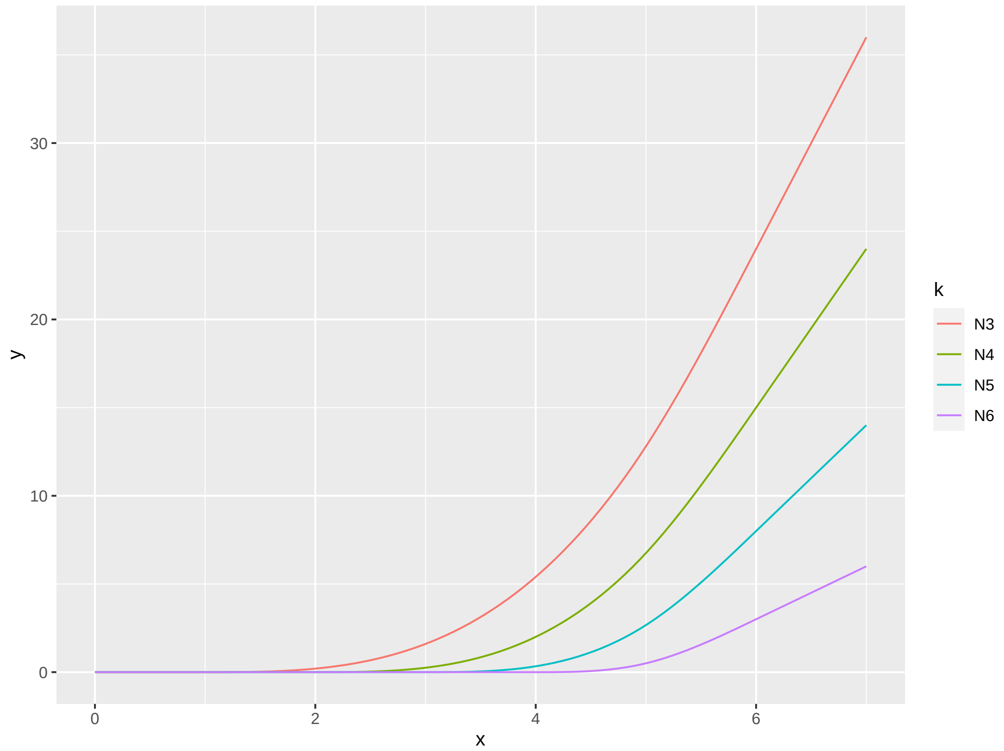
som løses ved at invertere $k \times k$ matricen B. Når de k y-værdier er fundet regnes matrix produkterne

$$\mathbf{V}(j_1,\ldots,j_n)^{-1}\mathbf{y}(j_1,\ldots,j_n) \tag{81}$$

og koefficienterne sættes ind i risk ratio filerne.

Multivariate splines

I stedet for eksplicit at modellere midtpunkterne mellem stykkerne, kan vi også modellere funktionen direkte. Hvis vi for eksempel tager den naturlige



Figur 0.5: For 6 punkter i $\{1, 2, 3, 4, 5, 6\}$ vises her de De 4 sidste basisfunktioner for natural cubic splines.

kubiske spline (se Figur 0.5) har den i en enkelt dimension basis

$$
\begin{aligned}
N_1(X) &= 1 \\
, N_2(X) &= X \\
N_{k+2}(X) &= \frac{(X - x_k)_+^3 - (X - x_K)_+^3}{x_K - x_k} - \frac{(X - x_{K-1})_+^3 - (X - x_K)_+^3}{x_K - x_{K-1}} \\
&= \mathbb{I}(x_K \geq X > x_k)\frac{X^3 - x_k^3 + 3Xx_k^2 - 3X^2x_k}{x_K - x_k} \\
&\quad -\mathbb{I}(x_K \geq X > x_{K-1})\frac{X^3 - x_{K-1}^3 + 3Xx_{K-1}^2 - 3X^2x_{K-1}}{x_K - x_{K-1}} \\
&\quad +\mathbb{I}(X > x_K)\frac{-x_k^3 + x_K^3 + 3X(x_k^2 - x_K^2) - 3X^2(x_k - x_K)}{x_K - x_k} \\
&\quad -\mathbb{I}(X > x_K)\frac{-x_{K-1}^3 + x_K^3 + 3X(x_{K-1}^2 - x_K^2) - 3X^2(x_{K-1} - x_K)}{x_K - x_{K-1}} \\
&= X^3\mathbb{I}(x_K \geq X > x_k)\Big(\frac{1}{x_K - x_k} - \frac{\mathbb{I}(X > x_{K-1})}{x_K - x_{K-1}}\Big) \\
&\quad +X^2\mathbb{I}(x_K \geq X > x_k)\Big(\frac{-3x_k}{x_K - x_k} - \mathbb{I}(X > x_{K-1})\frac{-3x_{K-1}}{x_K - x_{K-1}}\Big) \\
&\quad +X\mathbb{I}(X > x_k)\Big(\frac{\mathbb{I}(X \leq x_K)3x_k^2}{x_K - x_k} - \frac{\mathbb{I}(x_K \geq X > x_{K-1})3x_{K-1}^2}{x_K - x_{K-1}} \\
&\qquad\qquad +\mathbb{I}(X > x_K)\big(-3(x_K + x_k) + 3(x_K + x_{K-1})\big)\Big) \\
&\quad +\mathbb{I}(x_K \geq X > x_k)\frac{-x_k^3}{x_K - x_k} - \mathbb{I}(x_K \geq X > x_{K-1})\frac{-x_{K-1}^3}{x_K - x_{K-1}}
\end{aligned}
$$

