

TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Ralf Grossfeldt
Splainid ja üldistatud aditiivsed mudelid

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Raul Kangro

TARTU 2026

SPLAINID JA ÜLDISTATUD ADITIIVSED MUDELID

Bakalaureusetöö

Ralf Grossfeldt

Lühikokkuvõte

Antud bakalaureusetöö tutvustab splineide matemaatilist teooriat ja nende rakendamist üldistatud aditiivsetes mudelites (GAM). Töö eesmärk on pakkuda paindlikku alternatiivi polünoomregressioonile mittelineaarsete seoste modelleerimiseks, lahendades selliseid probleeme nagu ülesobitamine ja ebausaldusväärne ekstrapolatsioon. Käsitletakse B-splineide omadusi, siluvate splineide ja P-splineide põhimõtteid ning nende kasutust GAM-raamistikus. Selgitatakse karistusliikmete ja silumisparameetri valiku rolli mudeli sileduse tagamisel. Töö sisaldab ka ülevaadet interaktsioonide modelleerimist tensorsplineide abil ning juhiseid meetodite realiseerimiseks tarkvaras R.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Splainid, üldistatud aditiivsed mudelid, mittelineaarne regressioon, statistiline modelleerimine, masinõpe.

SPLINES AND GENERALIZED ADDITIVE MODELS

Bachelor thesis

Ralf Grossfeldt

Abstract

This bachelor's thesis introduces the mathematical theory behind splines and their application in generalized additive models (GAM). It aims to provide a flexible alternative to polynomial regression for modeling non-linear relationships, addressing issues such as overfitting and unreliable extrapolation. The work covers B-spline properties, smoothing splines, P-splines, and their integration into the GAM framework. The role of penalty terms and smoothing parameter selection is also explained. The thesis includes an overview of smooth interaction modeling via tensor product splines and practical implementation guidelines for the R software.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Splines, generalized additive models, non-linear regression, statistical modelling, machine learning.

Sisukord

Sissejuhatus	5
1 Lineaarsed regressioonimudelid	7
1.1 Lineaarsed mudelid	7
1.1.1 Olulised mõisted lineaarsete regressioonimudelite puhul	9
1.2 Lineaarne regressioon mittelineaarse sõltuvuse korral	11
1.3 Astmeteisenduste ja polünoomide puudujäägid	12
2 Splainid ja spline regressioon	15
2.1 Splainid	15
2.2 Splainide ruum ja selle baas	16
2.3 B-splainid	18
2.3.1 Vähimruutude meetod splineide leidmiseks	24
2.3.2 Ekstrapolatsioon	25
3 Üldistatud aditiivsed mudelid	27
3.1 Siluvad splineid	27
3.2 P-splineid	30
3.3 GAM-meetod	35
3.3.1 Silumisparameetri λ leidmine	37
3.3.2 Mõjumaatriks siluvate splineide puhul	37
3.3.3 Efektiivne vabadusastmete arv	38
3.3.4 Tensorsplineid	40
3.3.5 GAM-ide eripära	41
Kokkuvõte	43
Kasutatud allikad	45

Lisa 1. P-splainide diferentsimaatriksid	47
Lisa 2. GAM-ide rakendamine ja realiseerimine tarkvaras R	48

Sissejuhatus

Statistilises modelleerimises, eriti regressioonülesannete lahendamisel, tuleb sageli ette olukordi, kus uuritava tunnuse ja seletavate tunnuste vahelised seosed ei ole lineaarsed. Selliste seoste adekvaatseks kirjeldamiseks ei piisa klassikalisest lineaarsest regressioonist, vaid on tarvis paindlikumaid meetodeid. Käesolev bakalaureusetöö keskendubki mittelineaarsete seoste modelleerimisele, seades peamiseks fookuseks splineid ja nende rakendamise üldistatud aditiivsetes mudelites.

Sõna „spline“ (ingl *spline*) on inseneerias kasutusel olnud juba vähemalt 18. sajandi lõpust. Sel ajal tähistas see peent puidust või kummist elastset liistu, mida joonestajad kasutasid sujuvate kõverate tõmbamiseks, näiteks raudteetrasside planeerimisel (Oxford English Dictionary, 2026). Samuti on splaine joonestustöös kasutanud laevaehitajad, kes kinnitasid painduva liistu raskuste abil kindlatesse sõlmpunktidesse joonisel, et kõveraid paberile tõmmata. Füüsikaseadustest tulenevalt võttis liistu ülejäänud osa kuju, mis minimeeris selle paindeenergiat. See tagas kõvera visuaalse sujuvuse, mis on osutunud kasulikuks omaduseks ka regressioonülesannete lahendamisel. (Meijering, 2002)

Lõpuks leidis see algselt puhtpraktiline joonestusmeetod oma tee ka matemaatilisse statistikasse. Esimese täielikult splineidele keskenduva teadusartikli avaldas 1946. aastal Isaac Jacob Schoenberg (Oja, 2022, lk 5). Pärast seda on splineide teooria läbi teinud märkimisväärse arengu. Tänapäevaks on algsest ideest välja kasvanud üsna võimas meetod üldistatud aditiivsete mudelite näol.

Käesolev bakalaureusetöö koosneb kolmest peatükist. Esimeses peatükis antakse ülevaade klassikalisest lineaarsest regressioonist ja näidatakse, kuidas selle abil on võimalik modelleerida lihtsamaid mittelineaarseid seoseid. Ühtlasi defineeritakse edasiseks tööks vajalik terminoloogia ning tuuakse välja pakutud baaslahenduste puudujääke, mis annavadki motivatsiooni keerukamate meetodite kasutuselevõtuks.

Teine peatükk on pühendatud splineide matemaatilisele teooriale. Selles defineeritakse modelleerimise kontekstis kõige olulisemad mõisted ja tuuakse välja splineide matemaatilised omadused, mis teevad neist efektiivse tööriista andmete lähendamisel.

Kolmas peatükk seob eelneva teooria tervikuks, keskendudes siluvatele splineidele ja GAM-metodile. Selles peatükis näidatakse, kuidas luua paindlikke prognoosimudeleid nii, et andmete üle- või alasobitamine oleks kontrolli all. Töö lõpetab lisa 2, mis pakub

lühikese praktilise ülevaate GAM-mudelite rakendamisest statistikatarkvaras R.

1 Lineaarsed regressioonimudelid

Selles peatükis defineerime lineaarse regressioonimudeli ja sellega seotud terminoloogia ning kirjeldame parameetrite leidmist vähimruutude meetodil. Lisaks käsitleme puudujäärke, mis tekivad lineaarsete regressioonimudelite rakendamisel.

1.1 Lineaarsed mudelid

Iga statistilise prognoosimudeli saab esitada järgmiselt:

$$Y = f(X, \beta) + \epsilon,$$

kus Y on sõltuv ehk uuritav ehk sihttunnus,

- $f(\cdot)$ on funktsioon ehk eeskiri, mis seab igale sisendmuutujate komplektile vastavusse ühe ja ainult ühe väljundi, antud juhul sõltuva tunnuse väärtuse,
- X on argumenttunnuste ehk seletavate tunnuste ehk sõltumatute tunnuste vektor. Argumenttunnuseid võib mudelis olla mitu. Tuleb märkida, et mõned mudelid seavad kitsendusi argumenttunnuse tüübile ja jaotusele,
- β on mudeli parameetrite vektor,
- ϵ on juhuslike vigade vektor.

Praktikas on üks levinumaid statistilisi mudeleid lineaarne regressioonimudel. Lineaarset regressioonimudelit kasutatakse pideva tunnuse ennustamiseks arvuliste tunnuste kaudu. Mudelis võib kasutada ka mitte-arvulisi argumenttunnuseid, kuid nendele tuleb rakendada eeltötlust, mis viib nad arvulisele kujule. Selleks tötluseks võib olla kodeerimine või indikaatortunnuste loomine. Tulemuseks saame prognoosifunktsiooni $f : \mathbb{R}^k \rightarrow \mathbb{R}$.

Kirjanduses defineeritakse lineaarset regressioonimudelit vahel erinevalt. Käesolevas töös loeme regressioonimudeli lineaarseks, kui mudel $Y = f(X, \beta) + \epsilon$ täidab järgmised tingimused:

- uuritava tunnuse Y väärtused on pidevad;

- argumenttunnuste x_i , $i = 1, 2, \dots, n$ väärtused on arvulised;
- $f(X, \beta)$ on lineaarne parameetrite $\beta_0, \beta_1, \dots, \beta_k$ suhtes;
- mudeli parameetrid $\beta_0, \beta_1, \dots, \beta_k$ hinnatakse andmetelt vähimruutude meetodil.

Viimase kahe nõude olemust täpsustame peatselt. Oluline on aga mainida, et ülaltoodud definitsioon lubab meil kasutada argumenttunnuste x_i suhtes ka mittelineaarseid teisendusi, seni kuni prognoosifunktsioon $f(X, \beta)$ on jätkuvalt lineaarne parameetrite $\beta_0, \beta_1, \dots, \beta_k$ suhtes.

Regressioonimudeleid treenitakse andmestikel, kus võib olla tuhandeid või isegi miljoneid vaatlusi. Seetõttu on mugav kasutada mudeli töö kirjeldamiseks konkreetse andmemaatriksi korral maatriksnotatsiooni. Lineaarse regressiooni korral on see järgmine:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Siinkohal y_i tähistab i -nda vaatluse sihttunnuse väärtust, x_{ij} on i -nda vaatluse j -nda argumenttunnuse väärtus. Maatriksit $\mathbf{X}$ nimetatakse ka *disainimaatriksiks*. Maatriksnotatsioonis on mudeli töö seega esitatav kujul

$$y = \mathbf{X}\beta + \varepsilon.$$

Eelneva definitsiooni korral kehtib iga andmerea korral:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i,$$

kus β_0 nimetatakse vabaliikmeks, kordajaid $\beta_1, \dots, \beta_p$ regressioonikordajateks ning ϵ_i mudeli juhuslikuks veaks.

Mudeli jäägid e_i arvutatakse kujul

$$e_i = y_i - x_i \beta = y_i - \hat{y}_i,$$

kus x_i on argumenttunnuste maatriksi X i -s rida ning $\hat{y}_i$ tähistab mudeli prognoositud väärtust antud x_i ja parameetrivektori β korral. Mainime ka, et kui mudeli parameetrid on hinnatud vähimruutude meetodil ja mudelisse on kaasatud vabaliige β_0 , siis on mudeli jääkide keskvärtus 0 (Weisberg, 2014, lk 28).

Kuigi on mitmeid meetodeid kordajate suhtes lineaarsete prognoosifunktsioonide parameetrite hindamiseks, siis lineaarsest regressioonimudelist rääkides eeldatakse tavaliselt, et seda tehakse vähimruutude meetodil. Mitmed tulemused ja mõisted sõltuvad sellest eeldusest. Seega nõuame ka käesolevas töös parameetrite leidmist sellisel viisil. Vähimruutude meetod seisneb mudeli jääkide ruutude summade minimeerimises ehk lahendatakse parameetrite $\beta_0, \beta_1, \dots, \beta_k$ suhtes ekstreemumülesanne

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x_i \beta)^2.$$

Mainime ka ära, et jääkide ruutude summa minimeerimist võib vaadelda ka kui y ja $X\beta$ vahelise eukleidilise normi ruudu ehk $\|y - X\beta\|^2$ minimeerimist. Sellist kirjalpilti kasutame alapeatükis 3.2.

1.1.1 Olulised mõisted lineaarsete regressioonimodelite puhul

Modelleerimisel on tarvis mõõdupuud, mille alusel mudeleid võrrelda ja nende headust hinnata. Siinkohal tuleb lahti seletada mõned olulised mõisted, mis aitavad seda ülesannet täita.

Ennustavate mudelite puhul on oluline hinnata mudeli täpsust ka uutel andmetel. Selleks jagatakse modelleerimiseks kasutatav andmestik kaheks: esimest osa, *treeningandmeid*, kasutatakse mudeli parameetrite hindamiseks.

Teine osa, kõik ülejäänud andmed, hoitakse puutumata kuni modelleerimise lõppfaasini, kui kõik mudeli parameetrid on leitud ja soovitakse hinnata mudeli täpsust uutel andmetel. Seda osa nimetatakse *testandmeteks*.

Keskmine ruutviga (MSE) on laialdaselt kasutatav mõõdik prognoosivate regressioonimodelite ennustustäpsuse hindamiseks. Iga n vaatlusega andmestiku ja valitud regressiooni-

modeli korral arvutatakse see järgmise eeskirja järgi:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n},$$

kus y_i on tegelik väärtus ja $\hat{y}_i$ on mudeli poolt prognoositud väärtus.

Kuna MSE tulemus on ruutühikutes, mis raskendab selle sisulist interpreteerimist, defineeritakse sageli ka ruutkeskmise viga (RMSE, ingl *Root Mean Squared Error*):

$$RMSE = \sqrt{MSE}.$$

RMSE viib vea tagasi originaalsete andmete skaalale, muutes tulemuse tõlgendamise intuitiivsemaks. Tähtis on mõista, et ei MSE ega RMSE ei mõõda mudeli seletavust (erinevalt näiteks determinatsioonikordajast), vaid näitavad, kui suur on mudeli prognoosiviga keskmiselt.

Oluline on eristada, millisel andmestikul viga mõõdetakse. Kui mudeli lõplikuks hindamiseks kasutatakse täpselt sama andmestikku, millele mudel sobitati (treeningandmeid), siis on saadav veamõõdik sageli eksitavalt optimistlik. Treeningandmetel arvutatud MSE ei ole ka nihketa hinnang mudeli tegelikule ennustusveale, sest mudeli parameetrid on optimeeritud just selle konkreetse valimi põhjal.

Liialt treeningandmetele optimeerimine viib sageli nähtuseni, mida nimetatakse *ülesobitamiseks* (ingl *overfitting*), kus mudel hakkab lisaks tegelikule seosele õppima ka treeningandmetes sisalduvat juhuslikku müra. Sellisel juhul on MSE treeningandmetel väga madal, kuid uute andmete prognoosimisel teeb mudel suuri vigu.

