Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Karl-Joan Alesma

# Segmenteerimine läbi murdepunktide

Matemaatika

Bakalaureusetöö (9 EAP)

Juhendaja: prof. Jüri Lember

Tartu 2024

**SEGMENTEERIMINE LÄBI MURDEPUNKTIDE**

Bakalaureusetöö

Karl-Joan Alesma

**Lühikokkuvõte**

Varjatud Markovi ahelate segmenteerimiseks on olemas mitmeid algoritme nagu näiteks Viterbi ja PMAP, seejuures mõlemal on omad puudused. Me nimetame murdepunktideks indekseid, mille korral vaadeldava vektori väärtus muutub. Selles töös uurime segmenteerimist, kus kasutame ära praktikas teinekord esinevat informatsiooni murdepunktide kohta või selle puudumisel ennustame, millised need murdepunktid võiksid olla. Nende leidmiseks toome sisse uue kaofunktsiooni, mis põhineb Binderi kaofunktsionil, ning näitame, et sellele vastavat riski on võimalik minimiseerida kasutades dünaamilist planeerimist. Kirjeldame ära tulemused kaheastmeliseks segmenteerimiseks ning selle variatsioonide jaoks. Töö praktilises osas illustreerime mõningaid leitud algoritme ühe konkreetse mudeli korral.

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

**Märksõnad:** HMM, segmenteerimine, Viterbi algoritm, Binderi kaofunktsioon, klasterdamine, murdepunkt.

**SEGMENTATION WITH CHANGE POINTS**

Bachelor thesis

Karl-Joan Alesma

**Abstract**

There are many algorithms for the segmentation of hidden Markov models, such as Viterbi and PMAP, each with their own limitations. We call indices change points, if a vector changes value for those points. In this thesis, we study a variant of segmentation, where we wish to use the information we have about change points or in the absence of this, we try to predict where the change points might be. To find them, we define a new loss function that is based on Binder's loss and show how to minimize the risk of it using dynamical programming. We describe algorithms for two-step segmentation and its variants. In the practical part, we visualize some of the algorithms for one concrete model.

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

**Key Words:** HMM, segmentation, Viterbi algorithm, Binder's loss, clustering, change point.

# Sisukord

**Sissejuhatus** **4**

**1 Varjatud Markovi ahel** **5**

1.1 Definitsioon . . . . . 5

1.2 Omadused . . . . . 6

1.3 Edasi-tagasi valemid . . . . . 7

**2 Klastrid, murdepunktid ja risk** **10**

2.1 Klastrid ja murdepunktid . . . . . 10

2.2 Risk . . . . . 12

2.2.1 Murdepunktide kaofunktsioon . . . . . 13

2.2.2 Binderi kaofunktsioon . . . . . 15

2.2.3 Ersatz-Binderi kaofunktsioon . . . . . 18

2.2.4 Binderi ja Ersatz-Binderi kaofunktsiooni risk . . . . . 19

2.3 Optimiseerimisülesande lahendamine . . . . . 23

2.3.1 Arvutuslikud märkused . . . . . 26

**3 Segmenteerimine** **30**

3.1 Tähistused . . . . . 30

3.2 Viterbi algoritm murdepunktide prognoosi korral . . . . . 31

3.3 PMAP-joondus murdepunktide prognoosi korral . . . . . 34

3.4 Etteantud murdepunktide arv . . . . . 36

**4 Simulatsioonid** **41**

4.1 Kitsentatud Viterbi algoritm ette antud murdepunktide arvu korral 41

4.2 Kaheastmeline segmenteerimine . . . . . 42

**Kokkuvõte** **45**

**Kasutatud allikad** **46**

**Lisa 1. Simulatsiooni joonised** **47**

**Lisa 2. Korrigeeritud simulatsiooni joonised** **54**

# Sissejuhatus

Tihti on tarvis uurida või modelleerida mõnda ajas muutuvat protsessi (tähistame seda kui $Y$). Selle jaoks teostame mingisuguseid mõõtmisi ning saame selle tulemusel vaatluste vektori $x_1, \ldots, x_n$. Aga me kas teame või kahtlustame, et tegu ei ole päris väärtustega. Nimelt võivad meie mõõtmised sisaldada müra/mõõtmisvigu või meil hoopiski puudub võimalus mõõta uuritavat protsessi otse ning peame selle asemel leppima protsessist $Y$ sõltuva protsessi mõõtmisega (tähistame seda kui $X$). Kui paar $(Y, X)$ on varjatud Markovi ahel, siis vaatluste $x_1, \ldots, x_n$ põhjal protsessi $Y$ „parimas mõttes“ kirjeldava vektori $y_1, \ldots, y_n$ leidmist nimetatakse segmenteerimiseks.

Segmenteerimise probleemile leidub mitmeid lahendusi. Üheks võimalikuks lahendiks on Viterbi joondus, mille jaoks kõige parem hinnang on kõige tõenäolisem vektor. Aga igal hinnangul on teatavad puudused, Viterbi joonduse puhul on tema suur vigade arv. Selles töös läheneme segmenteerimisele teise nurga alt — nimelt uurime segmenteerimist, kus kasutame murdepunkte ja klasteranalüüsi. Murdepunktiks nimetame indekseid, mille korral vaadeldava vektori väärtus muutub. Näiteks vektori 11223 korral on murdepunktid 3 ja 5. Kui me suudame kuidagi hinnata või hoopis teame, kus murdepunktid paiknevad või mitu neid on või suudame anda nende koguarvule vahemiku, siis kas meil õnnestub kasutada ära seda informatsiooni segmenteerimiseks?

Antud töö eesmärk on kirjeldada ja tõestada ära vastavad algoritmid kaheastmelise segmenteerimise ja selle variatsioonide jaoks. Kaheastmeline segmenteerimine koosneb järgmistest sammudest:

1. esimesel sammul leiame teatavas mõttes „parima“ murdepunktide prognoosi ning
2. teisel sammul kasutame leitud prognoosi ja vaatlusi $x_1, \ldots, x_n$, et leida sobiv vektor $y_1, \ldots, y_n$.

Esimeses peatükis tutvume põgusalt varjatud Markovi ahelaga. Teises peatükis seletame ära, mida tähendab parim ning kuidas seda kirja panna. Lisaks näitame veel kuidas leida parimat murdepunktide hinnagut. Kolmandas osas tuletame algoritmid, mis kasutavad segmenteerimisel ära informatsiooni murdepunktide kohta. Viimases osas illustreerime mõnda leitud algoritmi ühe konkreetse mudeli jaoks.

# 1 Varjatud Markovi ahel

## 1.1 Definitsioon

Kogu töö vältel kasutame järgmisi tähistusi. Olgu $A$ mingi hulk, $m$ ja $n$ täisarvud, kus $m < n$, ning jada $a_1, a_2, \ldots \in A$. Siis tähistame $a_m^n = (a_m, \ldots, a_n)$. Kui $m = 1$, siis jätame selle märkimata, st $a^n = (a_1, \ldots, a_n)$. Tähistame kõigi seisundite hulka kui $\mathcal{Y} = \{1, \ldots, K\}$.

**Definitsioon 1.1.** Olgu $Y = \{Y_n\}_{t\in\mathbb{N}}$ homogeenne Markovi ahel seisundite hulgaga $\mathcal{Y}$. Olgu $X = \{X_n\}_{t\in\mathbb{N}}$ mingi juhuslik protsess, kus $X_t$ võtab väärtusi hulgast $\mathcal{X} \subseteq \mathbb{R}^d$, kus $d \in \mathbb{N}$.

Paar $(Y, X)$ on **varjatud Markovi ahel** (ingl *hidden Markov model*, edaspidi HMM), kui

1. juhuslikud suurused $X_t$ on tinglikult sõltumatud tingimusel $Y$,
2. suuruse $X_t$ jaotus sõltub protsessist $\{Y_t\}$ ainult läbi suuruse $Y_t$.

Igale seisundile $j \in \mathcal{Y}$ vastab tõenäosusmõõt $P_j$, mida nimetatakse **emissioonijaotuseks**, nii et iga Boreli hulga $A$ ja $t \geq 1$ korral kehtib

$$P_j(A) = \mathbf{P}(X_t \in A | Y_t = j).$$

Eeldame üldisust kitsendamata, et emissioonijaotusel leiduvad tihedused mingi ühise mõõdu $dx$ suhtes. Neid nimetatakse **emissioonitihedusdeks** ning tähistame neid kui $f_j(x_t)$.

Edaspidi kui räägime HMM-idest, kasutame järgmisi tähistusi. Markovi ahela $Y$ algtõenäosusi tähistame $\pi_j := \mathbf{P}(Y_1 = j)$ ning üleminekutõenäosusi, kui $p_{ij} := \mathbf{P}(Y_2 = j | Y_1 = i)$, kus $i, j \in \mathcal{Y}$. Juhuslike vektorite $Y^n$ ja $X^n$ realisatsioone tähistame $y^n \in \mathcal{Y}^n$ ja $x^n \in \mathcal{X}^n$. Tähistame tinglikke tõenäosusi järgmiselt:

$$\begin{aligned} p(y^n|x^n) &:= \mathbf{P}(Y^n = y^n | X^n = x^n), \\ p(x^n|y^n) &:= \mathbf{P}(X^n = x^n | Y^n = y^n), \\ p_t(y_t|x^n) &:= \mathbf{P}(Y_t = y_t | X^n = x^n). \end{aligned}$$

Vastavalt HMM-i definitsioonile

$$p(x^n, y^n) := p(x^n|y^n)p(y^n) = \pi_{y_1} f_{y_1}(x_1) \prod_{t=2}^{n} p_{y_{t-1}, y_t} f_{y_t}(x_t).$$

**Näide 1.1.** Järgmine näide pärineb Rabineri (1989) artiklist.

Olgu meil ruumis $n$ vaasi, kus igale vaasile on ilma korduseta määratud mingi arv vahemikust $1, \ldots, n$, mida nimetame sildiks. Olgu igas vaasis mingi arv värvitud kuule, seejuures eeldame, et kõikvõimalike värvide arv on fikseeritud.

Ruumis on ka Madis, kes hakkab nendest vaasidest kuule välja võtma. Kõigepealt valib Madis mingi juhusliku protsessi alusel välja esimese vaasi. Ta võtab sealt ühe kuuli, märgib ära selle värvi ja paneb tagasi vaasi. Seejärel valib ta mingi juhusliku protsessi põhjal, mis sõltub ainult tolle vaasi sildist, mille juures ta praegu seisab, välja uue vaasi. Nüüd kordab ta palli valiku protsessi, seejärel valib uue vaasi jne. Selle tulemusel saame lõpliku arvu vaatlusi, mida võime vaadelda kui HMM-i $(Y, X)$ väljundit.

Siin protsess $Y$ modelleerib valitud vaase ning $X$ nähtud palli värvi. Selle, millist emissioonijaotust kasutada, määrab ära vaasi silt, mille juures Madis on. Jaotuse enda määrab ära erinevat värvi pallide arv valitud vaasis. Järgmise vaasi valik sõltub Markovi ahela $Y$ üleminekumaatriksist.

## 1.2 Omadused

Toome selles peatükis välja mõned HMM-i kasulikud omadused, mida antud töös läheb vaja. Tõestused põhinevad tehisõppe konspekti eeskujul (Lember, 2017).

**Lemma 1.1.** Mistahes $x^n \in \mathcal{X}^n$ ja $y^n \in \mathcal{Y}^n$ ning $1 \leq t < n$ korral kehtib järgmine faktoriseering

$$p(x^n, y^n) = p(x^t, y^t)p(x_{t+1}^n, y_{t+1}^n | y_t).$$

*Tõestus.* Paneme tähele, et

$$\begin{aligned}
p(x^n, y^n) &= p(x^n|y^n)p(y^n) \\
&= p(x^t|y^n)p(x_{t+1}^n|y^n)p(y^n) && \text{(HMM-i esimene eeldus)} \\
&= p(x^t|y^t)p(x_{t+1}^n|y_t^n)p(y^n) && \text{(HMM-i teine eeldus)} \\
&= p(x^t|y^t)p(x_{t+1}^n|y_t^n)p(y^t)p(y_{t+1}^n|y_t) && \text{(Markovi omadus)} \\
&= p(x^t, y^t)p(x_{t+1}^n|y_t^n)p(y_{t+1}^n|y_t) && \text{(Ühistõenäosus)} \\
&= p(x^t, y^t)p(x_{t+1}^n, y_{t+1}^n|y_t). && \text{(Tinglik ühistõenäosus)}
\end{aligned}$$

□

Olgu meil fikseeritud vaatlused $x^n$ ja $1 \leq k \leq t \leq l < n$. Antud lemmast järeldub, et faktoriseering kehtib ka siis, kui $y^n$ asemel on $y_k^l$. Iga $y^n$ korral

$$p(x^n, y^n) = p(x^t, y^t)p(x_{t+1}^n, y_{t+1}^n | y_t).$$

Summeerides üle $y_{k-1}$ ja $y_{l+1}^n$, saame, et

$$p(x^n, y_k^l) = p(x^t, y_k^t)p(x_{t+1}^n, y_{t+1}^l|y_t).$$

Siit järeldub, et

$$p(x^n, y_k^l) = p(x^k, y_k)p(x_{k+1}^n, y_{k+1}^l|y_k) = p(x^l, y_k^l)p(x_{l+1}^n|y_l).$$

Erijuhul

$$p(x^n, y_k) = p(x^k, y_k)p(x_{k+1}^n|y_k).$$

**Lause 1.1.** Olgu meil vaatlused $x^n$. Siis juhuslik vektori $Y^n$ tinglik jaotus tingimusel $X^n = x^n$ on mittehomogeenne Markovi ahel. Nimelt mistahes $1 \leq t < n$ ja $y^{t+1} \in \mathcal{Y}^{t+1}$ korral kehtib

$$\mathbf{P}(Y_{t+1} = y_{t+1}|Y^t = y^t, X^n = x^n) = \mathbf{P}(Y_{t+1} = y_{t+1}|Y_t = y_t, X^n = x^n).$$

Veelgi enam

$$\mathbf{P}(Y_{t+1} = y_{t+1}|Y_t = y_t, X^n = x^n) = \mathbf{P}(Y_{t+1} = y_{t+1}|Y_t = y_t, X_{t+1}^n = x_{t+1}^n).$$

*Tõestus.* Fikseerime $y^{t+1} \in \mathcal{Y}^{t+1}$ ja näitame, et

$$p(y_{t+1}|y^t, x^n) = p(y_{t+1}|y_t, x_{t+1}^n).$$

Lemma 1.1 põhjal kehtivad järgmised võrdused

$$p(x^n, y^{t+1}) = p(x^t, y^t)p(x_{t+1}^n, y_{t+1}|y_t), \quad p(x^n, y^t) = p(x^t, y^t)p(x_{t+1}^n|y_t).$$

Seega

$$\begin{aligned} p(y_{t+1}|y^t, x^n) &= \frac{p(x^n, y^{t+1})}{p(x^n, y^t)} = \frac{p(x^t, y^t)p(x_{t+1}^n, y_{t+1}|y_t)}{p(x^t, y^t)p(x_{t+1}^n|y_t)} \\ &= \frac{p(x_{t+1}^n, y_{t+1}|y_t)}{p(x_{t+1}^n|y_t)} = p(y_{t+1}|y_t, x_{t+1}^n) \end{aligned}$$

nagu tarvis. □

## 1.3 Edasi-tagasi valemid

Toome sisse edasi-tagasi muutujad (ingl *forward-backward* muutujad):

$$\alpha(j, x^t) := p_t(j|x^t)p(x^t),$$

$$\beta(x_{t+1}^n|j) := \begin{cases} 1, & \text{kui } t = n, \\ p(x_{t+1}^n|Y_t = j), & \text{kui } t < n. \end{cases}$$

Muutuja $\alpha(j, x^t)$ kujutab endast tõenäosust, et meil on olemas osaline realisatsioon $x^t$ ning et ajahetkel $t$ oleme seisundis $j$. See-eest muutuja $\beta(x_{t+1}^n|j)$ kujutab endast osalise vaatluse $x_{t+1}^n$ tõenäosust tingimusel, et ajahetkel $t$ oleme seisundis $j$.

Nende muutujate sissetoomist motiveerib asjaolu, et me saame nende abil efektiivselt leida erinevaid tõenäosusi, näiteks

$$p_t(j|x^n) = \frac{\alpha(j, x^t)\beta(x_{t+1}^n|j)}{p(x^n)}.$$

Lisaks neid ennast saab leida rekursiivselt (Rabiner, 1989):

$$\alpha(j, x^{t+1}) = \sum_{i \in \mathcal{Y}} \alpha(i, x^t) p_{ij} f_j(x_{t+1}), \tag{1.1}$$

$$\beta(x_t^n|j) = \sum_{i \in \mathcal{Y}} \beta(x_{t+1}^n|i) f_i(x_t) p_{ji}. \tag{1.2}$$

Võrrand (1.1) kujutab endast järgmist protsessi. Me oleme ajahetkel $t$ seisundis $i$ ning oleme näinud vaatlusi $x^t$. Seejärel lähme üle seisundisse $j$. Kuna ajahetkel $t$ võime olla ka mõnes muus seisundis kui $i$, siis summeerime üle kõikvõimalike seisundite. Arvestame veel juurde vaatluse $x_{t+1}$ ja saamegi valemi (1.1).

Võrrand (1.2) on analoogne. Me leiame tõenäosused iga võimaliku seisundi jaoks ajahetkel $t$, emiteerime vaatluse $x_t$ ning liigume sealt seisundist tagasi seisundisse $j$.

Tasub märkida, et ajasammude kasvades kahanevad edasi-tagasi muutujad eksponentsiaalselt nulli ning võivad seetõttu minna arvuti ujukomaarvu esituse täpsusest välja. Seetõttu leiame ja kasutame arvutusdes edasi-tagasi muutujate logaritmitud variante $\ln \alpha(j, x^t)$ ja $\ln \beta(x_{t+1}^n|j)$, mis tagab arvutuslikult parema stabiilsuse. Nende leidmiseks kasutame logsumexp trikki (Gundersen, 2020). Nimelt kui meil on mingid arvud $x^n \in \mathbb{R}^n$ ja

$$y = \ln \sum_{i=1}^n e^{x_i},$$

siis

$$y = a^* + \ln \sum_{i=1}^n e^{x_i - a^*},$$

kus $a^* \in \mathbb{R}$. See trikk võimaldab meil nihutada reaaltelje keskpunkti nii, et tehtavad arvutused oleksid võimalikult soodsad. Nimelt võtame $a^* = \max_{i=1,\ldots,n} x_i$, mis

tagab, et kõige suurem suurus oleks null.

Nüüd kui me tahame leida logaritmitud edasi-tagasi muutujaid, siis kasutades seoseid (1.1) ja (1.2) ning logsumexp trikki, saame, et

$$\ln \alpha(j, x^{t+1}) = \alpha_t^*(j) + \ln \sum_{i \in \mathcal{Y}} \exp(\ln \alpha(i, x^t) + \ln p_{ij} - \alpha_t^*(j)) + \ln f_j(x_{t+1}), \quad (1.3)$$

$$\ln \beta(x_t^n | j) = \beta_{t+1}^*(j) + \ln \sum_{i \in \mathcal{Y}} \exp(\ln \beta(x_{t+1}^n | i) + \ln p_{ji} + \ln f_i(x_t) - \beta_{t+1}^*(j)), \quad (1.4)$$

kus

$$\alpha_t^*(j) = \max_{i \in \mathcal{Y}} \left( \ln \alpha(i, x^t) + \ln p_{ij} \right),$$

ja

$$\beta_{t+1}^*(j) = \max_{i \in \mathcal{Y}} \left( \ln \beta(x_{t+1}^n | i) + \ln p_{ji} + \ln f_i(x_t) \right).$$

# 2 Klastrid, murdepunktid ja risk

## 2.1 Klastrid ja murdepunktid

Olgu $y^n \in \mathcal{Y}^n$ vektor (edaspidi ka tee).

**Definitsioon 2.1.** Tee $y^n$ poolt tekitatud **klasterduseks** $C(y^n) = \{C_1, \ldots, C_m\}$ nimetatakse hulka, kus **klastrid** $C_1, \ldots, C_m$ on tee $y^n$ indeksite hulga $\{1, \ldots, n\}$ ekvivalentsiklassid, kus ekvivalentsusseoseks on

$$i \sim j \quad \text{parajasti, siis kui} \quad y_i = y_j.$$

*Märkus.* Igale teele vastavas klasterduses on ülimalt $K$ klastrit.

Edaspidi kui on antud hulga $\{1, \ldots, n\}$ klasterdus $C$, siis tähistame klasterdust $C$ tekitavat mingisugust teed kui $c^n = (c_1, \ldots, c_n)$, st $C = C(c^n)$.

**Definitsioon 2.2.** Klasterdust $C(y^n)$ nimetame **kanooniliseks**, kui iga klaster on intervall kujul $[i, j]$ (st sisaldab kõiki täisarve $i$ ja $j$ vahel).

**Näide 2.1.** Olgu $K = 4$, $y^{10} = 4111331411$ ja $a^{10} = 1111442223$. Siis

$$C(y^{10}) = \{\{1, 8\}, \{2, 3, 4, 7, 9, 10\}, \{5, 6\}\},$$
$$C(a^{10}) = \{\{1, 2, 3, 4\}, \{5, 6\}, \{7, 8, 9\}, \{10\}\}.$$

Klasterdus $C(a^{10})$ on kanooniline, aga $C(y^{10})$ ei ole.

