Ph.D. disszertáció kivonata

Néhány Előrelépés a Többváltozós Statisztikai Modellezésben

Baranyi Máté

Témavezető:

Bolla Marianna D.Sc.

egyetemi tanár
Sztochasztika Tanszék, Matematika Intézet,
Budapesti Műszaki és Gazdaságtudományi Egyetem

MŰEGYETEM 1782

BME

2022

Bevezetés

Ebben a disszertációban a többváltozós statisztikai modellezés különböző aspektusait tárgyaljuk. A többváltozós statisztikai módszereket széles körben kutatják és alkalmazzák a tudomány minden területén. Egy kísérlet vagy kutatási kérdés változói közötti kapcsolatok feltárása számos jól ismert témában elengedhetetlen: regresszió, dimenziócsökkentés, hipotézisvizsgálat és adatelemzés általában. Ez a mű a statisztika három fontos alterületére összpontosít, nevezetesen a grafikus modellekre, a nemparaméteres regresszióra és az idősorelemzésre.

Fejezet 1 egy új regressziós módszert mutat be, az Iterated Conditional Expectation (ICE) algoritmust, amely speciális grafikus modelleken, nevezetesen valószínűségi DAG-okon (irányított körmentes gráfok) és regressziós gráfokon működik. A fejezet hasonló című [J2] cikkünkön alapul, amelyet a témavezetőmmel közösen írtam. A téma a nemparaméteres statisztika és a valószínűségi grafikus modellek (vagy egyszerűen csak grafikus modellek) metszéspontjában helyezkedik el. Ezért a Szakasz 1.1-ben ez utóbbit vizsgáljuk általános megközelítésben, nagyrészt a diplomamunkámat [1] és a témavezetőmmel közösen készített cikkünk [J1] alapján. A grafikus modellekkel kapcsolatos bevezetés különböző részei az Fejezet 2-ben is felbukkannak, és további (a disszertációban nem kifejezetten használt) anyag található az Appendix A-ban. A bevezetett ICE algoritmus egy több kimenettel rendelkező regressziós technika, és hiányzó adatokkal rendelkező környezetben is alkalmazható. A kimeneti változókat egyenként, iteratív módon, nemparaméteres regressziós módszerekkel becsüljük. A módszer feltételezi a kísérletben szereplő változókkal kompatibilis grafikus modell meglétét. Leírjuk a technikát (Szakasz 1.2), és bizonyítjuk (négyzetes középben vett) konzisztenciáját (Szakasz 1.3) valószínűségi DAG-okon és regressziós gráfokon. Az algoritmus általánosabb, de ez a kétféle grafikus modell jobban összeegyeztethető a több bemenet, több kimenet felállással. Egyszerűen, ezekben a modellekben a fő bemeneti változók a DAG-ok esetében a gráf "szülő nélküli" csúcsai által reprezentált változók, a regressziós gráfok esetében pedig az úgynevezett "kontextus" változók. A fő kontribúció az, hogy kiterjesztettük az ilyen típusú grafikus modellek alkalmazási területét azáltal, hogy nemparaméteres simítást alkalmazunk rekurzívan a grafikus modell irányított folyamata mentén. A numerikus implementációt is részletesen tárgyaljuk. A Szakasz 1.4-ben, a választott kernelregressziós módszer megvalósítását tárgyaljuk, valamint a kernel és a sávszélesség kiválasztását a változók típusa és a minta szerint. Továbbá néhány egyéb gyakorlati megfontolás is szerepel. Az Appendix B-ben bizonyítjuk a választott kernelregressziós implementáció (négyzetes középben vett) konzisztenciáját. Az alkalmazások bemutatása a Szakasz 1.5-ben található.

Fejezet 2 a "regresszió szakadásos kísérlettervezés" (regression discontinuity design, RDD) elemzésével foglalkozik. Ez egy elő-/utóteszteléses kvázi-kísérleti módszer egy kezelés várható hatásának mérésére, amelyet olyan esetekben alkalmaznak, ahol a randomizáció nem lehetséges, de a kezelésben való részvételről egy további (úgynevezett futó) változó alapján döntenek. Ez a fejezet a Molontay Rolanddal [J3] közösen írt cikkemen alapul, amelyben az RDD-módszert alkalmazzuk a Budapesti Műszaki és Gazdaságtudományi Egyetemen (BME) két különböző kialakítású felzárkóztató matematika kurzus hatékonyságának mérésére. Ez a

cikk egy korábbi konferenciacikkünk kibővített változata [C2]. Ezekben RDD eszközökkel, különböző környezetben vizsgáljuk a matematikai felzárkóztató kurzusok közvetlen és hosszabb távú hatásait. Az RDD-t előzetesen a Szakasz 2.1-ben ismertetjük. Munkánk kontribúciója az, hogy rámutatunk néhány módszertani kérdésre a RDD-vel kapcsolatban, amelyeket más alkalmazott munkák elhanyagolnak. Módszertani újdonságként egy új alternatív módszert (az úgynevezett jitterelt RDD-t) vezetjük be a diszkrét futó változó problémájának kezelésére RDD környezetben, amelyet részletesen a Szakasz 2.2.2-ben mutatunk be. A fejezet végén, a a disszertációban, az RDD és a grafikus modellezés néhány összefüggését tárgyaljuk. A Szakasz 2.3-ban részletesebben bemutatjuk esettanulmányunkat. Vizsgáljuk a BME-n végzett matematikai felzárkóztatás hatását a hallgatók későbbi tanulmányi eredményeire mind az elsőéves matematika kurzusok, mind a hosszabb távú hatások tekintetében.

Fejezet 3-ban figyelmünk az idősorelemzés felé fordul. Itt a Szakasz 3.1-ben található néhány előzmény után stacionárius többdimenziós idősorok kovariancia struktúrájának és spektrális sűrűségeinek összefüggéseit vizsgáljuk. A témavezetőmmel [J4] egy másik közös cikkünk elméleti eredményei alapján egy új dinamikus főkomponens-analízis algoritmus (az úgynevezett dynPCA) is bemutatásra kerül. Ezt a [C4] és [C5] című munkákban tanulmányoztuk tovább. Konkrétabban többváltozós, gyengén stacionárius diszkrét idejű, valós értékű idősorokkal dolgozunk. Adott egy ilyen idősor, amelynek emellett abszolút összegezhetőek autokovarianciái, bebizonyítottunk egy aszimptotikus összefüggést az első n autokovariancia blokk Toeplitz-mátrixának sajátértékei és a spektrális sűrűségfüggvénymátrixok n Fourier-frekvencián vett spektrumainak uniója között. A bizonyításhoz a blokk-cirkuláns mátrixok tulajdonságait használjuk. A spektrum szót két különböző értelemben használjuk: a sajátértékekre és a spektrális reprezentációra. A bizonyított tétel különböző analógiákhoz vezet az idő- és frekvenciatartománybeli számítások között, valamint új technikákhoz is. Konkrétabban, bevezetünk egy módszert a folyamat alacsony rangú közelítésére, és ennek következményeként egy új (általunk dynPCA-nak nevezett) dinamikus főkomponens-analízis algoritmust is. Ez utóbbi algoritmus, amelynek több potenciális felhasználási lehetősége van, a dinamikus főkomponenseket becsüli meg a bemenetként kapott többdimenziós idősorból. Az eredmények illusztrálása és az alkalmazások bemutatása a Szakasz 3.6-ben történik.

Mint már említettük, a fejezetek a tárgyalt technikák, módszerek alkalmazásaival zárul, hogy bemutassuk az alkalmazások fontosságát a többváltozós statisztikai modellezésben.

A disszertáció három fő "tézispontja" a következő:

1. Bemutatjuk az ICE algoritmust, és bizonyítjuk (négyzetes középben vett) konzisztenciáját valószínűségi DAG-ok és regressziós gráfok felett (Fejezet 1). A numerikus implementációjához a felhasznált kernel regressziós módszer (négyzetes középben vett) konzisztenciáját is bizonyítjuk (Appendix B).

2. Elemezzük az RDD technikát, majd javasolunk és tárgyalunk egy új alternatív módszert (az úgynevezett jitterelt RDD-t) a diszkrét futó változó problémájának kezelésére RDD környezetben (Fejezet 2). Egy egyedi, részletes esettanulmányt is közlünk.

3. Vizsgáljuk a stacionárius többdimenziós idősorok kovariancia struktúrájának és spektrálsűrűségfüggvényeinek összefüggéseit, és bizonyítunk egy ezek közötti aszimptotikus összefüggést (Fejezet 3). Elméleti eredményeink alapján bevezetjük az új dynPCA algoritmust is, egyedi alkalmazásokkal.

Fejezet 1

Iterated Conditional Expectation algoritmus (ICE) valószínűségi DAG-okon és regressziós gráfokon

1.1. Bevezetés a grafikus modellekről

A disszertáció tartalmaz egy bevezetést a grafikus modellekről.

1.2. Az ICE algoritmus leírása

A fő gondolat a következő. A csúcsok DAG-gal vagy regressziós gráffal kompatibilis sorrendjében, a lineáris, linearizált vagy logisztikus regresszió helyett feltételes várható értéket veszünk nemparaméteres módon, hasonlóan Breiman és Fiedman [2] ACE (Alternating Conditional Expection) algoritmusához. Itt nem kell alternálni, csak "irányított" feltételes várható értékeket veszünk egymás után [3] nemparaméteres regressziós fogalmaival. Ezért rekurzív algoritmusunkat ICE-nek (Iterated Conditional Expectation) nevezzük. A konzisztenciabizonyításokban a feltételes várható értéket a simítóval hasonlítjuk össze.

A konstrukció alapját egy tanító minta, szakértői tudás és az élek statisztikai validálása, lásd [4], adja. A struktúratanulás messze nem triviális feladatát nem vettük figyelembe ebben a fejezetben. Itt a többszörös kimenetek előrejelzésén van a hangsúly a gráf alapján. Az ICE algoritmus megalapozottságát a konzisztenciájának bizonyítása is alátámasztja, ami azt jelenti, hogy nagy tanító minta esetén is közel optimális becsléseket kapunk (L^2 értelemben) a tesztmintára.

Az egyes simítók négyzetes középben vett konzisztenciája, lokális átlagolás alapján, C. J. Stone tételének feltételei mellett garantált, lásd [3] vagy Appendix B. Például a k legközelebbi szomszéd (k-NN), partícionáló, és kernelregressziós becslések négyzetes középben konzisztensek bizonyos feltételek mellett (lásd ismét [3]). Mi főként a [5] kernel-regressziós megközelítését használjuk, amely jól alkalmazkodik a különböző típusú változókhoz. Ez aszimptotikusan normális, de beláttuk a négyzetes középben vett konzisztenciáját is, lásd Appendix B. Összehasonlításképpen közlünk néhány más simítókkal kapott eredményt is, alkalmazásaink során.