Mudeli tegeliku ennustusvõime korrektseks hindamiseks MSE abil, tuleb see arvutada eraldiseisval testandmestikul, mida mudeli treenimisel ei kasutatud. Eeldusel, et prognoosivead on sõltumatud ja sama jaotusega (s.s.j.), annab just testandmetel arvutatud MSE meile nihketa hinnangu mudeli tegelikule keskmisele prognoosiveale uute andmete korral.

Ülesobitamisele on vastandiks *alasobitamine* (ingl *underfitting*). See kujutab endast olukorda, kus mudel on liiga lihtne, et püüda andmetes leiduvaid seoseid. Sellisel juhul on juba treeningandmetel mõõdetav viga suur. Näide sellest on klassikalise lineaarse regressiooni kasutamine mittelineaarse sõltuvuse modelleerimiseks, millest on juttu järgmises alapeatükis.

Statistilises modelleerimises on üks peamisi ülesandeid leida tasakaal üle- ja alasobitamise vahel. Praktikas tähendab see mudeli keerukuse valimist nii, et treeningandmetel mõõdetav viga (regressiooni puhul nt MSE või RMSE) oleks võimalikult väike, kuid seejuures vahe treening- ja testandmetel mõõdetud vigadel oleks samuti minimaalne. See tagab, et mudel suudab andmeid seletada, kuid ei õpi andmete müra. (Goodfellow *et al.*, 2016, lk 104)

1.2 Lineaarne regressioon mittelineaarse sõltuvuse korral

On üsna levinud väärarusaam, et lineaarseid regressioonimudeleid saab rakendada efektiivselt vaid siis, kui sihttunnuse ja seletavate tunnuste vahel eksisteeriv sõltuvus on lineaarne. See väide ei ole aga tõene. Nimelt, mittelineaarseid sõltuvusi saab edukalt modelleerida, kasutades sobivat eeltöötlust.

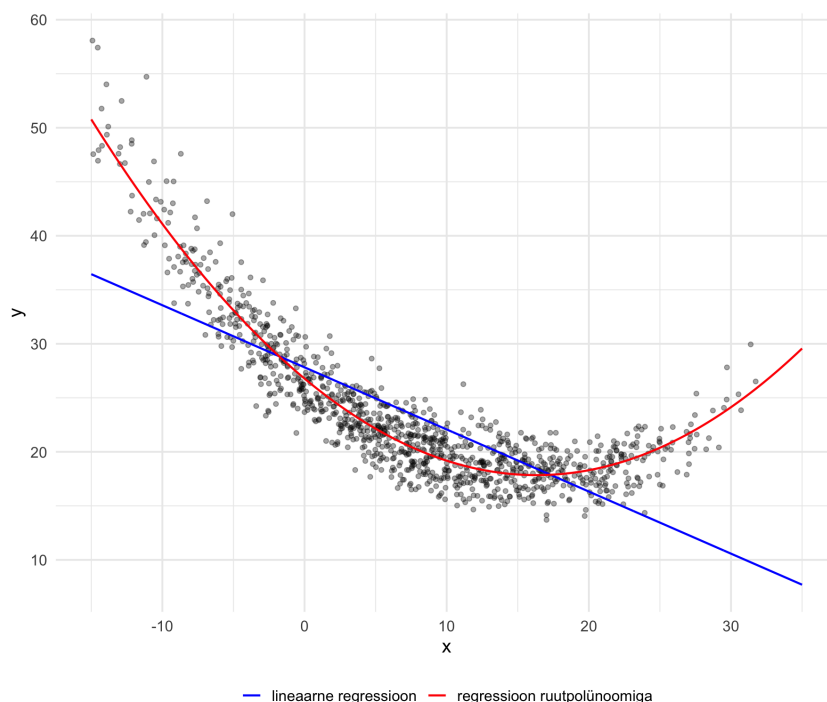
Üheks võimaluseks mittelineaarse käitumise modelleerimiseks on lisada mudelisse pidevate originaalsete argumentide astmeteisendusi või originaaltunnustest moodustatud polünoome. Säärane lähenemine võimaldab statistilise mudeliga püüda kõverat sõltuvust sihttunnuse ja seletava tunnuse vahel. Eksisteerib ka teisi transformatsioone, näiteks logaritmilise skaalaga muundamine, kuid neid me käesolevas töös ei käsitle.

Lihtsaim näide on ühe seletava muutuja korral ruutpolünoomi kasutamine ligikaudu U-kujulise seose modelleerimiseks. Sellisel juhul näeb funktsioon f välja järgmine:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Oluline on märkida, et säärane polünoomide kasutamine, olenemata polünoomi järgest, säilitab lineaarsuse regressioonikordajate $\beta_0, \beta_1, \beta_2, \dots$ suhtes (Weisberg, 2014, lk 57). Seega tegemist on endiselt lineaarse regressioonimudeliga.

Näidet polünoomi kasutamisest mittelineaarse sõltuvuse modelleerimiseks on näha joonisel 1. Joonisele on ka lisatud regressioonijoon, mis on saadud ilma eeltöötlust kasutamatta.

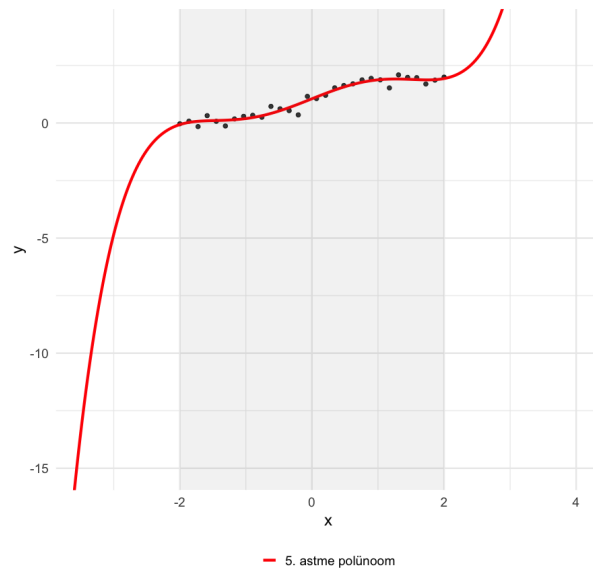


Joonis 1: Erinevus klassikalise ja polünoomi kasutava regressiooni vahel

1.3 Astmeteisenduste ja polünoomide puudujäägid

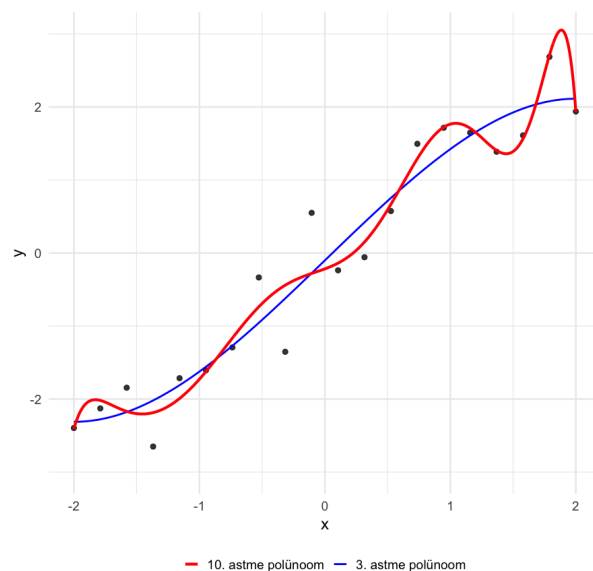
Nagu on näha jooniselt 1, siis on astmeteisendusi rakendades võimalik lineaarse regressiooniga modelleerida ka selgelt mittelineaarseid seoseid, saades seejuures mõistlikult häid tulemusi. Antud lähenemisel on aga mitu olulist puudujääki.

1. Kõrgema astme polünoomiga regressioonimudelid ekstrapoleerivad kehvasti. See tähendab, et väljaspool treeningandmestikku võivad mudeli vead äärmiselt kiiresti kasvada. See probleem eksisteerib juba klassikalise lineaarse regressiooni jaoks, kuid süveneb märgatavalt polünoomi järgu kasvades. Näidet sellest on näha jooniselt 2, kus regressioonikõvera järsk tõus ja langus väljaspool treeningvahemikku näitab, et mudel hakkab ekstrapoleerides tegema suuri vigu. (Hahn, 1977)



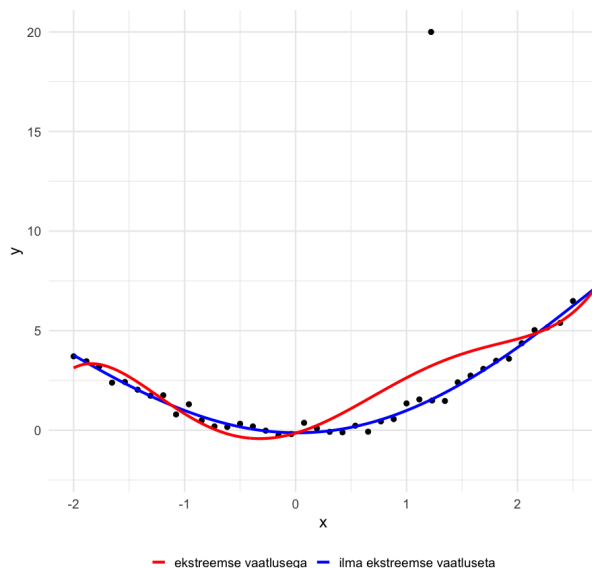
Joonis 2: Kujutus regressioonis kõrgema astme polünoomi rakendamisel tekkivast kehvast ekstrapolatsioonist. Halli varjundiga on tähistatud treeningandmete vahemik

2. Kõrgema astme polünoomide kasutamisel suureneb ka ülesobitamise oht, mis kasvab koos polünoomi järguga. Kõrgema astme polünoomid on väga paindlikud ning võivad hakata järgima ka andmetes esinevat juhuslikku müra. Näidet sellest on näha jooniselt 3. Tulemusena võib näiliselt paraneda mudeli sobivus treeningandmetel, kuid ennustusvõime uutel andmetel halveneb. (Hawkins, 2004)



Joonis 3: Polünoomi astme suurendamisest tingitud liigne müra sobitamine

3. Üksikutel ekstreemsetel vaatlustel võib olla ebaproportsionaalselt suur mõju regressioonikõverale, mis muudab kõvera kuju üle terve andmete vahemiku. Näidet sellest on näha joonisel 4, kus on võrdluseks esitatud kaks mudelit, mille polünoomide astmed on võrdsed, kuid üks mudel on sobitatud koos erindiga. Tulemusena on prognoosil märgatav nihe, ka üsna kaugel paiknevate x väärtuste korral.



Joonis 4: Ühe ekstreemaalse väärtuse mõju kogu regressioonikõverale

4. Klassikaliste polünoomide järku ei saa pidevalt muuta, kuna polünoomide astmed peavad olema mittenegatiivsed täisarvud. Seetõttu saab mudeli paindlikkust suurendada ainult diskreetsete sammudega ning üleminek ühelt järgult järgmisele võib olla liiga järsk. (Fan ja Gijbels, 1996, lk 3)

Kokkuvõtteks võib öelda, et lihtsate polünoomide kasutamisega regressioonis kaasneb mitmeid puudujääke. Üheks alternatiiviks on kasutada lokaalseid meetodeid, mille puhul mudel kohandub andmetele erinevates piirkondades erinevalt. See viib spline regressiooni ja üldistatud aditiivsete mudelite juurde. Need on ka peamised meetodid, mida käesolevas töös käsitleme, kuigi kirjanduses leidub ka teisi lähenemisi.

2 Splainid ja splinegressioon

Selles peatükis käsitleme splineide matemaatilist teooriat. Selgitame splineidega seotud terminoloogiat ning tutvustame nende olulisemaid omadusi statistilise modelleerimise aspektist, mis loovad tugeva aluse mittelineaarsete regressioonülesannete lahendamiseks.

2.1 Splainid

Olgu antud lõik $[a, b]$, kus $a, b \in \mathbb{R}$ ja $a < b$. Igal säärasel lõigul saame defineerida võrgu Δ :

$$\Delta = \{x_0, x_1, \dots, x_n : a = x_0 < x_1 < \dots < x_n = b\}.$$

Sedasi defineeritud võrk koosneb n osalõigust $[x_{i-1}, x_i]$, $i = 1, \dots, n$.

Definitsioon 1. Funktsiooni $S : [a, b] \rightarrow \mathbb{R}$ nimetatakse m astme splineiks defektiga k , $k = 1, \dots, m + 1$, kui on täidetud tingimused:

1. $S \in P^m[x_{i-1}, x_i]$, $i = 1, \dots, n$ ehk S on igal osalõigul ülimalt m astme polünoom,
2. $S \in C^{m-k}[a, b]$ ehk S on lõigul $[a, b]$ $m - k$ korda pidevalt diferentseeruv, kui $k < m + 1$.

Punkte $x_0, x_1, \dots, x_n$ nimetatakse splinei sõlmedeks ning punkte $x_1, x_2, \dots, x_{n-1}$ nimetatakse splinei sisesõlmedeks. Seega S on tükiti polünoom võrgu Δ suhtes: igal osalõigul $[x_{i-1}, x_i]$ langeb S kokku mõne ülimalt m -nda astme polünoomiga nii, et S tuletised kuni $m - k$ -nda järguni on pidevad kõikides sisesõlmedes.

Splinei S , mille aste on $m = 3$ ja defekt $k = 1$ ehk $S \in S^{3,1}$, nimetatakse *kuupsplineiks*. Definitsiooni kohaselt on S siis võrgu Δ igal osalõigul $[x_{i-1}, x_i]$ ülimalt kolmanda astme polünoom. Nõue, et $S \in C^2[a, b]$, tähendab, et funktsioonil S on lõigus $[a, b]$ pidev teist järku tuletis.

Naturaalseks kuupsplineiks nimetatakse kuupsplinei, mille teist järku tuletis lõigu $[a, b]$ otspunktides on null, s.t

$$S''(a) = S''(x_0) = 0 \quad \text{ja} \quad S''(b) = S''(x_n) = 0.$$

Kuigi teoreetiliselt on võimalik naturaalsplainide definitsiooni üldistada, siis statistilise modelleerimise vaatenurgast pole see vajalik. Seega piidrume antud töö kontekstis naturaalsete kuupsplainide defineerimisega.