Hvis vi har flere dimensioner dvs. $X_1, X_2, \dots X_N$ tager vi da basisen

$$\{1\} \cup \{N_k(X_i)\}_{i=1,\dots,N,k=2,\dots,K_i} \cup \Big\{N_k(X_i)N_l(X_j)\Big\}_{i\neq j\in\{1,\dots,N\},k=2,\dots,K_i,l=2,\dots,K_j} \tag{82}$$

Vi estimerer altså funktionen

$$g(X_1,\dots,X_N) = \beta_1 + \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k=2}^{K_i}\sum_{l=2}^{K_j}\beta_{ijkl}N_k(X_i)N_l(X_j) + \sum_{i=1}^{N}\sum_{k=1}^{K_i}\beta_{ik}N_k(X_i)$$

hvor vi som likelihoodfunktion har

$$r(f)|f| \sim N\left(\int_f g(X_1,\dots,X_N)\ \mathrm{d}(X_1,\dots,X_N), \sigma^2\right)$$

for alle $f \in \mathcal{F}$. For at reducere overfitningen, bruger vi regulariseringen

$$\lambda \sum_{f\in\mathcal{F}} \int_f \sum_{i=1}^{N}\sum_{j=1}^{N}\Big(\frac{\partial g}{\partial X_i \partial X_j}\Big)^2\ \mathrm{d}(X_1,\dots,X_N) \tag{83}$$

For $r \neq s$ er den inderste del på formen

$$\begin{aligned}\frac{\partial g}{\partial X_r \partial X_s} &= \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k=2}^{K_i}\sum_{j=2}^{K_j}\beta_{ijkl}\frac{\partial}{\partial X_r \partial X_s}N_k(X_i)N_l(X_j)\\ &= \sum_{k=2}^{K_r}\sum_{l=2}^{K_s}\beta_{rskl}N_k'(X_r)N_l'(X_s)\end{aligned}$$

og for $r = s$ på formen

$$\begin{aligned}\frac{\partial g}{\partial X_r \partial X_r} &= \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k=2}^{K_i}\sum_{j=2}^{K_j}\beta_{ijkl}\frac{\partial}{\partial X_r \partial X_r}N_k(X_i)N_l(X_j) \qquad (84)\\ &\quad + \sum_{k=1}^{K_r}\beta_{rk}N_k''(X_r)\\ &= \sum_{i=1}^{r-1}\sum_{k=2}^{K_r}\sum_{l=2}^{K_r}\beta_{irkl}N_k(X_i)N_l''(X_r) \qquad (85)\\ &\quad + \sum_{j=r+1}^{N}\sum_{k=2}^{K_r}\sum_{l=2}^{K_r}\beta_{rjkl}N_k''(X_r)N_l(X_j)\\ &\quad + \sum_{k=1}^{K_r}\beta_{rk}N_k''(X_r)\end{aligned}$$

Det vil sige at indmaden i (83) bliver for $r \neq s$

$$\sum_{k_1=2}^{K_r}\sum_{k_2=2}^{K_s}\sum_{l_1=2}^{K_r}\sum_{l_2=2}^{K_s}\beta_{rsk_1l_1}N'_{k_1}(X_r)N'_{k_2}(X_r)N'_{l_1}(X_s)N'_{l_2}(X_s)\beta_{rsk_2l_2} \tag{86}$$

og for $r = s$ bliver den

$$\sum_{j_1=r+1}^{N}\sum_{k_1=2}^{K_r}\sum_{l_1=2}^{K_{j_1}}\sum_{j_2=r+1}^{N}\sum_{k_2=2}^{K_r}\sum_{l_2=2}^{K_{j_2}}\beta_{rj_1k_1l_1}N''_{k_1}(X_r)N_{l_1}(X_{j_1})N''_{k_2}(X_r)N_{l_2}(X_{j_2})\beta_{rj_2k_2l_2} \tag{87}$$