**Definitsioon 2.3.** Vektori $y^n$ indeksit $i > 1$ nimetatakse **murdepunktiks** (ingl *change point*), kui $y_{i-1} \neq y_i$. Binaarset vektorit $w(y^n) = w_2^n \in \{0, 1\}^{n-1}$, mis on defineeritud seosega

$$w_i = 1 \iff i \text{ on murdepunkt},$$

nimetatakse vektorile $y^n$ vastavaks **murdepunktide vektoriks**.

*Märkus.* Vektor $w_2^n$ koosneb $n - 1$ elemendist, aga esimest elementi tähistame kui $w_2$ mitte $w_1$.

**Näide 2.2.** Vaatleme teed $y^{10} = 4111331411$. Siis selle murdepunktid on $2, 5, 7, 8, 9$ ning murdepunktide vektor on

$$w_2^{10} = 100101110.$$

Paneme tähele, et kaks erinevat teed võivad tekitada sama klasterduse, näiteks $C(4111331411) = C(1222332122)$, ning igale klasterdusele saab vastavusse seada vähemalt ühe tee. Seega on klasterduste arv hulgal $\{1, \ldots, n\}$ väiksem kui kõigi

teede arv $K^n$. Kõigi tükelduste arv, mis jagab $n$ elemendilise hulga $k$ tükiks, on teist järku Stirlingi arv $\left\{ {n \atop k} \right\}$ (Grassl ja Levin, 2018)

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n,$$

seega kõigi kõikvõimalike tükeltuste arv on

$$\sum_{k=1}^{K} \left\{ {n \atop k} \right\}.$$

Kui $K = n$, nimetatakse antud suurust Belli arvuks.

Kanoonilisi klasterdusi on see-eest vähem. Kui meil on $n$ elementi ja $k$ klastrit, siis meil on $n-1$ kohta, kus saab olla murdepunkt, ning meil saab kokku olla $k-1$ murdepunkti, seejuures murdepunktide järjekord ei ole oluline. Seega on

$$\binom{n-1}{k-1}$$

viisi $n$ elemendi tükeldamiseks $k$ klastriks ning kokku on kanoonilisi klasterdusi

$$\sum_{k=1}^{K} \binom{n-1}{k-1}.$$

Kui $K = n$, on see summa $2^{n-1}$.

**Näide 2.3.** Kui $n = 10$ ja $K = 4$, siis

$$\left\{ {10 \atop 1} \right\} = 1, \quad \left\{ {10 \atop 2} \right\} = 511, \quad \left\{ {10 \atop 3} \right\} = 9330, \quad \left\{ {10 \atop 4} \right\} = 34105$$

ja kõikvõimalikke tükeldusi on 43947. Kanoonilisi klasterdusi on $\binom{9}{0} + \binom{9}{1} + \binom{9}{2} + \binom{9}{3} = 130$ ning teede arv on $4^{10} = 1048576$.

Paneme tähele, et iga klasterdus defineerib üheselt murdepunktid, aga mitte vastupidi — murdepunktide vektor defineerib üheselt ainult kanoonilise klasterduse. Seega võime tähistada klasterdusele $C$ vastavat murdepunktide vektorit kui $w(C)$ ja murdepunktide vektorile $w$ vastavat kanoonilist klasterdust kui $\mathrm{C}_{\mathrm{kan}}(w)$.

**Näide 2.4.** Vaatleme kahte erinevat klasterdust $C_1 = \{\{1,4,5,6\}, \{2,3,10\}, \{7,8,9\}\}$ ja $C_2 = \{\{1,10\}, \{2,3\}, \{4,5,6\}, \{7,8,9\}\}$. Neile vastab sama murdepunktide vektor

$$w_2^n = w(C_1) = w(C_2) = 101001001.$$

Murdepunktide vektori $w_2^n$ põhjal saame üheselt leida ainult kanoonilise klasterduse, mis on

$$\mathrm{C}_{\mathrm{kan}}(w_2^{10}) = \{\{1\}, \{2,3\}, \{4,5,6\}, \{7,8,9\}, \{10\}\} = \{[1], [2,3], [4,6], [7,9], [10]\}.$$

## 2.2 Risk

Olgu $Y_1, \ldots, Y_n$ mingisugused juhuslikud suurused väärtuste hulgaga $\mathcal{Y}$. Me otsime vektorit $\hat{a}^n \in \mathcal{Y}^n$, mis mingis „parimas“ mõttes lähendaks juhuslikku vektorit $Y^n$.

**Definitsioon 2.4. Kaofunktsiooniks** nimetatakse funktsiooni

$$L\colon \mathcal{Y}^n \times \mathcal{Y}^n \to \mathbb{R}^+,$$

mis seab igale paarile $(y^n, a^n)$ vastavuse kahju, mis me saame, kui peame tõelist klassi $y^n$ ennustatud klassiks $a^n$.

Kaofunktsioon aitab meil konkreetse ülesande jaoks paika panna „headuse“ kriteeriumi, seejuures mida väiksem on kadu, seda parem on prognoositud vektor. Lisaks on mõistlik nõuda, et $L(y^n, y^n) = 0$ ehk õige hinnangu puhul viga pole. Kuna vektor $Y^n$ on juhuslik suurus, siis on seda ka $L(Y^n, a^n)$. Kogu vea leidmiseks võtame keskmise sellest.

**Definitsioon 2.5.** Keskmist kadu nimetatakse **riskiks**:

$$R(a^n) = \mathrm{E}\left[L(Y^n, a^n)\right].$$

Seega parim prognoos on jada, mis minimiseerib riski:

$$\hat{a}^n = \underset{a^n \in \mathcal{Y}^n}{\arg\min}\, R(a^n).$$

**Näide 2.5. Sümmeetriliseks kaofunktsiooniks** nimetatakse kaofunktsiooni

$$L_\infty(y^n, a^n) = \mathbb{I}(y^n \neq a^n) = \begin{cases} 0, & \text{kui } y^n = a^n, \\ 1, & \text{kui } y^n \neq a^n, \end{cases} \tag{2.1}$$

kus $\mathbb{I}$ on indikaatorfunktsioon. Selle risk on

$$R_\infty(a^n) = \mathrm{E}\left[L_\infty(Y^n, a^n)\right] = (1 - \mathrm{E}\left[\mathbb{I}(Y^n = a^n)\right]) = 1 - \mathbf{P}(Y^n = a^n)$$

ning parim prognoos on

$$\hat{a}^n = \underset{y^n \in \mathcal{Y}^n}{\arg\max}\, \mathbf{P}(Y^n = y^n).$$

**Näide 2.6.** Kaofunktsiooni

$$L_1(y^n, a^n) = \sum_{i=1}^{n} \mathbb{I}(y_i \neq a_i) \tag{2.2}$$

nimetatakse **Hammingu kauguseks**, mis on vigade arv kahe vektori vahel. Teatavasti on Hammingu kaugus meetrika ruumil $\mathcal{Y}^n$.

Keskmine vigade arv ehk risk on

$$\begin{aligned} R_1(a^n) &= \mathrm{E}\left[L_1(Y^n, a^n)\right] = \mathrm{E}\left[\sum_{i=1}^{n} \mathbb{I}(y_i \neq a_i)\right] = \sum_{i=1}^{n} (1 - \mathrm{E}\left[\mathbb{I}(y_i = a_i)\right]) \\ &= \sum_{i=1}^{n} (1 - \mathbf{P}(Y_i = a_i)) = n - \sum_{i=1}^{n} \mathbf{P}(Y_i = a_i) \end{aligned}$$

ning seega parim prognoos $\hat{a}^n$ on selline, mille korral

$$\hat{a}_t = \arg\max_{k \in \mathcal{Y}} \mathbf{P}(Y_t = k).$$

### 2.2.1 Murdepunktide kaofunktsioon

Olgu $y^n \in \mathcal{Y}^n$ vektor ja $u_2^n \in \{0,1\}^{n-1}$ murdepunktide vektor. Defineerime kaofunktsiooni

$$L\colon \mathcal{Y}^n \times \{0,1\}^{n-1} \longrightarrow \mathbb{R}^+,$$

kus

$$\begin{aligned} L(y^n, u_2^n) &:= \sum_{t=2}^{n} \left(A\,\mathbb{I}(y_{t-1} = y_t)\,\mathbb{I}(u_t = 1) + B\,\mathbb{I}(y_{t-1} \neq y_t)\,\mathbb{I}(u_t = 0)\right) \\ &= \sum_{t=2}^{n} \left(A\,\mathbb{I}(y_{t-1} = y_t) u_t + B\,\mathbb{I}(y_{t-1} \neq y_t)(1 - u_t)\right). \end{aligned} \tag{2.3}$$

Antud kaofunktsiooni puhul on $A > 0$ valesti prognoosimise hind (tegelikult murdepunkti ei ole) ning $B > 0$ on murdepunkti „mahamagamise“ hind (tegelikult meil on murdepunkt).

Samaväärselt saame selle kaofunktsiooni esitada veel kujul

$$L\colon \{0,1\}^{n-1} \times \{0,1\}^{n-1} \longrightarrow \mathbb{R}^+,$$

kus tee $y^n$ asemel on tema murdepunktide esitus $w_2^n = w(y^n)$ ning defineerime

$$\begin{aligned} L(y^n, u_2^n) = L(w_2^n, u_2^n) &:= \sum_{t=2}^{n} \left(A\,\mathbb{I}(w_t = 0)\,\mathbb{I}(u_t = 1) + B\,\mathbb{I}(w_t = 1)\,\mathbb{I}(u_t = 0)\right) \\ &= \sum_{t=2}^{n} \left(A(1 - w_t)u_t + Bw_t(1 - u_t)\right). \end{aligned}$$

Kui $A = B = 1$, siis on murdepunktide kadu näites 2.6 vaadatud Hammingu kaugus jadade $w_2^n$ ja $u_2^n$ vahel.

*Märkus.* Formaalselt on eespool toodud funktsioonid $L(y^n, u_2^n)$ ja $L(w_2^n, u_2^n)$ erinevad, kuna funktsioonid on määratud erinevatel hulkadel. Lihtsuse mõttes aga tähistame neid edaspidi sama tähisega.

Me saame kaofunktsiooni laiendada ka teede vahelise erinevuse mõõtmiseks, kui defineerime kao kahe tee $y^n, u^n$ vahel kui $L(y^n, u^n) := L(y^n, w(u^n))$. Samuti saame defineerida kao kahe klasterduse $C, C'$ korral kui

$$L(C, C') := L(w(C), w(C')). \tag{2.4}$$

**Näide 2.7.** Olgu $y^{12} = 111122211133, a^{12} = 111111111111, b^{12} = 111222222211$. Siis

$$L(y^{12}, a^{12}) = 3B \quad \text{ja} \quad L(y^{12}, a^{12}) = A + 2B.$$

Paneme tähele, et antud kaofunktsioon kui klasterduste omavaheline kaugus, ei ole invariantne klasterduse indeksite permutatsiooni suhtes. See tähendab, et kaofunktsioon sõltub vaatluste järjekorrast.

**Näide 2.8.** Olgu $K = 4$ ja $n = 7$. Vaatleme kahte klasterdust

$$\begin{aligned} C_1 &= \{\{1, 2, 3\}, \{4, 5\}, \{6\}, \{7\}\}, \\ C_2 &= \{\{1, 2\}, \{3, 4, 5\}, \{6, 7\}\} \end{aligned}$$

ning permutatsiooni $\sigma = (1\ 3\ 2\ 4\ 5\ 6\ 7)$. Siis pärast permutatsiooni rakendamist saame uued klastrid $C_1', C_2'$, kus

$$\begin{aligned} C_1' &= C_1, \\ C_2' &= \{\{1, 3\}, \{2, 4, 5\}, \{6, 7\}\} \end{aligned}$$

ning näeme, et

$$L(C_1, C_2) = A + 2B \quad \text{aga} \quad L(C_1', C_2') = 2A + B.$$

Murdepunktide kaofunktsioonile vastav risk on

$$\begin{aligned}
R(u_2^n) = \mathrm{E}\left[L(Y^n, u_2^n)\right] &= \sum_{t=2}^{n} \mathrm{E}\left[A\,\mathbb{I}(Y_{t-1} = Y_t)u_t + B\,\mathbb{I}(Y_{t-1} \neq Y_t)(1-u_t)\right] \\
&= \sum_{t=2}^{n} \left(A\,\mathrm{E}\left[\mathbb{I}(Y_{t-1} = Y_t)\right]u_t + B\,\mathrm{E}\left[\mathbb{I}(Y_{t-1} \neq Y_t)\right](1-u_t)\right) \\
&= \sum_{t=2}^{n} \left(A\,\mathbf{P}(Y_{t-1} = Y_t)u_t + B\,\mathbf{P}(Y_{t-1} \neq Y_t)(1-u_t)\right)
\end{aligned}$$

ning seega parim murdepunktide vektori prognoos $\hat{u}_2^n$ on

$$\begin{aligned}
\hat{u}_t &= \operatorname*{arg\,min}_{x\in\{0,1\}} \left(A\,\mathbf{P}(Y_{t-1} = Y_t)x + B\,\mathbf{P}(Y_{t-1} \neq Y_t)(1-x)\right) \\
&= \begin{cases} 1, & B\,\mathbf{P}(Y_{t-1} \neq Y_t) \geq A\,\mathbf{P}(Y_{t-1} = Y_t), \\ 0, & B\,\mathbf{P}(Y_{t-1} \neq Y_t) < A\,\mathbf{P}(Y_{t-1} = Y_t). \end{cases}
\end{aligned} \tag{2.5}$$

### 2.2.2 Binderi kaofunktsioon

Olgu $C$ hulga $\{1, \ldots, n\}$ klasterdus ning $c^n = (c_1, \ldots, c_n)$ mingi tee, mis selle klasterduse tekitab ehk $C = C(c^n)$.

**Definitsioon 2.6.** Klasterduse C **binaarseks esituseks** nimetatakse vektorit, mis esitub kujul

$$b(C) := \mathrm{vech}\left((\mathbb{I}(c_i = c_j))_{ij}\right),$$

kus vech annab sümmeetrilise maatriksi $(\mathbb{I}(c_i = c_j))_{ij}$ pool vektoriseeringu, st saadud vektor koosneb maatriksi esimesest veerust, teisest veerust, mis algab teisest reast, kolmandast veerust, mis hakkab kolmandast reast, jne kuni viimane element on $\mathbb{I}(c_n = c_n)$.

Paneme tähele, et binaarne esitus $b(C)$ ei sõltu tee $c^n$ valikust. Tõepoolest, binaarne esitus on tegelikult funktsioon ekvivalentsiseosest ning klasterdus $C$ on faktorhulk, mille ekvivalentsiseose maatriks on $(\mathbb{I}(c_i = c_j))_{ij}$. Aga kaks erinevat teed $c^n$ ja $\overline{c}^n$, mis tekitavad sama klasteduse $C$, peavad tekitama ka sama ekvivalentsiseose ehk mõlema tee puhul on leitud binaarne esitus sama. Samasugune argument õigustab ka all pool toodava Binderi kaofunktsiooni korrektsust.

**Näide 2.9.** Klasterduse $\{\{1,5\},\{2,3\},\{4\}\}$ ekvivalentsiseose maatriks on

$$(\mathbb{I}(c_i = c_j))_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

ning selle binaarne esitus on $(1,0,0,0,1,1,1,0,0,1,0,0,1,0,1)$.

**Definitsioon 2.7.** Olgu $C, C'$ kaks klasterdust. **Binderi kaofunktsiooniks** nimetatakse funktsiooni

$$B(C,C') := \sum_{t<s} \left( A\,\mathbb{I}(c_i = c_j)\,\mathbb{I}(c'_i \neq c'_j) + B\,\mathbb{I}(c_i \neq c_j)\,\mathbb{I}(c'_i = c'_j) \right), \tag{2.6}$$

kus iga paari korral tuleb maksta tasu, kui paar on samas klastris ühe klasterduse järgi aga erinevates klastrites teise klasterduse järgi, või vastupidi.

*Märkus.* Binderi kaofunktsioonis vaatleme summat üle paaride $\{(t,s) \mid 1 \leq t < s \leq n\}$.

Paneme tähele, et kui definitsioonis (2.6) summa üle kõigi paaride asendada summaga üle paaride $t = s-1, s$, siis saame kaofunktsiooni (2.4).

Kui $y^n, a^n \in \mathcal{Y}^n$ on kaks vektorit, siis me saame defineerida Binderi kao vektorite vahel (kuna iga vektor tekitab klasterduse)

$$\begin{aligned} B(y^n, a^n) &:= B(C(y^n), C(a^n)) \\ &= \sum_{t<s} \left( A\,\mathbb{I}(y_i = y_j)\,\mathbb{I}(a_i \neq a_j) + B\,\mathbb{I}(y_i \neq y_j)\,\mathbb{I}(a_i = a'_j) \right) \end{aligned}$$

või klasterduse ja vektori vahel

$$\begin{aligned} B(y^n, C) &:= B(C(y^n), C) \\ &= \sum_{t<s} \left( A\,\mathbb{I}(y_i = y_j)\,\mathbb{I}(c_i \neq c_j) + B\,\mathbb{I}(y_i \neq y_j)\,\mathbb{I}(c_i = c'_j) \right) \end{aligned}$$

või kahe klasterduse binaarsete esituste $b_1 = b(C_1), b_2 = b(C_2)$ vahel kui $B(b_1, b_2) := B(C_1, C_2)$.

Eelnevalt nägime, et murdepunktide kaofunktsion (2.3) ei ole indekside permutatsioonide suhtes invariantne. See-eest Binder jõudis kaofunktsioonini (2.6) just otsides kaofunktsiooni, mis oleks invariantne indeksite permutatsiooni suhtes (Binder, 1981). Indeksite invariantsus tähendab, et kaofunktsioon ei sõltu vaatluste järjekorrast. Sõnastame selle täpsemalt lausena.

**Lause 2.1.** Olgu $C, C'$ kaks klasterdust hulgal $\{1, \ldots, n\}$ ning olgu $c^n, c'^n \in \mathcal{Y}^n$ kaks teed, mille puhul $C = C(c^n)$ ja $C' = C(c'^n)$. Siis kaofunktsioon (2.6) on invariantne indeksite hulga permutatsiooni suhtes, st mistahes permutatsiooni $\pi\colon \{1, \ldots, n\} \to \{1, \ldots n\}$ korral

$$B(C, C') = B(C(c_{\pi(1)}, \ldots, c_{\pi(n)}), C(c'_{\pi(1)}, \ldots, c'_{\pi(n)}));$$

*Tõestus.* Indeksite permutatsioon ei tee midagi muud kui muudab liidetavate järjekorda summas (2.6). See aga summa väärtust ei muuda, mistõttu esimene väide kehtib. □

**Lause 2.2.** Binderi kaofunktsioon on meetrika klasterduste vahel, kui $A = B$.

*Tõestus.* Iga klasterduse põhjal saame konstrueerida tema binaarse esituse ning vastupidi. Kui $A = B$, võtab Binderi kaofunktsioon kuju

$$B(C, C') = A \sum_{t<s} \left( \mathbb{I}(c_i = c_j)\, \mathbb{I}(c'_i \neq c'_j) + \mathbb{I}(c_i \neq c_j)\, \mathbb{I}(c'_i = c'_j) \right),$$

mis on kordajaga A skaleeritud Hammingu kaugus klasterduste binaarsete esituste vahel. Tõepoolest, indeksite $\{(t, s) \mid t \leq s\}$ ja klasterduse binaarse esituse indeksite vahel on üksühene vastavus vech definitsiooni järgi, mistõttu käiakse iga binaarse esituse vektori element täpselt ühe korra läbi. Kui $t = s$, siis vastab sellele mõlema binaaresituse puhul element 1, mistõttu selle indeksi ära jätmine summast kaofunktsiooni väärtust ei muuda.

Niisiis, kuna Hammingu kaugus on meetrika, siis seda on ka Binderi kadu. □

**Näide 2.10.** Olgu $y^{12}, a^{12}, b^{12}$ nagu näites 2.7. Siis

$$\begin{aligned} C(y^{12}) &= \{\{1, 2, 3, 4, 8, 9, 10\}, \{5, 6, 7\}, \{11, 12\}\}, \\ C(a^{12}) &= \{\{1, 2, \ldots, 12\}\}, \\ C(b^{12}) &= \{\{1, 2, 3, 11, 12\}, \{4, 5, 6, 7, 8, 9, 10\}\} \end{aligned}$$

ja

$$\begin{aligned} B(y^{12}, a^{12}) &= \underbrace{4 \cdot 5B}_{t=1,\ldots,4} + \underbrace{3 \cdot 5B}_{t=5,6,7} + \underbrace{3 \cdot 2B}_{t=8,9,10} + \underbrace{0}_{t=11,12} = 41B, \\ B(y^{12}, b^{12}) &= \underbrace{3(4A + 2B)}_{t=1,2,3} + \underbrace{3B}_{t=4} + \underbrace{3 \cdot 3B}_{t=5,6,7} + \underbrace{0}_{t=8,\ldots,12} = 12A + 18B. \end{aligned}$$

### 2.2.3 Ersatz-Binderi kaofunktsioon

**Definitsioon 2.8.** Olgu $C, C'$ kaks klasterdust. **Ersatz-Binderi kaofunktsiooniks** nimetatakse funktsiooni

$$\begin{aligned}\widetilde{B}(C, C') &\coloneqq B(\mathrm{C}_{\mathrm{kan}}(w(C)), \mathrm{C}_{\mathrm{kan}}(w(C'))) \\ &= \sum_{t<s} \big(A\,\mathbb{I}(c_t = c_{t+1} = \ldots = c_s)(1 - \mathbb{I}(c'_t = c'_{t+1} = \ldots = c'_s)) \\ &+ B(1 - \mathbb{I}(c_t = c_{t+1} = \ldots = c_s))\,\mathbb{I}(c'_t = c'_{t+1} = \ldots = c'_s)\big).\end{aligned} \tag{2.7}$$

Niisiis, Ersatz-Binderi kao leidmisel esmalt klasterdus „kanoniseeritakse“ ning seejärel rakendatakse Binderi kadu. Kui klasterdused on juba kanoonilised, siis $\widetilde{B}$ ja $B$ on võrdsed ning võrrandi (2.7) summa all olev avaldis on võrdne võrrandi (2.6) summa all oleva avaldisega.