Mainime ka, et splaine astmega $m = 1$ ja defektiga $k = 1$ nimetatakse *lineaarsplainideks*.

Splainidega on võimalik kirjeldada palju erinevaid funktsioone. Paindlikkuse suurendamiseks on kolm peamist võimalust:

- Sõlmpunktide arvu suurendamine. Jaotades lõigu $[a, b]$ rohkemateks osalõikudeks saame anda splainile lokaalset paindlikkust juurde, lubades funktsioonil kiiremini ja täpsemalt muutuda.
- Defekti k suurendamine. Splaini siledus sisesõlmedes on määratud sellega, et funktsioon peab olema $(m - k)$ korda pidevalt diferentseeruv. Defekti k suurendamisel loobutakse osadest tuletiste pidevuse nõuetest, mis lubab vabamat käitumist sõlmpunktide ümbruses. See lokaalne paindlikkus saavutatakse aga funktsiooni sileduse arvelt sisesõlmedes.
- Astme m tõstmine. See võimaldab kirjeldada igas osalõigus keerukamaid kõveraid. Regressioonimudelite kontekstis tuleb aga arvestada, et sarnaselt polünoomregressioonile võib astme m liigne tõstmine põhjustada ülesobitamist.

Praktikas, eriti ennustavate regressioonimudelite koostamisel, piirdutakse sageli astmega $m = 3$. Seetõttu on modelleerimisel laialdaselt kasutusel kuupsplainid ja naturaalsed kuupsplainid (Fan ja Gijbels, 1996, lk 40).

Hulk $S_{\Delta}^{m,k}$ on vektorruum, sest splainide liitmine ja skalaariga korrutamine annab tulemuseks ikka splaini. Ruumi $S_{\Delta}^{m,k}$ nimetatakse *splainide ruumiks*, dimensioon on $m + 1 + k(n - 1)$ juhul, kui võrgus Δ on n osalõiku (vt Oja, 2022, lk 11, kus on toodud väite tõestuse üldine skeem). Samal võrgul defineeritud naturaalsete kuupsplainide korral on ruum $3 + n - 2 = n + 1$ mõõtmeline.

2.2 Splainide ruum ja selle baas

Nagu eelnevalt mainitud, moodustavad splainid lõplikumõõtmelise vektorruumi. Lineaaralgebrast on teada, et igal mittetriviaalsel lõplikumõõtmelisel vektorruumil on baas, mille

elementide lineaarkombinatsioonina saab esitada ruumi kõiki vektoreid. Kuna vektorruumi baas ei ole reeglina üheselt määratud, on ka ruumile $S_{\Delta}^{m,k}$ võimalik leida mitmeid erinevaid baase.

Näiteks kui vaadelda võrku $\Delta := x_0 < x_1 < \dots < x_n$ ja ruumi $S_{\Delta}^{m,1}$, siis üheks võimalikuks baasiks on hulk $\{g_1, g_2, \dots, g_{n+m}\}$, kus:

$$\begin{aligned} g_i &= x^{i-1}, \quad i = 1, 2, \dots, m+1; \\ g_{m+1+j} &= (x - x_j)_+^m, \quad j = 1, 2, \dots, n-1. \end{aligned}$$

Siinkohal tähistab $(x - x_j)_+^m$ lõigatud astmefunktsiooni (ingl *truncated power function*):

$$(x - x_j)_+^m = \begin{cases} (x - x_j)^m, & x \geq x_j \\ 0, & x < x_j. \end{cases}$$

On ilmne, et selliselt moodustatud baas on $(n+m)$ -mõõtmeline, mis on kooskõlas varasema väitega, et splineide ruumi dimensioon on $(m+1) + k(n-1)$. Lihtne on veenduda, et funktsioonid $g_1, \dots, g_{n+m}$ on ka lineaarselt sõltumatud, seega moodustavad nad korrektselt defineeritud baasi ruumis $S_{\Delta}^{m,1}$. Selle baasi kaudu saab esitada kõikvõimalikke võrgul Δ defineeritud m -astme splaine defektiga 1.

Kuigi eelkirjeldatud baas on matemaatiliselt korrektne, ei ole see arvutuslikult kuigi efektiivne. Nimelt olukorras, kus $x \in [x_{n-1}, x_n]$, on kõik lõigatud astmefunktsioonid positiivsed, mistõttu tuleb splaini esitamiseks selles vahemikus kasutada peaaegu kõiki baasifunktsioone. Arvutusprotsessi optimeerimiseks oleks kasulik defineerida baas, mille elemendid on nullist erinevad võimalikult väikeses piirkonnas.

Definitsioon 2. Funktsiooni f kandjaks nimetatakse hulka $\overline{\{x \mid f(x) \neq 0\}}$.

Teisisõnu on funktsiooni kandja sellise hulga sulund, kus funktsiooni väärtused erinevad nullist. Seega, et lihtsustada splineide arvutamist, tuleks leida baas, mille funktsioonide kandjad on minimaalsed. Tuleb välja, et sellisteks funktsioonideks on B-splained.

2.3 B-splainid

B-splaine ehk baassplaine saab defineerida mitmel moel, käesolevas töös kasutame selle jaoks Cox-de Boori valemeid. Tänu oma numbrilisele stabiilsusele eelistatakse praktikas just neid valemeid B-splainide arvutamiseks (Boor, 2001, lk 109).

Olgu antud võrk $\Delta := x_0 < x_1 < \dots < x_n$ ja olgu polünoomi aste $m \geq 0$. Laiendame võrku Δ lisapunktidega ja tähistame

$$\Delta' := x_{-m} \leq \dots \leq x_{-1} \leq x_0 < x_1 < \dots < x_n \leq x_{n+1} \leq \dots \leq x_{n+m}.$$

Tähistagu B_m^i i -ndat m astme B-splaini. Defineerime B_m^i rekurentselt järgneva eeskirja järgi:

$$B_0^i = \begin{cases} 1 & x_i \leq x < x_{i+1} \\ 0 & \text{muidu} \end{cases} \quad x \in \mathbb{R}, \quad i = -m, \dots, n-1;$$

$$B_m^i = \frac{x - x_i}{x_{i+m} - x_i} B_{m-1}^i(x) + \frac{x_{i+m+1} - x}{x_{i+m+1} - x_{i+1}} B_{m-1}^{i+1}(x).$$

Vaatame lihtsustatud näidet lineaarsete B-splainide leidmisest ühtlasel võrgul $\Delta := \{0, 1, 2, 3, 4\}$. Märgime, et täieliku baasi moodustamiseks tuleks võrku laiendada täiendavate sõlmedega väljaspool vaadeldavat lõiku $[x_0, x_4]$, või kasutada korduvaid otspunkte. Kuna tegemist on hetkel vaid illustratiivse näitega Cox-de Boori eeskirjast, siis jätame selle sammu vahele.

Olgu meil 5 sõlmega võrk, kus $x_0 = 0, x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$. Konstantsetel B-splainidel ($m = 0$) on kuju:

$$B_0^0(x) = \begin{cases} 1, & x \in [0, 1] \\ 0, & \text{muidu} \end{cases}, \quad B_0^1(x) = \begin{cases} 1, & x \in [1, 2] \\ 0, & \text{muidu} \end{cases}$$

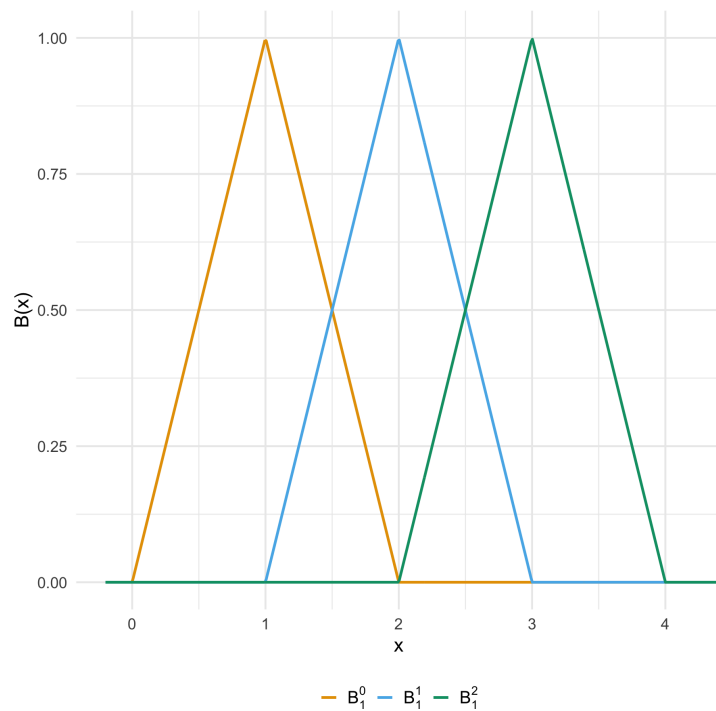
$$B_0^2(x) = \begin{cases} 1, & x \in [2, 3] \\ 0, & \text{muidu} \end{cases}, \quad B_0^3(x) = \begin{cases} 1, & x \in [3, 4] \\ 0, & \text{muidu.} \end{cases}$$

Kasutades Cox-de Boori valemit, saame lineaarsed B-splainid ($m = 1$):

$$B_1^0(x) = \frac{x-0}{1-0}B_0^0(x) + \frac{2-x}{2-1}B_0^1(x) = \begin{cases} x, & x \in [0, 1] \\ 2-x, & x \in [1, 2] \\ 0, & \text{muidu} \end{cases}$$

$$B_1^1(x) = \frac{x-1}{2-1}B_0^1(x) + \frac{3-x}{3-2}B_0^2(x) = \begin{cases} x-1, & x \in [1, 2] \\ 3-x, & x \in [2, 3] \\ 0, & \text{muidu} \end{cases}$$

$$B_1^2(x) = \frac{x-2}{3-2}B_0^2(x) + \frac{4-x}{4-3}B_0^3(x) = \begin{cases} x-2, & x \in [2, 3] \\ 4-x, & x \in [3, 4] \\ 0, & \text{muidu} \end{cases}$$



Joonis 5: Cox-de Boori valemitega leitud lineaarsed B-splainid

Jooniselt 5 on näha, et saadud lineaarsed B-splainid on nullist erinevad täpselt kahe osalõigu ulatuses. See ei ole juhuslik, vaid m -astme baassplainid on alati nullist erinevad ülimalt $m + 1$ järjestikusel osalõigul. Märkime täiendavalt, et baasfunktsioonid B_m^i on de-

fineeritud kõikjal reaalarvude hulgas $\mathbb{R}$. Tähistame kogu reaalteljel, võrgul Δ , defineeritud m -astme ja defektiga 1 splineide ruumi kui $S_{\Delta, \mathbb{R}}^{m,1}$.

Lause 2.1. *B-splineid on minimaalse kandjaga splineid hulka $S_{\Delta, \mathbb{R}}^{m,1}$ kuuluvate nullist erinevate splineide hulgas.*

Lause 2.1 tõestus on leitav allikast Oja, 2022, lk 21. Väitel on aga mitu praktilist järeldust. Eelkõige see, et B-splineide baas on splineide arvutamiseks optimaalne, mis muutub eriti oluliseks suure sõlmpunktide arvu korral. Näiteks N vaatluse ja K sõlmpunktiga vähimruutude meetodi arvutuste puhul, kus eesmärgiks on leida m -astme spline, on sorteeritud andmete korral võimalik saavutada ajaline keerukus $O(N)$ (Hastie *et al.*, 2009, lk 189).

Lause 2.2. *B-splineid B_m^i , $i = -m, \dots, n-1$, moodustavad baasi ruumis $S_{\Delta}^{m,1}[x_0, x_n]$.*

Lauset 2.2 me siinkohal ei tõesta, sest see läheb töö teemast kõrvale. Tõestus on aga leitav allikast Oja, 2022, lk 22. Täpsustame, et B-splineid sõltuvad ka võrgu Δ laiendi Δ' valikust, seega lause 2.2 kehtib eeldusel, et Δ' on välja valitud ja fikseeritud.

Järeldus 2.1. *Olgu võrk Δ ja selle laiend Δ' fikseeritud. Iga spline $S \in S_{\Delta}^{m,k}$ on üheselt esitatav kujul*

$$S(x) = \sum_{i=-m}^{n-1} \beta_i B_m^i(x), \quad x \in [x_0, x_n], \quad \beta_i \in \mathbb{R}. \quad (1)$$

Tegemist on vahetu järeldusega asjaolust, et B-splineid moodustavad vastava splineide ruumi baasi ning vektorruumi kõik elemendid on esitatavad baasivektorite ühese lineaarkombinatsioonina. Esituse (1) peamine eelis seisneb seega selles, et funktsiooni S lähendamine taandub vaid splineikordajate β_i leidmisele.

Märgime ka, et kuna B-splineid on minimaalse kandjaga, on fikseeritud x -i korral summas (1) ülimalt $m+1$ nullist erinevat liidetavat. Täpsemalt, kui $x \in [x_i, x_{i+1}]$, võivad nullist erineda vaid funktsioonide $B_m^{i-m}(x), \dots, B_m^i(x)$ väärtused. See asjaolu lihtsustab arvutusi märkimisväärselt, eriti arvestades, et praktikas enamasti ei kasutata modelleerimisel kõrgemat astet kui $m = 3$.

Lause 2.3. *Minimaalse kandjaga splineid ühtivad kordaja täpsuseni B-splineidega.*

Tõestus. Oletame, et meil on lõigul $[t_0, t_n]$ fikseeritud võrk $\Delta := x_0 = t_0 < t_1 < \dots < t_n = x_n$ ja selle laiend Δ' . Cox de Boori valemeid rakendades saame leida B-splineid $B_m^{-m}(x), \dots, B_m^{n-1}(x)$. Olgu $S(x)$ minimaalse kandjaga m astme spline defektiga 1.

Kuna B-splainid moodustavad baasi, siis saame $S(x)$ esitada nende lineaarkombinatsioonina:

$$S(x) = \sum_{i=-m}^{n-1} c_i B_m^i(x). \quad (2)$$

Kuna $S(x)$ on eelduse kohaselt minimaalse kandjaga, peab tema kandja olema täpselt $m+1$ osalõiku pikk. Olgu selleks kandjaks $[t_j, t_{j+m+1}]$, kusjuures üldisust kitsendamata, eeldame, et $j \geq 1$ ja $j \leq n-m-2$. Seega $S(x) = 0$, kui $x \notin (t_j, t_{j+m+1})$.