$$+\sum_{i_1=1}^{r-1}\sum_{k_1=2}^{K_{i_1}}\sum_{l_1=2}^{K_r}\sum_{i_2=1}^{r-1}\sum_{k_2=2}^{K_{i_2}}\sum_{l_2=2}^{K_r}\beta_{i_1rk_1l_1}N_{k_1}(X_{i_1})N''_{l_1}(X_r)N_{k_2}(X_{i_2})N''_{l_2}(X_r)\beta_{i_2rk_2l_2} \tag{88}$$

$$+\sum_{k_1=2}^{K_r}\sum_{k_2=2}^{K_r}\beta_{rk_1}N''_{k_1}(X_r)N''_{k_2}(X_r)\beta_{rk_2} \tag{89}$$

$$+2\sum_{k_1=2}^{K_r}\sum_{j=r+1}^{N}\sum_{k_2=2}^{K_r}\sum_{l_2=2}^{K_j}\beta_{rk_1}\beta_{rjk_2l_2}N''_{k_1}(X_r)N''_{k_2}(X_r)N_{l_2}(X_j) \tag{90}$$

$$+2\sum_{k_1=2}^{K_r}\sum_{i=1}^{r-1}\sum_{k_2=2}^{K_i}\sum_{l_2=2}^{K_r}\beta_{rk_1}\beta_{irk_2l_2}N''_{k_1}(X_r)N_{k_2}(X_i)N''_{l_2}(X_r) \tag{91}$$

$$+2\sum_{j=r+1}^{N}\sum_{k_1=2}^{K_r}\sum_{l_2=2}^{K_j}\sum_{i=1}^{r-1}\sum_{k_2=2}^{K_i}\sum_{l_2=2}^{K_r}\beta_{rjk_1l_1}\beta_{irk_2l_2}N''_{k_1}(X_r)N_{l_1}(X_j)N_{k_2}(X_i)N''_{l_2}(X_r) \tag{92}$$

Målet er nu at skrive (83) på formen

$$\lambda\beta^* W\beta.$$

Hvis vi betragter situationen udtrykket (86) ser vi at vi hvis vi definerer

$$\beta^T_{rs\bullet\bullet} = \begin{pmatrix}\beta_{rs2,2} & \beta_{rs2,3} & \dots & \beta_{rs2,K_s} & \beta_{rs3,2} & \dots & \beta_{rsK_rK_s}\end{pmatrix}$$

kan vi skrive (86) som

$$\beta^T_{rs\bullet\bullet}W_{rs}\beta_{rs\bullet\bullet}$$

hvor

$$W_{rs} = \begin{pmatrix} w_{rs2,2,2,2} & w_{rs2,2,2,3} & \dots & w_{rs2,2,2,K_s} & w_{rs2,2,3,2} & \dots & w_{rs2,2,K_r,K_s} \\ w_{rs2,3,2,2} & w_{rs2,3,2,3} & \dots & & & & w_{rs2,3,K_r,K_s} \\ \vdots & & \ddots & & & & \vdots \\ w_{rsK_rK_s,2,2} & & \dots & & & & w_{rsK_rK_sK_rK_s} \end{pmatrix} \tag{93}$$

hvor

$$w_{rsk_1,l_1,k_2,l_2} = N'_{k_1}(X_r)N'_{l_1}(X_s)N'_{k_2}(X_r)N'_{l_2}(X_s) \tag{94}$$

Lad nu

$$\tilde{w}_{rsk_1,l_1,k_2,l_2} = \sum_{f\in\mathcal{F}} \int_f w_{rsk_1,l_1,k_2,l_2} \; \mathrm{d}(X_1,\dots,X_N) \tag{95}$$

$$\tilde{W}_{rs} = \sum_{f\in\mathcal{F}} \int_f W_{rs} \; \mathrm{d}(X_1,\dots,X_N) \tag{96}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_{1,1} \\ \beta_{1,2} \\ \vdots \\ \beta_{N,K_N} \\ \beta_{1,2\bullet\bullet} \\ \vdots \\ \beta_{N-1,N\bullet\bullet} \end{pmatrix} \tag{97}$$

Dimensionen af β er

$$1 + \sum_{i=1}^{N}(K_i - 1) + \sum_{i=1}^{N}\sum_{j>i}^{N}(K_i - 1)(K_j - 1)$$

For at finde bidraget af $r \neq s$ leddene fra (86) i den samlede regularisering, (83), definerer jeg

$$\tilde{W}_1 = \left(\begin{array}{c|cccc} \mathbf{0} & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \tilde{W}_{12} & 0 & \dots & 0 \\ \cdot & 0 & \tilde{W}_{13} & & \vdots \\ \cdot & \vdots & & \ddots & \\ \cdot & 0 & \dots & & \tilde{W}_{N-1,N} \end{array}\right) \tag{98}$$

og da er regulariseringen fra $r \neq s$ leddene

$$\lambda\beta^T(2\tilde{W}_1)\beta \tag{99}$$

For $r = s$ bliver det mere kompliceret. Når man kigger på leddene i (87) ser man at der er for et fast j_1 og j_2 står tilbage

$$\sum_{k_1=2}^{K_r}\sum_{l_1=2}^{K_r}\sum_{k_2=2}^{K_r}\sum_{l_2=2}^{K_r} \beta_{rj_1k_1l_1} N''_{k_1}(X_r)N_{l_1}(X_{j_1})N''_{k_2}(X_r)N_{l_2}(X_{j_2})\beta_{rj_2k_2l_2}$$