Analoogselt Binderi kaoga, kui $y^n, a^n$ on kaks vektorit, siis saame defineerida kao vektorite vahel

$$\begin{aligned}\widetilde{B}(y^n, a^n) &\coloneqq \widetilde{B}(C(y^n), C(a^n)) \\ &= \sum_{t<s} \big(A\,\mathbb{I}(y_t = y_{t+1} = \ldots = y_s)(1 - \mathbb{I}(a'_t = a'_{t+1} = \ldots = a'_s)) \\ &+ B(1 - \mathbb{I}(y_t = y_{t+1} = \ldots = y_s))\,\mathbb{I}(a'_t = a'_{t+1} = \ldots = a'_s)\big)\end{aligned}$$

või murdepunktide vektorite vahel $\widetilde{B}(w, w') \coloneqq B(\mathrm{C}_{\mathrm{kan}}(w), \mathrm{C}_{\mathrm{kan}}(w'))$.

**Näide 2.11.** Olgu $y^{12}, a^{12}, b^{12}$ nagu näidetes 2.7 ja 2.10. Siis

$$\begin{aligned}\mathrm{C}_{\mathrm{kan}}(w(C(y^{12}))) &= \{\{1,2,3,4\}, \{5,6,7\}, \{8,9,10\}\,\{11,12\}\}, \\ \mathrm{C}_{\mathrm{kan}}(w(C(a^{12}))) &= \{\{1,2,\ldots,12\}\}, \\ \mathrm{C}_{\mathrm{kan}}(w(C(b^{12}))) &= \{\{1,2,3\}, \{4,5,6,7,8,9,10\}, \{11,12\}\}\end{aligned}$$

ja

$$\begin{aligned}\widetilde{B}(y^{12}, a^{12}) &= \underbrace{4 \cdot 8B}_{t=1,\ldots,4} + \underbrace{3 \cdot 5B}_{t=5,6,7} + \underbrace{3 \cdot 2B}_{t=8,9,10} + \underbrace{0}_{t=11,12} = 53B, \\ \widetilde{B}(y^{12}, b^{12}) &= \underbrace{3A}_{t=1,2,3} + \underbrace{6B}_{t=4} + \underbrace{3 \cdot 3B}_{t=5,6,7} + \underbrace{0}_{t=8,\ldots,12} = 3A + 15B.\end{aligned}$$

### 2.2.4 Binderi ja Ersatz-Binderi kaofunktsiooni risk

Olgu $Y_1, \ldots, Y_n$ mingid juhuslikud suurused hulgal $\mathcal{Y}$ ja $C$ klasterdus hulgal $\{1, \ldots, n\}$, mille tekitab tee $c^n \in \mathcal{Y}^n$. Võtame antud osas kasutusele vastavad tähistused $(1 \leq t < s \leq n)$

$$\begin{aligned} I_{ts} &\coloneqq \mathbb{I}(Y_t = Y_s), & I_{ts} &= \mathbb{I}(Y_t = Y_{t+1} = \ldots = Y_s), \\ i_{ts} &\coloneqq \mathbb{I}(c_t = c_s), & i_{ts} &= \mathbb{I}(c_t = c_{t+1} = \ldots = c_s), \\ p_{ts} &\coloneqq \mathbf{P}(Y_t = Y_s), & p_{ts} &\coloneqq \mathbf{P}(Y_t = Y_{t+1} = \ldots = Y_S), \end{aligned}$$

kus vasakus ja paremas tulbas olevaid tähistusi kasutame, kui peame vastavalt silmas Binderi või Ersatz-Binderi kaofunktsiooni.

Paneme tähele, et kasutades keskväärtuse lineaarsust, avaldub risk Binder ja Ersatz-Binder kaofunktsiooni korral järgmiselt

$$\begin{aligned} R(C) = \mathrm{E}\left[B(Y^n, C)\right] &= \mathrm{E}\left[\sum_{t<s} A I_{ts}(1 - i_{ts}) + B(1 - I_{ts}) i_{ts}\right] \\ &= \sum_{t<s} \left(A\,\mathrm{E}[I_{ts}](1 - i_{ts}) + B(1 - \mathrm{E}[I_{ts}]) i_{ts}\right) \qquad (2.8) \\ &= \sum_{t<s} \left(A(1 - i_{ts}) p_{ts} + B i_{ts}(1 - p_{ts})\right). \end{aligned}$$

**Järeldus 2.1.** Võttes $A = B$, saame, et

$$R(C) = A \sum_{t<s} |i_{ts} - p_{ts}|.$$

*Tõestus.* Teisendades saame

$$\begin{aligned} R(C) &= \sum_{t<s} \left(A(1 - i_{ts}) p_{ts} + B i_{ts}(1 - p_{ts})\right) \\ &= A \sum_{t<s} \left(p_{ts} - 2 i_{ts} p_{ts} + i_{ts}\right) = A \sum_{t<s} |i_{ts} - p_{ts}|, \end{aligned}$$

kus viimase võrduse saame, kui vaatame läbi juhud $i_{ts} = 0$ ja $i_{ts} = 1$ ning võtame saadud tulemused kokku. □

Lau ja Green näitasid järgmist (Lau ja Green, 2007).

**Järeldus 2.2.** Riski (2.8) minimiseerimine on samaväärne funktsiooni

$$\sum_{t<s} i_{ts}\left(p_{ts} - \frac{B}{A+B}\right)$$

maksimiseerimisega.

*Tõestus.* Paneme tähele, et

$$\begin{aligned}
R(C) &= \sum_{t<s} (A(1-i_{ts})p_{ts} + Bi_{ts}(1-p_{ts})) \\
&= \sum_{t<s} Ap_{ts} - \sum_{t<s} Ai_{ts}p_{ts} + \sum_{t<s} Bi_{ts}(1-p_{ts}) \\
&= \sum_{t<s} Ap_{ts} + \sum_{t<s} i_{ts}(B - Bp_{ts} - Ap_{ts}) \\
&= A\sum_{t<s} p_{ts} - (A+B)\sum_{t<s} i_{ts}\left(p_{ts} - \frac{B}{A+B}\right).
\end{aligned}$$

□

Märkame, et järelduse 2.2 saab teisendada järelduseks 2.1 ja vastupidi, kui $A = B$. Nimelt

$$\begin{aligned}
&-\arg\max \sum_{t<s} i_{ts}\left(p_{ts} - \frac{B}{A+B}\right) = \arg\min\left(-\sum_{t<s} i_{ts}p_{ts} + \frac{1}{2}\sum_{t<s} i_{ts}\right) \\
&= \arg\min\left(\frac{1}{2}\sum_{t<s} p_{ts} - \sum_{t<s} i_{ts}p_{ts} + \frac{1}{2}\sum_{t<s} i_{ts}\right) = \arg\min \sum_{t<s} (p_{ts} - 2i_{ts}p_{ts} + i_{ts}) \\
&= \arg\min \sum_{t<s} |i_{ts} - p_{ts}|,
\end{aligned}$$

kus arg min ja arg max võetakse üle klasterduste, kui on tegu Binderi kaofunktsiooniga, või üle kanooniliste klasterdute, kui on tegu Ersatz-Binderi kaofunktsiooniga.

Niisiis, Binderi kaofunktsiooni korral on parim klasterdus $\widehat{C}$ selline, mille korral

$$\widehat{C} = \underset{C\in\mathcal{C}(n,K)}{\arg\min} \sum_{t<s} (A\,\mathbb{I}(c_t \neq c_s)\,\mathbf{P}(Y_t = Y_s) + B\,\mathbb{I}(c_t = c_s)\,\mathbf{P}(Y_t \neq Y_s)) \tag{2.9}$$

või

$$\widehat{C} = \underset{C\in\mathcal{C}(n,K)}{\arg\max} \sum_{t<s} \mathbb{I}(c_t = c_s)\left(\mathbf{P}(Y_t = Y_s) - \frac{B}{A+B}\right) \tag{2.10}$$

või $A = B$ korral

$$\widehat{C} = \underset{C \in \mathcal{C}(n,K)}{\arg\min} \sum_{t<s} |\, \mathbb{I}(c_t = c_s) - \mathbf{P}(Y_t = Y_s)|,$$

kus $\mathcal{C}(n, K)$ on kõikide ülimalt $K$ elemendiliste klasterduste hulk hulgal $\{1, \ldots, n\}$.

See-eest Ersatz-Binderi korral on parim klasterdus $\widehat{C}$ selline, kus

$$\begin{aligned}\widehat{C} = \underset{C \in \mathcal{C}_{\text{kan}}(n)}{\arg\min} \sum_{t<s} \big(&A(1 - \mathbb{I}(c_t = \ldots = c_s))\, \mathbf{P}(Y_t = \ldots = Y_s) \\ &+ B\, \mathbb{I}(c_t = \ldots = c_s)(1 - \mathbf{P}(Y_t = \ldots = Y_s))\big)\end{aligned} \tag{2.11}$$

või

$$\widehat{C} = \underset{C \in \mathcal{C}_{\text{kan}}(n)}{\arg\max} \sum_{t<s} \mathbb{I}(c_t = \ldots = c_s) \left( \mathbf{P}(Y_t = \ldots = Y_S) - \frac{B}{A+B} \right) \tag{2.12}$$

või $A = B$ korral

$$\widehat{C} = \underset{C \in \mathcal{C}_{\text{kan}}(n)}{\arg\min} \sum_{t<s} |\, \mathbb{I}(c_t = \ldots = c_s) - \mathbf{P}(Y_t = \ldots = Y_s)|,$$

kus $\mathcal{C}_{\text{kan}}(n)$ on kõikide kanooniliste klasterduste hulk hulgal $\{1, \ldots, n\}$, seejuures klastrite arvul ülempiiri ei ole.

Märgime, et kordajate $A, B$ väärtuste valimisel ei ole konkreetsed väärtused olulised vaid hoopis nende suhe. Nimelt, kui $A > B$, siis leitakse potentsiaalselt vähem murdepunkte, kui juhul $A = B$. See-eest kui $A < B$, siis on leitud murdepunktide arv potentsiaalselt suurem, kui juhul $A = B$.

Eelnevalt näitasime, et Binderi kaofunktsioon kätkeb endas ka murdepunktide kaofunktsiooni. Seega kui võrrandis (2.9) või (2.11) vaadata summat üle paaride $t = s - 1, s$, peaksime saama murdepunktide prognoosi hinnangu (2.5).

Vaatame summat üle paaride $t = s - 1, s$. Siis võtab optimiseerimisülesanne (2.9) kuju

$$\widehat{C} = \underset{C \in \mathcal{C}(n,K)}{\arg\min} \sum_{t=2}^{n} \left( A\, \mathbb{I}(c_t \neq c_{t-1})\, \mathbf{P}(Y_t = Y_{t-1}) + B\, \mathbb{I}(c_t = c_{t-1})\, \mathbf{P}(Y_t \neq Y_{t-1}) \right).$$

Siin $\mathbb{I}(c_t \neq c_{t-1})$ tähendab, et indeks $t$ on murdepunkti, ning $\mathbb{I}(c_t = c_{t-1})$ tähendab, et $t$ ei ole murdepunkt. Seega kui tähistada $\hat{u}_2^n = w(\widehat{C})$ ja otsida optimaalset

lahendit üle murdepunktide, saame, et

$$\begin{aligned}\hat{u}_2^n &= \underset{u_2^n \in \{0,1\}^{n-1}}{\arg\min} \sum_{t=2}^{n} \left(A\,\mathbf{P}(Y_t = Y_{t-1})\,\mathbb{I}(u_t = 1) + B\,\mathbf{P}(Y_t \neq Y_{t-1})\,\mathbb{I}(u_t = 0)\right) \\ &= \underset{u_2^n \in \{0,1\}^{n-1}}{\arg\min} \sum_{t=2}^{n} \left(A\,\mathbf{P}(Y_t = Y_{t-1})u_t + B\,\mathbf{P}(Y_t \neq Y_{t-1})(1-u_t)\right),\end{aligned}$$

mis ongi (2.5).

Kui nüüd võrrandites (2.10) või (2.12) toimida analoogselt, saame sama tingimuse murdepunktide prognoosi leidmiseks aga teisel kujul:

$$\begin{aligned}\hat{u}_2^n &= \underset{u_2^n \in \{0,1\}^{n-1}}{\arg\max} \sum_{t=2}^{n} (1-u_t)\left(\mathbf{P}(Y_t = Y_{t-1}) - \frac{B}{A+B}\right) \\ &= \underset{u_2^n \in \{0,1\}^{n-1}}{\arg\min} \sum_{t=2}^{n} u_t\left(\mathbf{P}(Y_t = Y_{t-1}) - \frac{B}{A+B}\right)\end{aligned} \tag{2.13}$$

ehk

$$\begin{aligned}\hat{u}_t &= \underset{x \in \{0,1\}}{\arg\min}\, x\left(\mathbf{P}(Y_t = Y_{t-1}) - \frac{B}{A+B}\right) \\ &= \begin{cases} 1, & \frac{B}{A+B} \geq \mathbf{P}(Y_t = Y_{t-1}), \\ 0, & \frac{B}{A+B} < \mathbf{P}(Y_t = Y_{t-1}). \end{cases}\end{aligned} \tag{2.14}$$

Teine võimalus seoste (2.13) ja (2.14) näitamiseks on tõestuse 2.2 skeemi matkida murdepunktide kaofunktsiooni (2.3) jaoks.

## 2.3 Optimiseerimisülesande lahendamine

Osutub, et optimiseerimisülesande (2.9) lahendamine on NP-raske (Bansal, Blum ja Chawla, 2004). See on ka peamine motivatsioon, miks me tõime sisse Ersatz-Binderi kaofunktsiooni mõiste. Nimelt näitame, et optimiseerimisülesannet (2.11) on võimalik lahendada kasutades dünaamilist planeerimist.

Võtame kasutusele tähistused:

$$p_{t,s} := \mathbf{P}(Y_t = Y_{t+1} = \ldots = Y_s), \quad i_{ts} := \mathbb{I}(c_t = \ldots = c_s), \quad t < s,$$

$$f(l,k) := A \sum_{t \le l < s \le k} p_{t,s}, \quad f(0,k) := 0, \quad 1 \le l < k \le n, \tag{2.15}$$

$$h(l,k) := B \sum_{l \le t < s \le k} (1 - p_{t,s}), \quad h(k,k) := 0, \quad 1 \le l < k \le n, \tag{2.16}$$

$$\delta(0) = \delta(1) := 0,$$

$$\delta(k) := \min_{C \in \mathcal{C}_{kan}(k)} \sum_{t < s \le k} \left( A(1 - i_{ts}) p_{ts} + B i_{ts}(1 - p_{ts}) \right), \quad k = 2, \ldots, n.$$

**Teoreem 2.1.** Kehtib järgmine rekursioon:

$$\delta(k+1) = \min_{l \in \{0,\ldots,k\}} \left\{ \delta(l) + f(l, k+1) + h(l+1, k+1) \right\}.$$

*Tõestus.* Vaatleme kõiki klasterdusi hulgas $\mathcal{C}_{kan}(k+1)$, mille viimane murdepunkt on $i$, kus $i \in \{2, \ldots, k+1\}$, ning leiame nende seast sellise, mis minimiseerib Ersatz-Binder riski. Olgu see riski minimiseeriv klasterdus $C$ ja ning olgu $y^{k+1}$ tee, mis tekitab selle klasterduse, st $C = C(y^{k+1})$.

Et tegu on kanoonilise klasterdusega, siis

1. peavad kõik elemendid $y_i, \ldots, y_{k+1}$ kuuluma samasse klastrisse ja
2. ükski element $y_1, \ldots, y_{i-1}$ ei kuulu klastrisse $\{y_i, \ldots, y_{k+1}\}$.

Kuna klasterdus $C$ minimiseerib riski üle kõigi kanooniliste klasterduste, mille viimane mudepunkt on $i$, siis peab elementide $y_1, \ldots, y_{i-1}$ klasterdus $\widetilde{C}$ (see on klasterdus, mille me saame, kui klasterdusest $C$ jätame klastri $\{y_i, \ldots, y_{k+1}\}$ välja) olema selline, mis annab riski $\delta(i-1)$, kuna vastasel korral saaksime koostada

elementide $y_1, \ldots, y_{k+1}$ klasterduse, mille risk oleks väiksem kui $R(C)$. Seega

$$\begin{aligned}
R(C) &= \sum_{t<s} (A(1-i_{ts})p_{t,s} + Bi_{ts}(1-p_{t,s})) \\
&= \sum_{t<s\leq i-1} (A(1-i_{ts})p_{t,s} + Bi_{ts}(1-p_{t,s})) \\
&+ \sum_{t\leq i-1<s\leq k+1} Ap_{t,s} + \sum_{i\leq s<t\leq k+1} B(1-p_{t,s}) \\
&= \delta(i-1) + f(i-1,k+1) + h(i,k+1),
\end{aligned}$$

kus $h(i,k+1)$ ja $f(i-1,k+1)$ on riski osad, mis on tingitud vastavalt tingimustest 1 ja 2.

Olgu $C \in \mathcal{C}_{kan}(k+1)$, mille korral $R(C) = \delta(k+1)$. Kuna me ei tea, mis on klasterduse viimane murdepunkt või kas klasterdus koosneb ainult ühest klastrist, siis võtame miinimumi üle kõikvõimalike murdepunktide. Seega

$$\delta(k+1) = \min_{l\in\{0,\ldots,k\}} \{\delta(l) + f(l,k+1) + h(l+1,k+1)\}.$$

□

*Märkus.* Kui rekursioon lahti kirjutada, siis võtab see kuju

$$\begin{aligned}
\delta(k+1) = \min\{&\delta(k) + f(k,k+1), \\
&\delta(k-1) + f(k-1,k+1) + h(k,k+1), \\
&\delta(k-2) + f(k-2,k+1) + h(k-1,k+1), \\
&\ldots \\
&\delta(1) + f(1,k+1) + h(2,k+1), \\
&\delta(0) + h(1,k+1)\}.
\end{aligned}$$

**Näide 2.12.** Olgu $A = B = 1$. Leiame $\delta(2)$ ja $\delta(3)$ (all on kirjas, mis klasterdusele antud kadu vastab):

$$\begin{aligned}
\delta(2) &= \min\{\delta(0) + h(1,2), \delta(1) + f(1,2)\} = \min\{\underbrace{1 - p_{12}}_{\{1,2\}}, \underbrace{p_{12}}_{\{1\},\{2\}}\} \\
\delta(3) &= \min\{\delta(0) + h(1,3), \delta(1) + f(1,3) + h(2,3), \delta(2) + f(2,3)\} \\
&= \min\{\delta(0) + h(1,3), \delta(1) + f(1,3) + h(2,3), \min\{1 - p_{12}, p_{12}\} + p_{13} + p_{23}\} \\
&= \min\{\underbrace{(1 - p_{12}) + (1 - p_{13}) + (1 - p_{23})}_{\{1,2,3\}}, \underbrace{p_{12} + p_{13} + (1 - p_{23})}_{\{1\},\{2,3\}}, \\
&\qquad \underbrace{(1 - p_{12}) + p_{13} + p_{23}}_{\{1,2\},\{3\}}, \underbrace{p_{12} + p_{13} + p_{23}}_{\{1\},\{2\},\{3\}}\}.
\end{aligned}$$

Lause 2.1 annab meile viisi, kuidas rekursiivselt lahendada optimiseerimisülesannet (2.11). Olgu $\widehat{C}$ võrrandit (2.11) minimiseeriv klasterdus. Leiame järjest $\delta(1), \delta(2)$ jne kuni lõpuks leiame $\delta(n)$, mille puhul kehtib

$$\delta(n) = \mathrm{E}\,\widetilde{B}(Y^n, \widehat{C}) = R(\widehat{C}).$$

Tähistame nüüd

$$\psi(k+1) := \underset{l \in \{0,\ldots,k\}}{\arg\min}\{\delta(l) + f(l, k+1) + h(l+1, k+1)\}$$

ning rekursiooni igal sammul salvestame $\psi(k+1)$. Selle tulemusel saame arvud $\psi(1) := 0, \psi(2), \psi(3), \ldots, \psi(n)$, kus $\psi(k) \in [0, k-1]$. Kasutades neid arve, saame tagant ettepoole tulles leida murdepunktid

$$r_1 = \psi(n), \quad r_2 = \psi(r_1), \quad \ldots, \quad r_{j+1} = \psi(r_j), \quad \ldots, \quad r_m = \psi(1) = 0,$$

kus $r_1 + 1$ on klasterduse viimane murdepunkt, mis jaotab klasterduse kaheks osaks $[1, r_1]$ ja $[r_1 + 1, n]$; $r_2 + 1$ on elementide $1, \ldots, r_1$ klasterduse viimane murdepunkt, mis jaotab klasterduse kaheks osaks $[1, r_2]$ ja $[r_2 + 1, r_1]$ jne. Järelikult tulevad murdepunktid

$$\tau_1 = r_{m-1} + 1, \quad \tau_2 = r_{m-2} + 1, \quad \cdots, \quad \tau_{m-1} = r_1 + 1$$

ning vastav klasterdus on

$$\widehat{C} = \{[1, r_{m-1}], [r_{m-1} + 1, r_{m-2}] \cdots [r_3 + 1, r_2], [r_2 + 1, r_1], [r_1 + 1, n]\}.$$

Kõik sammud on kokku võetud algoritmis 2.1.