Vaatame 1. osalõiku $[t_0, t_1]$, eeldades lihtsuse huvides, et $t_0 < t_j$. Lõigul $[t_0, t_1]$ saavad nullist erinevad olla ainult need B-splainid, mille kandja ulatub siia lõiku. Neid on $m+1$ tükki: $B_m^{-m}, \dots, B_m^0$.

Kuna oleme hetkel väljaspool $S(x)$ kandjat, peab kehtima:

$$S(x) = c_{-m} B_m^{-m}(x) + c_{-m+1} B_m^{-m+1}(x) + \dots + c_0 B_m^0(x) = 0, \quad x \in [t_0, t_1].$$

Lause 2.2 kohaselt moodustavad $B_m^{-m}, \dots, B_m^0$ ruumi $S_{\{t_0, t_1\}}^{m,1}$ baasi. Märgime, et splineide definitiooni kohaselt on ühel osalõigul $[t_0, t_1]$ defineeritud splineide ruum lihtsalt kuni m astme polünoomide ruum sellel lõigul. Järelikult peavad $B_m^{-m}, \dots, B_m^0$ olema lineaarselt sõltumatud. Kui nüüd nende baasielementide lineaarkombinatsioon võrdub nulliga, siis ainuke variant on, et:

$$c_{-m} = c_{-m+1} = \dots = c_0 = 0.$$

Vaatleme järgmist lõiku $[t_1, t_2]$. Ka siin saavad korraga olla nullist erinevad vaid $m+1$ B-splaini, seega lisandub B_m^1 , kuid B_m^{-m} on samaselt null sellel lõigul. Kui me oleme endiselt väljaspool piirkonda $[t_j, t_{j+m+1}]$, siis järelikult ka $c_1 = 0$. Seda sammu saame korrata, kuni jõuame sõlmeni t_j , millest järeldeb, et

$$c_i = 0, \quad \text{kui } i < j.$$

Sisuliselt sama argumentatsiooni saame kasutada vaadates ka tagumisi osalõike, täpsemalt piirkonda $x \geq t_{j+m+1}$. Seal kehtib samuti $S(x) = 0$. Alustades viimasest osalõigust $[t_{n-1}, t_n]$, saame, et igal osalõigul kuni sõlmeni t_{j+m+1} peavad kordajad $c_i = 0$. Seega:

$$c_i = 0, \quad \text{kui } i > j.$$

Kokkuvõtteks jääb summast (2) alles ainult üks liidetav, kus $i = j$:

$$S(x) = c_j B_m^j(x).$$

Sellega on väide tõestatud. □

Lause 2.3 selgitab, miks võib kirjandusest leida erinevaid eeskirju B-splainide arvutamiseks. Näiteks diferentssuhete (ingl *divided differences*) kaudu tuletatud B-splainid on samuti minimaalse kandjaga ning ühtivad Cox-de Boori rekurentse eeskirja abil leitud baasfunktsioonidega kordaja täpsuseni.

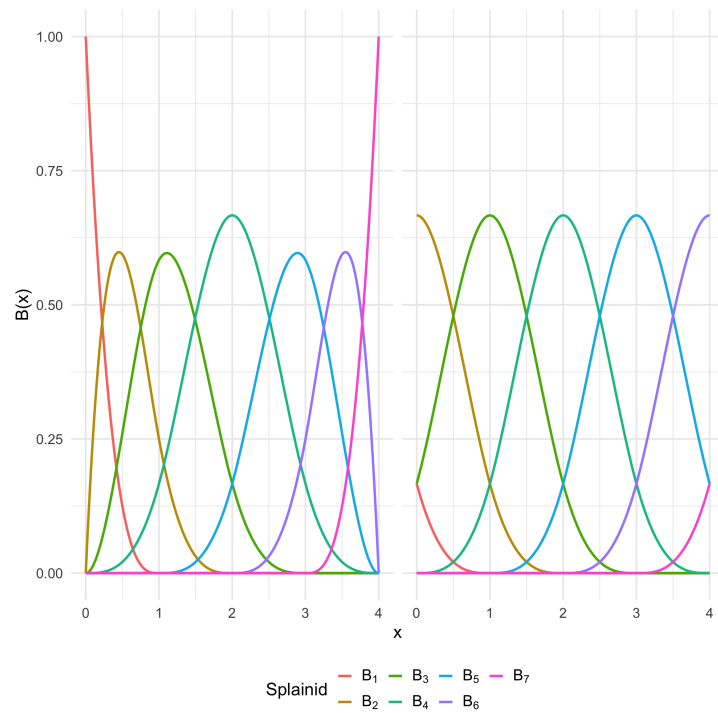
B-splainide definitsiooni kohaselt eeldame, et võrku Δ laiendatakse mõlemalt poolt m täiendava sõlmega. Praktikas kasutatakse aga võrgu laiendamiseks sageli korduvaids otspunkte (ingl *clamped knots*), kusjuures laiendatud võrk defineeritakse järgmiselt:

$$x_{-m} = \cdots = x_{-1} = x_0 \quad \text{ja} \quad x_n = x_{n+1} = \cdots = x_{n+m}.$$

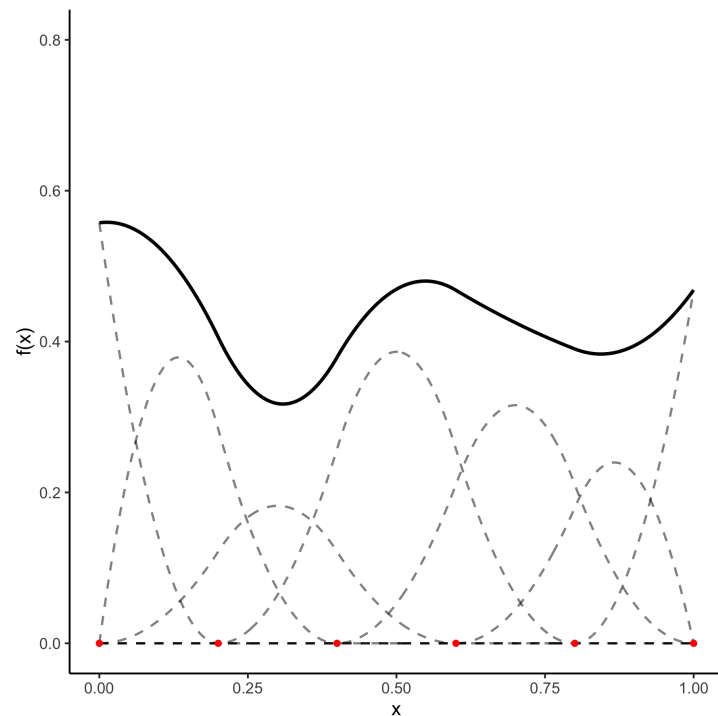
Kordsete otspunktide kasutamine tõstab neis sõlmedes B-splainide defekti. Mistõttu väheneb splineide siledus, kui neid vaadata kogu reaalteljel, sest need on vähem kordi kõikjal pidevalt diferentseeruvad. Küll aga säilib nõutud siledus vahemikus (x_0, x_n) .

Märgime, et korduvate sõlmede korral tekivad Cox-de Boori valemities mõned jagatised, kus nii lugeja kui nimetaja on nullid. Kokkuleppeliselt loetakse kõik sellised liikmed nulliks ($0/0 := 0$), seda järgivad ka standardsed tarkvarapaketid baassplainide arvutamisel.

Joonisel 6 on võrdlevalt kujutatud 3. astme B-splainid võrgul $\{0, 1, 2, 3, 4\}$, kus ühel juhul on kasutatud otspunktides korduvaids sõlmi ja teisel juhul mitte.



Joonis 6: Vasakpoolsel joonisel on kujutatud korduvate sõlmedega leitud B-splaine. Parempoolsel on laiendatud võrku punktidega $\{-3, -2, -1\}$ ja $\{5, 6, 7\}$



Joonis 7: Kujutus sellest, kuidas lineaarne kombinatsioon B-splainidest moodustab sujuva kõvera. Sõlmpunktid on märgitud punasega

Jooniselt 7 on näha, kuidas teist järku B-splainide lineaarkombinatsioon moodustab uue

splainkõvera. Keskmised kolm baassplaini on ootuspäraselt nullist erinevad vaid nelja sõlme ulatuses, äärmised on korduvate sõlmede kasutuse tõttu suurema defektiga. Näha on ka seda, et juba ruutsplainide ($m = 2$) korral on saadud kõver niivõrd sujuv, et silmaga on üsna keeruline näha üleminekuid sõlmpunktides. Väidetakse, et kuupsplainide korral on see inimsilmale juba praktiliselt võimatu (Hastie *et al.*, 2009, lk 144).

2.3.1 Vähimruutude meetod splineide leidmiseks

Meenutame järeldust 2.1, mille kohaselt esituvad kõik splineid üheselt B-splineide lineaarkombinatsioonina, eeldusel, et võrk ja selle laiend on fikseeritud. Oletame, et soovime prognoosida sihttunnust Y argumenttunnuse X põhjal ning andmetes esineb mittelineaarne seos, mida tahame modelleerida splineide abil. Oletame, et sõlmpunktid on eelnevalt valitud ja need moodustavad võrgu Δ , millel on n osalõiku. (Järgmises peatükis tutvustame, kuidas sõlmpunkte praktikas paigutatakse ning kuidas seda protsessi automatiseerida.)

Tulemuseks saame mudeli:

$$Y = S(X) + \varepsilon.$$

Etteantud andmestiku $\{(x_i, y_i)\}_{i=1}^n$ korral saame, et:

$$y_i = S(x_i) + \epsilon_i.$$

Järelduse 2.1 põhjal võime mudeli esitada kujul:

$$y_i = \sum_{j=-m}^{n-1} \beta_j B_m^j(x_i) + \varepsilon_i. \quad (3)$$

Kordajate vektori $\boldsymbol{\beta} = (\beta_{-m}, \dots, \beta_{n-1})^T$ saame nüüd leida vähimruutude meetodil, lahendades ekstreemumülesande:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^N \left(y_i - \sum_{j=-m}^{n-1} \beta_j B_m^j(x_i) \right)^2.$$

Seega vastab mudel kujul (3) kõigile I peatükis defineeritud lineaarse regressioonimudeli eeldustele. Kuigi selline mudel üksi ei lahenda veel kõiki eelmises peatükis mainitud polünoomregressiooni puudujääke, annab see üsna kasuliku alusraamistiku. Enne võimsamate

meetodite juurde asumist pöörame pilgu splineidega ekstrapoleerimisele.

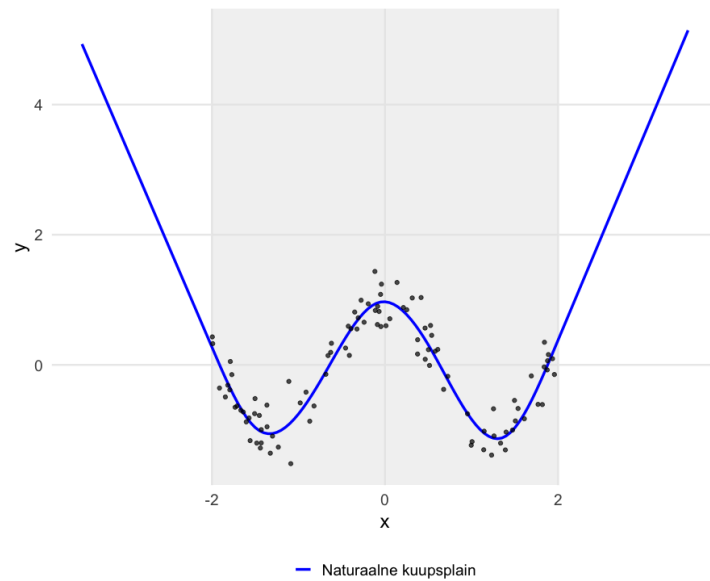
2.3.2 Ekstrapolatsioon

Meenutame, et üks probleeme polünoomide kasutamisega lineaarses regressioonis on ekstrapoleerimine: mudeli vead võivad treeningandmete piirkonnast väljutes kiiresti kasvada. Spline kasutades tuleb samuti mõelda, kuidas arvutada funktsiooni väärtusi väljaspool esialgset sõlmede võrku $[x_0, x_n]$.

Kuigi B-splineid on formaalselt defineeritud kogu reaalteljel, muutuvad nad otspunktides asuvate korduvsõlmede korral piirkonnast väljutes nulliks. Seetõttu annab nende lineaarkombinatsioon väljaspool seda vahemikku samuti nulli, mis ei ole enamasti mõistlik prognoos. Seega jätkatakse ekstrapoleerimisel splinei esimesel ja viimasel osalõigul defineeritud polünoomidega (Boor, 2001, lk 69). Tavaliste kuupsplineide korral tähendab see ekstrapoleerimist kuupfunktsiooniga, mis säilitab ohu prognoosivigade suureks kasvuks.

Naturaalsete kuupsplineidega saab seda probleemi maandada, sest nende puhul toimub ekstrapolatsioon lineaarsete funktsioonidega. See tuleneb naturaalsele kuupsplineidele seatud rajatingimustest, mis nõuavad, et teist järku tuletised otspunktides võrduksid nulliga. Kuna lineaarse funktsiooni teine tuletis on samuti null, võimaldab see rajatingimus jätkata splinei väljaspool lõiku $[x_0, x_n]$ lineaarsete funktsioonidega nii, et tulemuseks on C^2 pidev funktsioon kogu reaalteljel.

Seetõttu ühtib lineaarse jätku kalle viimases sõlmpunktis kuupsplinei kaldega, tagades sujuva ja murdumisteta ülemineku. Kuna me modelleerides tegelikkuses ei tea, mis juhtub väljaspool treeningandmeid, on üldjuhul kõige turvalisem eeldada, et trend jätkub lineaarselt. Näidet ekstrapolatsioonist on näha joonisel 8.



