LATVIJAS UNIVERSITĀTE

PARAMETRISKĀS UN NEPARAMETRISKĀS BEIJESA METODES UN TO PIELIETOJUMI

DIPLOMDARBS

Autors: **Aleksis Jurševskis**

Stud. apl. aj06048

Darba vadītājs: doc. Dr. math. Jānis Valeinis

RĪGA 2013

Anotācija

Diplomdarbā tiek apskatītas parametriskās un neparametriskās Beijesa metodes. Tiek salīdzinātas Beijesa un klasiskās, neparametriskās metodes maiņas punktu noteikšanā un laikrindu prognozēšanā.

Atslēgas vārdi: Parametriskā Beijesa statistika, neparametriskā Beijesa statistika, maiņas punktu noteikšana, laikrindu prognozēšana

Abstract

This diploma thesis investigates parametrical and nonparametrical Bayesian methods. A comparison is made between Bayesian and classical methods in changepoint detection and time series prediction.

Keywords: Parametrical Bayesian statistics, nonparametric Bayesian statistics, Changepoint detection, Times series prediction

Saturs

Ievads

Ja vajadzētu klasificēt visas statistikas metodes, viens no veidiem, kā to izdarīt, varētu būt 1. tabulā uzskicētais.

1. tabula: Statistikas metožu klasifikācija.

Klasiskās metodes	Klasiskās, neparametriskās metodes
Beijesa metodes	Beijesa neparametriskās metodes

Beijesa (*Bayes*)[1] metodes pazīstamas jau kopš K. Gausa (*C. F. Gauss*) laikiem, kurš ieguva normālo sadalijumu divos dažādos veidos: ar klasisko, sauktu arī par frekventistu (*frequentist*), un ar Beijesa [1, 57. lpp.] metodi. Neparametriskās Beijesa metodes savukārt parādījās salīdzinoši nesen, 70. gadu sākumā, kad T. Fērgusons (*T. S. Ferguson*) konstruēja pirmo sadalījumu funkciju telpā [2]. Vadošie nozares pētnieki S. Gošals (*S. Ghoshal*) un A. van der Vārts (*A. van der Vaart*) prognozē [3], ka līdz šīs dekādes beigām neparametriskās Beijesa statistikas metodoloģija būs lielā mērā izpētīta.

Beijesa (*Bayes*) metodes, gan parametriskās, gan neparametriskās, tiek plaši pielietotas visdažādākajās zinātnes nozarēs. Īpaši populāras tās ir sociālajās zinātnēs, piemēram, politikā [4], socioloģijā [5], jurisprudencē [6], kā arī ekonomikā [7]. Taču metodes tiek pretrunīgi vērtētas, lai arī daži zinātnieki tās uzskata par fundamentālu pētnieka rīku bez kura nevar iztikt [8], citi argumentē, ka no Beijesa statistikas zinātnē ir jāizvairās, cik vien iespējams [9].

Diplomdarba mērķi ir sekojoši:

1. iepazīties gan ar parametrisko, gan neparametrisko Beijesa metožu teorētisko pamatojumu;

2. izpētīt Beijesa pieeju maiņas punktu noteikšanā, salīdzināt to ar empīriskās ticamības funkcijas metodi, ko izpētījis A. Vaselāns [10];

[1] Šeit un turpmāk gan pētnieku uzvārdi, gan mazāk pazīstami termini doti kursīvā angliski.

3. izprast neparametrisko Beijesa metožu pamatojumu un praktisko pielietojamību laikrindu prognozēšanā.

Darbs sastāv no trīs nodaļām. Pirmajā tiks aprakstīti pamatjēdzieni, ilustrētas Beijesa metožu idejas un problemātika. Otrā nodaļa tiks veltīta maiņaspunktu tematikai, salīdzināta parametriskā Beijesa metode ar procedūru no klasiskās, neparametriskās statistikas procedūru klāsta. Trešā daļa ir veltīta Gausa procesiem, kas ir viens no daudzsološākajām neparametriskās Beijesa statistikas izpētes virzieniem. Tiek salīdzināta to un ARIMA modeļu laikrindu prognožu kvalitāte.

1. Beijesa statistikas pamati

1.1. Pamatprincipi

Klasiskajā statistikā parametri tiek modelēti kā konstantes θ. Tiek konstruēti to novērtējumi $\hat{\theta}$, veidotas hipotēžu pārbaudes par to vērtību, piemēram, $H_0 : \theta = c$ pret $H_1 : \theta \neq c$. Frekventistu mērķis ir radīt metodes ar labām īpašībam, novērtēšanas procedūrām jābūt nenovirzītām, efektīvām, ātri konverģējošām uz patieso vērtību, savukārt hipotēžu testiem jābūt ar augstu jaudu.

Beijesa metožu piekritējiem jeb beijesiešiem šīs īpašības arīdzan ir svarīgas, taču otršķirīgas. Tiek pieņemts, ka parametri ir nevis konstantes, bet gan gadījuma lielumi, par kuriem *a priori* tiek izdarīti pieņēmumi. Beijesiešiem pirmajā plānā ir jautājums par šo pieņēmumu ticamību, proti, tiek mēģināts atbildēt uz jautājumu: vai novērojumi apstiprina vai noraida sākotnējos pieņēmumus, un, ja noraida, tad kā būtu jāmaina sākotnējie uzskati? L. Vasermans (*L. Wasserman*) uzskata [11], ka ir veltīgi salīdzināt, kura pieeja ir pārāka, jo to mērķi ir fundamentāli atšķirīgi.