**Algoritm 2.1: Ersatz-Binderi kaofunktsioonile vastava riski minimiseerimine.**

1. Fikseerime algväärtused: $\delta(0) = \delta(1) := 0$.
2. Iga $k = 1, \ldots, n-1$ korral leiame
$$\delta(k+1) = \min_{l \in \{0,\ldots,k\}} \{\delta(l) + f(l, k+1) + h(l+1, k+1)\}$$
ja
$$\psi(k+1) = \arg\min_{l \in \{0,\ldots,k\}} \{\delta(l) + f(l, k+1) + h(l+1, k+1)\}.$$
3. Tagant ettepoole tulles leiame punktid
$$r_1 = \psi(n),\ r_2 = \psi(r_1),\ \ldots,\ r_{j+1} = \psi(r_j),\ \ldots,\ r_m = \psi(1) = 0.$$
4. Tagastame murdepunktid
$$\tau_1 = r_{m-1} + 1, \quad \tau_2 = r_{m-2} + 1, \quad \cdots, \quad \tau_{m-1} = r_1 + 1$$
ning klasterduse
$$\widehat{C} = \{[1, r_{m-1}], [r_{m-1}+1, r_{m-2}] \cdots [r_3+1, r_2], [r_2+1, r_1], [r_1+1, n]\}.$$

### 2.3.1 Arvutuslikud märkused

Osutub, et suuruste $f$ ja $h$ leidmine otse definitsioonide (2.15) ja (2.16) põhjal ei ole mõistlik. Nimelt peab igal sammul tegema arvutusi, mida on tehtud eelmisel sammul, ning seetõttu võib algoritm joosta suurema sisendi korral ootatust aeglasemalt. Toome ära mõningad rekursiivsed seosed, mida saab $f$ ja $h$ leidmiseks kasutada.

**Lause 2.3.** Olgu $1 \le l < k$. Kehtivad järgmised seosed:

1. $f(l,k) + A\sum_{t:t\le l} p_{t,k+1} = f(l,k+1)$, (2.17)
2. $f(l,k) + A(\sum_{s:l+1<s\le k} p_{l+1,s} - \sum_{t\le l} p_{t,l+1}) = f(l+1,k)$, (2.18)
3. $h(l,k) + B\sum_{t:l\le t\le k}(1 - p_{t,k+1}) = h(l,k+1)$, (2.19)
4. $h(l+1,k) + B\sum_{s:l<s\le k}(1 - p_{ls}) = h(l,k)$. (2.20)

*Tõestus.* Tähistame $h_t(k) = B\sum_{s:t<s\le k}(1-p_{ts})$. Siis

$$h(l,k) = B\sum_{t,s:\ l\le t<s\le k}(1-p_{ts}) = \sum_{t:\ l\le t<k} h_t(k).$$

1. Kehtib

$$\begin{aligned} f(l,k+1) &= A\sum_{t,s:\ t\le l<s\le k+1} p_{ts} = A\left(\sum_{t,s:\ t\le l<s\le k} p_{ts} + \sum_{t:\ t\le l} p_{t,k+1}\right) \\ &= f(l,k) + A\sum_{t:\ t\le l} p_{t,k+1}. \end{aligned}$$

2. Näeme, et

$$\begin{aligned} f(l+1,k) &= A\sum_{t,s:\ t\le l+1<s\le k} p_{ts} = A\left(\sum_{t,s:\ t\le l<l+1<s\le k} p_{ts} + \sum_{s:\ l+1<s\le k} p_{l+1,s}\right) \\ &= A\left(\sum_{t,s:\ t\le l<s\le k} p_{ts} + \sum_{s:\ l+1<s\le k} p_{l+1,s} - \sum_{t:\ t\le l} p_{t,l+1}\right) \\ &= f(l,k) + A\left(\sum_{s:\ l+1<s\le k} p_{l+1,s} - \sum_{t:\ t\le l} p_{t,l+1}\right). \end{aligned}$$

3. Märkame, et

$$\begin{aligned} h(l,k+1) &= \sum_{t:\ l\le t<k+1} h_t(k+1) \\ &= \sum_{t:\ l\le t<k+1} (h_t(k) + B(1-p_{t,k+1})) \\ &= \sum_{t:\ l\le t<k+1} h_t(k) + B\sum_{t:\ l\le t<k+1}(1-p_{t,k+1}) \\ &\overset{(*)}{=} \sum_{t:\ l\le t<k} h_t(k) + B\sum_{t:\ l\le t\le k}(1-p_{t,k+1}) \\ &= h(l,k) + B\sum_{t:\ l\le t\le k}(1-p_{t,k+1}), \end{aligned}$$

kus võrdus $(*)$ tuleb sellest, et $h_k(k) = 0$.

4. Paneme tähele

$$h(l,k) = \sum_{t:\ l\le t<k} h_t(k) = \sum_{t:\ l+1\le t<k} h_t(k) + h_l(k)$$
$$= h(l+1,k) + h_l(k) = h(l+1,k) + B \sum_{s:\ l<s\le k} (1-p_{ls}).$$

□

Oletame, et meil on tõenäosused $p_{ts}$ olemas. Siis paneme tähele, et $h$ arvutamiseks definitsiooni (2.16) põhjal tuleb teha $^{(k-l)(k-l+1)}/_2 - 1$ tehet, mis tähendab, et $h(1,2), h(2,3), \ldots$ leidmisel ei peagi ühtegi tehet tegema. Lause 2.3 põhjal tuleb seosete (2.19) ja (2.20) kasutamisel teha vastavalt $k-l+1$ ja $k-l$ tehet.

Niisiis, $h$ arvutamiseks on kaks varianti. Alguses leiame kõigepealt $h(1,2)$, $h(2,3)$, $h(3,4)$ jne. Seejärel kas kasutame seost (2.19) ning leiame $h(1,2) \to h(1,3) \to h(1,4)$, $h(2,3) \to h(2,4)$ jne (liigume ülevalt alla) või kasutame seost (2.20) ja leiame $h(2,3) \to h(1,3)$, $h(3,4) \to h(2,4) \to h(1,4)$ jne (liigume paremalt vasakule). Mõlemal juhul teeme kokku sama palju liitmistehteid. Ülevaate skeemist annab joonis 1.

Funktsiooni $f$ arvutamiseks on mõistlik kasutada seost (2.17), kuna nähtavasti seos (2.18) sisaldab rohkem arvutustehteid. Mõistlik on toimida järgmiselt. Igal sammul $k$ leiame ja jätame meelde suurused $f(1,k), \ldots, f(k-1,k)$. Kasutades seost (2.17) leiame $f(1,k+1), \ldots, f(k-1,k+1)$ ning arvutame definitsiooni põhjal suuruse $f(k,k+1)$. Nüüd võib $f(1,k), \ldots, f(k-1,k)$ ära unustada ning hoida meeles hoopis suuruseid $f(1,k+1), \ldots, f(k,k+1)$.

Oletame nüüd, et $(Y,X)$ on varjatud Markovi ahel ning meil on olemas vaatlused $x^n \in \mathbb{R}^n$. Uurime, kuidas arvutada tõenäosusi $\mathbf{P}(Y_t = \ldots = Y_s | X^n = x^n)$, kus $1 \le t < s \le n$.

Paneme tähele, et kui me soovime, siis me võime tõenäosuse $\mathbf{P}(Y_t = \ldots = Y_s | X^n = x^n)$ asemel leida tõenäosuse $\mathbf{P}(Y_t = \ldots = Y_s, X^n = x^n)$, kuna vaatlused on fikseeritud. Olgu $y_t^s \in \mathcal{Y}^{s-t+1}$. Siis kasutades lemmat 1.1, saame

$$\begin{aligned} p(x^n, y_t^s) &= p(x^s, y_t^s) p(x_{s+1}^n \mid y_s) = p(x^t, y_t) p(x_{t+1}^s, y_{t+1}^s | y_t) p(x_{s+1}^n \mid y_s) \\ &= p(x^t, y_t) p(y_{t+1}^s | y_t) p(x_{t+1}^s | y_t^s) p(x_{s+1}^n \mid y_s) \\ &= \alpha(y_t, x^t) \left( \prod_{i=t}^{s-1} p_{y_i, y_{i+1}} \right) \left( \prod_{i=t+1}^{s} f_{y_i}(x_i) \right) \beta(x_{s+1}^n | y_s). \end{aligned}$$

$$
\begin{tikzcd}
h(1,2),[0] \arrow[d, "2"] & & & & \\
h(1,3),[2] \arrow[d, "3"] & h(2,3),[0] \arrow[l, "2"'] \arrow[d, "2"] & & & \\
h(1,4),[5] \arrow[d, "4"] & h(2,4),[2] \arrow[l, "3"'] \arrow[d, "3"] & h(3,4),[0] \arrow[l, "2"'] \arrow[d, "2"] & & \\
h(1,5),[9] \arrow[d, "5"] & h(2,5),[5] \arrow[l, "4"'] \arrow[d, "4"] & h(3,5),[2] \arrow[l, "3"'] \arrow[d, "3"] & h(4,5),[0] \arrow[l, "2"'] \arrow[d, "2"] & \\
h(1,6),[14] & h(2,6),[9] \arrow[l, "5"'] & h(3,6),[5] \arrow[l, "4"'] & h(4,6),[2] \arrow[l, "3"'] & h(5,6),[0] \arrow[l, "2"']
\end{tikzcd}
$$

Joonis 1: Ülevaade $h$ arvutamisest. Kantsulgude sees on kujutatud liitmistehete arvu, mida tuleb $h(l,k)$ leidmiseks teha. Nooltega on kujutatud võimalikku arvutusskeemi, kus noole peal on ära toodud üleminukul tehtavate arvutuste arv. Märgime, et kõik nooled on ümberpööratavad.

Seega

$$
\begin{aligned}
\mathbf{P}(Y_t = \ldots = Y_s, X^n = x^n) &= \sum_{j \in \mathcal{Y}} \mathbf{P}(Y_t = j, \ldots, Y_s = j, X^n = x^n) \\
&= \sum_{j \in \mathcal{Y}} \alpha(j, x^t)(p_{jj})^{s-t} \left( \prod_{i=t+1}^{s} f_j(x_i) \right) \beta(x_{s+1}^n | j).
\end{aligned}
$$

# 3 Segmenteerimine

Eelmises peatükis uurisime, kuidas leida erinevate kaofunktsioonide ja nende poolt tekitatud riskide korral parimat murdepunktide prognoosi, mis asetab meid kahesammulisel segmenteerimisel poolele teele. Selles peatükis näitame lähemalt, kuidas leitud prognoosi ära kasutada

Kui $(Y, X)$ on HMM, siis **segmenteerimiseks** nimetatakse vaatluste $x^n$ põhjal „parima“ tee $y^n$ leidmist. Parimuse kriteeriumi määrab ära kaofunktsioon ja tema poolt tekitatud risk (nagu peatükis 2.2). Ainuke erinevus on see, et meil on olemas ka vaatlused $x^n$, mistõttu vaatleme riski asemel tinglikut riski

$$R(a^n|x^n) = \mathrm{E}\, L(Y^n, a^n|X^n = x^n) = \sum_{y^n \in \mathcal{Y}^n} L(y^n, a^n)\, \mathbf{P}(Y^n = y^n|X^n = x^n),$$

kus $a^n$ on mingisugune prognoos ja $L$ on mingisugune kaofunktsioon. Eesmärk on uurida lähemalt olukordi, kus kaofunktsiooniks on sümmeetriline kaofunktsioon (näide 2.5) või Hammingu kaugus (näide 2.6).

Lause 1.1 põhjal teame, et $Y^n$ tinglik jaotus tingimusel $X^n = x^n$ on mittehomogeenne Markovi ahel. Seetõttu võime vaadata järgmist probleemi: kui $Y^n$ on mittehomogeenne Markovi ahel, siis kuidas leida mingit riski minimiseerivat teed. Kui oskame seda probleemi lahendada, siis leitud lahenduse saab lause 1.1 põhjal teisendada ümber HMM-i jaoks.

## 3.1 Tähistused

Olgu $u_2^n$ murdepunktide vektori prognoos, mis on saadud siis kas (2.5), (2.9) või (2.11) lahendamisel. Olgu $Y_1, Y_2, \ldots, Y_n$ mittehomogeenne Markovi ahel seisundite hulgaga $\mathcal{Y}$. Iga $j \in \mathcal{Y}$ korral tähistame algtõenäosusi järgnevalt

$$\pi_j = \mathbf{P}(Y_1 = j).$$

Kui $1 \le u < v$, $1 \le t$, $y^n \in \mathcal{Y}^n$ ja $i, j \in \mathcal{Y}$, siis võtame kasutusele järgmised tähistused

$$p_t(y_t) := \mathbf{P}(Y_t = y_t), \qquad p_t(y_t|y_u^v) := \mathbf{P}(Y_t = y_t \mid Y_u^v = y_u^v),$$
$$p(y_u^v) := \mathbf{P}(Y_u^v = y_u^v), \qquad p_{i,j}^{(t+1)} := \mathbf{P}(Y_{t+1} = j \mid Y_t = i).$$

Meenutame, et murdepunktide kaofunktsioon oli defineeritud kui

$$\begin{aligned} L(y^n, u_2^n) &:= \sum_{t=2}^{n} \left( A\,\mathbb{I}(y_{t-1} = y_t)\,\mathbb{I}(u_t = 1) + B\,\mathbb{I}(y_{t-1} \neq y_t)\,\mathbb{I}(u_t = 0) \right) \\ &= \sum_{t=2}^{n} \left( A\,\mathbb{I}(y_{t-1} = y_t) u_t + B\,\mathbb{I}(y_{t-1} \neq y_t)(1 - u_t) \right). \end{aligned} \tag{2.3}$$

Võtame selles peatükis kasutusele tähise

$$m_{ij}(t) := A\,\mathbb{I}(i = j) u_t + B\,\mathbb{I}(i \neq j)(1 - u_t)$$

ehk $m_{ij}(t)$ on üks liidetav murdepunktide kaofunktsioonist. Seega esitub (2.3) kujul

$$L(y^n, u_2^n) = \sum_{t=2}^{n} m_{y_{t-1}, y_t}(t).$$

## 3.2 Viterbi algoritm murdepunktide prognoosi korral

Näites 2.5 vaatasime sümmeetrilist kaofunktsiooni

$$L_\infty(y^n, a^n) = \begin{cases} 0, & \text{kui } y^n = a^n, \\ 1, & \text{kui } y^n \neq a^n, \end{cases} \tag{2.1}$$

mille korral oli riski

$$R_\infty(a^n) = \mathrm{E}\, L_\infty(Y^n, a^n) = 1 - P(Y^n = a^n)$$

minimiseerimine samaväärne suuruse $\mathbf{P}(Y^n = a^n)$ maksimiseerimisega. Sellist sobivat teed ehk kõige tõenäolisemat teed

$$v^n = \arg\max_{y^n \in \mathcal{Y}^n} p(y^n) \tag{3.1}$$

nimetatakse **Viterbi joonduseks** ning vastavat algortimi selle leidmiseks **Viterbi algoritmiks**.

Vaatame ülesande (3.1) kitsentatud varianti, kus me otsime teed, mille murdepunktide vektor kattub meie prognoositud murdepunktide vektoriga $u_2^n$:

$$v^n = \arg\max_{y^n : L(y^n, u_2^n) = 0} p(y^n). \tag{3.2}$$

Siin $L$ on murdepunktide kaofunktsioon (2.3) ning me otsime kõige tõenäolisemat teed, mille murdepunktid oleksid täpselt $u_2^n$. Selle ülesande lahendamise eesmärk

on ära kasutada informatsiooni, mida me saime murdepunktide prognoosimisega.

Tähistame

$$\delta_t(j) := \max_{\substack{y^t, y_t = j \\ L(y^t, u_2^t)=0}} p(y^t).$$

**Lause 3.1.** Kehtib rekursioon

$$\delta_{t+1}(j) = \max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} p_{ij}^{(t+1)} \delta_t(i).$$

*Tõestus.* Kui $y^{t+1} \in \mathcal{Y}^{t+1}$ on tee, mis lõpeb punktis $j$ (st $y_{t+1} = j$), siis saame teisendada

$$p(y^{t+1}) = p(y^t)p(y_{t+1} \mid y^t) = p(y^t)p(y_{t+1} \mid y_t) = p(y^t)p_{y_t,j}^{(t+1)}.$$

Vaatleme teesid $s^t \in \mathcal{Y}^t$, mille lõpp-punkt on $i$ ning mille murdepunktide vektor on $u_2^t$ (st $L(s^t, u_2^t) = 0$), ning valime neist sellise, mille korral on suurus $p(s^t)p_{s_t,j}^{(t+1)}$ maksimaalne. Sel juhul maksimiseerides

$$\max_{\substack{y^t, y_t = i \\ L(y^t, u_2^t)=0}} p(y^{t+1}) = \max_{\substack{y^t, y_t = i \\ L(y^t, u_2^t)=0}} p_{ij}^{(t+1)} p(y^t) = p_{ij}^{(t+1)} \max_{\substack{y^t, y_t = i \\ L(y^t, u_2^t)=0}} p(y^t) = p_{ij}^{(t+1)} p(s^t)$$

ning seega on tee $s^t$ selline, mille korral $\delta(i) = p(s^t)$. Järelikult

$$\delta_{t+1}(j) = \max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} p_{ij}^{(t+1)} \max_{\substack{y^t, y_t = i \\ L(y^t, u_2^t)=0}} p(y^t) = \max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} p_{ij}^{(t+1)} \delta_t(i).$$

□

Lause 3.1 põhjal saame koostada algoritmi ülesande (3.2) lahendamiseks (vt algoritm 3.1).

Olgu murdepunktide prognoosi $u_2^n$ murdepunktide indeksid $\tau_1 < \tau_2 < \ldots < \tau_k$, kus $\tau_i = j$ kui indeks $j$ on $i$-ndas murdepunkt. Sõltuvalt kordajate $A, B$ väärtustest sõltub uuendamise tingimus. Kui $A, B > 0$, siis on leitud joonduse $v^n$ kõik murdepunktid ette antud (st murdepunktid on täpselt $\tau_1, \ldots, \tau_k$)

$$\delta_{t+1}(j) = \max_{\substack{i \in \mathcal{Y} \\ i=j \text{ kui } u_{t+1}=0 \\ i \neq j \text{ kui } u_{t+1}=1}} \delta_t(i) p_{ij}^{(t+1)};$$

kui $A > 0; B = 0$, siis on kõik indeksid $\{\tau_1, \ldots, \tau_k\}$ joonduse $v^n$ murdepunktid aga

**Algoritm 3.1: Viterbi joonduse leidmine murdepunktide vektori prognoosi korral.**

1. Fikseerime rekursiooni alguse: $\forall j \in \mathcal{Y} : \delta_1(j) = \pi_j$.
2. Iga $j \in \mathcal{Y}$ ja $t = 1, \ldots, n-1$ korral leiame uue $\delta_{t+1}(j)$ lause 3.1 põhjal ning salvestame valitud seisundi muutujasse $i_{t+1}(j)$:

$$\delta_{t+1}(j) = \max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} \delta_t(i) p_{ij}^{(t+1)},$$
$$i_{t+1}(j) = \arg\max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} \delta_t(i) p_{ij}^{(t+1)}.$$

3. Tagant ettepoole tulles leiame kitsentatud Viterbi joonduse:

$$v_n = \arg\max_j \delta_n(j), \quad v_t = i_{t+1}(v_{t+1}) \quad t = n-1, n-2, \ldots, 1.$$

4. Tagastame kitsentatud Viterbi joonduse $v^n = (v_1, v_2, \ldots, v_n)$.

neid võib-olla rohkem

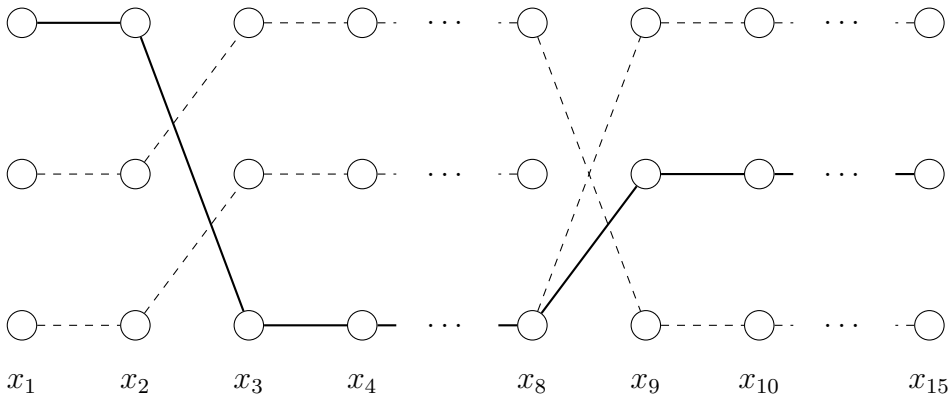
$$\delta_{t+1}(j) = \max_{\substack{i \in \mathcal{Y} \\ i \neq j \text{ kui } u_{t+1}=1}} \delta_t(i) p_{ij}^{(t+1)};$$

kui $A = 0; B > 0$, siis sisalduvad $v^n$ murdepunktid hulgas $\{\tau_1, \ldots, \tau_k\}$ aga neid võib-olla vähem

$$\delta_{t+1}(j) = \max_{\substack{i \in \mathcal{Y} \\ i = j \text{ kui } u_{t+1}=0}} \delta_t(i) p_{ij}^{(t+1)}.$$

**Näide 3.1.** Vaatleme olukorda, kus on meil $t = 15$ ajaseisundit ning ainsad murdepunktid on $t = 3$ ja $t = 9$. Lisaks olgu $A > 0$ ja $B > 0$. Antud olukorda on kujutada joonisel 2, kus punktiirjooned kujutavad väärtusi $i_t(j)$ ning jämedam joon tähistsab leitud joondust.