Joonis 8: Ekstrapolatsioon naturaalse kuupsplainiga. Treeningvahemik on märgitud halliga

Lineaarse funktsiooniga jätkamine on loomulik ka teoreetilisest vaatepunktist, kuna see vastab kogu reaalteljel sellise funktsiooni otsimisele, mis minimeerib teise tuletise ruudu integraaliga trahvitud vähimruutude kriteeriumit. Et kriteeriumis karistusliige üle kogu reaaltelje oleks lõplik, peab splaini teine tuletis väljaspool andmepunkte olema samaselt null. See ei jäta muud võimalust, kui jätkata prognoosikõverat väljaspool andmete vahemikku lineaarsete funktsioonidega. Säärasest trahviliikmest ning sellest tulenevast funktsiooni silumisest räägime lähemalt järgmises peatükis.

3 Üldistatud aditiivsed mudelid

Selles peatükis käsitleme karistusliikmega splaine ja üldistatud aditiivseid mudeleid ning nendega seonduvat terminoloogiat.

3.1 Siluvad splainid

Regressioonülesannete puhul on statistilise modelleerimise eesmärgiks lähendada tõelist, ent tundmatut seosefunktsiooni andmete põhjal hinnatud prognoosifunktsiooniga. Eeldatakse enamasti, et uuritavaid protsesse kirjeldav teoreetiline funktsioon on sile (pidev ja mitmeid kordi diferentseeruv). Seda seetõttu, et naturaalselt kujunevates seostes ei esine reeglina järsku võnkumist, vaid see on pigem tingitud andmetega kaasnevast müra. Sellest tulenevalt võiks olla üsna kasulik, kui ka hinnatav lähendusfunktsioon on adekvaatse siledusastmega, et vältida mudeli ülesobitamist.¹

Ülesobitamine tekib splainide kontekstis tüüpiliselt olukorras, kus kasutatakse väga tihedat sõlmede võrku. Kui näiteks iga andmepunkt määrata sõlmpunktiks (põhjendame peatüki lõpu poole, miks nii üldse teha), võivad isegi madala astme splainid hakata liigselt võnkuma. Tulemuseks võib splainfunktsioon hakata andmeid pigem interpoleerima ja oma kujult meenutama esimeses peatükis esitatud joonist 3. Selline käitumine pole enamasti kasulik, sest see pärsib mudeli üldistusvõimet uutel andmetel. Antud probleemi üheks lahenduseks on siluvad splainid (ingl *smoothing splines*), mis võimaldavad leida optimaalse tasakaalu andmetega sobivuse ja funktsiooni sileduse vahel.

Kui soovime funktsioonide siledust formaalselt käsitleda, tuleb defineerida vastav matemaatiline mõõdik. Üks võimalus siledust analüüsida on vaadelda selle puudumist, defineerides funktsiooni karedust (ingl *roughness*) hindava trahvifunktsiooni. Kuigi karedust saab defineerida mitmeti, mõõdame käesolevas töös funktsiooni f karedust järgmise avaldisega:

$$K(f) = \int_{\mathbb{R}} (f''(x))^2 dx \quad (4)$$

Loogika selle taga on üsna intuitiivne. Funktsiooni teine tuletis näitab funktsiooni tõusu muutumise kiirust ning on seetõttu seotud funktsiooni kõverusega punktis – mida järsem

¹Märgime, et tegemist pole raudse reegluga. Eksisteerib ka teisi väga efektiivseid ennustusmudeleid, näiteks otsustuspuudel põhinevaid, kus prognoosifunktsiooni ei saa siledaks pidada.

on paine, seda suurem on ka teise tuletise absoluutväärtus. Teise tuletise ruutu tõstmine täidab kahte eesmärki. Esiteks muudab see kõik väärtused mittenegatiivseks. Teiseks karistab see tugevaid paindeid ebaproportsionaalselt rohkem kui väikese kõverusega paindeid. Viimaks toimib integraal pideva summana, koondades kõik need karistused üle kogu vaadeldava lõigu.

Teoreem 1. *Olgu $S(x)$ andmepunkte $\{(x_i, y_i)\}_{i=0}^n$ interpoleeriv naturaalne kuupsplain sõlmedega x_i , mis on väljaspool andmevahemikku jätkatud lineaarsete funktsioonidega kogu reaalteljele. Kõikidest pidevatest funktsioonidest, millel on absoluutselt pidev esimene tuletis ja mis interpoleerivad andmepunkte $\{x_i, y_i\}$, on K minimeerimise mõistes siledaimaks funktsioon $S(x)$, kusjuures $S(x)$ on üheselt määratud.*

Seda teoreemi me siinkohal ei tõesta, kuid detailse tõestuse leiab allikast Green ja Silverman, 1993, lk 16. Kuigi me ei soovi statistiliste mudelite puhul enamasti tegeleda interpolatsiooniga, siis sellest hoolimata järeldub teoreemist 1, et kõiki unikaalseid andmepunkte sõlmedena kasutavad naturaalsed kuupsplainid on heaks aluseks siledate prognoosifunktsioonide konstrueerimisel.

Olgu antud andmepunktid $\{(x_i, y_i)\}_{i=0}^n$. Eeldame, et need on järjestatud kasvavalt x -i suhtes. Kui soovime nüüd I peatükis toodud vähimruutude meetodil splainfunktsiooni andmetele sobitada, siis sisuliselt üritame leida funktsiooni S , mis minimeeriks avaldist

$$\sum_{i=0}^n (y_i - S(x_i))^2.$$

Sellise lähenemise tulemuseks on alati andmeid võimalikult täpselt interpoleeriv funktsioon. Nagu eelnevalt mainitud, siis ei ole statistiliste mudelite puhul see enamasti eesmärgiks, sest täpse interpoleerimisega kaasneks paratamatult ka andmetes leiduva juhusliku müra modelleerimine ehk ülesobitamine. Et seda vältida, defineerime uue minimeerimiskriteeriumi:

$$\sum_{i=0}^n (y_i - S(x_i))^2 + \lambda \int_{\mathbb{R}} S''(x)^2 dx. \quad (5)$$

Seda kriteeriumi minimeerivat funktsiooni $S(x)$ nimetatakse *siluvaks splainiks*. Avaldisel (5) on kolm peamist komponenti:

- Esimene osa ehk $\sum_{i=0}^n (y_i - S(x_i))^2$ tähistab hälvete ruutude summat, mis on tun-

tud juba klassikalisest lineaarsest regressioonist. Selle eesmärk on tagada prognoosifunktsiooni võimalikult täpne sobitumine andmetega.

- Teine osa ehk $\int_{\mathbb{R}} (S''(x))^2 dx$ kujutab endast kareduse trahvi. Kuna naturaalsplaini teine tuletis on väljaspool andmepunkte null, toimub integraali põhiline arvutamine endiselt lõigul $[x_0, x_n]$.
- Kolmas osa ehk λ on *silumisparameeter*. See parameeter omandab mittenegatiivseid väärtusi ($\lambda \geq 0$). On ilmne, et kui $\lambda = 0$ on tegu klassikalise vähimruutude meetodiga. Samas, kui $\lambda \rightarrow \infty$, on funktsioon $S(x)$ sunnitud võnkumisest täielikult loobuma ning taandub sisuliselt sirgjooneliseks lineaarseks prognoosifunktsiooniks.

Põhjendame ka ära, miks kriteeriumit (5) minimeeriv funktsioon on just naturaalne kuupsplain, mis on jätkatud lineaarselt kogu reaalteljele.

Olgu meil järjestatud andmepunktid $x_0 < x_1 < \dots < x_n$ ja olgu kriteeriumit (5) minimeerivaks funktsiooniks S . Konstrueerime naturaalse kuupsplaini S_n , mis ühtib sõlmedes funktsiooniga S (ehk interpoleerib S väärtusi punktides x_i , nii et $S_n(x_i) = S(x_i)$) ja on väljaspool andmevahemikku jätkatud lineaarselt.

Vaatame nüüd kriteeriumi (5) väärtust uue funktsiooni S_n korral:

- Avaldise esimene liige (hälvete ruutude summa) on funktsioonide S_n ja S korral täpselt sama, sest andmepunktides langevad nende väärtused kokku;
- Avaldise teine liige $\int_{\mathbb{R}} (S''(x))^2 dx$ on teoreemi (1) kohaselt funktsiooni S_n korral väiksem või võrdne funktsiooni S vastava integraaliga.

Mistõttu saab kriteeriumi (5) koguväärtus funktsiooni S_n korral olla vaid sama suur või väiksem kui avaldist minimeeriva funktsiooni S korral. Järelikult kriteeriumit (5) minimeeriv funktsioon on naturaalne kuupsplain, mis on jätkatud lineaarsete funktsioonidega kogu reaalteljele.

Avaldises (5), kontrollib seega parameeter λ tasakaalu mudeli ülesobitamise ja liiga lihtsast (lineaarsest) mudelist tingitud alasobitamise vahel, olles siluvate splainide puhul peamiseks tuunimisparameetriks.

Iga unikaalse andmepunkti sõlmena kasutamisel on veel üks praktiline lisaeelis. Nimelt see võimaldab modelleerijal mudeli sobitamist sisuliselt automatiseerida, sest tal pole enam

tarvis sõlmpunktide valimisega vaeva näha ning võimalik ülesobitamine lahendatakse automaatselt trahvifunktsiooniga $K(f)$. Kogu protsess taandub vaid silumisparameetri λ valimisele, mille leidmist on omakorda võimalik näiteks ristvalideerimise kaudu delegeerida arvutile. See on ka põhjus, miks siluvate splineide rakendamist peetakse üheks masinõppe meetodiks.

Säärase peene võrgu kasutamise peamiseks puuduseks on aga see, et sõlmede arv kasvab proportsionaalselt vaatluste arvuga. Seega võib mudeli sobitamine muutuda arvutuslikult üsna kulukaks, eriti just suurte andmemahtude puhul. Üks viis selle probleemi maandamiseks avaldub P-splineide näol.

3.2 P-splainid

P-splainid (ingl *Penalized B-splines*) on paindlik silumismeetod, mis kombineerib kaks põhilist ideed. Esiteks kasutatakse regressiooniks suurt hulka ühtlaste vahedega võrgul defineeritud B-splaine. Teiseks nendele B-splainidele, täpsemalt nende kordajatele, rakendatakse diskreetne karedustrahv.

Olgu antud vaatlusandmed, kus argumenttunnuste väärtused on $x_1, x_2, \dots, x_N$ ning jagagu x -telge ühtlane võrk $\Delta := t_0 < t_1 < \dots < t_n$. Lähtudes m -astme B-splineide definitsioonist, saame defineerida $N \times (n + m)$ mõõtmelise baasimaatriksi $\mathbf{B}$ järgmiselt:

$$\mathbf{B} = \begin{pmatrix} B_m^{-m}(x_1) & B_m^{-m+1}(x_1) & \dots & B_m^{n-1}(x_1) \\ B_m^{-m}(x_2) & B_m^{-m+1}(x_2) & \dots & B_m^{n-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_m^{-m}(x_N) & B_m^{-m+1}(x_N) & \dots & B_m^{n-1}(x_N) \end{pmatrix}. \quad (6)$$

Maatriksis (6) vastab iga rida ühele konkreetsele vaatluspunktile x_i ($i = 1, \dots, N$) ja iga veerg ühele baasifunktsioonile B_m^j ($j = -m, \dots, n - 1$). Seega maatriksi $\mathbf{B}$ iga üksik element tähistab vastava baasfunktsiooni väärtust kindlas vaatluspunkti. Oluline on silmas pidada, et P-splainid eeldavad ühtlast sõlmede võrku. Seetõttu ei kasutata võrgu laiendamiseks korduvaid otspunkte, vaid täiendavad sõlmed lisatakse ühtlase sammuga väljapoole andmete piirkonda (Eilers ja Marx, 2021, lk 132).

Statistilise mudeli saame maatrikskujul esitada järgmiselt:

$$Y = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

kus $\boldsymbol{\beta}$ on baasfunktsioonide kordajate vektor pikkusega $n + m$.

Klassikalise vähimruutude meetodi puhul leiame kordajate vektori hinnangu $\hat{\boldsymbol{\beta}}$, kui minimeerime $\|Y - \boldsymbol{\beta}\mathbf{B}\|^2$. P-splainide minimeerimiskriteeriumiga kaasneb aga lisaks ka trahviliige, mille eesmärk on karistada suuri võnkumisi $\boldsymbol{\beta}$ elementide vahel.

Intuitsioon seal taga seisneb selles, et kuna B-splainide lineaarnekombinatsioon moodustab uue sujuva kõvera, siis selle kuju on defineeritud kordajate β_i poolt. Kui nüüd lähetikku olevate B-splainide kordajad tugevalt võnguvad, siis hakkab ka graafikul esinema ka võnkumist, mis on teadupärast see käitumine, mida me modelleerides vältida tahame. Kui selliseid võnkeid karistada, siis saame tulemuseks siledama prognoosifunktsiooni. Seda ülesannet täitvat trahviliiget nimetame *diferentsitrahviks*.

P-splainide trahvifunktsiooni aluseks olevad diferentsid on defineeritud järgmiste eeskirjade kaudu, kus Δ tähistab diferentsioperaatorit:

- Esimest järku diferents: $\Delta\beta_j = \beta_j - \beta_{j-1}$,
- Teist järku diferents: $\Delta^2\beta_j = \Delta(\beta_j - \beta_{j-1}) = \Delta\beta_j - \Delta\beta_{j-1} = (\beta_j - \beta_{j-1}) - (\beta_{j-1} - \beta_{j-2}) = \beta_j - 2\beta_{j-1} + \beta_{j-2}$,
- Üldine k -ndat järku diferents $\Delta^k\beta_j$ saadakse, rakendades operaatorit Δ kordajale β_j järjestikku k korda.

Märgime, et $\Delta^k\beta_j$ ei eksisteeri, kui $j < 1 + k$.

Ka diferentstrahvide puhul on mugav edaspidi kasutada maatriksnotatsiooni, mis on lahti seletatud lisas 1.