For $r < j_1, r < j_2$ ses det at være

$$\beta^T_{rj_1\bullet\bullet} W_{(r,j_1),(r,j_2)} \beta_{rj_2\bullet\bullet} \tag{100}$$

hvor $W_{(r,j_1),(r,j_2)}$ er matricen

$$\begin{pmatrix} w_{rj_1rj_22,2,2,2} & w_{rj_1rj_22,2,2,3} & \cdots & w_{rj_1rj_22,2,2,K_{j_2}} & \cdots & w_{rj_1rj_22,2,K_r,K_{j_2}} \\ w_{rj_1rj_22,3,2,2} & w_{rj_1rj_22,3,2,3} & \cdots & & & w_{rj_1rj_22,3,K_r,K_{j_2}} \\ \vdots & & \ddots & & & \vdots \\ w_{rj_1rj_2K_rK_{j_1},2,2} & & \cdots & & & w_{rj_1rj_2K_rK_{j_1}K_rK_{j_2}} \end{pmatrix} \tag{101}$$

hvor

$$w_{rj_1rj_2k_1l_1k_2l_2} = N''_{k_1}(X_r)N_{l_1}(X_{j_1})N''_{k_2}(X_r)N_{l_2}(X_{j_2}) \tag{102}$$

Når fx $r > j_1$ er problemet, at der ikke findes en $\beta_{rj_1\bullet\bullet}$. Der findes dog
Betragt nu (**??**) for et fast j_1.

$$2\sum_{k_1=2}^{K_r}\sum_{l_1=2}^{K_{j_1}}\sum_{k_2=2}^{K_r} \beta_{rj_1kl} N''_{k_1}(X_r)N_{l_1}(X_{j_1})N''_{k_2}(X_r)\beta_{rk_2} \tag{103}$$

Lad nu

$$\beta^T_{r\bullet} = \begin{pmatrix} \beta_{r2} & \ldots & \beta_{rK_r} \end{pmatrix} = \begin{pmatrix} \beta_{rj2,1} & \beta_{rj3,1} & \ldots & \beta_{rjK_r1} \end{pmatrix} \tag{104}$$

og definer

$$W_{(r,0),(r,j_2)} \\ = \begin{pmatrix} w_{r0rj_22,1,2,2} & w_{r0rj_22,1,2,3} & \cdots & w_{r0rj_22,1,2,K_{j_2}} & \cdots & w_{r0rj_22,1,K_r,K_{j_2}} \\ w_{r0rj_23,1,2,2} & w_{r0rj_22,1,2,3} & \cdots & & & w_{r0rj_22,1,K_r,K_{j_2}} \\ \vdots & & \ddots & & & \vdots \\ w_{r0rj_2K_r1,2,2} & & \cdots & & & w_{r0rj_2K_r1K_rK_{j_2}} \end{pmatrix} \tag{105}$$

$$W_{(r,j_1),(r,0)} = W^T_{(r,0),(r,j_1)} \tag{106}$$

hvor

$$w_{r0rjk_1,1,k_2,l_2} = N''_{k_1}(X_r)N_{l_1}(X_{j_1})N''_{k_2}(X_r) \tag{107}$$

Da kan vi skrive (102) som

$$2\beta^T_{rj_1\bullet\bullet} W_{(r,0),(r,j_1)}\beta_{r\bullet} = \beta^T_{rj_1\bullet\bullet} W_{(r,0),(r,j_1)}\beta_{r\bullet} + \beta^T_{r\bullet} W_{(r,j_1),(r,0)}\beta_{rj_1\bullet\bullet} \tag{108}$$

Det sidste led, (**??**) ses at være

$$\beta_{r\bullet}^T \left\{ N''_{k_1}(X_r) N_{k_2}(X_r) \right\}_{k_1,k_2=2,\ldots,K_r} \beta_{r\bullet} \tag{109}$$

Vi kalder matricen $\left\{ N''_{k_1}(X_r) N_{k_2}(X_r) \right\}_{k_1,k_2=2,\ldots,K_r}$ for W_0.

For at udtrykke den fulde regularisering, (83), i formen $\beta^T W \beta$ mangler vi kun at få sat