1.2.1. ICE DAG-on

Legyen az $X_1, \ldots, X_d$ együttes eloszlása (ebben a topológiai sorrendben) Markov-kompatibilis egy DAG-gal. Mondjuk, hogy legalább a $X_{k+1}, \ldots, X_d$ változók értékei ismertek; ezek közé tartoznak legalább azok is, amelyeknek nincsenek szülei a DAG-ban. Ekkor először X_k, majd

$X_{k-1}, \ldots,$ végül X_1-t jelezzük előre, egy n elemű minta alapján: $x_1^{(i)}, \ldots, x_d^{(i)}$, $i = 1, \ldots, n$ teljes tanító minta. Bevezetünk néhány jelölést még: $\boldsymbol{X}_{>j} := \{X_{j+1}, \ldots, X_d\}$, $\boldsymbol{x}_{>j} := \{x_{j+1}, \ldots, x_d\}$ és $\boldsymbol{X}_{\mathrm{Par}(j)}$ az X_j szüleinek halmazát jelöli, ahonnan irányított élek mutatnak felé. $j = k, k-1, \ldots, 1$ esetén definiáljuk a következőket

$$P_j(\boldsymbol{x}_{>j}) := \mathbb{E}\left(X_j \mid \boldsymbol{X}_{>j} = \boldsymbol{x}_{>j}\right) = \mathbb{E}\left(X_j \mid \boldsymbol{X}_{\mathrm{Par}(j)} = \boldsymbol{x}_{\mathrm{Par}(j)}\right) \tag{1.1}$$

mint a regressziós feladat elméleti megoldása; ez egy vetítés és argumentumainak mérhető függvénye. Az utolsó egyenlőség a DAG feletti Markovitásból következik.

Ha adataink többváltozós Gauss-eloszlásból származnak, akkor a fenti feltételes várható értékek lineáris függvényei a változóknak, és lineáris regresszióval kaphatóak meg, ahol az együtthatókat a tanító adatokból becsüljük. Ellenkező esetben, ismeretlen eloszlás esetén a feltételes várható értéket nemparaméteres módon, a simítási eljárásokkal, becsüljük a [2, 3] című könyvben tárgyalt eljárásokkal. Ezért definiáljuk az

$$S_j^{(n)}\left(\boldsymbol{x}_{>j}\right) := \sum_{i=1}^{n} x_j^{(i)} W_j^{(n,i)}\left(\boldsymbol{x}_{>j}\right) = \sum_{i=1}^{n} x_j^{(i)} W_j^{(n,i)}\left(\boldsymbol{x}_{\mathrm{Par}(j)}\right) \tag{1.2}$$

simító, lokális átlagoláson alapuló és a X_j mintatér konvex burkába képző leképezést. A Markovitás értelmében a $\boldsymbol{x}_{>j}$ halmaz helyett az $\boldsymbol{x}_{\mathrm{Par}(j)}$ a szülőhalmazt vesszük figyelembe.

Itt a $W_j^{(n,i)}(\boldsymbol{x}_{>j})$ nemnegatív súlyokat az n elemű tanító mintából számítjuk ki, és általában a úgy normalizálják, hogy

$$\sum_{i=1}^{n} W_j^{(n,i)}\left(\boldsymbol{x}_{>j}\right) = 1.$$

Maga a simító $S_j^{(n)}$ a $x_j^{(i)}, x_{j+1}^{(i)}, \ldots, x_d^{(i)}$ szegmenstől függ a $i = 1, \ldots, n$ tanító minta esetén, és az argumentumainak mérhető függvénye is. Az általunk implementált simítók főként Nadaraya–Watson típusú kernelbecslők, ahol

$$W_j^{(n,i)}\left(\boldsymbol{x}_{>j}\right) = \frac{K\left(\boldsymbol{x}_{>j}, \boldsymbol{x}_{>j}^{(i)}\right)}{\sum_{i'=1}^{n} K\left(\boldsymbol{x}_{>j}, \boldsymbol{x}_{>j}^{(i')}\right)}.$$

A lehetséges K kernelekről lásd a kernel megközelítést és más lokális átlagoló becslőket tárgyaló Szakasz 1.4 és Appendix B részeket.

Ezután az iterált feltételes várható értékeket

$$P_j\left(P_{j+1}\left(P_{j+2}\left(\ldots\left(P_k\left(\boldsymbol{X}_{>k}\right), \ldots\right), \boldsymbol{X}_{>k}\right), \boldsymbol{X}_{>k}\right), \boldsymbol{X}_{>k}\right)$$

az iterált simításokkal imitáljuk:

$$S_j^{(n)}\left(S_{j+1}^{(n)}\left(S_{j+2}^{(n)}\left(\ldots\left(S_k^{(n)}\left(\boldsymbol{X}_{>k}\right), \ldots\right), \boldsymbol{X}_{>k}\right), \boldsymbol{X}_{>k}\right), \boldsymbol{X}_{>k}\right).$$

A Szakasz 1.3-ban bebizonyítjuk, hogy az átlagos négyzetben vett különbség az iterált simítások és az iterált feltételes várható értékok között 0-hoz tart, ahogy $n \to \infty$, $j = k-1, \ldots, 1$ esetén. Ezért a az egymást követő becslések közel vannak az optimumhoz (L^2 értelemben).

Ami az algoritmus konzisztenciáját jelenti. Részletes magyarázatért lásd: Szakasz 1.3.1.

1.2.2. ICE regressziós gráfon

Itt az $X_1, \dots X_d$ együttes sűrűsége faktorizálható a $C_1, \dots, C_\ell$ lánckomponensek mentén, mint ahogy azt a disszertációban tárgyaljuk, így a feltételes várható értékeket láncösszetevőről láncösszetevőre vesszük, kezdve az utolsó, C_ℓ lánckomponenssel. Ismét posztuláljuk hogy legalább a C_ℓ-ban szereplő kontextusváltozók ismertek.

Az általános lépésben, $k = \ell - 1, \ell - 2, \dots, 2, 1$ esetén, legyen

$$\boldsymbol{P}_k\left(\boldsymbol{x}_{C_{>k}}\right) := \mathbb{E}\left(\boldsymbol{X}_{C_{C_k}} \mid \boldsymbol{X}_{C_{>k}} = \boldsymbol{x}_{C_{C_{>k}}}\right) \tag{1.3}$$

a regressziós feladat elméleti megoldása. Az eredmény egy vektor az alábbi komponensekkel

$$P_{k,j}\left(\boldsymbol{x}_{C_{>k}}\right) := \mathbb{E}\left(X_j \mid \boldsymbol{X}_{C_{>k}} = \boldsymbol{x}_{C_{>k}}\right) = \mathbb{E}\left(X_j | \boldsymbol{X}_{\mathrm{Par}(j)} = \boldsymbol{x}_{\mathrm{Par}(j)}\right), \quad X_j \in C_k. \tag{1.4}$$

Ez a párhuzamos regresszió egy példája, ahol kihasználjuk, hogy az összes $X_j \in C_k$ szülője $C_{>k}$-ban van. Azt is felhasználtuk, hogy egy véletlen vektor feltételes várható értéke egy másik vektorra a vektor összetevőinek a feltételes várható értékei ugyanarra a kondicionáló halmazra. A regressziós gráf Markov-tulajdonságai alapján még úgy is, hogy a feltételezés ugyanarra a változóhalmazra vonatkozik, az egyes komponenseknek lehetnek különböző szülőhalmazai, amelyekre a kondicionálások korlátozódnak.

A simítókat szintén komponensenként definiáljuk:

$$S_{k,j}^{(n)}\left(\boldsymbol{x}_{C_{>k}}\right) := \sum_{i=1}^{n} x_j^{(i)} W_{k,j}^{(n,i)}\left(\boldsymbol{x}_{C_{>k}}\right) = \sum_{i=1}^{n} x_j^{(i)} W_{k,j}^{(n,i)}\left(\boldsymbol{x}_{\mathrm{Par}(j)}\right), \quad X_j \in C_k \tag{1.5}$$

a közelítő megoldás egy simító segítségével, lokális átlagolás alapján. A simító $\boldsymbol{S}_k^{(n)}(\boldsymbol{x}_{C_{>k}})$ minden X_j C_k-re $S_{k,j}^{(n)}(\boldsymbol{x}_{C_{>k}})$ koordinátákból áll.

Ekkor a következő Szakasz 1.3.2 alapján becsléseink közel vannak az optimálishoz, ha az átlagos négyzetes különbség a az iterált komponensenkénti feltételes várható értékek

$$\boldsymbol{P}_k\left(\boldsymbol{P}_{k+1}\left(\boldsymbol{P}_{k+2}\left(\dots\left(\boldsymbol{P}_{\ell-1}\left(\boldsymbol{X}_{C_\ell}\right), \dots\right), \boldsymbol{X}_{C_\ell}\right), \boldsymbol{X}_{C_\ell}\right), \boldsymbol{X}_{C_\ell}\right)$$

és az analóg simítók

$$\boldsymbol{S}_k^{(n)}\left(\boldsymbol{S}_{k+1}^{(n)}\left(\boldsymbol{S}_{k+2}^{(n)}\left(\dots\left(\boldsymbol{S}_{\ell-1}^{(n)}\left(\boldsymbol{X}_{C_\ell}\right), \dots\right), \boldsymbol{X}_{C_\ell}\right), \boldsymbol{X}_{C_\ell}\right), \boldsymbol{X}_{C_\ell}\right)$$

között 0-hoz tart, ahogy $n \to \infty$, $k = \ell - 1, \ell - 2, \dots, 2, 1$ esetén.

1.3. Az ICE algoritmus konzisztenciája

1.3.1. Konzisztencia egy DAG mentén

Ha a célváltozó abszolút folytonos

Most tegyük fel, hogy $X_1, \ldots, X_d$ (a DAG topológiai sorrendjében) véges szórással rendelkeznek és engedelmeskednek egy abszolút folytonos együttes eloszlásnak.

A [3] szellemében a nemparaméteres becslésünket négyzetes középben konzisztensnek nevezzük, ha

$$\mathbb{E} \int_{\mathbb{R}^{d-k}} \left[S_k^{(n)}(\boldsymbol{x}_{>k}) - P_k(\boldsymbol{x}_{>k}) \right]^2 \mu(d\boldsymbol{x}_{>k}) \to 0, \quad n \to \infty,$$

ahol μ a mögötte álló változók együttes eloszlását jelöli. Itt mind az n-elemű minta $X_k, X_{k+1}, \ldots, X_d$, amelyet a simító használ, valamint a változók $X_{k+1}, \ldots, X_d$ egy újonnan érkező megfigyelésének értékei véletlenszerűek, ezért választottuk szét ezt a két fogalmat a várható érték vevésekor. Röviden,

$$\mathbb{E} \left[S_k^{(n)}(\boldsymbol{X}_{>k}) - P_k(\boldsymbol{X}_{>k}) \right]^2 \to 0, \quad n \to \infty,$$

ahol $\mathbb{E}$ most a várható értéket veszi figyelembe a véletlenszerűség mindkét fenti fogalmának függvényében.

Először az iteráció első lépésében bizonyítjuk a konzisztenciát (lásd Lemma 1.1), majd a teljes ICE konzisztenciáját bizonyítjuk DAG mentén indukcióval (lásd Tétel 1.1).