Diferentside kaudu saame nüüd defineerida uue kareduse mõõdiku $R(\boldsymbol{\beta})$:

$$R(\boldsymbol{\beta}) = \sum_{j=1+k}^{n+m} (\Delta^k\beta_j)^2 = \boldsymbol{\beta}^T D_k^T D_k \boldsymbol{\beta} = \|D_k \boldsymbol{\beta}\|^2. \quad (7)$$

Eelnevale toetudes defineerime P-splainide minimeerimiskriteeriumi maatrikskujul järg-

miselt:

$$||\mathbf{y} - \mathbf{B}\boldsymbol{\beta}||^2 + \lambda ||D_k \boldsymbol{\beta}||^2. \quad (8)$$

Sarnaselt eelmises alapeatükis käsitletud siluvatele splineidele sõltub ka kriteerium (8) mittenegatiivsest silumisparameetrist λ . Parameetrit λ muutes saame kontrollida ühtlasel võrgul defineeritud B-splineide lineaarkombinatsiooni siledust. Kui $\lambda = 0$, siis eraldi silumist ei toimu ning prognoosifunktsiooni lähendamiseks rakendatakse standardset vähimruutude meetodit.

P-splineide meetodi puhul soovitatakse vaikimisi kasutada 50 osalõiguga võrku, millest peaks enamike praktiliste ülesannete puhul olema rohkem kui küllalt (Eilers ja Marx, 2021, lk 15). Oluline on aga mainida, et P-splineide puhul on sisuliselt võimatu kasutada liiga palju sõlmi.

Mida tihedamat võrku kasutada, seda suurem on baasifunktsioonide arv. Trahvimata B-splineide puhul põhjustab liigne sõlmede hulk tihti ülesobitamist, sest tekkiv prognoosifunktsioon on äärmiselt paindlik ning vähimruutude meetod viib andmeid interpoleeriva kõverani. P-splineide puhul on see oht välditud. Selleks, et nende lineaarkombinatsioon saaks tugevalt võnkuda ning hakata müra püüdma, peab esinema võnkumine juba korrajate vektori $\boldsymbol{\beta}$ elementide vahel. Seda takistab aga takistab karistusliige $\lambda ||D_k \boldsymbol{\beta}||^2$, sõltumata osalõikude arvust.

Tasub ka märkida, et kuna P-splineide puhul rakendatakse ühtlast sõlmede võrku, on ekstreemumülesandes (8) teist järku diferentsitrahvi kasutamine mõnes mõttes diskreetne analoog klassikalisele teise tuletise ruudu integraali ehk eelnevalt defineeritud trahvifunktsiooni $K(f)$ kasutamisele.

See tuleneb sellest, et m -astme B-splineide tuletised avalduvad astme võrra madalamate ehk $(m - 1)$ astme B-splineide kaudu (Boor, 2001, lk 116).

Ühtlasel võrgul, kus sõlmede vahe on h , saame seega tuletiseks:

$$\frac{d}{dx} \sum_{j=-m+1}^{n-1} \beta_j B_m^j(x) = \frac{1}{h} \sum_{j=-m+1}^{n-1} (\beta_j - \beta_{j-1}) B_{m-1}^j(x).$$

Selle teine tuletis avaldub omakorda kujul:

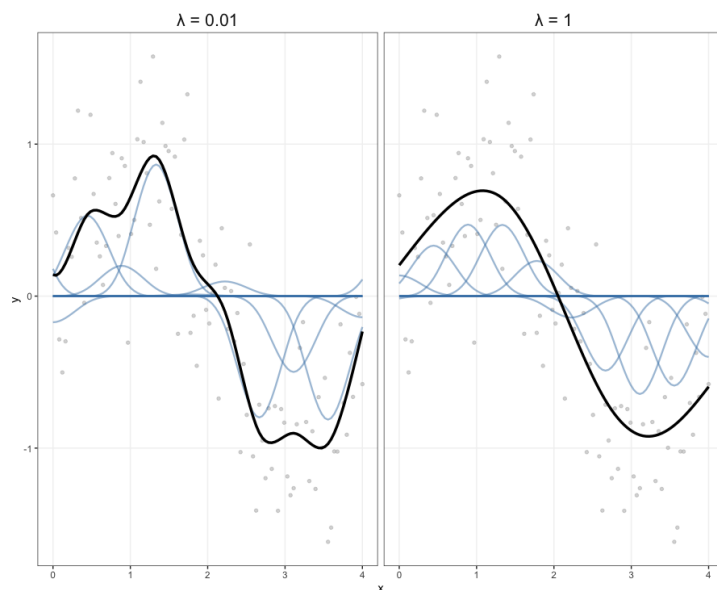
$$\frac{1}{h^2} \sum_{j=-m+2}^{n-1} (\Delta\beta_j - \Delta\beta_{j-1}) B_{m-2}^j(x) = \frac{1}{h^2} \sum_{j=-m+2}^{n-1} \Delta^2\beta_j B_{m-2}^j(x).$$

Niisiis, paneme tähele, et splainkõvera teine tuletis avaldub teist järku diferentside kaudu. Kuna sõlmede võrk ja splainide aste on eelnevalt fikseeritud, on baasifunktsioonid $B_{m-2}^j(x)$ ning seega ka nende väärtused iga x korral üheselt määratud. See tähendab, et teise tuletise suuruse määravad sisuliselt $\Delta^2\beta_j$.

Seega, kui P-splainide puhul minimeeritakse avaldist $\|D_2\boldsymbol{\beta}\|^2 = \sum_{j=3}^{n+m} (\Delta^2\beta_j)^2$, siis nõutakse ühtlasi, et ka kogu splainkõvera teise tuletise ruut oleks väike, mis on mõnes mõttes sarnane $K(f)$ karistusliikme kasutamisele.

Järelikult dikteerib mudeli tegelikku paindlikkust taaskord peamiselt parameeter λ , kuid mitte sõlmede koguarv. See garanteerib, et isegi suure arvu baasifunktsioonide korral püsib kõver sobiva λ väärtuse juures siledana ning liigsete sõlmede kasutamise peamiseks miinuseks jääb vaid arvutusliku keerukuse mõningane kasv. Trahviikmete $K(f)$ ja $R(\boldsymbol{\beta})$ vahel on võimalik tuua veelgi sügavamaid matemaatilisi seoseid, kuid see jääb käesoleva töö skoobist välja.

Joonisel 9 on näha, kuidas parameetri λ muutmine mõjutab B-splainide kordajaid ja see läbi ka nende lineaarkombinatsiooni kuju. Kui λ on väike, siis on ekstreemumülesandest (8) leitud kordajad vägagi varieeruvad, ning kordajaga β_j skaleeritud B-splainid käituvad üsna metsikult. Tulemuseks on selge ülesobitamine andmetele. Silumisparameetrit suurendades on aga näha, kuidas kordajate erinevusi on trahvitud, mistõttu võtavad skaleeritud B-splainid korrapärasema kuju, andes tulemuseks siledama kõvera.



Joonis 9: Silumisparameetri mõju P-splainidele. Helesinisega on tähistatud skaleeritud B-splainid ning mustaga nende lineaarkombinatsioon

P-splainidel on oma olemuse tõttu mitmeid väga häid praktilisi omadusi:

- Puudub vajadus käsitsi valida sõlmpunktide asukohta, kuna kõik sõlmed paiknevad ühtlasel võrgul. Samuti puudub vajadus sõlmpunktide arvu täpselt optimeerimiseks, sest tänu trahviliikmele ei too liiga tiheda võrgu kasutamine kaasa ülesobitamist.
- Trahv ja baasifunktsioonid on üksteisest täielikult eraldatud. Kui $K(f)$ karistusliige põhines teisel tuletisel ja nõudis seetõttu, et baasi moodustavate B-splainide aste oleks vähemalt 3, siis P-splainide puhul sellist piirangut ei ole. Seega saab vabalt kasutada näiteks kolmandat järku diferentsitrahvi 1. astme B-splainide baasil. (Eilers ja Marx, 2021, lk 3)
- P-splainid on arvutuslikult väga efektiivsed. See tuleneb B-splainide minimaalse kandja omadusest ja diskreetsest diferentsitrahvist, mistõttu on arvutustes kasutatavad maatriksid $\mathbf{B}$ ja D_k hõredad (sisaldavad palju nulle). See kiirendab oluliselt mudeli sobitamist. P-splainidega silumine on seetõttu üldjuhul märgatavalt kiirem kui alapeatükis 3.1 arutatud klassikaliste siluvate splineide meetod (Eilers ja Marx, 2021, lk 162).

P-splainidel on lisaks veel mitmeid häid omadusi, mis jäävad käesoleva töö skoobist välja, kuid on kirjeldatud allikas (Eilers ja Marx, 2021).

3.3 GAM-meetod

GAM (ingl *generalized additive model*) ehk üldistatud aditiivne mudel on statistiline raamistik, mis võtab kogu eelnevalt käsitletu kokku. Ülevaade meetodi praktilisest rakendamisest tarkvara R abil on leitav lisas 2.

Sõna „üldistatud“ tähendab siin kontekstis sisuliselt sedasama, mida üldistatud lineaarsete mudelite (GLM) puhul. Nimelt ei ole me GAM-meetodit rakendades piiratud vaid ühe kindla jaotuse eeldusega, mistõttu saame mudeli abil analüüsida näiteks binaarseid andmeid ja loendusandmeid, aga lahendada ka tavalisi regressiooniülesandeid. Kuna käesoleva töö fookus on aga ainult andmete suhtes mittelineaarsete regressiooniülesannete lahendamisel, läheneme üldistatud aditiivsetele mudelitele edaspidises käsitluses vaid sellest vaatepunktist.

Üldistatud aditiivsete mudelite idee on sobitada lineaarne mudel, kus regressioonikordajad asendatakse siledate funktsioonidega. Regressiooni GAM-i üldkuju avaldub järgmiselt:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \quad (9)$$

Kus:

- y_i on sihttunnuse väärtus i -ndal vaatlusel,
- β_0 on mudeli vabaliige,
- x_{ip} on p -nda seletava tunnuse väärtus i -ndal vaatlusel,
- $f_p(\cdot)$ on sile funktsioon p -nda seletava tunnuse jaoks,
- ϵ_i on mudeli juhuslik viga.

Seega peamine erinevus I peatükis defineeritud lineaarse regressiooni mudeliga seisneb selles, et GAM-meetod modelleerib iga tunnuse mõju paindlikku siluvat funktsiooni f_j kasutades. Sellest tuleneb ka sõna „aditiivne“ mudeli nimes, sest iga seletava tunnuse mõju - olgu ta lineaarne või mitte - liidetakse vabaliikmele, mille tulemusel saadakse ennustus.

Kui tahame, et mudel kujul (9) oleks kooskõlas meie I peatükis toodud lineaarse mudeli definitsiooniga, siis tuleb f_j esitada läbi baasifunktsioonide lineaarkombinatsiooni. See

tähendab, et kasutame juba varasemalt nähtud esitust:

$$f(x) = \sum_{i=-m}^{n-1} \beta_i B_m^i(x) \quad (10)$$

. Kui nüüd iga f_j puhul kasutada esitust (10), siis mudel (9) on vastavuses lineaarse regressioonimudeli nõuetega. Märkime ka ära, et iga argumenttunnuse puhul eeldame, et n osalõiguga võrk Δ ja selle laiend on fikseeritud. Kordajad β_i leitakse eelnevates alapeatükkides mainitud silumismeetodeid kasutades.

Üldistatud aditiivsete mudelite puhul kerkib üks matemaatiline probleem. Nimelt mudeli parameetrid ei pruugi olla üheselt määratavad. Selle mõistmiseks vaatame lihtsat mudelit:

$$y = \beta_0 + f_1(x_1) + f_2(x_2).$$

Lisapiiranguid rakendamata on võimalik liita funktsioonile f_1 suvaline konstant C ja lahutada see maha funktsioonist f_2 ilma, et mudeli prognoos muutuks:

$$y = \beta_0 + [C + f_1(x_1)] + [f_2(x_2) - C].$$

Seega tekib lõpmatu arv lahendeid vähimruutude meetodile, sest me võime ühte funktsiooni liigutada lõputult y teljel üles ja teist samapalju alla, saades tulemuseks sama prognoosi. Et mudeli parameetrid jääksid üheselt määratavaks, lisame tingimuse:

$$\sum_{i=1}^n f_p(x_{ip}) = 0 \quad \forall p \text{ korral.} \quad (11)$$

Ehk teisisõnu nõuame, et iga siluva funktsiooni väärtuste summa üle andmepunktide peab olema 0. Oluline on märkida, et see tingimus ei muuda ühegi f_j kuju, vaid nihutab igat f_j vertikaalselt nii, et keskmine oleks 0.

Sellise tingimuse seadmine tagab ühtlasi ka mudeli parema tõlgendatavuse, sest keskmistatud funktsioonide korral langeb mudeli vabaliige β_0 kokku uuritava tunnuse oodatava keskväärtusega ning iga funktsioon f_j väljendab üksiku tunnuse lokaalset kõrvalekallet sellest üldisest keskmisest.

Tasub ka märkida, et üldistatud aditiivsed mudelid võimaldavad kombineerida nii mitte-

lineaarset silutud funktsioone kui ka klassikalisi arvulisi argumenttunnuseid. Lineaarsete seoste või kategooriliste tunnuste kaasamisel lisatakse need disainimaatriksisse tavalisel kujul ja vastav silumistrahv defineeritakse nullina. See tähendab, et lineaarsete liikmete kordajad leitakse sisuliselt klassikalise vähimruutude meetodi abil ülejäänud silutud funktsioonidega samaaegselt. Mudeli kuju on siis:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + f_1(x_{i,k+1}) + \cdots + f_p(x_{ip}) + \epsilon_i.$$

3.3.1 Silumisparameetri λ leidmine

Hoolimata sellest, kas me kasutame GAM-mudelil siluvaid splaine trahviga $K(f)$ või P-splaine karistusliikmega $R(\beta)$, on mõlemal juhul vaja leida optimaalne silumisparameeter λ . Kui λ valida liiga väike, hakkab mudel treeningandmetele üle sobitama. Samas, kui λ valida liiga suur, siis hakkame funktsioone f_j üle siluma. Sellega kaasneb tihti alasobitamine, sest liialt sile funktsioon ei pruugi suuta andmetes leiduvat tõelist seosefunktsiooni lähendada.