Esimesel sammul murdepunkte ei ole, mistõttu ainuke võmalus on samas seisus edasi olla. Sammul kaks on murdepunkt, mistõttu peab seisundit vahetama ning ei saa samasse seisundisse jääda. Peale seda ei saa enam seisundit vahetada, kuni jõutakse murdepunktini $t = 9$, mille korral peab seisundit muutma. Jõudes lõppu, käijakse mööda tuldud teed tagasi ning leitakse parim tee, mis on märgitud joonisel jämedalt.




Joonis 2: Näide kitsendatud Viterbi algoritmi jooksutamisest.

## 3.3 PMAP-joondus murdepunktide prognoosi korral

Sümmeetrilise kaofunktsiooni (2.1) puhul ei mängi vigade arv rolli (st indeksite arv, kus kaks teed erinevad teineteisest) — kadu kahe tee vahel on sama, hoolimata sellest kas erinevus on ühes kohas või mitmes kohas. Seetõttu tasub vaadelda punktiviisilist kaofunktsiooni, mis arvestab ka vigade arvuga.

Näites 2.6 vaadatud Hammingu kaugus

$$L_1(y^n, a^n) = \sum_{i=1}^{n} \mathbb{I}(y_i \neq a_i) \tag{2.2}$$

loeb paarikaupa kokku vigade arvu kahe vektori vahel. Parim prognoos $v^n$ on selline, mille korral

$$v_t = \arg\max_{k \in \mathcal{Y}} \mathbf{P}(Y_t = k), \tag{3.3}$$

seejuures joondust $v^n$ nimetatakse **PMAP-joonduseks** (ingl *pointwise maximum a posteriori alignment*).

Vaatleme kitsendatud ülesannet, kus otsime PMAP-joonduse (3.3) analoogi, mis kasutaks ka murdepunktide prognoosi $u_2^n$

$$v^n = \arg\max_{y^n : L(y^n, u_2^n) = 0} \sum_{t=1}^{n} p_t(y_t).$$

Tähistame

$$\delta_t(j) := \max_{\substack{y^t, y_t = j \\ L(y^t, u_2^t) = 0}} \sum_{i=1}^{t} p_i(y_i).$$

**Lause 3.2.** Kehtib rekursioon

$$\delta_t(j) = \max_{i \in \mathcal{Y}:\ m_{ij}(t)=0} \delta_{t-1}(i) + p_t(j), \quad t \geq 2.$$

*Tõestus.* Paneme tähele, et

$$\begin{aligned}\delta_t(j) &= \max_{\substack{y^t, y_t = j \\ L(y^t, u_2^t)=0}} \sum_{i=1}^{t} p_i(y_i) = \max_{\substack{y^{t-1}, y_t = j \\ L(y^t, u_2^t)=0}} \sum_{i=1}^{t-1} p_i(y_i) + p_t(j) \\ &= \max_{i \in \mathcal{Y}:\ m_{ij}(t)=0} \delta_{t-1}(i) + p_t(j).\end{aligned}$$

□

Antud lause põhjla saame koostada algoritmi 3.2. Sõltuvalt kordajate $A, B$ väärtustest saab kitsendatud PMAP joonduse uuendamise tingimuse lahti kirjutada nagu sai tehtud Viterbi joonduse puhul.

**Algoritm 3.2: PMAP joonduse leidmine murdepunktide vektori prognoosi korral.**

1. Fikseerime rekursiooni alguse: $\forall j \in \mathcal{Y} : \delta_1(j) = \pi_j$.
2. Iga $j \in \mathcal{Y}$ ja $t = 1, \ldots, n-1$ korral leiame $\delta_{t+1}(j)$ ning salvestame valitud seisundi muutujasse $i_{t+1}(j)$:

   $$\delta_{t+1}(j) = \max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} \delta_t(i) + p_{t+1}(j)$$

   ja

   $$i_{t+1}(j) = \arg\max_{i \in \mathcal{Y}:\ m_{ij}(t+1)=0} \delta_t(i) + p_{t+1}(j).$$

3. Tagant ettepoole tulles leiame joonduse:

   $$v_n = \arg\max_{j \in \mathcal{Y}} \delta_n(j), \quad v_t = i_{t+1}(v_{t+1}), \quad t = n-1, n-2, \ldots, 1.$$

4. Tagastame joonduse $v^n = (v_1, v_2, \ldots, v_n)$.

## 3.4 Etteantud murdepunktide arv

Praktikas on teinekord teada murdepunktide arv või selle ülemine/alumine piir ning me sooviksime seda informatsiooni ära kasutada. Seega kui on teada, et on (täpselt/vähemalt/ülimalt) $k$ murdepunkti, siis me otsime teatavas mõttes parimat teed nende teede hulgast, kus on (täpselt/vähemalt/ülimalt) $k$ murdepunkti.

Üks variant on leida kõik murdepunktide vektorid, kus on $k$ murdepunkti, ning jooksutada ühte eelnevat algoritmi. Paraku on see üsna ressursikulukas ja ei ole efektiivne. Näiteks kui on vaja leida parim prognoos, kus on täpselt $k$ murdepunkti, siis selliseid murdepunktide vektoreid on $\binom{n-1}{k}$ ning see arv on suurim, kui $k = \lfloor {}^{n-1}/_{2} \rfloor$ (halvim juht). Sellise $k$ korral

$$\frac{2^{n-1}}{n} \leq \binom{n-1}{k}.$$

Siin aitab tähelepanek, et kaks erinevat murdepunktide vektorit võvad omavahel kattuda (näiteks 01101 ja 01110 kattuvad esimese kolme murdepunkti osas) ning me ei pea mõlema jaoks rakendama eelnevat algoritmi täies ulatuses.

Olgu $\mu(y^n)$ tee $y^n$ murdepunktide arv. Otsime lahendust järgmistele optimismeerimisülesannetele:

$$\max_{y^n\colon\ \mu(y^n)=k} p(y^n) \tag{3.4}$$

ja

$$\max_{y^n\colon\ \mu(y^n)=k} \sum_{t=1}^{n} p_t(y_t), \tag{3.5}$$

kus märk = maksimumi tingimuses võib olla nii $\geq$ kui ka $\leq$.

Lahendame neist (3.4). Siis ülesande (3.5) lahendamine on analoogne. Tähistame

$$\delta_t^m(j) \coloneqq \max_{\substack{y^t, y_t=j \\ \mu(y^t)=m}} p(y^t).$$

**Lause 3.3.** Kehtib rekursioon

$$\delta_{t+1}^m(j) = \max\left(\delta_t^m(j)p_{jj}^{(t+1)}, \max_{i\in\mathcal{Y}\setminus\{j\}} \delta_t^{m-1}(i)p_{ij}^{(t+1)}\right), \quad t \geq 1,\ m \geq 1.$$

*Tõestus.* Kasutades Markovi omadust ja arvestades, et uuel sammul kas on murdepunkt või ei ole, saame

$$\begin{aligned}
\delta_{t+1}^m(j) &= \max\left(\max_{\substack{y^t, y_t=j\\ \mu(y^t)=m}} p(y^t)p_{y_t,j}^{(t+1)}, \max_{\substack{y^t, y_t\neq j\\ \mu(y^t)=m-1}} p(y^t)p_{y_t,j}^{(t+1)}\right)\\
&= \max\left(\delta_t^m(j)p_{jj}^{(t+1)}, \max_{i\in\mathcal{Y}\setminus\{j\}} \max_{\substack{y^t, y_t=i\\ \mu(y^t)=m-1}} p(y^t)p_{y_t,j}^{(t+1)}\right)\\
&= \max\left(\delta_t^m(j)p_{jj}^{(t+1)}, \max_{i\in\mathcal{Y}\setminus\{j\}} \delta_t^{m-1}(i)p_{ij}^{(t+1)}\right).
\end{aligned}$$

□

Maksimumi esimene element kirjeldab juhtu, kus on juba eelnevalt läbitud tees $m$ murdepunkti ning seega ainuke variant on jääda samasse seisundisse. Teine element vaatleb olukorda, kus varasema tee puhul on $m-1$ murdepunkti, mistõttu tuleb antud ajahetkel hüpata teise seisundisse.

Saame algoritmi.

**Algoritm 3.3: Viterbi algoritm murdepunktide arvu korral.**

1. Fikseerime rekursiooni alguse:

$$\forall j \in \mathcal{Y} : \delta_1^0(j) := \pi_j,\ \delta_1^{-1}(j) := 0;$$
$$\forall t = 1, \ldots, n : \delta_t^t := 0.$$

2. Iga $t = 1, \ldots, n-1$ ja $m = \alpha(t+1), \ldots, \beta(t+1)$ korral leiame

$$\delta_{t+1}^m(j) = \max\left(\delta_t^m(j)p_{jj}^{(t+1)}, \max_{i\in\mathcal{Y}\setminus\{j\}} \delta_t^{m-1}(i)p_{ij}^{(t+1)}\right),$$

$$k_{t+1}^m(j) = \begin{cases} 1, & \text{kui } \delta_t^m(j)p_{jj}^{(t+1)} < \max_{\substack{i=1,\ldots,n\\ i\neq j}} \delta_t^{m-1}(i)p_{ij}^{(t+1)},\\ 0, & \text{kui } \delta_t^m(j)p_{jj}^{(t+1)} \geq \max_{\substack{i=1,\ldots,n\\ i\neq j}} \delta_t^{m-1}(i)p_{ij}^{(t+1)}, \end{cases}$$

$$i_{t+1}^m(j) = \begin{cases} j, & \text{kui } k_{t+1}^m(j) = 0, \\ \arg\max_{\substack{i=1,\dots,n \\ i \neq j}} \delta_t^{m-1}(i) p_{ij}^{(t+1)}, & \text{kui } k_{t+1}^m(j) = 1. \end{cases}$$

kus $\alpha(t) = \max(0, t-(n-k))$ ja $\beta(t) = \min(t-1, k)$. Siin $\alpha$ tagab, et prognoosi leidmise lõpuks oleks olemas vähemalt $k$ murdepunkti, ning $\beta$ tagab, et prognoosis ei oleks rohkem kui $k$ murdepunkti (vt joonist 3).

3. Tagant ettepoole tulles leiame joonduse, samal ajal pidades arvet tekkinud murdepunktide üle:

$$\begin{aligned} v_n &:= \arg\max_{j \in \mathcal{Y}} \delta_n^k(j), & l_n &:= k, \\ v_{n-1} &= i_n^{l_n}(v_n), & l_{n-1} &= l_n - \mathbb{I}(v_n \neq v_{n-1}), \\ &\dots & &\dots \\ v_t &= i_{t+1}^{l_{t+1}}(v_{t+1}), & l_t &= l_{t+1} - \mathbb{I}(v_{t+1} \neq v_t), \\ &\dots & &\dots \\ v_1 &= i_2^{l_2}(v_2). \end{aligned}$$

4. Tagastame joonduse $v^n = (v_1, \dots, v_n)$.

Vaatame ka juhtu, kui märk $=$ võrrandi (3.4) maksimumi tingimuses on $\geq$ või $\leq$ (vt joonis 4). Kui on vaja, et $\mu(y^n) \leq k$, siis võtame iga $t$ korral $\alpha(t) = 0$ ja

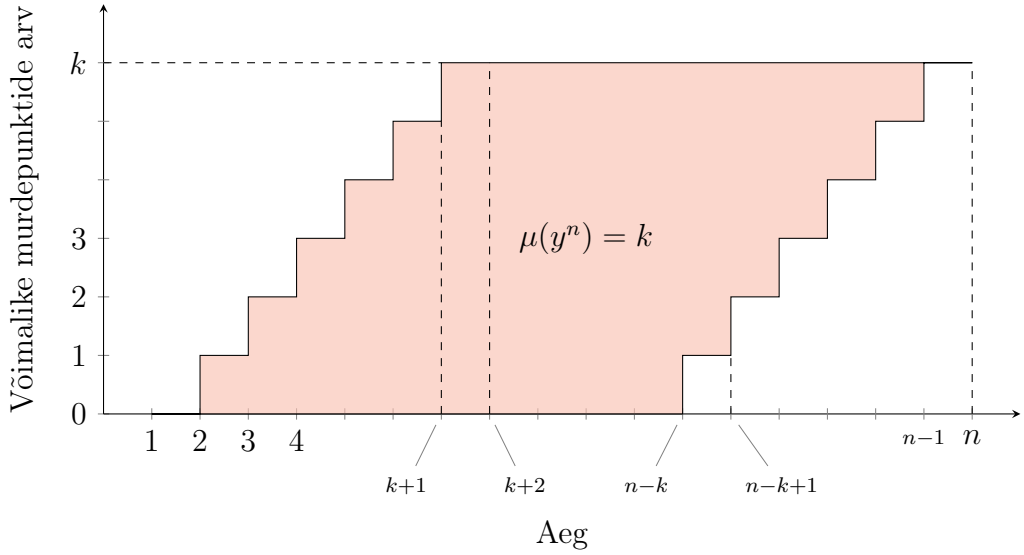
$$(v_n, l_n) := \arg\max_{j \in \mathcal{Y},\ m=0,\dots,k} \delta_n^m(j).$$

Kui on vaja, et $\mu(y^n) \geq k$, siis võtame iga $t$ korral $\beta(t) = t-1$ ja

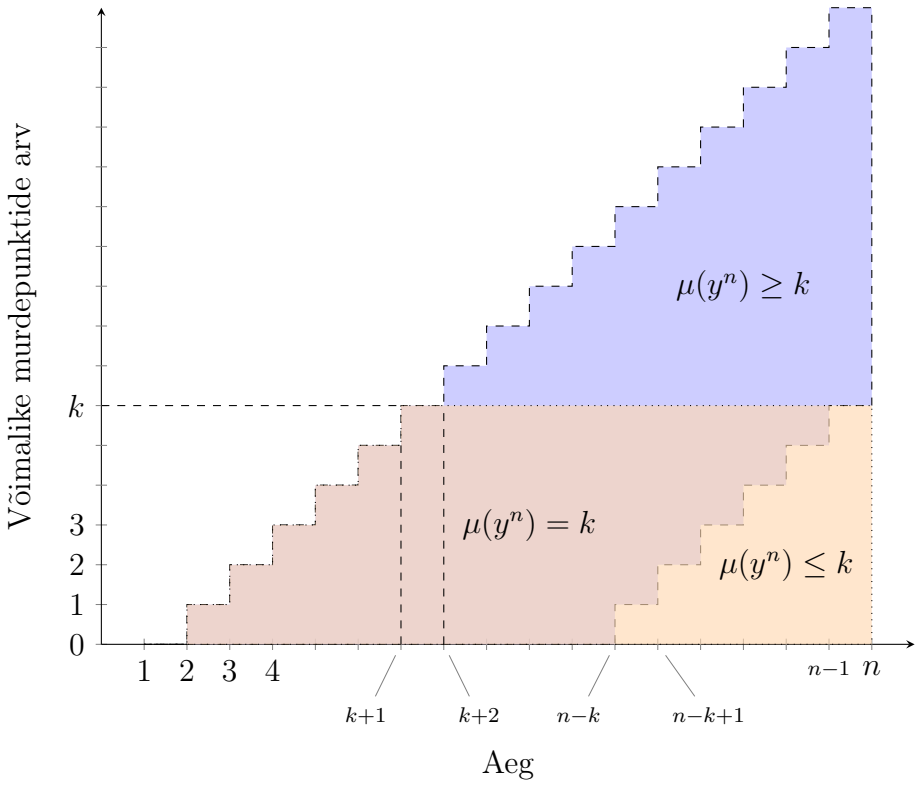
$$(v_n, l_n) := \arg\max_{j \in \mathcal{Y},\ m=k,\dots,n-1} \delta_n^m(j).$$

Kui on vaja, et $k_1 \leq \mu(y^n) \leq k_2$, siis saame kombineerida eelmised variandid võttes $\alpha(t) = \max(0, t-(n-k_1))$, $\beta(t) = \min(t-1, k_2)$ ja

$$(v_n, l_n) := \arg\max_{j \in \mathcal{Y},\ m=k_1,\dots,k_2} \delta_n^m(j).$$




Joonis 3: Tingimus $\mu(y^n) = k$, kus värvitud ala näitab võimalike murdepunktide arvu sõltuvust ajahetkest. Ülesannete (3.4) ja (3.5) lahendi $a^n$ iteratiivsel leidmisel, on $a_1$ korral 0 murdepunkti, $a_2$ korral võib olla kuni 1 murdepunkt, $a_3$ korral kuni 2 murdepunkti ja nii edasi, kuni $a_{k+1}$ korral ja edasi ei saa olla rohkem kui $k$ murdepunkti. Hetkel $n - k$ ei pea olema veel ühtegi murdepunkti, aga hetkel $n - k + 1$ peab olema vähemalt üks murdepunkt, kuna vastasel korral ei saa järgmise $n-(n-k+1) = k-1$ saamu jooksul vajalikku $k$ murdepunkti kokku. Hetkel $n-2$ võib olla kaks murdepunkti puudu, $n-1$ korral võib olla üks murdepunkt puudu ning hetkel $n$ peab olema täpselt $k$ murdepunkti.




Joonis 4: Võimalike murdepunktide arv tingimusel $\mu(y^n) \geq k$ on toodud välja sinise värvi ja kriipsjoonega. Võimalike murdepunktide arv tingimusel $\mu(y^n) \leq k$ on toodud välja oranži värvi ja punktjoonega. Seal, kus need piirkonnad kattuvad, saame murdepunktide arvu tingimusel $\mu(y^n) = k$ nagu näidatud joonisel 3.

# 4 Simulatsioonid

Selle peatüki eesmärk on veenduda töö käigus leitud algoritmide korrektsuses ning illustreerida neid ühe konkreetse mudeli jaoks. Me ei proovi kuidagi öelda kas üks algoritm on parem kui teine, kuna iga leitud joondus on mingi kriteeriumi mõttes parim.

Nimelt vaatleme järgmist HMM-i, kus üleminekumaatriks on

$$P = \begin{pmatrix} 0.97 & 0.03 \\ 0.02 & 0.98 \end{pmatrix}$$

ning emissioonijaotused on $\mathcal{N}(0,1)$ ja $\mathcal{N}(1,1)$. Üleminekumaatriks on valitud nii, et pika ajaperioodi peale oleks vähe murdepunkte. Me genereerisime sellest mudelist ühe vaatluste tee $x^{400}$ ja sellele vastava varjatud seisundite tee $y^{400}$.

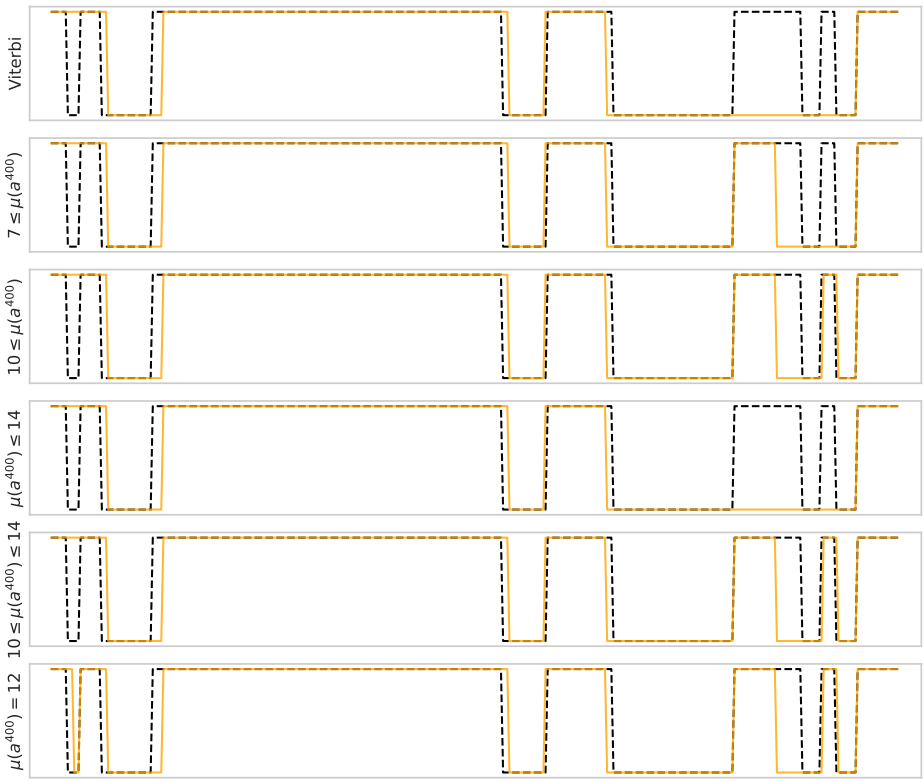
Esimeses osas vaatleme segmenteerimist, kus me anname ette murdepunktide arvu, ning teises osas vaatleme kaheastmelist segmenteerimist. Mõlemal juhul eeldame, et meie mudel on õige, st mudeli parameetrid vastavad tegelikult kasutatud parameetritele.