1.1. Lemma. *A Képlet* (1.1) *és Képlet* (1.2) *jelöléssel legyen* $P_k(\boldsymbol{x}_{>k})$ *és* $S_k^{(n)}(\boldsymbol{x}_{>k})$ *a regressziós feladat egzakt megoldása és egy simító (lokális átlagoláson alapuló és a* X_k *mintatér konvex burkába képezően). Tegyük fel, hogy a simító a feltételes várható értékre egy négyzetes középben konzisztens becslést ad, azaz,*

$$\mathbb{E} \left[S_k^{(n)}(\boldsymbol{X}_{>k}) - P_k(\boldsymbol{X}_{>k}) \right]^2 \to 0, \quad n \to \infty.$$

Ekkor az egylépéses előre simítás is négyzetes középben konzisztens, azaz,

$$\mathbb{E} \left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(P_k(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) \right]^2 \to 0, \quad n \to \infty.$$

Most már meg tudjuk fogalmazni az általános tételt az ICE négyzetes középben vettkonzisztenciájáról egy DAG mentén.

1.1. Tétel. *Legyen* $X_1, \ldots, X_d$ *egy abszolút folytonos együttes eloszlású változóhalmaz, véges szórással, amely Markov-kompatibilis egy DAG-gal (ebben a sorrendben). Úgy indexeljük őket, hogy* $\boldsymbol{X}_{>k}$ *a DAG-ban a szülők nélküli változókat jelöli, a többit pedig egy* n *elemű minta alapján becsüljük. A Képlet* (1.1) *és Képlet* (1.2) *jelöléssel,* $j = k, k-1, \ldots, 1$ *esetén legyen* $P_j(\boldsymbol{x}_{>j})$ *és* $S_j^{(n)}(\boldsymbol{x}_{>j})$ *a regressziós feladat egzakt megoldása és egy simító (lokális átlagoláson alapuló és a* X_j *mintatér konvex burkába képező). Tegyük fel, hogy az egyes simítók a feltételes várható értékekre négyzetes középben konzisztens becslést adnak, azaz,*

$$\mathbb{E} \left[S_j^{(n)}(\boldsymbol{X}_{>j}) - P_j(\boldsymbol{X}_{>j}) \right]^2 \to 0, \quad n \to \infty, \quad j = k, k-1, \ldots, 1.$$

Ekkor az iterált simítások is négyzetes középben konzisztensek, azaz,

$$\mathbb{E}\Big[S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}) \\ - P_j(P_{j+1}(P_{j+2}(\dots(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2 \to 0, \quad n\to\infty,$$

minden $j = k, k-1, \dots, 1$ *esetén.*

Megjegyezzük, hogy a főként kernel-alapú simítóink paramétereit és a szorzatkernel paramétereit az adatokhoz igazítjuk; és a kernel-alapú lokális átlagolású becslőkön kívül más becslőket is figyelembe veszünk a Szakasz 1.5-ban. Feltételezések, amelyek alapján az egyes simítók négyzetes középben konzisztens becslést adnak, a Szakasz 1.4 és Appendix B részek tárgyalják. Továbbá a Markovitás miatt a z$\boldsymbol{X}_{>j}$ halmaz helyett a $\boldsymbol{X}_{\mathrm{Par}(j)}$ szülőhalmazt vesszük figyelembe az algoritmus során.

Fontos, hogy a célváltozó véges szórással rendelkezik és abszolút folytonos. A köztes kimenetek is legyenek folytonosak, de diszkrét rendezett változók is elég jól működhetnek, ha kellően nagy mintatérrel rendelkeznek. Ha a célváltozó diszkrét, kategorikus, akkor ebben a lépésben a Szakasz 1.3.1-ban bemutatandó módszert használjuk.

Ha a célváltozó kategorikus

Amikor X_k kategorikus, és c különböző értéket vesz fel, akkor az előrejelzés a Bayes-szabály alapján történik a módusz-maximalizáláson keresztül, lásd [3]:

$$P_k(\boldsymbol{x}_{>k}) := \operatorname*{argmax}_{1\le i\le c} \mathbb{P}\left(X_k = i \mid \boldsymbol{X}_{>k} = \boldsymbol{x}_{>k}\right).$$

A $\mathbb{I}_{\{X_k=i\}}$ $(i = 1, \dots, c)$ bináris változók bevezetésével, a az posterior valószínűségek a feltételes várható értékek, azaz, regressziós függvények:

$$P_{k,i}(\boldsymbol{x}_{>k}) := \mathbb{P}\left(X_k = i \mid \boldsymbol{X}_{>k} = \boldsymbol{x}_{>k}\right) = \mathbb{E}\left(\mathbb{I}_{\{X_k=i\}} \mid \boldsymbol{X}_{>k} = \boldsymbol{x}_{>k}\right).$$

Adott tanító adatok mellett $\mathbb{I}_{\{X_k=i\}}, \boldsymbol{X}_{>k}$-re, összerakjuk a (simító) becslést $S_{k,i}^{(n)}(\boldsymbol{x}_{>k})$ minden egyes $i = 1, \dots, c$ esetében, és a plug-in becslés pedig

$$S_k^{(n)}(\boldsymbol{x}_{>k}) := \operatorname*{argmax}_{1\le i\le c} S_{k,i}^{(n)}(\boldsymbol{x}_{>k}).$$

A disszertációban megmutatjuk, hogy a plug-in becslés hibája közel van a $P_k(\boldsymbol{x}_{>k})$ optimális hibájához.

1.3.2. Konzisztencia a regressziós gráf mentén

Most az négyzetes középben vett konzisztenciát a vektorváltozók négyzetes középben vett konzisztencia szempontjából fogalmazzuk meg, ami viszont a koordinátáik négyzetes középben

vett konzisztenciájára redukálódik. Mivel

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{X}_{C_{\ell-1}}\right\|^2 \\
&= \mathbb{E}\left\|\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell})\right\|^2 + \mathbb{E}\left\|\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{X}_{C_{\ell-1}}\right\|^2 \\
&= \sum_{j:\, X_j \in C_{\ell-1}} \mathbb{E}\left|S_{\ell-1,j}^{(n)}(\boldsymbol{X}_{C_\ell}) - P_{\ell-1,j}(\boldsymbol{X}_{C_\ell})\right|^2 + \mathbb{E}\left\|\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{X}_{C_{\ell-1}}\right\|^2,
\end{aligned}$$

ahol az utolsó tag csak az alapul szolgáló eloszlástól függ, és az első az a kifejezés, amely közvetlenül csökkenthető a mintavételi és simítási eljárással. Bizonyítani fogjuk, hogy az összegzés utáni véges számú tagok 0-hoz tartanak, ahogy $n \to \infty$, ha az egyes simítók négyzetes közép konzisztenciája garantált. Ily módon a következő konzisztencia-tételt fogalmazhatjuk meg a Tétel 1.1-hez hasonló feltételek mellett.

1.2. Tétel. *Legyen $X_1, \dots, X_d$ abszolút folytonos együttes eloszlású változóhalmaz véges szórással, amely Markov-kompatibilis egy regressziós gráffal. A változók $C_1, \dots, C_\ell$ lánckomponensekbe rendeződnek úgy, hogy a közös eloszlásukra teljesül a láncgráfok eloszlás-faktorizációja; továbbá, úgy vannak indexelva, hogy az utolsó C_ℓl áncösszetevő tartalmazza a kontextusváltozókat, a többit pedig visszafelé, komponensről komponensre, egy n elemű minta alapján becsüljük.*

A feltételes várható érték és a simítókra vett Képlet (1.4) *és Képlet* (1.5) *vektoros kiterjesztését véve, és feltételezve, hogy az egyes simítók komponensenként négyzetes középben konzisztens becslést adnak, az iterált simítások is négyzetes középben konzisztensek, azaz,*

$$\begin{aligned}
\mathbb{E}\Big\|&\boldsymbol{S}_k^{(n)}(\boldsymbol{S}_{k+1}^{(n)}(\boldsymbol{S}_{k+2}^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}), \dots), \boldsymbol{X}_{C_\ell}), \boldsymbol{X}_{C_\ell}), \boldsymbol{X}_{C_\ell}) \\
&- \boldsymbol{P}_k(\boldsymbol{P}_{k+1}(\boldsymbol{P}_{k+2}(\dots(\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}), \dots), \boldsymbol{X}_{C_\ell}), \boldsymbol{X}_{C_\ell}), \boldsymbol{X}_{C_\ell})\Big\|^2 \to 0, \quad n \to \infty,
\end{aligned}$$

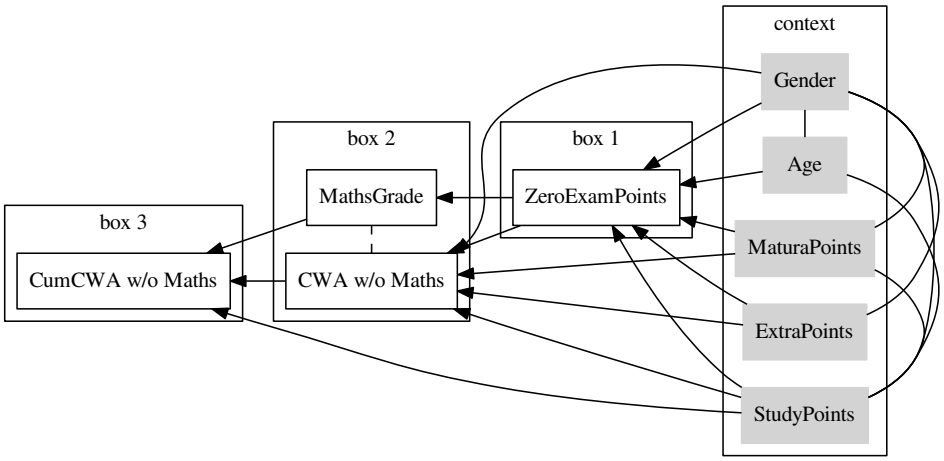
minden $k = \ell - 1, \ell - 2, \dots, 2, 1$ esetén.