Klassikaline k -kordne ristvalideerimine, kuigi universaalne ja intuiitiivne meetod, osutub siinkohal ebaefektiivseks, sest mitme seletava tunnusega GAM-mudelite korral toob iga lisanduv silutav splain kaasa uue hindamist vajava silumisparameetri, mis nõuab modelleerijalt kandidaatväärtuste valimist iga λ jaoks ja teeb arvutused keerukaks.

Seetõttu kasutatakse kaasaegses statistilises tarkvaras (näiteks R-i paketi `mgcv`) silumisparameetrite leidmiseks teisi meetodeid, mis automatiseerivad valikuprotsessi. Praktikas on neist eelistatuim piiratud suurima tõepära meetod ehk REML (ingl *restricted maximum likelihood*). Antud meetodi tööpõhimõtteid me siinkohal üksikasjalikult ei käsitle, kuna see osutuks liialt mahukaks, kuid põhjalik kirjeldus on leitav allikast (Wood, 2017a).

3.3.2 Mõjumaatriks siluvate splainide puhul

Selleks, et hiljem GAM-mudelite efektiivsetest vabadusastmetest rääkida, peame sisse tooma mõjumaatriksi (ingl *hat matrix*) definitsiooni siluvate splainide puhul. Selle esitus erineb klassikalise mõju maatriksi esitusest, mille korral mudeli prognoositud väärtuste

vektor $\hat{y}$ on algsete vaatluste vektori y lineaarkujutus, mis avaldub järgmiselt:

$$\hat{y} = \mathbf{H}y,$$

kus maatriksit $\mathbf{H}$ nimetatakse mudeli mõjumaatriksiks. Kusjuures

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T,$$

kus $\mathbf{X}$ on mudeli disainimaatriks.

Kui aga kasutatakse silumistrahviga mudeleid, parameetriga λ , võtab mõjumaatriks kuju:

$$\mathbf{H} = \mathbf{B}(\mathbf{B}^T \mathbf{B} + \lambda \Omega)^{-1} \mathbf{B}^T,$$

kus Ω on trahvimaatriks, mis sisaldab näiteks baasifunktsioonide teiste tuletiste ruutude integraale (Hastie *et al.*, 2009, lk 153). P-splainide puhul $\Omega = D_k^T D_k$.

Mõjumaatriksil peadiagonaali elemendid h_{ii} iseloomustavad iga üksiku vaatluse y_i mõju ehk kaalu omaenda prognoosväärtusele $\hat{y}_i$.

3.3.3 Efektiivne vabadusastmete arv

Üks oluline nüanss, mida üldistatud aditiivsete mudelite kontekstis mõista, on siluvate funktsioonide efektiivne vabadusastmete arv. Klassikalise lineaarse regressiooni puhul on mudeli vabadusastmete arv täpselt võrdne hinnatavate parameetrite β_j arvuga. Näiteks lisades mudelisse kolm uut seletavat tunnust, kulutatakse täpselt kolm vabadusastet.

GAM-ides kasutatavate siluvate funktsioonide puhul see loogika aga ei kehti. Kui kasutada P-splainidega silumist ja valida konkreetse siluva funktsiooni f_j jaoks baasi mõõtmeks (baasifunktsioonide arvuks) näiteks $k = 20$, tekib esialgu 20 vaba kordajat, millest tsentreerimistingimuse (11) tõttu jääb järele 19. Kuna kordajatele rakendatakse diferentsitrahvi, mis piirab nende varieeruvust, ei kasuta mudel tegelikkuses ära kogu järelejääva 19 vabadusastme jagu paindlikkust. Seetõttu kasutatakse GAM-mudelite puhul efektiivsete vabadusastmete arvu (EDF, ingl *effective degrees of freedom*) mõistet. Oluline on ka märkida, et EDF ei ole täisarvuline diskreetne näitaja, vaid pidev suurus.

Matemaatiliselt defineeritakse EDF mudeli mõjumaatriksi $\mathbf{H}$ (kus $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$) jäljena. Iga

üksiku funktsiooni f_j efektiivsete vabadusastmete leidmiseks on aga otstarbekas vaadelda alamülesannet, kus fikseerime kõik ülejäänud mudeli liikmed ja lahutame nende prognoosid sihttunnuse tegelikest väärtustest, saame nn osalised jäägid.

Funktsiooni f_j sobitamine nendele osalistele jääkidele kujutab endast klassikalist ühe tunnuse siluva splaini leidmise ülesannet, millel on vastavalt omaenda mõjumaatriks $\mathbf{H}_j$. Üksiku funktsiooni f_j efektiivse vabadusastmete arvu saame seega leida tema vastava mõjumaatriksi $\mathbf{H}_j$ jälje ehk peadiagonaali elementide summa kaudu:

$$\text{EDF} = \text{tr}(\mathbf{H}_j).$$

Praktikas saab EDF-i väärtust tõlgendada siluva funktsiooni paindlikkuse indikaatorina. Kui $\text{EDF}(f_j) \approx 1$, on trahv väga suur ning seos silutakse sisuliselt lineaarseks. See tähendab, et selle j -nda argumenttunnuse suhtes mittelineaarset käitumist sisuliselt ei eksisteeri.

Kui EDF on oluliselt suurem kui 1, on tegemist mittelineaarse seosega. Kui aga EDF on väga lähedal väärtusele $k - 1$, kasutab mudel ära praktiliselt kogu lubatud paindlikkuse. Sel juhul on oht, et mudeli tegelikku keerukust on kunstlikult piiratud ning silumiseks kasutatava baasi suurust k tuleks tõsta.

Sarnane loogika kehtib ka klassikaliste siluvate splainide korral, kus trahvifunktsiooniks on teise tuletise ruudu integraal $K(f)$. Vaatleme mudelit

$$y_i = \beta_0 + f_1(x_i) + \epsilon_i.$$

Ka seal leitakse EDF vastava silumismaatriksi jäljena, kuid selgelt joonistuvad välja kaks äärmust. Kuna klassikalise siluva splaini puhul asub sõlmpunkt igas unikaalses andmepunktis, sõltub mudeli paindlikkus täielikult parameetrist λ (P-splainide puhul ka mingil määral k -st).

Kui trahv sisuliselt puudub ehk $\lambda \rightarrow 0$, hakkab mudel andmeid interpoleerima (eeldusel, et kõik x väärtused on erinevad) ja mudeli $\text{EDF} \rightarrow n$, kus n on vaatluste koguarv. Kui aga trahvi lõputult suurendada ehk $\lambda \rightarrow \infty$, surutakse funktsiooni teine tuletis nulli ehk kõver muutub sirgeks. Sellisel juhul kogu mudeli $\text{EDF} \rightarrow 2$, kuna üks vabadusaste kulub vabaliikme ja teine sirge tõusu jaoks.

3.3.4 Tensorsplainid

Üldistatud aditiivsete mudelite standardkuju eeldab, et silutavate tunnuste mõju sihttunnusele on aditiivne, mis tähendab, et ühe tunnuse mõju ei sõltu teiste tunnuste väärtustest. Paljudes praktilistes ülesannetes on aga tarvis arvestada tunnustevahelisi interaktsioone ehk koosmõjusid. GAM-ide raamistikus saab seda teha tensorsplainide (ingl *tensor product splines*) abil.

Selle mõistmiseks vaatleme esmalt aditiivset mudelit kujul:

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \epsilon.$$

Et modelleerida tunnuste x_1 ja x_2 koosmõju, on meil vaja moodustada uus kahemõõtmeline siluv funktsioon $f(x_1, x_2)$, mis modelleerib joone asemel tervet pinda. Mõlema ühemõõtmelise funktsiooni saame esitada baasifunktsioonide lineaarkombinatsiooniga:

$$f_1(x_1) = \sum_{i=1}^{K_1} \alpha_i B_i(x_1), \quad f_2(x_2) = \sum_{j=1}^{K_2} \gamma_j B_j(x_2),$$

kus α_i ja γ_j on vastavad kordajad ning $B_i(x_1)$ ja $B_j(x_2)$ tähistavad üheselt määratud baasifunktsioone.

Et minna üle kahemõõtmelisele pinnale $f(x_1, x_2)$, eeldame, et funktsiooni $f_1(x_1)$ kordajad α_i ei ole enam konstandid, vaid muutuvad sujuvalt vastavalt tunnuse x_2 väärtustele. Seega saame kordaja α_i asendada siledalt muutuva funktsiooniga $\alpha_i(x_2)$, mida me saame omakorda esitada f_2 baasifunktsioonide $B_j(x_2)$ kaudu:

$$\alpha_i(x_2) = \sum_{j=1}^{K_2} \delta_{ij} B_j(x_2),$$

kus δ_{ij} tähistavad uusi hinnatavaid parameetreid. Asendades selle seose tagasi algsesse $f_1(x_1)$ valemisse, saamegi tensorsplaini üldkuju:

$$f(x_1, x_2) = \sum_{i=1}^{K_1} \alpha_i(x_2) B_i(x_1) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \delta_{ij} B_i(x_1) B_j(x_2).$$

Seda tulemust saab mugavalt lühendada, defineerides uue ühise baasi:

$$f(x_1, x_2) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \delta_{ij} B_{ij}(x_1, x_2),$$

kus $B_{ij}(x_1, x_2) = B_i(x_1)B_j(x_2)$. Sellest tuleneb ka meetodi nimi, sest iga baasifunktsioon B_{ij} on funktsioonide B_i ja B_j tensorkorrutis ehk $B_{ij} = B_i \otimes B_j$. Seega uue pinna baasifunktsioonid tekivad algsete marginaalsete baasifunktsioonide omavahelisel korrutamisel ning mudelis on nüüd hindamiseks kokku $K_1 \cdot K_2$ parameetrit δ_{ij} .

Olles leidnud baasi tensorsplainide esitamiseks, oleks nüüd silumiseesmärkidel vaja defineerida ka korrektne karistusliige. Oletame, et f_1 ja f_2 silumiseks kasutati P-splaine. Meenutame hetkeks P-splainide puhul kasutatavat karistusliiget $R(\boldsymbol{\beta})$, mis avaldus kui $\boldsymbol{\beta}^T D_k^T D_k \boldsymbol{\beta}$. Tähistame f_1 puhul $S_1^* = (D_k^1)^T D_k^1$ ja f_2 puhul $S_2^* = (D_k^2)^T D_k^2$. Nüüd saame defineerida uue karistuse $f(x_1, x_2)$ -le:

$$\lambda_1 \boldsymbol{\delta}^T S_1 \boldsymbol{\delta} + \lambda_2 \boldsymbol{\delta}^T S_2 \boldsymbol{\delta},$$

kus $S_1 = S_1^* \otimes I_{K_2}$ ja $S_2 = I_{K_1} \otimes S_2^*$ ning $\boldsymbol{\delta}$ on $K_1 \cdot K_2$ pikkune vektor, mis sisaldab kõiki kordajaid δ_{ij} . Parameetrid λ_1 ja λ_2 on eraldiseisvad silumisparameetrid.

See, et λ_1 ja λ_2 on eraldiseisvad, on praktikas väga kasulik omadus, sest see lubab kasutada igal teljel erinevat silumisparameetrit, mis tagab, et silumine ei ole tundlik tunnuste skaalade erinevuste suhtes (Wood, 2017a, lk. 227).

Tensorsplainide puhul saab tegelikkuses kasutada ka klassikalist silumist (teise tuletise ruudu integraal), mille matemaatilistesse detailidesse me siinkohal ei lasku. Põhjaliku ülevaate leiab allikast (Wood, 2017a), kus vaadeldakse ka kolme tunnuse tensorsplaine. Tuleb aga märkida, et praktikas võib kolme ja enama tunnuse kasutamine tensorsplainide puhul osutuda arvutuslikult väga kulukaks (Fahrmeir *et al.*, 2021, lk 551).

3.3.5 GAM-ide eripära

Eelnevalt arutletust joonistub tegelikult välja veel üks oluline nüanss, mille poolest üldistatud aditiivsed mudelid erinevad mitmetest teistest (näiteks vähimruutude meetodit rakendavatest) klassikalise lineaarse regressiooni ja masinõppe mudelitest. Nimelt ei väldita

GAM-ide puhul ülesobitamist mudeli keerukuse ehk liikmete arvu otsese vähendamisega.

Näiteks klassikalises regressioonis, kui polünoome kasutatav mudel õpib liialt andmetes esinevat müra, lahendatakse seda polünoomi astme vähendamisega või teatud argument-tunnuste eemaldamisega, mis omakorda kaasab mudelisse vähem liikmeid ja kordajaid.

Klassikalise masinõppe paradigmas rakendatakse sageli teist ideed: ei proovita kohe ära arvata, milline oleks optimaalne parameetrite arv, vaid mudelile antakse tahtlikult liiga palju paindlikkust, et olla kindel, et ükski oluline seos ei jääks tabamata. Selleks, et sellised üleparametriseeritud mudelid ei hakkaks treeningandmete müra püüdma, lisatakse vähimruutude karistusliige, mille eesmärk on tõkestada ülesobitamist mudeli liigse keerukuse trahvimise kaudu. Teisisõnu, mudelil lubatakse tehniliselt kasutada kõiki parameetreid, aga seatakse piirangud kordajate käitumisele. Seda protsessi nimetatakse laiemalt regulariseerimiseks ning seda kasutavad näiteks Ridge, Lasso ja elastse võrgu (ingl *elastic net*) regressioon. Märkime, et need meetodid rakendavad sageli kogu mudeli peale vaid ühte (elastse võrgu puhul kahte) tuunitavat trahviparameetrit λ .

Siinkohal tulebki mängu üldistatud aditiivsete mudelite unikaalne lähenemine: igat komponenti (lihtsamal juhul tunnust) silutakse ja seega ka trahvitakse eraldi. Veelgi enam, kui eelnevalt mainitud regulariseerivad meetodid sunnivad liikmete kordajaid nulli poole või lausa täiesti nulli, siis standardsete karistustega GAM-id seda ei tee. Nemad suruvad funktsioonid äärmisel juhul lihtsalt sirgeks ehk lineaarseks.

See võimaldab ühe tunnuse mõjul olla väga keeruline, samal ajal kui teise tunnuse mõju on lineaarne. Mudelist ei kao ära ükski parameeter, vaid teise tuletise kaudu karistatakse funktsiooni karedust. Selline kareduse vältimine ongi andmetest tegeliku signaali eraldamise võti.