Simulatsioonideks loodud kood ning genereeritud andmed asuvad GitLabis (Alesma, 2024). Koodi koostamisel on osaliselt kasutatud koodi teegist `hmmlearn` (2014).

## 4.1 Kitsentatud Viterbi algoritm ette antud murdepunktide arvu korral

Vaatleme selles peatükis Viterbi joonduse (3.1) ja kitsentatud Viterbi joonduse (3.4) toimist andmete $x^{400}$ puhul. Kui tegeliku tee murdepunktide arv on $d$, siis ette antud murdepunktide arvud on valitud järgmiselt: murdepunkte on vähemalt üks rohkem, kui Viterbi joonduse murdepunktide arv; murdepunkte on vähemalt $d-2$; murdepunkte on ülimalt d+2; murdepunktide arv on vahemikus $[d-2, d+2]$ ja murdepunkte on täpselt $d$.

Joonisel 5 on kujutatud saadud tulemusi. Näeme, et Viterbi joondus magab maha kolm erinevat seisundi vahetust, mis on Viterbi joonduse puhul tavaline. Kui me nõuame, et oleks vähemalt 7 murdepunkti, siis leiame üles ühe seisundi vahetuse, mida Viterbi algoritm ei leidnud. Nõudes, et oleks vähemalt 10 murdepunkti, leiame Viterbi algoritmist kaks seisundi vahetust rohkem üles. Kui murdepunktide arv on ülimalt 14, siis leiame Viterbi joonduse, kuna Viterbi joonduses on 6 murdepunkti ning sisaldub võimalike teede hulgas. Juhul kui murdepunktide arv on 10 ja 14 vahel, saame tagasi seisundi nagu kolmandas reas. Kui tahame, et murdepunktide




Joonis 5: Andmete $x^{400}$ segmenteerimisel saadud joondused. Musta punktiirjoonega on kujutatud tee $y^{400}$ ning oranži joonega prognoositud joondus $a^{400}$. Joonise vasakul on välja toodud kas tegu on tavalise Viterbi joondusega või on võimalikele murdepunktide arvule seatud piirang.

arv oleks täpselt 12 (see on tõelise tee $y^{400}$ murdepunktide arv), siis õnnestub enam-vähem edukalt leida kõik seisundi vahetused.

## 4.2 Kaheastmeline segmenteerimine

Vaatleme selle peatükis juhtu, kus teostame vaatluste $x^{400}$ peal kaheastmelist segmenteerimist kasutates kitsentatud Viterbi algoritmi — nimelt leiame kõigepealt murdepunktide vektori prognoosi kas lahendades ülesande murdepunktide kaofunktsiooni (2.5) või Ersatz-Binderi kaofunktsiooni (2.11) puhul ning seejärel leiame sellele vastava joonduse lahendades ülesande (3.2). Ülesannet (2.5) ja (2.11) lahendamisel fikseerime $A = 1$ ning muudame $B$ väärtust nii, et käime läbi nende korda-

jate erinevad suhted. Murdepunktide kaofunktsiooni korral vaatleme suuremaid $B$ väärtusi, kui teeme seda Ersatz-Binderi kaofunktsiooni puhul. See on tingitud asjaolust, et murdepunktide kaofunktsiooni korral leiame tõenäosused, mis sõltuvad ainult kahest järjestikusest ajasammust, aga Ersatz-Binderi puhul on vaadeldavad ajavahemikud pikemad; lisaks on Markovi ahela üleminekumaatriksi diagonaal peaaegu üks. See tähendab, et murdepunktide kaofunktsiooni puhul leitud tõenäosused on suuremad kui on need Ersatz-Binderi korral.

Kitsendatud Viterbi joonduse (3.2) leidmisel vaatame kolme juhtu: leitud joonduse murdepunktid peavad täpselt kattuma murdepunktide vektori prognoosiga, leitud joonduses võib olla vähem murdepunkte ja leitud joonduses võib olla rohkem murdepunkte. Saadud tulemused on toodud lisas joonistel 6, 8, 10 ja 7, 9, 11. Prognoositud murdepunktid on märgitud katkentliku püstkriipsuga.

Kui otsitava joonduse puhul lubada rohkem murdepunkte, kui on murdepunktide vektori prognoosis, siis on mõlema kaofunktsiooni korral saadud joonised sarnased.

Kui otsitavas joonduses võib murdepunkte olla vähem kui murdepunktide vektori prognoosis, siis me saame mõlemal juhul midagi Viterbi joonduse sarnast või täpselt Viterbi joonduse. Seejuures saame täpselt Viterbi joonduse siis, kui murdepunktide vektor sisaldab Viterbi joonduse murdepunkte.

Jooniselt näeme, et mõlema kaofunktsiooni puhul esineb tihti järgmine muster — minnakse ühest seisundist üle teise ning seejärel hüppatakse kohe mõne ajasammu järel tagasi (nendes punktides tekivad joonisel teravikud). Seejuures selline hüppamine tehakse umbes ajahetke kandis, kus tegelikult esineb ka murdepunkt, ning kui neid hüppamisi on mitu, siis need on tihti koos. Plokkide sees, kus tegelik joondus on pikalt paigal, ei esine sellist hüppamist. Problemaatilised on kohad, kus tee $y^{400}$ vahetab tihti seisundit.

Spekuleerime siin kohal miks selline nähtus esineb. Mõlema kaofunktsiooni puhul sõltuvad leitavad tõenäosused vaatlustest. Kui olla pika ploki keskel, mis koosneb samadest seisunditest, siis jääb antud ajahetkest mõlemale poole palju vaatlusi, mis on umbes tegeliku seisundi lähedal. Need vaatlused aitavad leitavat tõenäosust siluda, isegi siis kui antud ajahetkel vastav vaatlus ei vasta üldse tegelikule seisundile. Kui vaadelda hetki plokkide otstes, siis seal on nende vaatluste arv, mis aitavad tõenäosusi siluda, väiksem. Kui mudeli varjatud osa vahetab sageli seisundit (nagu on joonise lõpus), siis on siluvaid vaatlusi veel vähem.

Ersatz-Binderi kaofunktsiooni puhul võib murdepunktide kuhjumist lisaks veel põhjustada asjaolu, et seda on kaofunktsioonil odavam teha. Vaatleme sellist olukorda, kus tegelik klasterdus on $\{\{1,\ldots,50\},\{51,\ldots,150\}\}$ ning nõuame, et meil oleks kaks murdepunkti. Paigutame esimese murdepunkti indeksile 51. Kui me paigutaksime teise murdepunkti klastri $\{51,\ldots,150\}$ keskele ning vaatleksime Ersatz-Binderi kaofunktsiooni tegeliku klasterduse ja klasterduse $\{\{1,\ldots,50\},\{51,\ldots,100\},$

$\{101, \ldots, 150\}\}$ vahel, siis me teeksime vea $50^2$ indeksite paari korral. Kui paigutaksime teise murdepunkti klastri $\{1, \ldots, 50\}$ keskele, siis teeksime vea $25^2$ indeksite paari korral. Võttes indeksi 52 teiseks murdepunktiks, tekib klasterdus $\{\{1, \ldots, 50\}, \{51\}, \{52, \ldots, 150\}\}$ ning teeksime vea ainult 99 paari korral.

Teravike tekkimine murdepunktide läheduses, nende kuhjumine ja asjaolu, et vaadeldav protsess tahab pikalt ühes seisundis olla, motiveerib katsetama järgmist — leiame prognoosis üles murdepunktid $x_1$ ja $x_2$, kus $|x_1 - x_2| \leq 5$, ning asendame need kaks elementi arvuga $\left\lfloor \frac{x_1+x_2}{2} \right\rfloor$; kordame seda seni, kuni enam sellised murdepunkte ei leia. Selle tulemusel saadud joondused on lisas joonistel 12, 14, 16 ja 13, 15, 17. Joonistelt võib näha, et sellise protsessi käigus saadud joondused on mõistlikumad kui need, mis me enne nägime; st et ei ole täheldada teravike joonistel. Veel tasub märkida, et leitud joondused on nii murdepunktide kaofunktsiooni kui ka Ersatz-Binderi kaofunktsiooni korral üsna sarnased.

# Kokkuvõte

Selles töös uurisime kaheastmeliset segmenteerimist varjatud Markovi ahelate korral. Tõime välja erinevad kaofunktsioonid, mida saab kasutada murdepunktide vektori prognoosimiseks, ning neile vastavad riskid. Erilist huvi pakub Binderi kaofunktsioonile vastava riski minimiseerimine, mis on muidu NP-raske probleem. Selle jaoks tõime sisse uue kaofunktsiooni, mis põhineb Binderi kaofunktsioonil ning mille puhul näitasime, kuidas minimiseerida sellele vastavat riski kasutades dünaamilist planeerimist.

Kolmandas peatükis kirjeldasime ära algoritmid kaheastmeliseks segmenteerimiseks. Need algoritmid on saadud Viterbi ja PMAP optimiseerimisülesannete kitsentatud variantide lahendamisel. Lisaks vaatasime juhtu, kui meil on olemas täpne või ligikaudne hinnang võimalikule murdepunktide arvule.

PMAP joonduse leidmisel ei arvestata üleminekutõenäosustega ning seetõttu võibolla saadud joonduse tõenäosus väga väike. Uurimata jäi kas kaheastmeline segmenteerimine kitsentatud PMAP-i korral võib leida joondusi, mille tõenäosused on väga väiksed, või aitab kitsendav optimeerimistingimus seda vältida. Veel jäi vaatamata juht, kus me otsime kitsentatud Viterbi või PMAP joondust, mis vastab mingile ette antud klasterdusele.

Praktilises osas illustreerisime kitsentatud Viterbi algoritme ühe konkreetse mudeli jaoks. Nimelt vaatasime juhte, kus meil on murdepunktide arvule ette antud mingi vahemik ning vaatasime kaheastmelist segmenteerimist nii murdepunktide kui ka Ersatz-Binderi kaofunktsiooni korral.

Et saada kaheastmelisest segmenteerimisest parem ülevaade, tuleks tulevikus läbi viia põhjalikud ja süstemaatilised simulatsioonid. Nimelt tuleks vaadelda kaheastmelist segmenteerimist rohkem kui kahe seisundi korral ning ühtlasi vaadelda ka segmenteerimist kitsentatud PMAP-i korral. Lahtiseks jäi veel kuidas valida konstantide väärtusi erinevates kaofunktsioonides ning kuidas käituvad algoritmid, kui me ei segmenteeri õige mudeli korral.

# Kasutatud allikad


Alesma, Karl-Joan (2024). *Segmenteerimine läbi murdepunktide kood.* URL: https://gitlab.com/karl-joan/segmentation-with-change-points (vaadatud 15.05.2024).

Bansal, Nikhil, Avrim Blum ja Shuchi Chawla (2004). "Correlation Clustering". *Machine Learning* 56, lk. 89–113.

Binder, David A. (1981). "Approximations to Bayesian Clustering Rules". *Biometrika* 68.1, lk. 275–285.

Grassl, Richard ja Oscar Levin (2018). *More Discrete Mathematics via Graph Theory.* URL: https://discrete.openmathbooks.org/more/mdm/sec_adv-stirling.html (vaadatud 14.03.2024).

Gundersen, Gregory (2020). *The Log-Sum-Exp Trick.* URL: https://gregorygundersen.com/blog/2020/02/09/log-sum-exp/ (vaadatud 01.05.2024).

Lau, John W. ja Peter J. Green (2007). "Bayesian Model-Based Clustering Procedures". *Journal of Computational and Graphical Statistics* 16.3, lk. 526–558.

Lember, Jüri (2017). *Tehisõpe I — loengukonspekt.* URL: https://courses.ms.ut.ee/MTMS.02.046/2017_spring/uploads/Main/tehispe17.pdf (vaadatud 03.03.2024).

Mitmed autorid (2014). *hmmlearn: Hidden Markov Models in Python, with scikit-learn like API.* URL: https://github.com/hmmlearn/hmmlearn (vaadatud 14.05.2024).

Rabiner, L.R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE* 77.2, lk. 257–286.

Wade, Sara ja Zoubin Ghahramani (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls". *Bayesian Analysis* 13.2, lk. 559–626.

# Lisa 1. Simulatsiooni joonised

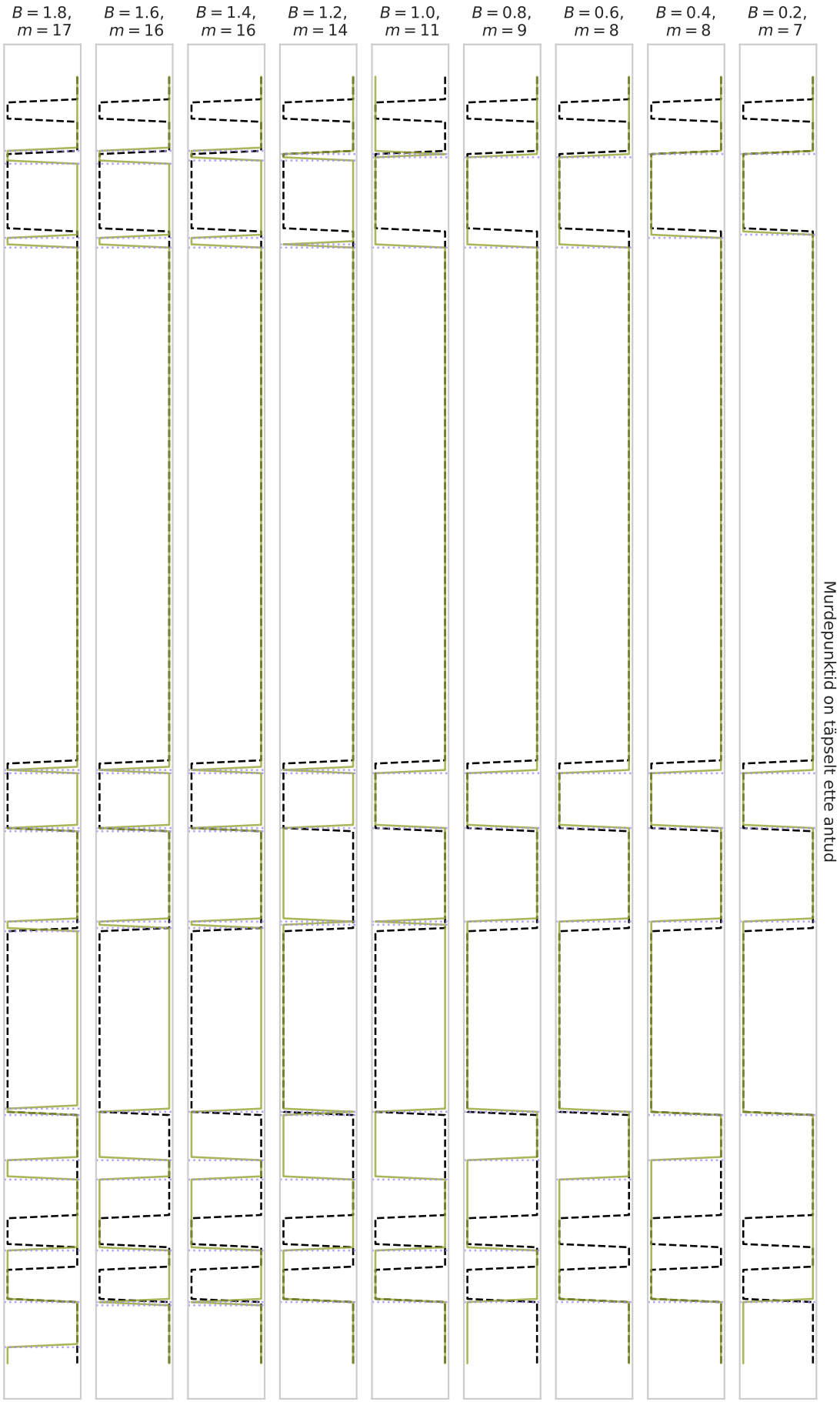
Antud lisas on toodud välja erinvad joondused, mis on leitud kaheastmelise segmenteerimise tulemusel. Rohelise joonega on märgitud prognoos, mis on saadud Ersatz-Binderi kaofunktsiooni korral, ning sinise joonega on märgitud prognoos, mis on saadud murdepunktide kaofunktsiooni korral. Prognoositud murdepunktid on märgitud katkentliku püstkriipsuga. Iga rea ees on kirjas, mis oli kordaja $B$ väärtus murdepunktide vektori prognoosi leidmisel. Lisaks on muutujaga $m$ tähistatud murdepunktide vektor prognoosi pikkus. Allpool olevates tabelites on eraldi välja toodud leitud murdepunktid, seejuures märgime, et juhuslike suuruste indeksid hakkavad nullist.

| $B$ | Murdepunktid |
|---|---|
| 0.2 | 24, 49, 216, 233, 262, 322, 380 |
| 0.4 | 24, 50, 216, 233, 262, 322, 336, 380 |
| 0.6 | 25, 53, 216, 233, 262, 321, 342, 380 |
| 0.8 | 25, 53, 216, 233, 262, 321, 336, 364, 380 |
| 1.0 | 24, 25, 53, 216, 233, 262, 263, 321, 342, 364, 380 |
| 1.2 | 24, 26, 52, 53, 215, 216, 233, 262, 263, 321, 322, 342, 364, 380 |
| 1.4 | 23, 26, 50, 53, 215, 216, 233, 234, 262, 264, 321, 336, 342, 364, 380, 381 |
| 1.6 | 23, 27, 50, 53, 215, 216, 233, 234, 262, 264, 321, 336, 342, 364, 380, 381 |
| 1.8 | 23, 27, 50, 53, 215, 216, 233, 234, 262, 265, 320, 322, 336, 342, 364, 380, 394 |

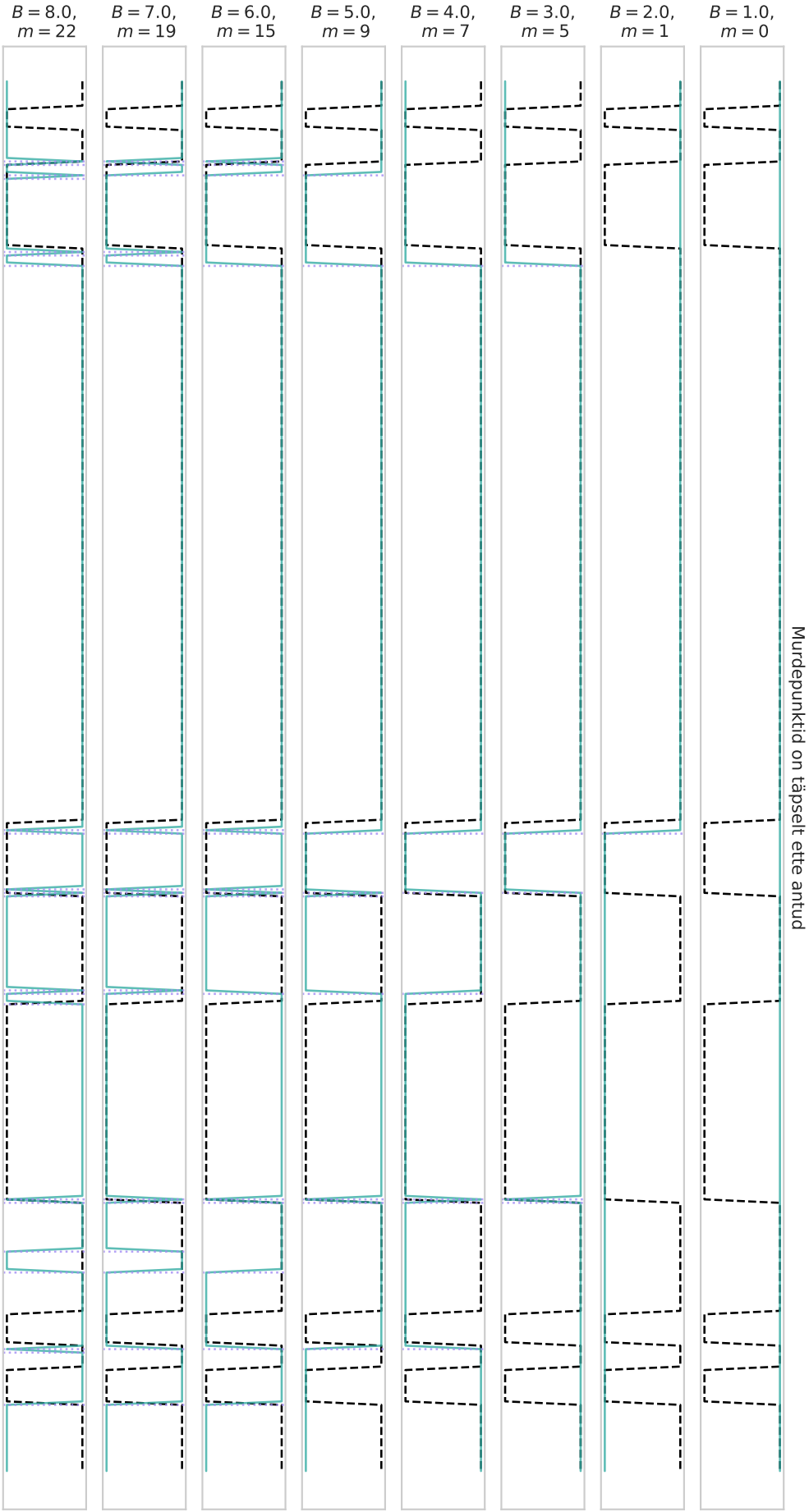
Tabel 1: Ersatz-Binderi kaofunktsioonile vastavad murdepunktid.