1.4. A numerikus algoritmus

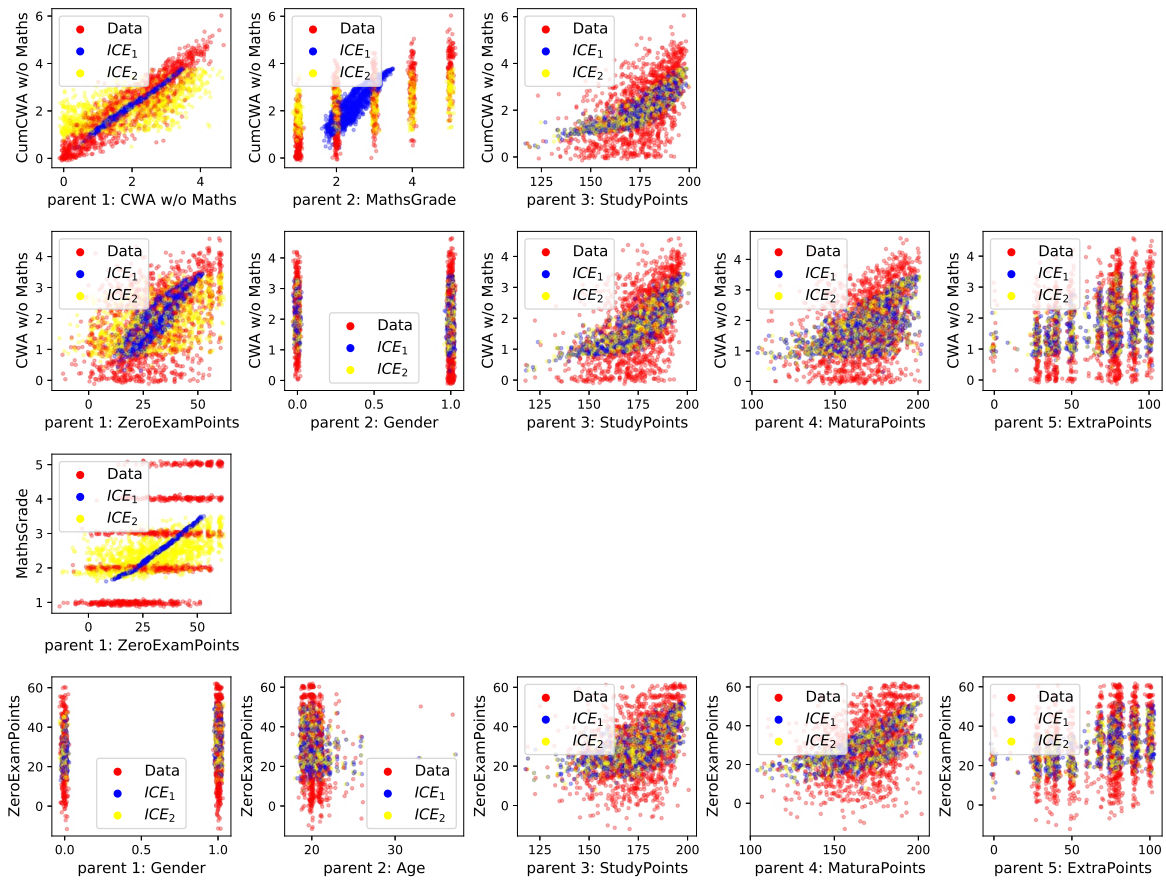
A disszertációban az ICE numerikus megvalósítását tárgyaljuk.A választott kernelregressziós módszer átlagos négyzet konzisztenciáját bizonyítjuk Appendix B-ban.

1.5. Az ICE alkalmazásai

Három alkalmazási példát mutatunk be. Az első véletlenszerű generált adatokon alapul. A második a 2014-es egyiptomi demográfiai és egészségügyi felmérés adataival (EDHS 2014) foglalkozik. A harmadik a Budapesti Műszaki és Gazdaságtudományi Egyetemen (BME) gyűjtött adatokat veszi figyelembe, hogy megvizsgálja a hallgatók első két félévben elért tanulmányi eredményeit.



1.1. ábra. Az oktatási példa regressziós gráfja



1.2. ábra. ICE az oktatási adatokon: minden nem kontextusváltozóhoz tartozik egy sor pontdiagram. Minden egyes részdiagramban egy nem kontextusváltozót mutatunk be az egyik szülőváltozójával szemben. A piros pontok megfelelnek a tényleges adatpontoknak, míg a kék és a sárga pontok az ICE által adott becsléseknek a szülő változó becsült értékével, illetve tényleges értékével szemben.

Fejezet 2

Regression Discontinuity Design (RDD), jittereléssel

2.1. Bevezetés az RDD módszertanba

Az RDD akkor használható egy program vagy *kezelés* hatásának becslésére, ha a jelölteket azzal a feltétellel választják ki a kezelésre, hogy egy numerikus endogén változó, az úgynevezett *futó változó*, meghalad (vagy alulmúl, a kontextustól függően) egy kijelölt küszöbértéket vagy *vágási pontot*.

Dióhéjban a módszer úgy tekinthető, mint két regressziós görbe összehasonlítása egy kijelölt ponton. Van egy Y kimeneti változó, amelyre a kezelés és a futó X változó hatását szeretnénk megbecsülni. A T bináris kezelési változót a futó változó alapján rendeljük hozzá (meghalad-e egy bizonyos határértéket). E módszertan alapváltozatát Thistlethwaite és Campbell [6] vezette be, mint Y lineáris regresszióját X változóra T függvényében.

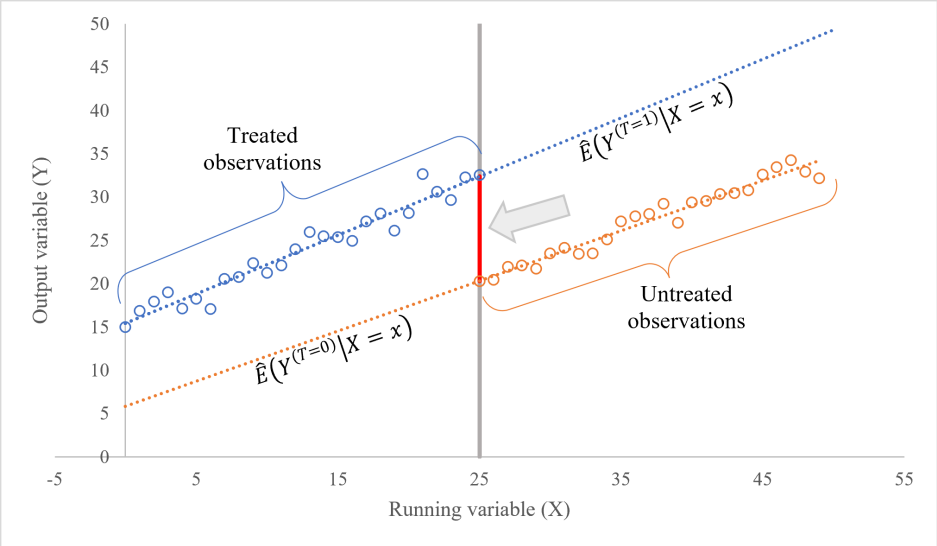
Tipikusan úgy képzeljük el, hogy minden egyes i egyénre vonatkozóan létezik egy eredménypár: $Y_i^{(1)}$ és $Y_i^{(0)}$, ha a személy ki volt téve a kezelésnek (1) , illetve ha nem volt kitéve a kezelésnek (0). A kezelés hatását a $Y_i^{(1)} - Y_i^{(0)}$ különbséggel mérnénk. A valóságban az egyén számára a két kimenet közül csak az egyik megfigyelhető, ezért ennek a különbségnek a populációra vonatkozó átlagait vizsgáljuk a feltételes várható értékek különbségével:

$$\mathbb{E}\left[Y_i^{(1)} \mid X_i = x\right] - \mathbb{E}\left[Y_i^{(0)} \mid X_i = x\right]. \tag{2.1}$$

Ez a különbség érdekel minket a X futó változó c vágási értékénél, így – attól függően, hogy mi figyelhető meg a vágási pont két oldalán – becslést végzünk:

$$\lim_{\varepsilon \to 0} \left\{ \mathbb{E}\left[Y_i^{(1)} \mid X_i = c \pm \varepsilon\right] - \mathbb{E}\left[Y_i^{(0)} \mid X_i = c \mp \varepsilon\right] \right\}. \tag{2.2}$$

Ez az átlagos kezelési hatás (ATE) a vágási ponton, amelyet a hatás értékelésére használunk. Ezek a feltételes várható értékek a regressziós problémák elméleti megoldásai, amelyeket nemparaméteres vagy parametrikus módszerekkel közelítünk.



2.1. ábra. Dummy példa RDD-re: a kezelést a 25-ös vágási pont alatt adjuk be, a feltételes várható értékek itt lineáris regresszióval egyértelműen becsülhetők, a látható kezelési hatást a piros vonal és a nyíl jelzi.

2.2. A diszkrét futó változó problémája

Számos RDD-alkalmazásban a futó változó diszkrét. Ezért az eredmények összehasonlítása a megfigyelések közvetlenül a kezelési küszöbérték felett és közvetlenül az alatt nem lehetséges, és/vagy kérdéseket vet fel. A vágási pont közelében nincsenek adatok, így a Képlet (2.2) használata a valódi hatás túlbecsléséhez vezetne. Legyen például egy olyan futó változó, amely a $\{x_1, x_2, \ldots, x_j, \ldots, x_k\}$ megszámlálható halmazból veszi fel értékeit, amelynek vágási pontja $c = x_j$, és a kezelés akkor adott, ha $X \geq c$. Ebben az esetben valójában a következő becslést adható

$$\mathbb{E}\left[Y_i^{(1)} \mid X_i = c\right] - \mathbb{E}\left[Y_i^{(0)} \mid X_i = x_{j-1}\right].$$

Képlet (2.1) helyett c-nél. Belátható, hogy az extrapoláció miatt alá/túlbecslés történik.

2.2.1. Specifikációs hiba

A diszkrét futó változó problémájának kezelésére Lee és Card [7] úgy módosítja modellt, hogy egy a futó változó értékétől függő s_X véletlen tag bevezetésével, amelyet *specifikációs hiba*-nak nevezünk:

$$Y = \alpha_0 + \beta_0 T + f(X - c) + s_X + \epsilon.$$

2.2.2. Diszkrét futó változó kezelése jittereléssel

Itt a specifikációs hiba helyett egy újszerű alternatív módszert javaslunk. A regressziós becslésben az ortogonális hibatag helyett közvetlenül a futó változóhoz adunk egy független zaj komponenst. A független zaj hozzáadását a diszkrét változókhoz általában jitterelésnek nevezik. Ez egy gyakori megoldás, amikor valaki egy érthetőbb ábrát szeretne kapni szórásdiagramok esetén, vagy például olyan algoritmusok esetén, amelyek az adatpontok szigorú sorrendiségét igénylik. A javasolt módszerünk Nagler [8] ötletén alapul, aki elméletileg is

alátámasztja a diszkrét változók jitterelését a nemparaméteres regressziós becsléshez. Ezért a diszkrét X helyett a $\tilde{X} = X + \delta$ futó változót használjuk, ahol δ egy független folytonos zajtényező. Feltételezve, hogy X egy egyenlő távolságokat tartalmazó diszkrét értékkészletből veszi fel az értékét, δ-t eszerint a távolság szerint kell kalibrálni.