Kokkuvõte

Käesolev bakalaureusetöö andis ülevaate splineide matemaatilisest teooriast ja nende rakendamise üldistatud aditiivsetes mudelites. Töö alguses lükati ümber levinud väärarusaam, et lineaarsete regressioonimudelitega ei ole võimalik modelleerida mittelineaarseid seoseid. Mittelineaarsuse lisamise lihtsaima meetodina käsitleti polünoomregressiooni, millel on aga mitmeid praktilisi puudujääke, mida saab edukalt lahendada splineide ehk tükiti defineeritud polünoomide abil.

Esimese olulise probleemina käsitleti polünoomidele iseloomulikke ebausaldusväärset ekstrapolatsiooni. Töös näidati, et seda puudujääki saab maandada naturaalse splineide kasutamisega, mis suruvad prognoosifunktsiooni väljaspool treeningandmete piirkonda lineaarseks, hoides ära kiiresti kasvavad prognoosivead uute andmete korral.

Teiseks murekohaks on kõrge astmega polünoomide kalduvus ülesobitamisele. Kuigi ka klassikaliste (karistamata) splineide puhul säilib ülesobitamise oht, vähendavad siluvad splineid ja karistatud B-splineid seda probleemi märkimisväärselt. Nende mudelite efektiivsus sõltub korrektselt valitud silumisparameetrist λ , mille optimeerimine on tänapäevases tarkvaras täielikult automatiseeritav.

Kolmandaks toodi välja üksikute ekstreemsete vaatluste ebaproportsionaalselt suur globaalne mõju kogu polünoomregressiooni kõverale. Seda probleemi maandavad splineid (eriti B-splineide baasil defineeritud splineid) just tänu oma lokaalsele kandjale. Ehkki vähimruutude meetodi korral on parameetrite hindamine jätkuvalt ühine, mistõttu kandub üksikute vaatluste mõju endiselt üle kogu kordajate vektori, tagab B-splineide struktuur, et see mõju sumbub kauguse kasvades. Seega muudab ekstreemne väärtus prognoosifunktsiooni kuju peamiselt oma lokaalses ümbruses, jättes regressioonikõvera kaugemad osad suuresti häirimata.

Neljandaks probleemiks on globaalsete polünoomide jäikus. Kuna polünoomi aste m peab olema mittenegatiivne täisarv, siis keerukuse tõstmiseks tuleb minna astmele $m + 1$. See võib aga muuta funktsiooni kuju liiga drastiliselt. Splineid lahendavad selle probleemi, võimaldades paindlikkust tõsta lokaalselt, näiteks sõlmpunkte lisades, mis annab modelleerijale täpsema kontrolli funktsiooni kuju üle.

Kõik loetletud eelised tulevad korraga mängu üldistatud aditiivsete mudelite raamistikus. GAM-id pakuvad modelleerijale vabadust valida, milliseid tunnuseid siluda ja milliste

mõju jätta lineaarseks. Lisaks ühemõõtmeliste seoste modelleerimisele võimaldavad tensorsplainid arvestada ka tunnustevahelisi interaktsioone, hoides silumisprotsessi erinevate tunnuste skaalade suhtes invariantse. Erinevalt paljudest teistest regulariseerivatest masinõppemeetoditest, mis suruvad kordajaid nulli poole või lausa nulli, eemaldades seeläbi mudelist parameetreid, karistavad GAM-id hoopis funktsiooni teist tuletist. See tähendab, et kõik parameetrid säilivad, kuid vajadusel surutakse funktsioonid siledamaks või äärmisel juhul lineaarseks.

Kuigi käesolev töö andis põhjaliku ülevaate GAM-ide matemaatilisest vundamendist pideva sihttunnuse kontekstis, on meetodi rakendusala tegelikkuses palju laiem. Aditiivseid mudeleid saab edukalt üldistada sobiva seosefunktsiooni abil binaarseks klassifitseerimiseks või loendusandmete modelleerimiseks. Samuti on tensorsplainide kõrval võimalik modelleerida interaktsioone ka näiteks õhukese plaadi regressioonsplainidega (ingl *thin plate regression splines*). Seega üheks asjakohaseks töö edasiarenduseks oleks käsitleda ka neid teemasid.

Kasutatud allikad

- Boor, Carl de (2001). *A Practical Guide to Splines*. Revised. Applied Mathematical Sciences. New York: Springer-Verlag.
- Eilers, Paul H.C. ja Brian D. Marx (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge University Press.
- Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang ja Brian D. Marx (2021). *Regression: Models, Methods and Applications*. 2nd. Berlin, Heidelberg: Springer.
- Fan, Jianqing ja Irène Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Kõide 66. Monographs on Statistics and Applied Probability. London: Chapman ja Hall.
- Goodfellow, Ian, Yoshua Bengio ja Aaron Courville (2016). *Deep Learning*. MIT Press.
- Green, Peter J ja Bernard W Silverman (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman ja Hall/CRC. ISBN: 978-0412300400.
- Hahn, Gerald J. (1977). “The Hazards of Extrapolation in Regression Analysis”. *Journal of Quality Technology* 9.4, lk. 159–165. DOI: [10.1080/00224065.1977.11980791](https://doi.org/10.1080/00224065.1977.11980791).
- Hastie, T. J., R. J. Tibshirani ja J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. second. Heidelberg: Springer-Verlag.
- Hawkins, Douglas M. (2004). “The Problem of Overfitting”. *Journal of Chemical Information and Computer Sciences* 44.1, lk. 1–12. DOI: [10.1021/ci0342472](https://doi.org/10.1021/ci0342472).
- Meijering, Erik (2002). “A chronology of interpolation: from ancient astronomy to modern signal and image processing”. *Proceedings of the IEEE* 90.3, lk. 319–342. DOI: [10.1109/JPROC.2002.993399](https://doi.org/10.1109/JPROC.2002.993399).
- Oja, Peeter (2022). *Splainid. Loengukonspekt*. Tartu Ülikool, Matemaatika ja statistika instituut. Tartu.
- Oxford English Dictionary (2026). *spline*, *n.* https://www.oed.com/dictionary/spline_n. (Vaadatud 10.05.2026).
- Weisberg, Sanford (2014). *Applied Linear Regression*. 4th. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.
- Wood, Simon N. (2017a). *Generalized Additive Models: An Introduction with R*. 2nd. Boca Raton, Florida: Chapman ja Hall/CRC.

Wood, Simon N. (2017b). *mgcv.package: Mixed GAM Computation Vehicle with GCV/AIC/-REML smoothness estimation and GAMMs by REML/PQ*. R paketi versioon 1.8-18.
URL: <https://www.rdocumentation.org/packages/mgcv/versions/1.8-19/topics/mgcv.package> (vaadatud 10.05.2026).

Lisa 1. P-splainide diferentsimaatriksid

Olgu meil $(n+m)$ -mõõtmeline baasfunktsioonide kordajate vektor $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{n+m})^T$. Esimest järku diferentsimaatriks D_1 on mõõõtmega $(n+m-1) \times (n+m)$ ning selle korrutis vektoriga $\boldsymbol{\beta}$ avaldub maatrikskujul järgmiselt:

$$D_1 \boldsymbol{\beta} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n+m-1} \\ \beta_{n+m} \end{pmatrix} = \begin{pmatrix} \beta_2 - \beta_1 \\ \beta_3 - \beta_2 \\ \vdots \\ \beta_{n+m} - \beta_{n+m-1} \end{pmatrix} = \begin{pmatrix} \Delta \beta_2 \\ \Delta \beta_3 \\ \vdots \\ \Delta \beta_{n+m} \end{pmatrix}.$$

Teist järku diferentsimaatriks D_2 on vastavalt mõõõtmega $(n+m-2) \times (n+m)$ ning selle rakendamine näeb välja järgmine:

$$D_2 \boldsymbol{\beta} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{n+m-1} \\ \beta_{n+m} \end{pmatrix} = \begin{pmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \vdots \\ \beta_{n+m-2} - 2\beta_{n+m-1} + \beta_{n+m} \end{pmatrix} = \begin{pmatrix} \Delta^2 \beta_3 \\ \Delta^2 \beta_4 \\ \vdots \\ \Delta^2 \beta_{n+m} \end{pmatrix}.$$

Seega on trahvimaatriksid leitavad rekursiivselt. Eeskiri n -mõõõtmelise kordajate vektori puhul on:

$$D_k = D_1^{(n-k+1)} D_{k-1}^n,$$

kus $D_1^{(n-k+1)}$ on eelnevalt toodud maatriks D_1 , aga seekord mõõõtmega $(n-k) \times (n-k+1)$ ja D_{k-1}^n on $(n-k+1) \times n$ maatriks. Tulemuseks saame seega $(n-k) \times n$ maatriksi. Praktikas on sellised diferentsimaatriksid aga R-i funktsiooniga `diff(diag(n), diff=k)` üsna lihtsalt leitavad.

Lisa 2. GAM-ide rakendamine ja realiseerimine tarkvaras R

Siinkohal anname kiire ülevaate üldistatud aditiivsete mudelite rakendamisest tarkvaras R. R-keskkonnas on üks levinumaid pakette GAM-raamistiku kasutamiseks `mgcv`. GAM-ide defineerimine R-is on süntaksi poolest sarnane tuntud funktsioonidele `lm()` ja `glm()`. Peamine erinevus seisneb silumisfunktsiooni `s()` kasutamises mudeli valemis, mis annab programmile korralduse antud tunnust splineidega siluda.

Tüüpiline GAM-mudeli spetsifitseerimine R-is näeb välja midagi sellist:

```
library(mgcv)
mudel <- gam(y ~ s(x1) + x2, data = andmestik, method = "REML")
```

Siinkohal y on sihttunnus, x_1 on silutav ja x_2 tavaline argumenttunnus. Ehkki käsklus `gam()` kasutab vaikimisi silumisparameetrite otsimiseks üldistatud ristvalideerimist, saab seda soovi korral mudeli sees muuta argumentiga `method=`. Näites valiti meetodiks REML, sest see on praktikas tihti eelistatav (Wood, 2017a, lk 266).

Käsu `s()` kasutamisel saab veel mitut parameetrit täpsustada. Peamised on argument `k`, mis dikteerib kasutatava baasi dimensiooni. Vaikimisi seatakse väärtuseks 10. Lisaks saab valida splineide koostamiseks rakendatavaid baasfunktsioone argumentiga `bs`, mis laseb valida näiteks 3. astme B-splineide baasi (käsuga `bs="cr"`), mille korral silumine toimub trahviga $K(f)$ või P-splineide baasi (käsuga `bs="ps"`). Kogu valikuga saab tutvuda allikas (Wood, 2017b).

Tulemuste tõlgendamine ja mudeli diagnostika

Pärast mudeli sobitamist saab näha tulemusi, kasutades funktsiooni `summary(mudel)`. See annab ülevaate mudeli parameetritest. Olulisim näitaja silutud argumentide puhul on EDF. Nagu varasemalt mainitud, kui silutud tunnuse EDF on ligikaudu 1, on tunnus surutud lineaarseks, mis tähendab, et mittelineaarset käitumist selle tunnuse suhtes tõenäoliselt ei eksisteeri. Mida suurem on EDF, seda keerukam on seos sõltuva ja sõltumatu tunnuse vahel. Samuti kuvatakse ka iga muutuja p-väärtus.

Üks kasulik diagnostika käsklus on veel `gam.check(mudel)`. See genereerib automaatselt mudeli jääkide graafikud ja testib, kas silumiseks kasutatavate baasifunktsioonide arv on piisav. Võimalik väljund on:

```
##           k'   edf k-index p-value
## s(x2) 9.00 8.98      0.8 <2e-16 ***
```

Toodud näite puhul näeme, et vaikinisi kasutatav baasifunktsioonide arv $k=10$ (1 vabadusaste kaob tsentreerimistingimuse tõttu) on liiga väike, sest mudel kasutab ära praktiliselt kõik saadaval olevad vabadusastmed ($EDF = 8.98$). Seega peaks käsus `s(x2)` käsitsi valima näiteks $k=20$ või rohkem. K-indeks näitab, kas mudeli jääkides on veel alles mingeid mustreid. K-indeks < 1 tähendab, et mudeli jäägid pole päris juhuslikud ja siluv funktsioon alasobitab andmetele. Taaskord aitab baasi dimensiooni tõstmine. Ideaalis peaks k-indeks olema 1 lähedal. Märkime, et siinkohal ei väljenda `p-value` silutud tunnuse olulisust mudelis, vaid testib nullhüpoteesi jääkide juhuslikkuse kohta. Seega madal `p`-väärtus lükkab nullhüpoteesi ümber, viidates, et mudel tõenäoliselt alasobitab.

Kahe pideva tunnuse koosmõju

Kahe pideva tunnuse vahelise (sileda) koosmõju modelleerimiseks kasutatakse sageli ka eelnevalt mainitud `tensorsplaine`. Pakett `mgcv` pakub selleks kahte peamist lähenemist.

Esimene variant kasutab funktsiooni `te()`, mis kaasab mudelisse samaaegselt nii mõlema tunnuse peamõjud kui ka nende interaktsiooni:

```
mudel1 <- gam(y ~ te(x1, x2), data = andmestik)
```

Teine, sageli eelistatum ja analüütiliselt paindlikum variant on eraldada peamõjud ja puhas koosmõju, kasutades funktsiooni `ti()` (ingl *tensor product interaction*):

```
mudel2 <- gam(y ~ s(x1) + s(x2) + ti(x1, x2), data = andmestik)
```

Selle lähenemise peamine eelis seisneb tõlgendatavuses. Eraldades peamõjud `s(x1)` ja `s(x2)` puhtast koosmõjust `ti(x1, x2)`, on võimalik mugavalt testida, kas interaktsioonitermini lisamine mudelisse parandab sobitust või piisab andmete kirjeldamiseks üksnes aditiivsete peamõjude kasutamisest.

Paketis `mgcv` sisalduva kohta leiab täpsemat teavet allikast (Wood, [2017b](#)).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Ralf Grossfeldt,

1. annan Tartu Ülikoolile tasuta loa minu loodud teose „Splainid ja üldistatud adiitiivsed mudelid“, mille juhendaja on Raul Kangro, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Ralf Grossfeldt

11.05.2026