| $B$ | Murdepunktid |
|---|---|
| 1.0 | |
| 2.0 | 216 |
| 3.0 | 53, 216, 233, 321, 322 |
| 4.0 | 53, 216, 233, 262, 321, 322, 364 |
| 5.0 | 27, 53, 216, 233, 234, 262, 321, 322, 364 |
| 6.0 | 23, 24, 27, 53, 215, 216, 232, 233, 234, 262, 321, 322, 342, 364, 380 |
| 7.0 | 23, 24, 27, 49, 50, 53, 215, 216, 232, 233, 234, 261, 262, 321, 322, 336, 342, 364, 380 |
| 8.0 | 23, 24, 27, 28, 49, 50, 53, 215, 216, 232, 233, 234, 261, 262, 265, 321, 322, 336, 342, 364, 365, 380 |

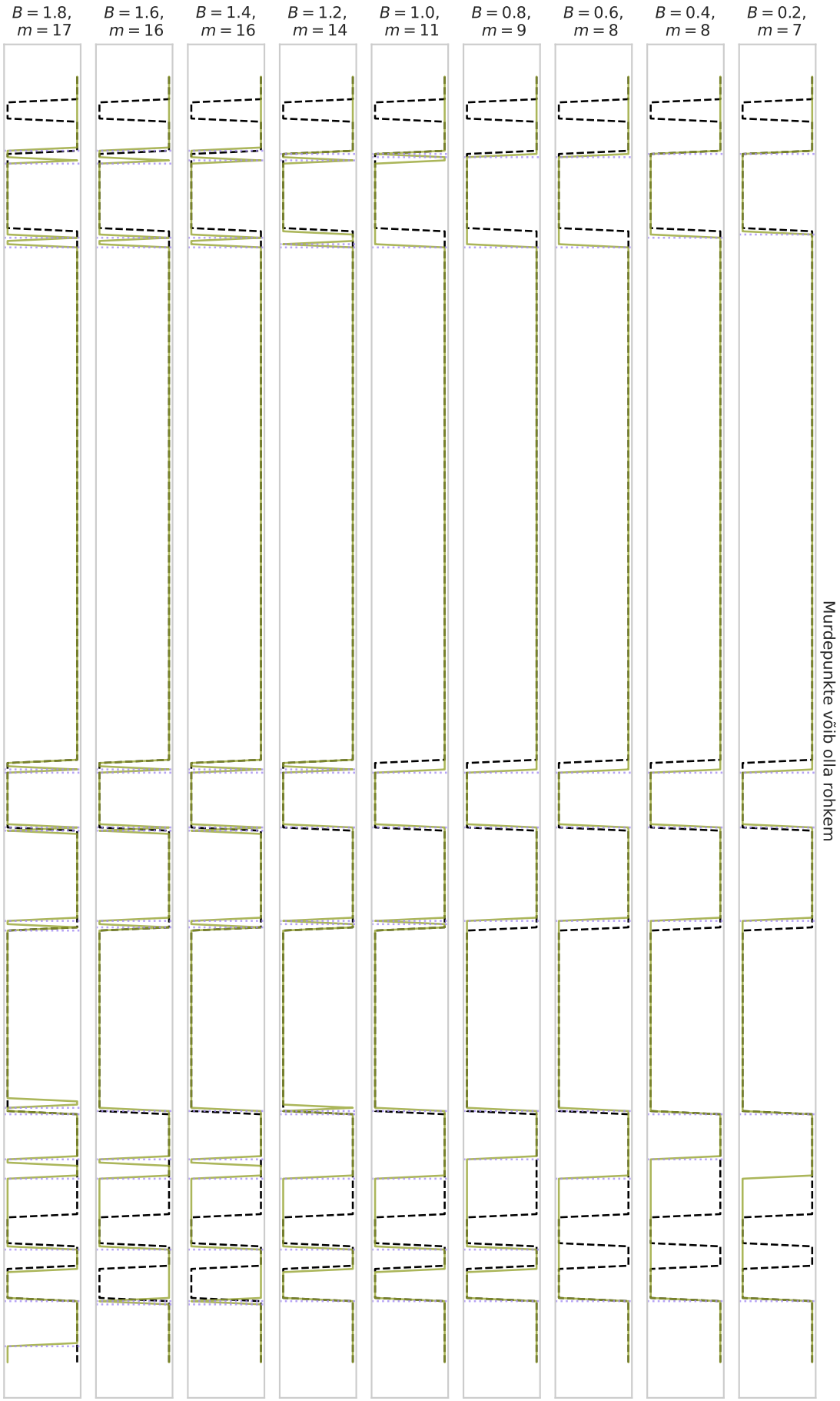
Tabel 2: Murdepunktide kaofunktsioonile vastavad murdepunktid.



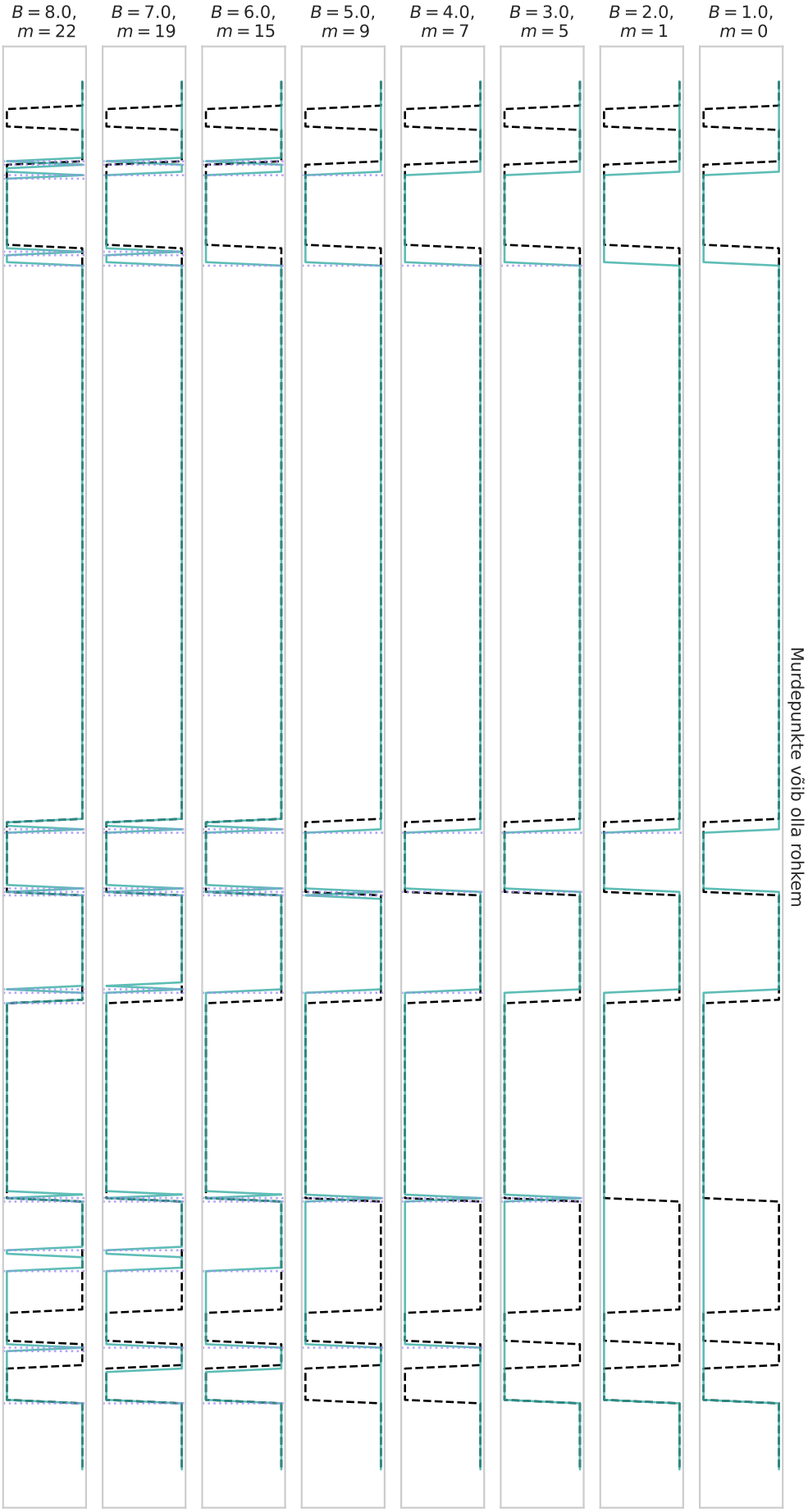

Joonis 6: Kaheastmeline segmenteerimine Ersatz-Binderi kaofunktsiooni korral.



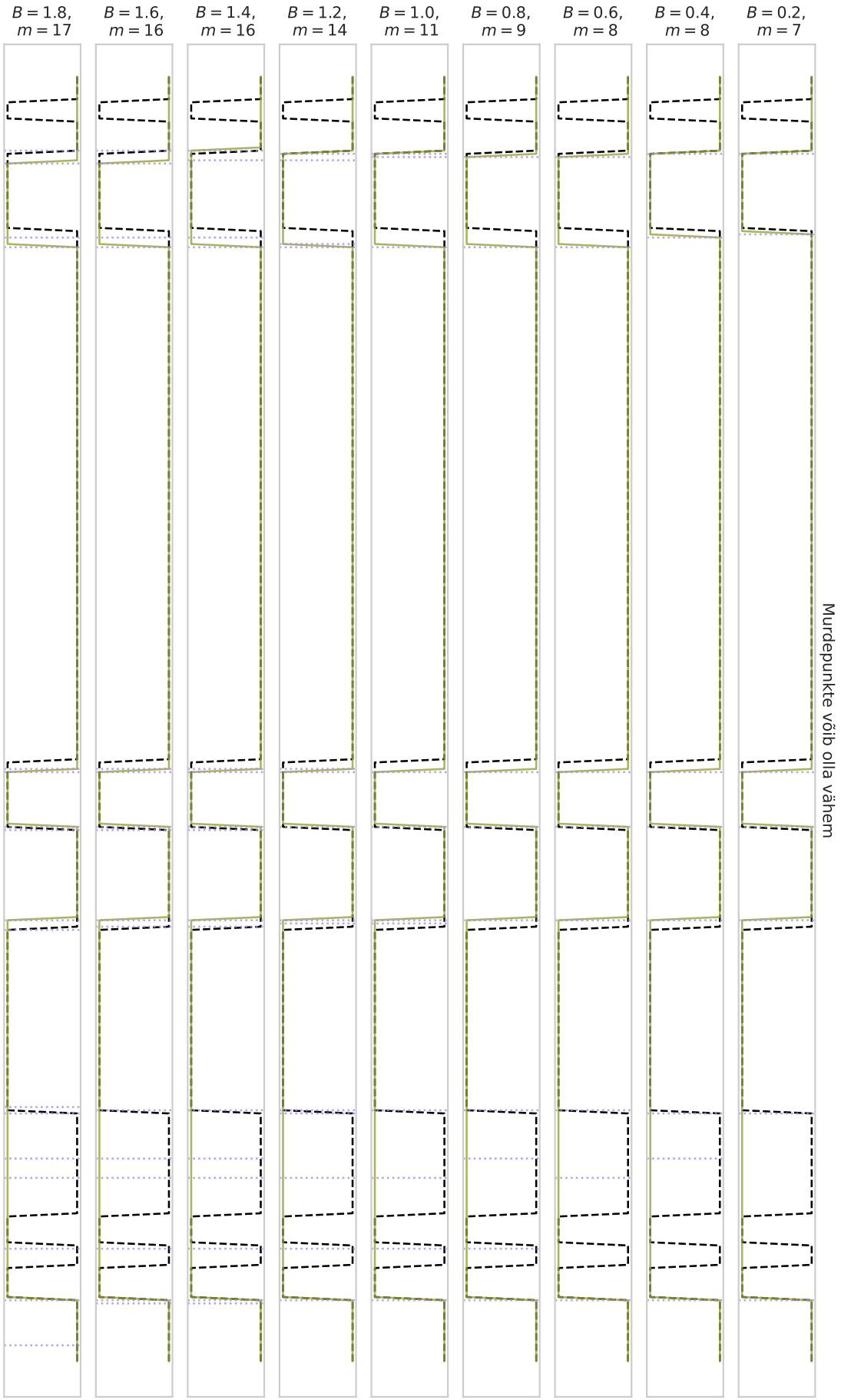

Joonis 7: Kaheastmeline segmenteerimine murdepunktide kaofunktsiooni korral.



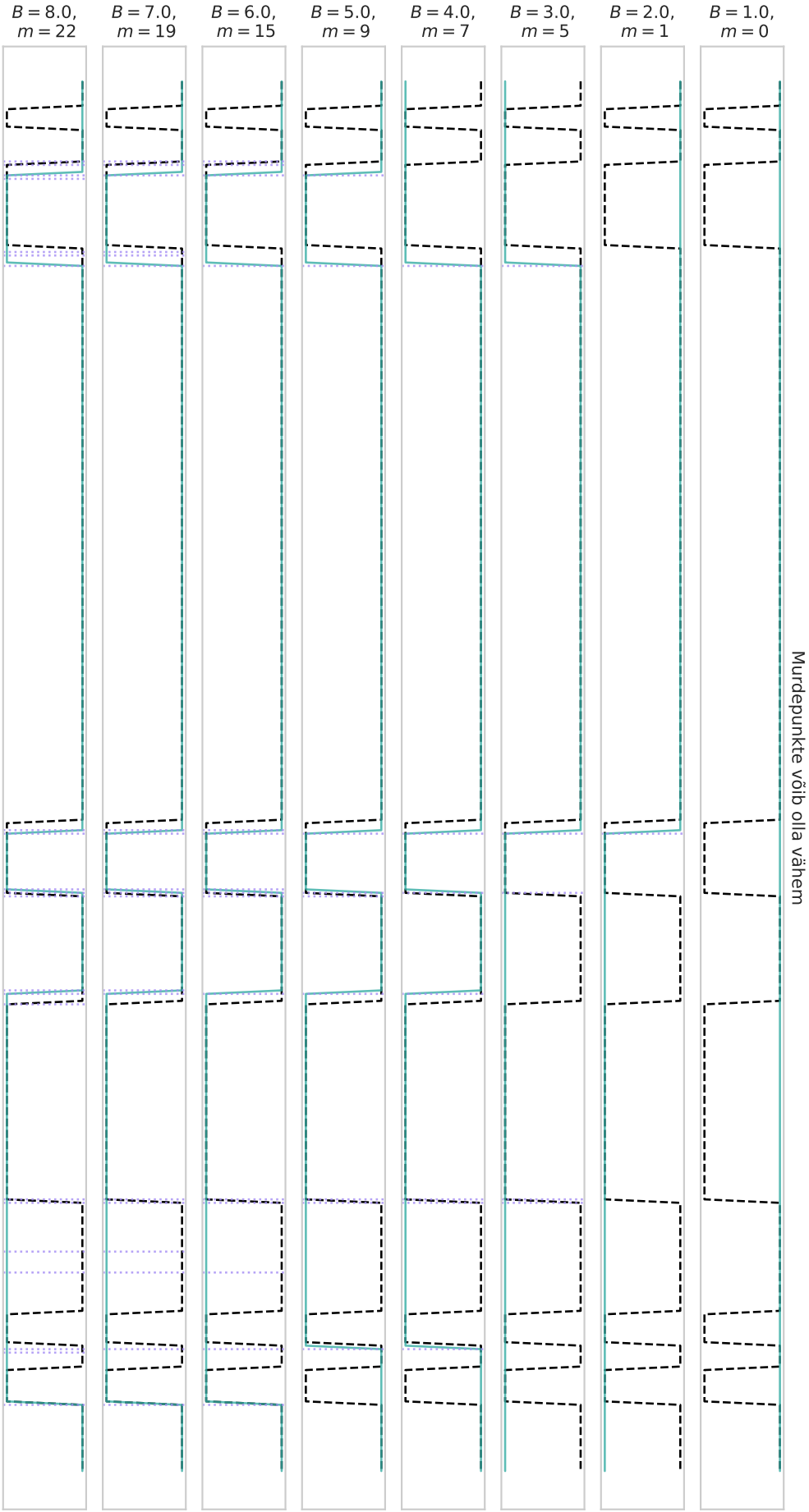

Joonis 8: Kaheastmeline segmenteerimine Ersatz-Binderi kaofunktsiooni korral.




Joonis 9: Kaheastmeline segmenteerimine murdepunktide kaofunktsiooni korral.




Joonis 10: Kaheastmeline segmenteerimine Ersatz-Binderi kaofunktsiooni korral.




Joonis 11: Kaheastmeline segmenteerimine murdepunktide kaofunktsiooni korral.

# Lisa 2. Korrigeeritud simulatsiooni joonised

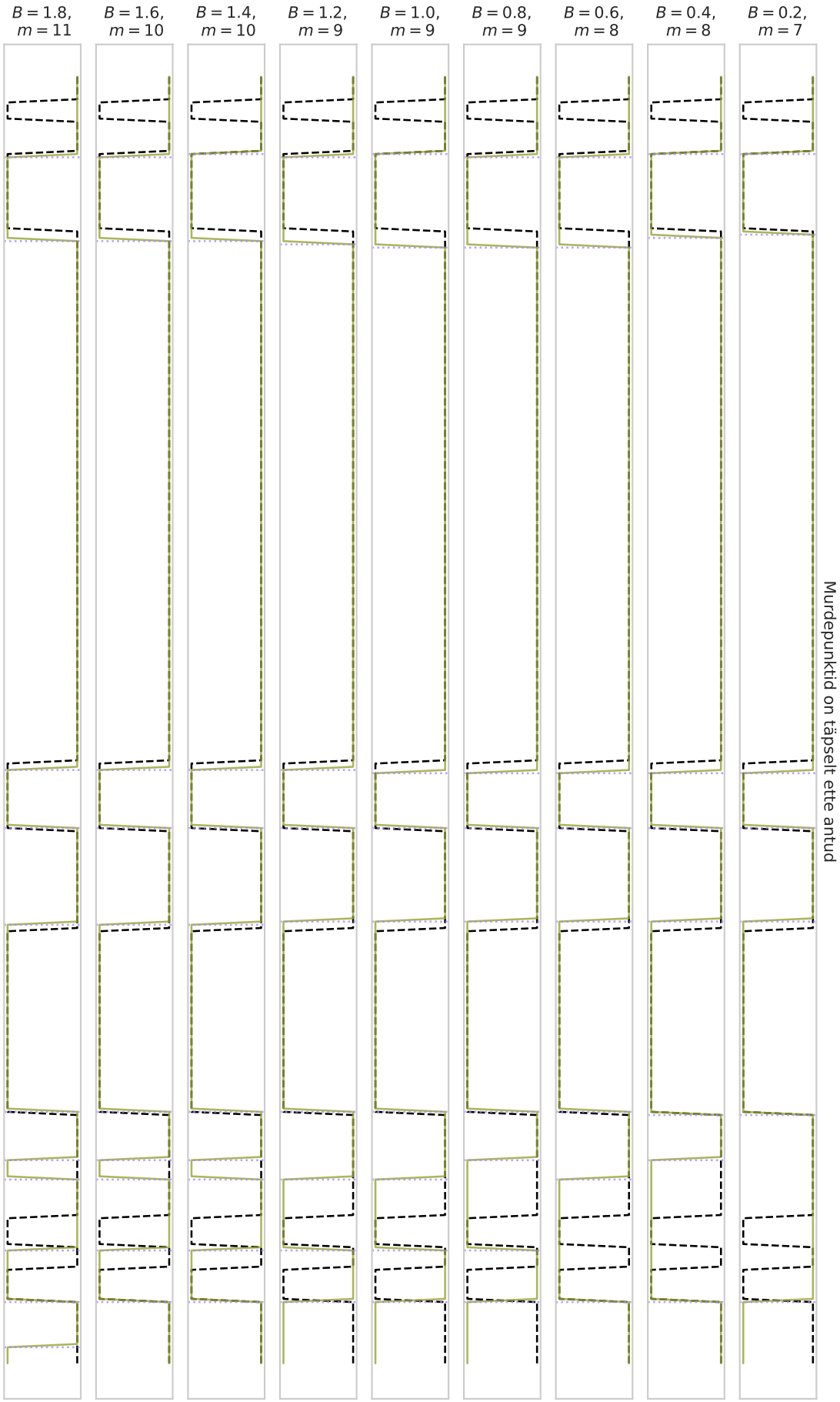
Antud lisas on joondused leitud järgmise protsetuuri järgi. Võtame lisas 1 leitud murdepunktide vektori prognoosi, leiame prognoosis üles murdepunktid $x_1$ ja $x_2$, kus $|x_1 - x_2| \leq 5$, ning asendame need kaks elementi arvuga $\left\lfloor \frac{x_1+x_2}{2} \right\rfloor$; kordame seda seni, kuni enam sellised murdepunkte ei leia. Jooniste skeem on sama nagu lisas 1. Leitud murdepunktid on esitatud allpool olevates tabelites.