Pontosabban, feltételezve, hogy a távolság 1, pl. mint az $1, 2, 3, \ldots, 100$ skálán; a zaj f_δ valószínűségi sűrűségfüggvényének meg kell felelnie a Nagler [8] által írt feltételnek, miszerint létezik egy olyan pozitív valós számpár γ_1 és γ_2, amely szerint

$$0 < \gamma_1 \leq 0.5 \leq \gamma_2 < 1 \quad \text{és} \quad f_\delta(x) = \begin{cases} 1 & \text{ha} \quad x \in [-\gamma_1, \gamma_1], \\ 0 & \text{ha} \quad x \in \mathbb{R} \setminus (-\gamma_2, \gamma_2). \end{cases}$$

A $[-0.5, 0.5]$ intervallumon való egyenletes eloszlás triviálisan kielégíti Nagler [8] feltételét $\gamma_1 = \gamma_2 = 0.5$ mellett. Szükség lehet egy új $\tilde{c}$ vágási pontra is, ami az eredeti c a fenti γ_2 értékkel eltolva. Ellenkező esetben a mérések a jittering miatt a vágási pont rossz oldalán végződnének. Egyenletes zaj esetén a $[-0.5, 0.5]$ tartományban az RDD új vágási pontja $c \pm 0.5$-nél van, attól függően, hogy a vágási pont melyik oldala szigorú. A kezelés hatása (Képlet (2.1)) ebben a jitterelt esetben a következő lesz

$$\mathbb{E}\left[Y^{(1)} \mid \tilde{X} = \tilde{c}\right] - \mathbb{E}\left[Y^{(0)} \mid \tilde{X} = \tilde{c}\right] =$$

$$\mathbb{E}\left[Y^{(1)} \mid X + \delta = c \pm \gamma_2\right] - \mathbb{E}\left[Y^{(0)} \mid X + \delta = c \pm \gamma_2\right].$$

A nemparaméteres RDD környezetben a lokális lineáris becslés jitteringgel a következő lesz

$$Y = \alpha_0 + \beta_0 T + \beta_1 (X - \tilde{c} + \delta) + \beta_2 T (X - \tilde{c} + \delta) + \epsilon, \tag{2.3}$$

ahol a becslést azokkal az adatokkal végezzük, amelyekre $\tilde{c} - h \leq \tilde{X} \leq \tilde{c} + h$. Megjegyezzük, hogy a jitterelés miatt a becsült hatás mértéke véletlenszerűvé válik. A hozzáadott perturbáció kicsi, és a minta véletlenszerű súlyozásának tekinthető.

Ennek a módszernek megvan az az előnye, hogy a hozzáadott zajtényező folytonossá teszi a futó változót, így könnyebb lenne általánosítani néhány elméleti eredményt az RDD-re vonatkozóan ezzel a felállással, mint a specifikációs hibát feltételezve. További előny, hogy a vágási pont közvetlen környezete ebben a felállásban adatpontokkal lesz töltve, így a nemparaméteres becslés megalapozottabb. Diszkrét futó változó esetén a specifikációs hiba használata jobban megfelel a parametrikus becslésnek, ahogy arra Lee és Card [7] rámutatott. Úgy véljük, hogy a javasolt módszerünk jobban alkalmas a hatás nemparaméteres becslésére.

2.3. A RDD alkalmazása, egy esettanulmány a BME-n

Egy hosszabb, részletes esettanulmányt mutatunk be. A RDD-t alkalmazunk két különböző módon tervezett korrekciós intézkedés hatékonyságának mérésére. matematika tantárgyak különböző, a Budapesti Műszaki és Gazdaságtudományi Egyetemen Gazdaságtudományi Egyetemen (BME).

Fejezet 3

Többváltozós stacionárius idősorok spektruma és időtartománybeli összefüggések

3.1. Bevezetés a többváltozós idősorokról

Mint már említettük, a többváltozós idősorokra összpontosítunk. Appendix C részben az egydimenziós idősorokra vonatkozó néhány előzmény szerepel.

Először is, *az időtartományban*, legyen $\{\boldsymbol{X}_t\}$ egy gyengén stacionárius d-dimenziós idősor, $t \in \mathbb{Z}$ és az állapottér $\mathbb{R}^d$. Legyen $\mathbb{E}(\boldsymbol{X}_t) = \boldsymbol{0}$ és a $d \times d$ méretű autokovariancia mátrixok sorozata $\boldsymbol{C}(0), \boldsymbol{C}(1), \boldsymbol{C}(2), \ldots$ abszolút összegezhető (elemenként), ami azt jelenti, hogy $\sum_{h=0}^{\infty} |c_{pq}(h)| < \infty$ $p, q = 1, \ldots, d$ esetén, ahol $\boldsymbol{C}(-h) = \boldsymbol{C}^T(h)$ és $\boldsymbol{C}(0)$ a $\boldsymbol{X}_t$ szokásos kovariancia mátrixa. Ekkor a $nd \times nd$ méretű $\mathfrak{C}_n$ összetett kovarianciamátrixa az $(\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$ véletlen vektorra a szimmetrikus, pozitív szemidefinit blokk-Toeplitz mátrix, amely a következő módon definiálható

$$\mathfrak{C}_n\left(\text{block}_i, \text{block}_j\right) = \begin{cases} \boldsymbol{C}(j-i), & j \geq i; \\ \boldsymbol{C}^T(i-j), & i \leq j. \end{cases}$$

Ugyanekkor *a frekvenciatartományban*, tekintjük a a folyamat $\boldsymbol{f}$ spektrális sűrűségmátrixát, amely az előző feltételek mellett létezik. Ez egy önadjungált, pozitív szemidefinit, $d \times d$ mátrixs értékű függvény, ami tulajdonképpen az autokovariancia mátrixok Fourier-transzformáltja:

$$\boldsymbol{f}(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \boldsymbol{C}(h) e^{-ih\omega}, \quad \omega \in [0, 2\pi]. \tag{3.1}$$

Valós diagonális, de általában komplex off-diagonális elemekkel rendelkezik.

A disszertáció további bevezetést tartalmaz a többváltozós idősorokról.

3.2. A spektrum jellemzése az időtartományban

A $\mathfrak{C}_n$ blokk-Toeplitz mátrix sajátértékeinek jellemzésére szükségünk van a $\mathfrak{C}_n^{(s)}$ szimmetrikus blokkcirkuláns mátrixra, amelyet most páratlan n esetén nézünk, amire legyen $n = 2k+1$ (páros n esetén a a számítások hasonlóak); a definícióért lásd [9]. Lényeges, hogy k éppen $\frac{n}{2}$ egészrésze, mivel a konstruált blokkcirkuláns mátrix csak ezzel a választással szimmetrikus.

(Páros n esetén a $\frac{n}{2}$ választása tökéletes, és csak kisebb változásokat okoz a spektrumban, amint azt később tárgyalni fogjuk.) $\mathfrak{C}_n^{(s)}$ ugyanaz, mint a $\mathfrak{C}_n$ a $|j-i| \leq k := \lfloor n/2 \rfloor$ blokkindexekre, és a következő módon definiáljuk

$$\mathfrak{C}_n^{(s)}\left(\text{block}_i, \text{block}_j\right) = \begin{cases} \boldsymbol{C}(j-i), & 0 \leq j-i \leq k; \\ \boldsymbol{C}(n-(j-i)), & 0 \leq j-i > k; \\ \boldsymbol{C}^T(i-j), & 0 < i-j \leq k; \\ \boldsymbol{C}^T(n-(i-j)), & 0 < i-j > k. \end{cases}$$

Ez egy szimmetrikus blokk-Toeplitz mátrix, mint a $\mathfrak{C}_n$, de a $\mathfrak{C}_n^{(s)}$ blokkcirkuláns mátrix is, amely számunkra most jobb. [1] A blokkcirkuláns mátrixok spektrumai jól jellemezhetőek, de $\mathfrak{C}_n$ általában nem blokkcirkuláns; ezért konstruáltuk meg a $\mathfrak{C}_n^{(s)}$-t. A cirkuláns mátrix sorai az előző sorok ciklikus permutáltjai, míg a blokkcirkuláns esetben a permutáció után a blokkok transzpontját vesszük, ha azok maguk nem szimmetrikusak. A $\mathfrak{C}_n^{(s)}$-ban a $\frac{n}{2}$-nél nagyobb rendű autokovarianciákat figyelmen kívül hagyjuk. Ezt csak akkor lehet megtenni, ha feltételeztük, hogy az autokovarianca mátrixok sorozata (elemenként) abszolút összegezhető.

$\mathfrak{C}_n^{(s)}$ sajátértékei

Ha a $\boldsymbol{C}(h)$-k $d \times d$ méretű mátrixok, akkor az inflációs technikák és a Kronecker-szorzatok alkalmazásával, blokkokat használunk az elemek helyett, és a sajátvektorok is blokkstruktúrát követnek. [9, 10]-ben, egy általános szimmetrikus blokkcirkuláns mátrix sajátértékei és sajátvektorai vannak karakterizálva. Ezt az eredményt alkalmazzuk a mi helyzetünkben, az $n = 2k+1$ páratlan esetre. Ennek fényében a $\mathfrak{C}_n^{(s)}$ spektruma az alábbiak mátrixok spektrumainak uniója:

$$\begin{aligned} \boldsymbol{M}_j &= \boldsymbol{C}(0) + \sum_{h=-k}^{k} \boldsymbol{C}(h)\rho_j^h \\ &= \boldsymbol{C}(0) + \sum_{h=1}^{k} [\boldsymbol{C}(h)\rho_j^h + \boldsymbol{C}^T(h)\rho_j^{-h}] \\ &= \boldsymbol{C}(0) + \sum_{h=1}^{k} [\boldsymbol{C}(h)e^{i\omega_j h} + \boldsymbol{C}^T(h)e^{-i\omega_j h}] \\ &= \boldsymbol{C}(0) + \sum_{h=1}^{k} [(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h) + \cos(\omega_j h) + i(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h) \end{aligned}$$

$j = 0, 1, \ldots, n-1$ esetén; míg a sajátvektorokat a ezen $d \times d$ méretű mátrixok sajátvektorainak összetoldásával kapjuk meg. Itt $\rho_j := e^{i\omega_j}$ az 1 j-edik (komplex) n-edik primitív (komplex) gyöke, és $\omega_j = \frac{2\pi j}{n}$ a j-edik Fourier-frekvencia. Vagyis, szükségünk van a $\boldsymbol{M}_j$ mátrixok spektrális dekompozíciójára. minden $j = 1, 2, \ldots, n-1$ esetén. Mivel $\boldsymbol{C}(h) + \boldsymbol{C}^T(h)$ szimmetrikus és $\boldsymbol{C}(h) - \boldsymbol{C}^T(h)$ antiszimmetrikus, 0-s átlóval, $\boldsymbol{M}_j$ önadjungált minden j esetében és valós sajátértékekkel rendelkezik, amelyekhez található komplex koordinátájú ortonormális

[1] Páros n esetén minden sorban csak egy $\boldsymbol{C}(K)$ van.

sajátvektorrendszer. Valóban, $\boldsymbol{M}_j$-nak komplex értékei is lehetnek, ha $j \neq 0$; pontosabban, $\sum_{h=1}^{k}(\boldsymbol{C}(h)+\boldsymbol{C}^T(h))\cos(\omega_j h)$ a valós és $\sum_{h=1}^{k}(\boldsymbol{C}(h)-\boldsymbol{C}^T(h))\sin(\omega_j h)$ az $\boldsymbol{M}_j$ képzetes része.

Könnyű belátni, hogy $\boldsymbol{M}_{n-j} = \overline{\boldsymbol{M}_j}$ (elemenkénti konjugált), ezért ugyanazokkal a (valós) sajátértékekkel rendelkezik, mint $\boldsymbol{M}_j$, de a (komplex) sajátvektorai az $\boldsymbol{M}_j$ sajátvektorainak (komponensek szerinti) komplex konjugáltjai.

Összefoglalva, páratlan $n = 2k+1$ esetén a $\mathfrak{C}_n^{(s)}$ sajátértékeit (nd db) a következő sajátértékek egyesítésével kapjuk meg: $\boldsymbol{M}_0$ (valós) sajátértékei és az $\boldsymbol{M}_j$ ($j = 1, \ldots, k$) sajátértékeinek kettőzésével. Megjegyezzük, hogy páros n esetén hasonló eredmények érvényesek, azzal a különbséggel, hogy ott a $\mathfrak{C}_n^{(s)}$ spektruma az $\boldsymbol{M}_0$ és $\boldsymbol{M}_{n-1}$ sajátértékeinek egyesítésével, és az $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{\frac{n}{2}-1}$ sajátértékeinek kettőződéséből adódik.

$\mathfrak{C}_n^{(s)}$ sajátvektorai

A sajátvektorok megkaphatóak az $\boldsymbol{M}_0, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_{n-1}$ $d \times d$ méretű önadjungált mátrixok d db (általában komplex) ortonormális sajátvektorainak összetételével a következőképpen. Bármely $j = 1, \ldots, k$ esetén: ha $\boldsymbol{v}$ egy egységnyi hosszúságú sajátvektor az $\boldsymbol{M}_j$ λ sajátértékéhez, akkor [9]-ben belátták, hogy az összetett vektor

$$(\boldsymbol{v}^T, \rho_j \boldsymbol{v}^T, \rho_j^2 \boldsymbol{v}^T, \ldots, \rho_j^{n-1} \boldsymbol{v}^T)^T \in \mathbb{C}^{nd}$$

a $\mathfrak{C}_n^{(s)}$ sajátvektora is azonos λ sajátértékkel. Normanégyzete

$$\boldsymbol{v}^* \boldsymbol{v}(1 + \rho_j \rho_j^{-1} + \rho_j^2 \rho_j^{-2} + \cdots + \cdots + \rho_j^{n-1} \rho_j^{-(n-1)}) = n.$$

Ezért a vektor

$$\boldsymbol{w} = \frac{1}{\sqrt{n}}(\boldsymbol{v}^T, \rho_j \boldsymbol{v}^T, \rho_j^2 \boldsymbol{v}^T, \ldots, \rho_j^{n-1} \boldsymbol{v}^T)^T \in \mathbb{C}^{nd} \tag{3.2}$$

egy egységnormájú sajátvektora $\mathfrak{C}_n^{(s)}$-nek (komplex koordinátákkal).

Továbbá, ha

$$\boldsymbol{z} = \frac{1}{\sqrt{n}}(\boldsymbol{t}^T, \rho_\ell \boldsymbol{t}^T, \rho_\ell^2 \boldsymbol{t}^T, \ldots, \rho_\ell^{n-1} \boldsymbol{t}^T)^T \in \mathbb{C}^{nd}$$

a $\mathfrak{C}_n^{(s)}$ egy másik egységnormájú sajátvektora, amely egy másik $\boldsymbol{M}_\ell$ ($\ell \neq j$) mátrix egységnormájú $\boldsymbol{t}$ sajátvektorából áll össze, akkor $\boldsymbol{w}$ és $\boldsymbol{z}$ ortogonálisak, függetlenül attól, hogy $\boldsymbol{M}_\ell$-nek ugyanaz a λ sajátértéke-e, mint $\boldsymbol{M}_j$-nek vagy sem. Hasonló konstrukció érvényes a $\boldsymbol{M}_0$ sajátvektoraiból kiindulva.

Itt minden $j = 0, 1, \ldots, n-1$ esetén az $\boldsymbol{M}_j$-nek van d páronként ortonormális sajátvektora (potenciális $\boldsymbol{v}$-k) , és az így kapott $\boldsymbol{w}$-k szintén páronként ortonormálisak. Tegyük fel, hogy a $\boldsymbol{M}_j$ sajátvektorai a (valós) sajátértékeinek nem növekvő sorrendjében felsorolhatók, és a kapott $\boldsymbol{w}$-k is ezt a sorrendet követik, $j = 0, 1, \ldots, n-1$ esetén.

Válasszunk egy egységnormájú sajátvektort $\boldsymbol{v} \in \mathbb{C}^d$ az $\boldsymbol{M}_j$-ben, amelynek (valós) sajátértéke a λ. Ekkor $\overline{\boldsymbol{v}} \in \mathbb{C}^d$ az $\boldsymbol{M}_{n-j}$ megfelelő egységnormájú sajátvektora, amelynek sajátértéke megegyezik, λ. Tekintsük a $\boldsymbol{w} \in \mathbb{C}^{nd}$ és $\overline{\boldsymbol{w}} \in \mathbb{C}^{nd}$ vektorokat, amelyeket Képlet (3.2) segítségével kapunk. Láttuk, hogy ezek a $\mathfrak{C}_n^{(s)}$ ortonormális sajátvektorai, amelyek a (legalább) kettes multiplicitású λ sajátértéknek felelnek meg. Ezekből, az ennek a kettős λ sajátértéknek

megfelelő, új ortonormális sajátvektorpár

$$\frac{\boldsymbol{w}+\overline{\boldsymbol{w}}}{\sqrt{2}} \quad \text{és} \quad -i\frac{\boldsymbol{w}-\overline{\boldsymbol{w}}}{\sqrt{2}} \tag{3.3}$$

konstruálódik, de ők ebben a sorrendben a $\boldsymbol{w}$ és $\overline{\boldsymbol{w}}$ eredeti helyét foglalják el. Ezek valós koordinátákkal és egységnyi normával rendelkeznek. (Valójában a koordinátáik a $\boldsymbol{w}$ koordinátáinak valós és képzetes részének $\sqrt{2}$-szeresei). Ez összhangban van azzal a ténnyel, hogy egy valós szimmetrikus mátrixnak, mint $\mathfrak{C}_n^{(s)}$, kell, hogy legyen egy ortogonális sajátvektorrendszerük valós koordinátákkal. Nem megyünk bele a részletekbe, és nem tárgyaljuk a szélsőséges eseteket.

Tekintsük az $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{nd}$ vektorokat, az így kapott (valós koordinátájú) ortonormális sajátvektorhalmazát a $\mathfrak{C}_n^{(s)}$-nek (a fenti sorrendben), és jelöljük a következővel $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{nd})$ az $nd \times nd$ méretű (valós) ortogonális mátrixot, amely ezeket oszloponként tartalmazza. Legyen

$$\mathfrak{C}_n^{(s)} = \boldsymbol{U}\boldsymbol{\Lambda}^{(s)}\boldsymbol{U}^T \tag{3.4}$$

a megfelelő spektrális dekompozíció. Ezt az előkészítést követően, be tudjuk bizonyítani a következő tételt.

3.3. A fő tétel a spektrumról az idő- és frekvenciatartományban

Az előző megfigyelések alapján megfogalmazható a következő tétel.

3.1. Tétel. *Legyen $\{\boldsymbol{X}_t\}$ egy d-dimenziós gyengén stacionárius idősor valós komponensekkel. Jelöljük $\boldsymbol{C}(h) = [c_{ij}(h)]$-val a $d \times d$-s autokovariancia mátrixokat ($\boldsymbol{C}(-h) = \boldsymbol{C}^T(h)$, $h \in \mathbb{Z}$) az időtartományban. Feltételezzük, hogy a elemenként abszolút összegezhetőek, azaz $\sum_{h=0}^{\infty} |c_{pq}(h)| < \infty$ $p, q = 1, \ldots, d$ esetén. Ekkor az önadjungált, pozitív szemidefinit spektrális sűrűségmátrix $\boldsymbol{f}$ létezik a frekvenciatartományban, amelyet a Képlet (3.1) definícióval határozunk meg. Páratlan $n = 2k+1$ esetén tekintsük az $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ és a $\mathfrak{C}_n$ blokk-Toeplitz mátrixot; továbbá, Tekintsük az $\omega_j = \frac{2\pi j}{n}$ $j = 0, \ldots, n-1$ Fourier-frekvenciákat. Továbbá, legyen $\boldsymbol{D}_n$ az a $dn \times dn$-s diagonális mátrix, amely az $\boldsymbol{f}(0), \boldsymbol{f}(\omega_1), \boldsymbol{f}(\omega_2), \ldots, \boldsymbol{f}(\omega_k)$, $\boldsymbol{f}(\omega_k), \ldots, \boldsymbol{f}(\omega_2), \boldsymbol{f}(\omega_1)$ mátrixok spektrumait tartalmazza [2] a főátlójában [3], azaz,*

$$\boldsymbol{D}_n = \operatorname{diag}(\operatorname{spec} \boldsymbol{f}(0), \operatorname{spec} \boldsymbol{f}(\omega_1), \ldots, \operatorname{spec} \boldsymbol{f}(\omega_k), \operatorname{spec} \boldsymbol{f}(\omega_k), \ldots, \operatorname{spec} \boldsymbol{f}(\omega_1)).$$

Ezután a Képlet (3.4) spektrális dekompozíciójával,

$$\boldsymbol{U}^T\mathfrak{C}_n\boldsymbol{U} - 2\pi\boldsymbol{D}_n \to \boldsymbol{O}, \quad n \to \infty,$$

azaz a $\boldsymbol{U}^T\mathfrak{C}_n\boldsymbol{U} - 2\pi\boldsymbol{D}_n$ mátrix értékei 0-hoz tartanak egyenletesen.

A Tétel 3.1 bizonyítása azt sugallja, hogy a konvergencia sebessége attól függ, hogy az abszolút összegezhető autokovarianciák farka milyen gyorsan közelíti a nullát. Például

[2] Az érintett mátrix sajátértékei (spec $(\cdot)$) nem növekvőben vannak sorrendben, ha nincs másképp megadva.
[3] A duplikáció annak a ténynek köszönhető, hogy $\boldsymbol{f}(\omega_j) = \boldsymbol{f}(\omega_{n-j})$, $j = 1, \ldots, k$, valós idősorok esetén.

a VARMA-folyamatok esetében az autokovarianciák exponenciálisan csökkennek, és ez azt jelenti, hogy a konvergencia sebessége geometriai rátájú.

3.4. A fő tétel összefüggései és következményei

3.4.1. Korlátok $\mathfrak{C}_n$ sajátértékeire

Analóg módon az 1D állítással (lásd [11]), a $\mathfrak{C}_n$ blokk-Toeplitz mátrix sajátértékeihez a következő alsó és felső korlátok kaphatóak.

3.1. Állítás. *Tegyük fel, hogy a a d-dimenziós gyengén stacionárius valós koordinátájú $\{X_t\}$ folyamat $\boldsymbol{f}$ spektrális sűrűségfüggvényére a következők érvényesek:*

$$\lambda_{\inf} := \inf_{\omega\in[0,2\pi],q\in\{1,\dots,d\}} \lambda_q(\boldsymbol{f}(\omega)) > 0,$$
$$\lambda_{\sup} := \sup_{\omega\in[0,2\pi],q\in\{1,\dots,d\}} \lambda_q(\boldsymbol{f}(\omega)) < \infty.$$

(Megjegyezzük, hogy a Tétel 3.1 feltételei szerint $\boldsymbol{f}(\omega) > 0$ és majdnem mindenhol folytonos a $[0, 2\pi]$ tartományban, így a fenti feltételek könnyen teljesülnek.)