| $B$ | Murdepunktid |
|---|---|
| 0.2 | 24, 49, 216, 233, 262, 322, 380 |
| 0.4 | 24, 50, 216, 233, 262, 322, 336, 380 |
| 0.6 | 25, 53, 216, 233, 262, 321, 342, 380 |
| 0.8 | 25, 53, 216, 233, 262, 321, 336, 364, 380 |
| 1.0 | 24, 53, 216, 233, 262, 321, 342, 364, 380 |
| 1.2 | 25, 52, 215, 233, 262, 321, 342, 364, 380 |
| 1.4 | 24, 51, 215, 233, 263, 321, 336, 342, 364, 380 |
| 1.6 | 25, 51, 215, 233, 263, 321, 336, 342, 364, 380 |
| 1.8 | 25, 51, 215, 233, 263, 321, 336, 342, 364, 380, 394 |

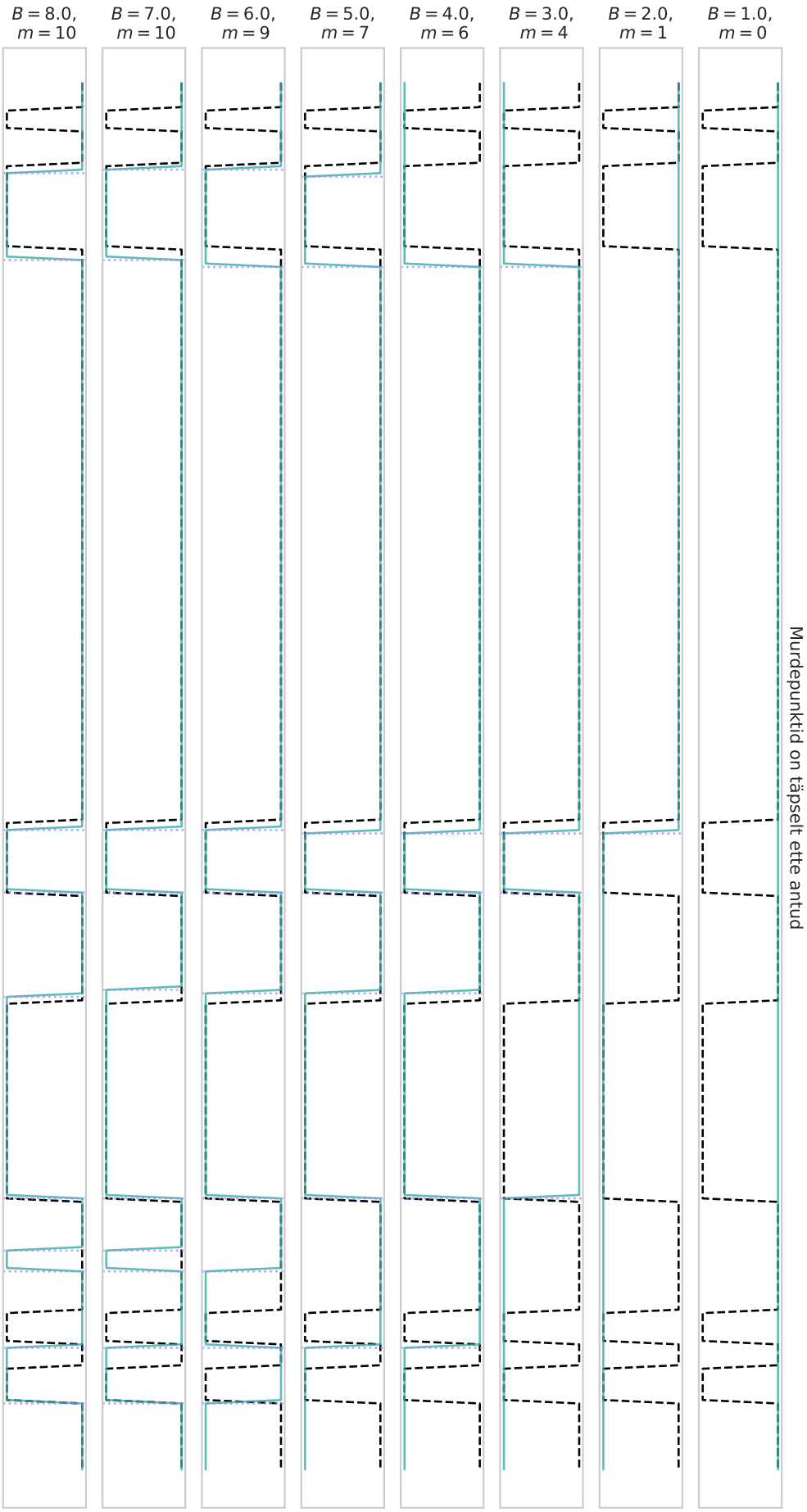
Tabel 3: Ersatz-Binderi kaofunktsioonile vastavad korrigeeritud murdepunktid.

| $B$ | Murdepunktid |
|---|---|
| 1.0 | |
| 2.0 | 216 |
| 3.0 | 53, 216, 233, 321 |
| 4.0 | 53, 216, 233, 262, 321, 364 |
| 5.0 | 27, 53, 216, 233, 262, 321, 364 |
| 6.0 | 25, 53, 215, 233, 262, 321, 342, 364, 380 |
| 7.0 | 25, 51, 215, 233, 261, 321, 336, 342, 364, 380 |
| 8.0 | 26, 51, 215, 233, 263, 321, 336, 342, 364, 380 |

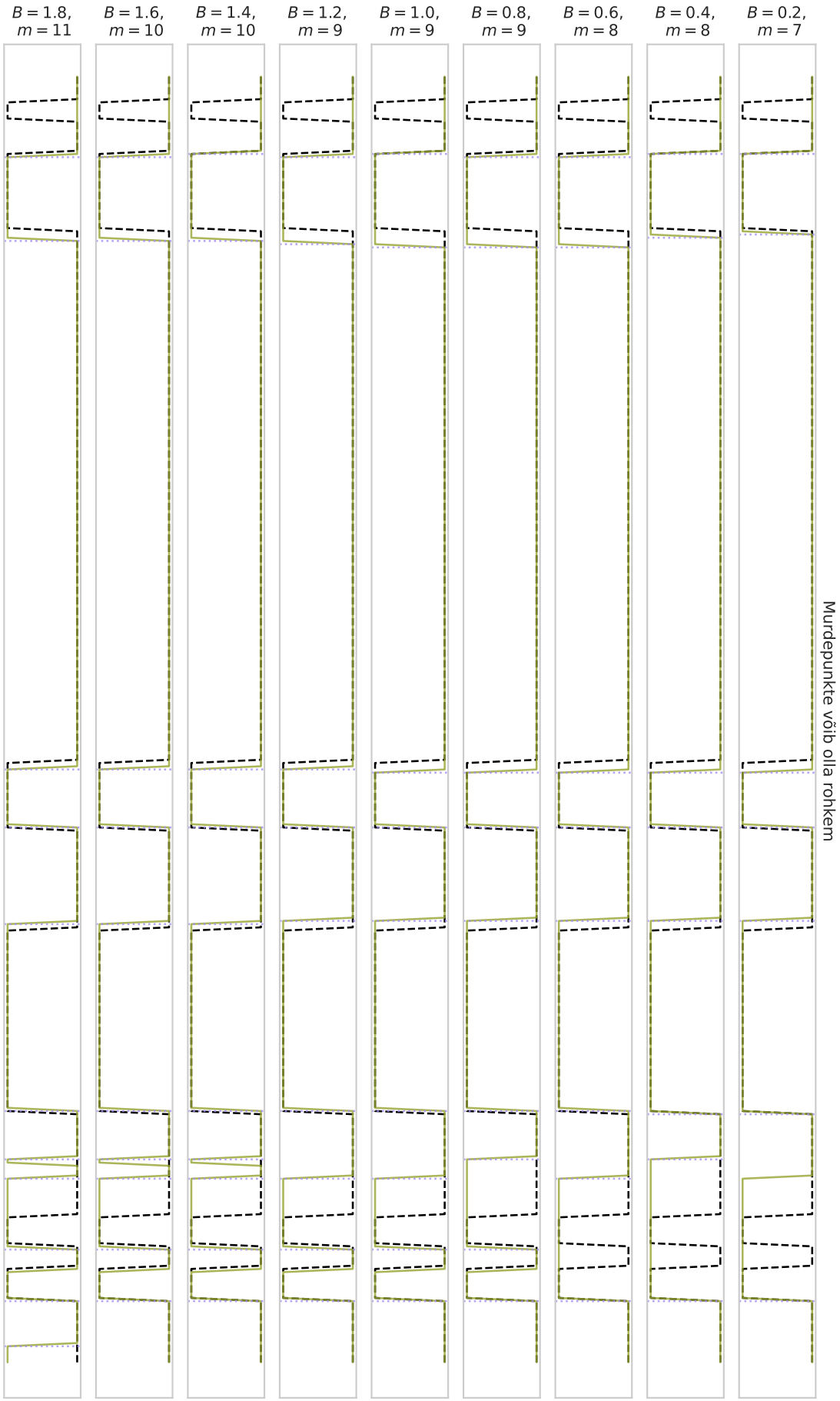
Tabel 4: Murdepunktide kaofunktsioonile vastavad korrigeeritud murdepunktid.



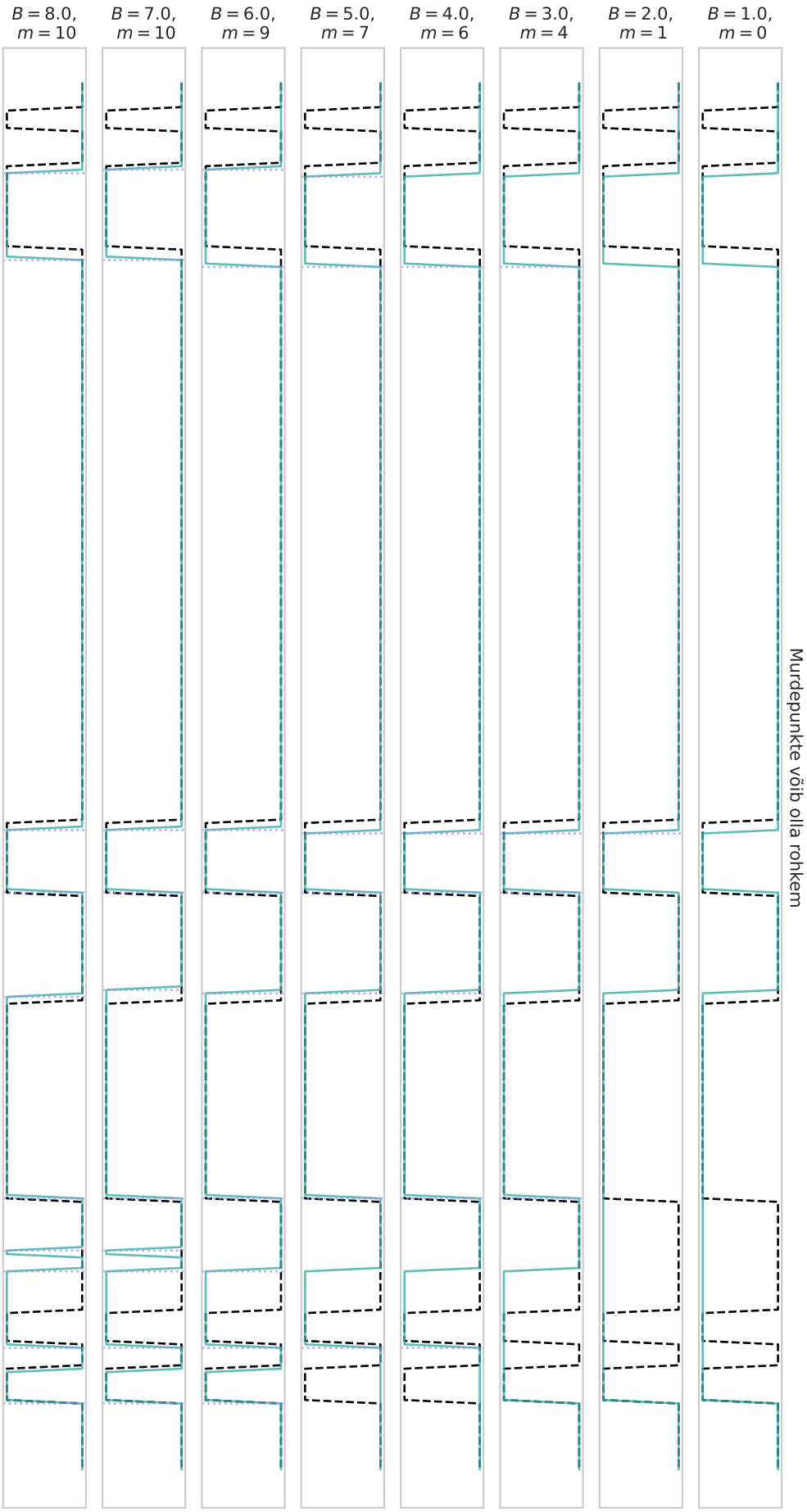

Joonis 12: Kaheastmeline segmenteerimine Ersatz-Binderi korral koos korrektsiooniga.



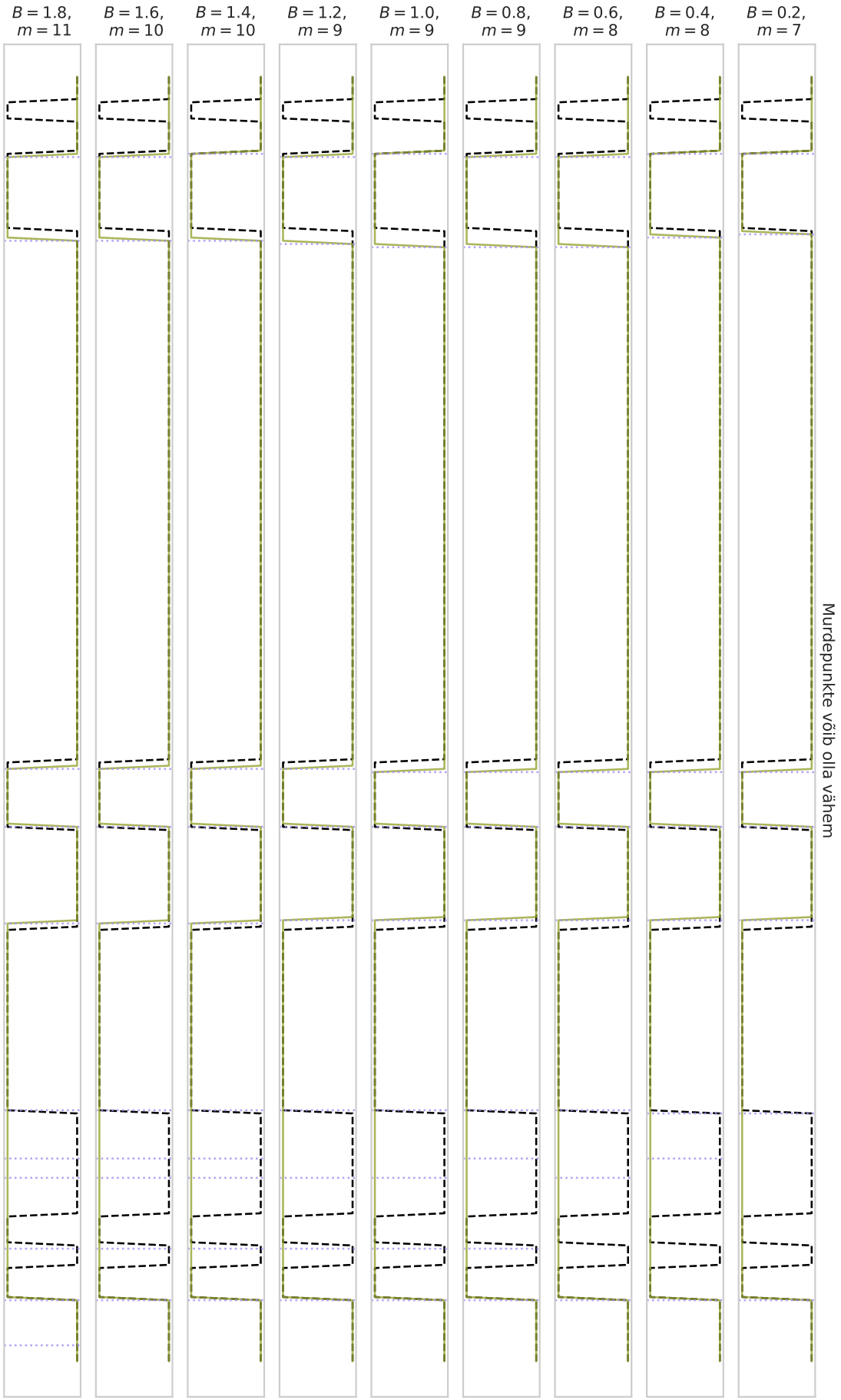

Joonis 13: Kaheastmeline segmenteerimine murdepunktide kaofunktsiooni korral koos korrektsiooniga.



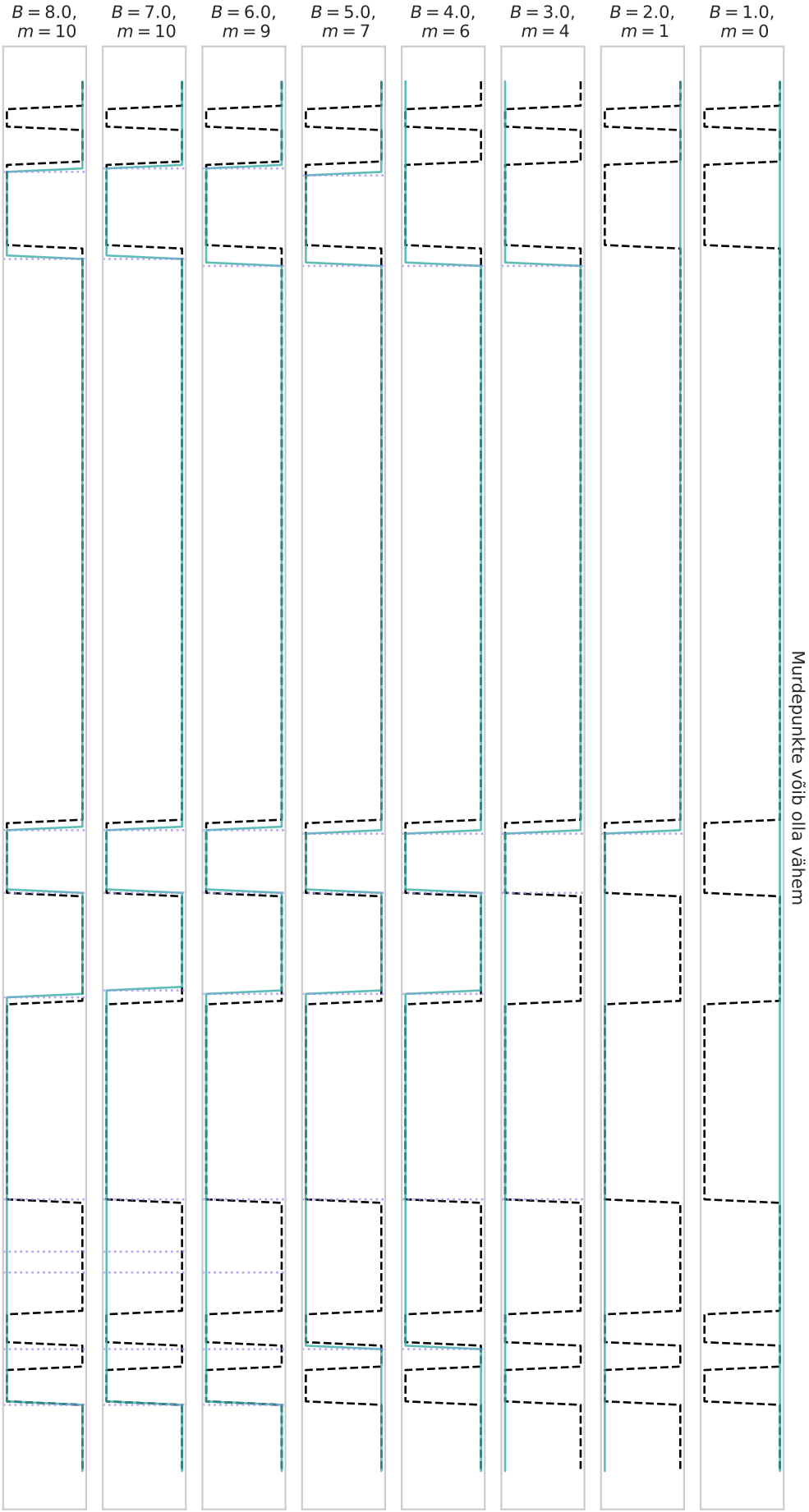

Joonis 14: Kaheastmeline segmenteerimine Ersatz-Binderi korral koos korrektsiooniga.




Joonis 15: Kaheastmeline segmenteerimine murdepunktide kaofunktsiooni korral koos korrektsiooniga.




Joonis 16: Kaheastmeline segmenteerimine Ersatz-Binderi korral koos korrektsiooniga.




Joonis 17: Kaheastmeline segmenteerimine murdepunktide kaofunktsiooni korral koos korrektsiooniga.

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Karl-Joan Alesma,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Segmenteerimine läbi murdepunktide“, mille juhendaja on Jüri Lember, 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 Commonsi litsentsiga CC BY NC ND 4.0, 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.

Karl-Joan Alesma
15.05.2024