Ekkor a $\mathfrak{C}_n$ blokk-Toeplitz mátrix sajátértékeire $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_{nd}$ a következők érvényesek:

$$2\pi\lambda_{\inf} \le \lambda_1 \le \lambda_{nd} \le 2\pi\lambda_{\sup}.$$

3.4.2. Komplex PC transzformáció

A valós koordinátájú $\boldsymbol{X} = (\boldsymbol{X}_1^T, \dots \boldsymbol{X}_n^T)^T$ véletlen vektorok gyűjteményére komplex koordinátájú PC-transzformáció $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \dots, \boldsymbol{Z}_n^T)^T$ írható fel, amelyet a következő módon kapunk

$$\boldsymbol{Z} = \boldsymbol{W}^*\boldsymbol{X}.$$

Itt, az Képlet (3.4)-vel analóg módon, $\mathfrak{C}_n^{(s)}$ a következő spektrál dekompozícióval rendelkezik

$$\mathfrak{C}_n^{(s)} = \boldsymbol{W}\boldsymbol{\Lambda}^{(s)}\boldsymbol{W}^* \tag{3.5}$$

komplex sajátvektorokkal: $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_{nd})$, ahol a $\boldsymbol{w}$-k a $\mathfrak{C}_n^{(s)}$ ortonormális (komplex koordinátájú) sajátvektorai az Képlet (3.2) által bevezetetten.

Az Képlet (3.2) alapján a $\boldsymbol{Z}_j$ blokk a következőképpen írható fel

$$\boldsymbol{Z}_j = \frac{1}{\sqrt{n}} \left(\boldsymbol{V}_j^* \otimes \boldsymbol{r}^*\right) \boldsymbol{X},$$

ahol $\boldsymbol{r}^* = (1, \rho_j^{-1}, \rho_j^{-2}, \dots, \rho_j^{-(n-1)})$ és $\boldsymbol{V}_j$ a $d \times d$-s unitér mátrix az $\boldsymbol{M}_j = \boldsymbol{V}_j\boldsymbol{\Lambda}_j\boldsymbol{V}_j^*$ spektrál dekompozícióban.

3.4.3. Kapcsolat a diszkrét Fourier-transzformációhoz

Ezt a komplex PC-transzformációt a diszkrét Fourier-transzformációhoz kapcsolhatjuk. Mivel $\mathbb{E}\boldsymbol{Z}_j\boldsymbol{Z}_j^* = \boldsymbol{\Lambda}_j$, a Tétel 3.1 bizonyításából az következik, hogy

$$\mathbb{E}(\boldsymbol{V}_j\boldsymbol{Z}_j)(\boldsymbol{V}_j\boldsymbol{Z}_j)^* = \boldsymbol{V}_j\boldsymbol{\Lambda}_j\boldsymbol{V}_j^* = \boldsymbol{M}_j.$$

Ugyanakkor,

$$\boldsymbol{V}_j\boldsymbol{Z}_j = \frac{1}{\sqrt{n}}\boldsymbol{V}_j\left(\boldsymbol{V}_j^* \times \boldsymbol{r}^*\right)\boldsymbol{X} = \frac{1}{\sqrt{n}}\left(\mathbb{I}_d \otimes \boldsymbol{r}^*\right)\boldsymbol{X} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{X}_t e^{-it\omega_j}, \quad j = 1, \ldots, n.$$

Ez a $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ diszkrét Fourier-transzformációja.

3.4.4. A folyamat alacsony rangú közelítése

Legyen $\{\boldsymbol{X}_t\}_{t=1}^n$ egy d dimenziós folyamat véges szelete, valós koordinátákkal és állandó $1 \leq r \leq d$ ranggal. Diszkrét Fourier-transzformáltja, amelyet a Szakasz 3.4.3-ban tárgyalunk, a következő

$$\boldsymbol{T}_j := \boldsymbol{V}_j\boldsymbol{Z}_j = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{X}_t e^{-it\omega_j}, \quad j = 0, \ldots, n-1.$$

Pontosabban, $\boldsymbol{T}_0 = \frac{1}{\sqrt{n}}\sum_{t=1}^n \boldsymbol{X}_t$, $\boldsymbol{T}_j = \frac{1}{\sqrt{n}}\sum_{t=1}^n \boldsymbol{X}_t[\cos(t\omega_j) - i\sin(t\omega_j)]$, és $\boldsymbol{T}_{n-j} = \overline{\boldsymbol{T}}_j$, $j = 1, \ldots, k$ $(n = 2k+1)$ esetén. Ezért,

$$\boldsymbol{Z}_j = \boldsymbol{V}_j^{-1}\boldsymbol{T}_j = \boldsymbol{V}_j^*\boldsymbol{T}_j, \quad j = 0, \ldots, n-1;$$

és könnyen belátható, hogy $\boldsymbol{Z}_{n-j} = \overline{\boldsymbol{Z}}_j$ is teljesül $j = 1, \ldots, k$ esetén.

A folyamat legjobb m-rangú közelítőjének megtalálásához $(1 < m \leq r)$, a d-dimenziós $\boldsymbol{T}_j$ vektort vetítjük ki a $\boldsymbol{V}_j$ m vezető sajátvektora által kifeszített altérre. Fontos, hogy a $\boldsymbol{\Lambda}_j$ sajátértékei nem-növekvő sorrendben vannak. Legyen $\boldsymbol{v}_{j1}, \ldots, \boldsymbol{v}_{jm}$ az m legnagyobb sajátértéknek megfelelő sajátvektorok, amit a $\boldsymbol{V}_j^{(m)}$ mátrixba foglalunk. Ekkor

$$\widehat{\boldsymbol{T}}_j^{(m)} := \mathtt{Proj}_{\,\mathtt{Span}\,\{\boldsymbol{v}_{j1},\ldots,\boldsymbol{v}_{jm}\}}\boldsymbol{T}_j = \left[\boldsymbol{V}_j^{(m)}\left(\boldsymbol{V}_j^{(m)}\right)^*\right]\boldsymbol{V}_j\boldsymbol{Z}_j = \sum_{\ell=1}^{m}\left(\boldsymbol{v}_{j\ell}^*\boldsymbol{V}_j\boldsymbol{Z}_j\right)\boldsymbol{v}_{j\ell} = \sum_{\ell=1}^{m} Z_{j\ell}\boldsymbol{v}_{j\ell},$$

és $\widehat{\boldsymbol{T}}_{n-j} = \overline{\widehat{\boldsymbol{T}}}_j$, $j = 1, \ldots, k$ esetén (az előző megfontolások szerint) $n = 2k+1$ esetén. Tehát minden j esetében, a kapott vektor a $\boldsymbol{v}_{j\ell}$ vektorok és a megfelelő $\boldsymbol{Z}_j$ vektor $\boldsymbol{Z}_{j\ell}$ koordinátáinak lineáris kombinációja, $\ell = 1, \ldots, m$.

Végül az $\boldsymbol{X}_t$ m-rendű közelítését az időtartományban inverz Fourier-transzformálással találjuk meg:

$$\widehat{\boldsymbol{X}}_t^{(m)} := \underset{j}{\mathrm{IDFT}}\left[\widehat{\boldsymbol{T}}_j^{(m)}\right] = \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\widehat{\boldsymbol{T}}_j^{(m)}e^{it\omega_j}.$$

Ily módon egy alacsonyabb rangú folyamatot kapunk, amelynek spektrálsűrűségfüggvénye $m \leq r$ rangú. Megjegyezzük, hogy ha a folyamat szabályos (pl. racionális spektrálsűrűségfüggvénnyel rendelkezik), akkor az alacsony rangú közelítése is az.

3.5. dynPCA: alacsony dimenziós közelítés a komplex PC-kből

Az előző *rang*csökkentésből adódik egy *dimenzió*csökkentési technika is. A frekvenciatartományban a $\boldsymbol{V}_j$ m vezető sajátvektorai által kifeszített d-dimenziós altérre való vetítés helyett, abba képezünk le bele. Ezt a transzformációt már láttuk korábban, de a teljes $\boldsymbol{V}_j$ mátrixszal:

$$\boldsymbol{Z}_j^{(m)} = \left[\boldsymbol{V}_j^{(m)}\right]^* \boldsymbol{T}_j, \quad j = 0, \ldots, n-1.$$

Vagy komponensenként: $Z_{j\ell} = \boldsymbol{v}_{j\ell}^* \boldsymbol{T}_j$, $(\ell = 1, \ldots m)$ a $\boldsymbol{T}_j$ egy m-dimenziós közelítésének koordinátáit határozza meg, $m \leq r \leq d$. $\boldsymbol{Z}_j^{(m)} = (Z_{j1}, \ldots, Z_{jm})^T$. Vagyis minden blokkban az első m komplex PC-t vesszük. Megjegyezzük ismét, hogy az minden egyes $\boldsymbol{\Lambda}_j$ átlójában az értékek nem-növekvő sorrendben vannak. A frekvenciatartomány PC-folyamatának visszatranszformálásához az időtartományba inverz Fourier-transzformációt használunk, ahogyan a rangcsökkentésnél is tettük:

$$\boldsymbol{Q}_t^{(m)} = \underset{j}{\mathrm{IDFT}}\left[\boldsymbol{Z}_j^{(m)}\right] = \underset{j}{\mathrm{IDFT}}\left[\left[\boldsymbol{V}_j^{(m)}\right]^* \underset{t}{\mathrm{DFT}}\left(\boldsymbol{X}_t\right)\right] = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[\left[\boldsymbol{V}_j^{(m)}\right]^* \boldsymbol{T}_j\right] e^{it\omega_j},$$

ahol $\boldsymbol{Q}_t^{(m)}$ a kapott m-dimenziós idősor. Ez hasonlít Brillinger [12] *idősor PCA a frekvenciatartományban*[4] módszeréhez, és a Szakasz 3.5.1-ban szereplő numerikus megfontolásainkkal *dynPCA*-nak nevezzük. Komponensenként az m-dimenziós PC-folyamat a következő:

$$\boldsymbol{Q}_t^{(m)} = \frac{1}{\sqrt{n}} \left(\sum_{j=0}^{n-1} Z_{j1} e^{it\omega_j}, \ldots, \sum_{j=0}^{n-1} Z_{jm} e^{it\omega_j} \right)^T. \tag{3.6}$$

Rekonstrukció az alacsonyabb dimenziós térből

Az előző transzformáció egy veszteséges dimenziócsökkentő technika is. A $\boldsymbol{Q}_t$-ból rekonstruálhatunk egy, az eredetihez közeli idősort, hasonlóan a statikus PCA működéséhez. A rekonstruált idősor valójában a Szakasz 3.4.4 alacsony rangú közelítése:

$$\tilde{\boldsymbol{X}}_t := \underset{j}{\mathrm{IDFT}}\left[\left[\boldsymbol{V}_j^{(m)}\right] \underset{t}{\mathrm{DFT}}\left(\boldsymbol{Q}_t^{(m)}\right)\right] = \widehat{\boldsymbol{X}}_t^{(m)}.$$

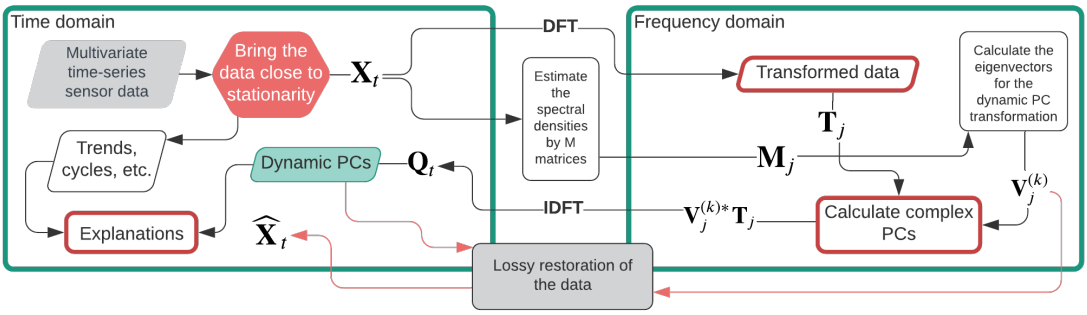
Minél nagyobb az átlagos különbség a spektrumban a λ_k és λ_{k+1} sajátérték-folyamatok között, annál kisebb a veszteség. Az algoritmus sematikus működési folyamata az Ábra 3.1-ban látható.

3.5.1. A dynPCA numerikus megfontolásai

Elméletileg egy n hosszúságú idősor esetében n ponton (Fourier-frekvenciákon) kell kiszámítanunk a $\boldsymbol{V}_j^{(k)}$ mátrixokat (és így az $\boldsymbol{M}_j$ mátrixokat). A $\boldsymbol{C}(h)$ mátrixokat ($\widehat{\boldsymbol{C}}(h)$) is meg kell becsülnünk $m = \lfloor n/2 \rfloor$ lagig. Ez n növekedésével gyorsan megoldhatatlanná válik.

A gyakorlatban két egyszerűsítést alkalmazunk [C4]. A kovarianciákat csak egy bizonyos lagig becsüljük, mivel azok "csökkenésre hajlamosak"; és az $\boldsymbol{M}_j$ mátrixokat kevesebb Fourier-

[4] Ezt később dinamikus főkomponens-elemzésnek nevezték el a szakirodalomban.



3.1. ábra. A dynPCA algoritmus sematikus működési folyamata

frekvencián számoljuk ki.

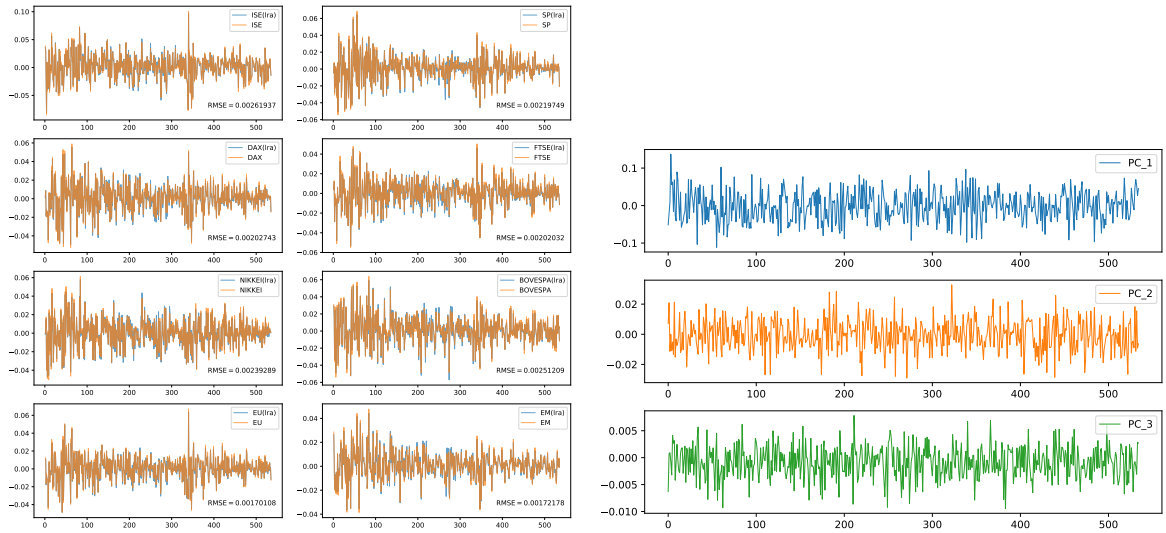
$$\boldsymbol{M}_j = \sum_{h=-k}^{k} \boldsymbol{C}(h)e^{i\omega_j h} \quad \rightarrow \quad \widehat{\boldsymbol{M}}_{j'} = \sum_{h=-l}^{l} \widehat{\boldsymbol{C}}(h)e^{i\omega_{j'} h}$$

$$\text{for } j = 1, \dots n; \qquad \rightarrow \quad \text{for } j' = 1, \dots s;$$

ahol $s \ll n$ a felhasznált egyenlő távolságú Fourier-frekvenciák száma, és $l \ll k$ a kovarianciák becsléséhez tartozó lag. Ezt követően a $\widehat{\boldsymbol{M}}_{j'}$-k inter-/extrapolálhatóak az n Fourier-frekvenciára, a spektrálsűrűségfüggvény-mátrix folyamat folytonossága miatt.

3.6. A dynPCA alkalmazásai

A dynPCA algoritmus két különböző alkalmazását mutatjuk be. Az elsőt egy pénzügyi adatkészletre (lásd [J4]), a másodikat elektrokardiográfiás (EKG) adatkészletekre (lásd [C4]) alkalmazzuk. Mindegyik alkalmazás az algoritmus kimenetének különböző aspektusaira összpontosít, megmutatva az algoritmus sokoldalúságát.



(a) 3 rangú közelítés (kék) az eredeti adatokon (narancssárga)

(b) A 3 vezető főkomponens az idő tartományban

3.2. ábra. Példa alkalmazásra tőzsdei idősorokon

Appendix A

Grafikus modellek

A grafikus modellekkel kapcsolatos további bevezető anyagokat tartalmaz a disszertáció.

Appendix B

Nemparaméteres regresszió

A nemparaméteres regresszióval kapcsolatos további bevezető anyagokat tartalmaz a disszertáció; és bizonyítottuk az implementált kernel regressziós becslés négyzetes értelemben vett konzisztenciáját.

Appendix C

Idősorelemzés

Az idősorelemzéssel kapcsolatos további bevezető anyagokat tartalmaz a disszertáció.

Publikációim listája

Folyóiratcikkek

[J1] M. Bolla, F. Abdelkhalek és M. Baranyi. „Graphical models, regression graphs, and recursive linear regression in a unified way". *Acta Scientiarum Mathematicarum* 85.12 (2019), 9–57. old. ISSN: 0001-6969. DOI: 10.14232/actasm-018-331-4.

[J2] M. Baranyi és M. Bolla. „Iterated Conditional Expectation Algorithm on DAGs and Regression Graphs". *Econometrics and Statistics* 20 (2021), 131–152. old. ISSN: 2452-3062. DOI: 10.1016/j.ecosta.2020.05.003.

[J3] M. Baranyi és R. Molontay. „Comparing the effectiveness of two remedial mathematics courses using modern regression discontinuity techniques". *Interactive Learning Environments* 29.2 (2021), 247–269. old. DOI: 10.1080/10494820.2020.1839506.

[J4] M. Bolla, T. Szabados, M. Baranyi és F. Abdelkhalek. „Block circulant matrices and the spectra of multivariate stationary sequences". *Special Matrices* 9.1 (2021. jan.), 36–51. old. ISSN: 2300-7451. DOI: 10.1515/spma-2020-0121.

Konferenciacikkek

[C1] M. Baranyi, K. Gál, R. Molontay és M. Szabó. „Modeling Students' Academic Performance Using Bayesian Networks". *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*. IEEE, 2019. nov., 42–49. old. DOI: 10.1109/ICETA48886.2019.9040067.

[C2] M. Baranyi és R. Molontay. „Effect of Mathematics Remediation on Academic Achievement – A Regression Discontinuity Approach". *2019 International Symposium on Educational Technology (ISET)*. IEEE, 2019. júl., 29–33. old. DOI: 10.1109/ISET.2019.00016.

[C3] M. Baranyi, M. Nagy és R. Molontay. „Interpretable Deep Learning for University Dropout Prediction". *Proceedings of the 21st Annual Conference on Information Technology Education*. New York, NY, USA: Association for Computing Machinery, 2020, 13–19. old. ISBN: 9781450370455. DOI: 10.1145/3368308.3415382.

[C4] M. Baranyi, M. Bolla és Gy. Kocsisné Szilágyi. „A Novel Dynamic Principal Component Analysis Method, Applied to ECG Signals". *2021 55th Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, CA, USA: IEEE, 2021. okt., 265–269. old. ISBN: 978-1-66545-828-3. DOI: 10.1109/IEEECONF53345.2021.9723158.

[C5] M. Bolla, M. Baranyi és F. Abdelkhalek. „Dimension Reduction of High Frequency and High Dimensional Data in Time and Space". *HU-MATHS-IN Success Stories of Mathematical Short-Term Projects for Industry in 2017-2021*. Szerk. Z. Horváth. Györ, Hungary: Széchenyi István University, 2021, 60–66. old. ISBN: 978-615-5837-99-9.

Irodalom

[1] M. Baranyi. „Graphical Models and Some Related Algorithms”. MSc Thesis. Budapest: Budapest University of Technology és Economics, 2018.

[2] L. Breiman és J. H. Friedman. „Estimating optimal transformations for multiple regression and correlation”. *Journal of the American Statistical Association* 80 (1985), 580–619. old.

[3] L. Györfi, M. Kohler, A. Krzyzak és H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002, I–XVI., 1–647. old. ISBN: 978-0-387-95441-7.

[4] N. Wermuth és K. Sadeghi. „Sequences of regressions and their independences”. *TEST* 21 (2012), 215–279. old.

[5] J. Racine és Q. Li. „Nonparametric estimation of regression functions with both categorical and continuous data”. *Journal of Econometrics* 119.1 (2004), 99–130. old.

[6] D. L. Thistlethwaite és D. T. Campbell. „Regression-discontinuity analysis: An alternative to the ex post facto experiment”. *Journal of Educational Psychology* 51.6 (1960), 309–317. old. ISSN: 1939-2176(Electronic),0022-0663(Print). DOI: `10.1037/h0044319`.

[7] D. S. Lee és D. Card. „Regression discontinuity inference with specification error”. *Journal of Econometrics* 142.2 (2008), 655–674. old.

[8] T. Nagler. „A generic approach to nonparametric function estimation with mixed data”. *Statistics & Probability Letters* 137 (2018), 326–330. old. ISSN: 0167-7152. DOI: `https://doi.org/10.1016/j.spl.2018.02.040`.

[9] G. J. Tee. „Eigenvectors of block circulant and alternating circulant matrices”. *New Zealand Journal of Mathematics* 36.8 (2007), 195–211. old.

[10] B. Friedman. „Eigenvalues of composite matrices”. *Mathematical Proceedings of the Cambridge Philosophical Society* 57.1 (1961), 37–49. old.

[11] P. J. Brockwell, R. A. Davis és S. E. Fienberg. *Time series: Theory and methods*. Springer Science & Business Media, 1991.

[12] D. R. Brillinger. *Time Series: Data Analysis and Theory*. Classics in Applied Mathematics 36. SIAM, 2001.