Aprakstīsim Beijesa procedūras, ilustrējot tās ar pāris piemēriem. Tālāk sekojošā teorija lielā mērā balstīta uz Vasermaņa grāmatas [12] 11. nodaļu.

Jebkuras Beijesa metodes pamatā ir trīs posmi:

I Izvēlas aprioro *(prior)* blīvuma funkciju $f(\theta)$. Tā var būt gan specifiski sadalījumi, gan veselas funkciju klases. Ja pētnieka rīcībā ir kāda informācijas par iespējamajām θ vērtībām, tad šajā solī var mēģināt to izmantot, izvēloties aprioro sadalījumu.

II Izvēlas statistisko modeli $f(x|\theta)$, kas ataino pētnieka uzskatus par datiem, ja ir zināma θ vērtība.

III Novēro datus $X_1, ..., X_n$, realizācijas ievieto modelī $f(x|\theta)$. Kombinējot modeli ar

aprioro θ sadalījumu $f(\theta)$, tiek iegūts aposteriorais sadalījums $f(\theta|X_1, ..., X_n)$, atspoguļojot uzskatu par θ sadalījumu, ņemot vērā datu sniegto informāciju.

Apriorās un datu sniegtās informācijas kombinēšanai izmanto Beijesa teorēmu, kā dēļ Beijesa metodes nosauktas angļu matemātiķa Tomasa Beijesa vārdā. Ilustrācijai pieņemsim, ka Θ ir ar diskrētu blīvuma funkcija un esam novērojuši vienu gadījuma lieluma X realizāciju. Varam iegūt, ka

$$\mathbb{P}(\Theta = \theta|X = x) = \frac{\mathbb{P}(X = x, \Theta = \theta)}{\mathbb{P}(X = x)} = \\ = \frac{\mathbb{P}(X = x|\Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x|\Theta = \theta)\mathbb{P}(\Theta = \theta)}.$$

Pārejot uz nepārtrauktām blīvuma funkcijām, iegūsim

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}.$$

Ja dotas $X^n \equiv (X_1, ..., X_n)$ realizācijas $x^n \equiv (x_1, ..., x_n)$, tad, lai iegūtu $f(x^n|\theta)$, konstruē ticamības funkciju

$$f(x^n|\theta) = \prod_{i=1}^{n} f(x_i, \theta) = \mathcal{L}_n(\theta).$$

Savukārt aposteriorais sadalījums iegūst formu

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f(\theta), \tag{1.1}$$

kur $c_n = \int f(x^n|\theta)f(\theta)d\theta$ sauc par normalizējošo konstanti. Varam ievērot, ka $\mathcal{L}_n$ tiek noteikta ar precizitāti līdz konstantei.

Viens no Beijesa statistikas svarīgākajiem pieņēmumiem ir ticamības funkcijas principa *(likelihood principle)* patiesums, ko 1962. gadā postulēja Alans Birnbaums (*Alan Birnbaum*) [13]. Proti, tas apgalvo, ka visu nepieciešamo informāciju par datiem var iegūt no ticamības funkcijas $\mathcal{L}_n$. Varam to novērot vienādojumā (1.1), jo tas ir pēdējais solis Beijesa metožu analītiskajā daļā, viss tālākais ir skaitliskie rēķini.

Parametra novērtējumu var iegūt aprēķinot integrāli

$$\hat{\theta} = \int \theta f(\theta|x^n)d\theta.$$

Ja parametri ir vairāki, tad

$$f(\theta_1|x^n) = \int ... \int f(\theta_1, ..., \theta_p|x^n) d\theta_2 ... d\theta_p$$

$$\hat{\theta}_1 = \int \theta_1 f(\theta_1|x^n) d\theta_1.$$

Lai iegūtu parametra $1-\alpha\%$ ticamības intervālu, atrod tādus a un b, ka $\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_b^{\infty} f(\theta|x^n)d\theta = \alpha/2$. Tad

$$\mathbb{P}(\theta \in [a,b]|x^n) = \int_a^b f(\theta|x_n)d\theta = 1-\alpha. \tag{1.2}$$

Beijesieši ticamības intervāla apzīmēšanai angļu valodā reizēm izmanto vārdu *credible*, nevis *confidence*. Šeit parādās interesanta atšķirība starp beijesiešu un klasiskajām metodēm, proti, tā kā konstante var intervālam tikai piederēt vai nepiederēt, frekventistiem $\mathbb{P}(\theta \in [a,b])$ var pieņemt tikai bināras vērtības, un (1.2) ir jēga tikai tad, ja $\alpha = 0$ vai $\alpha = 1$.

Aprakstīsim vienu no pirmajiem piemēram Beijesa statistikas vēsturē.

Piemērs 1. Pieņemsim, ka $X_1, ..., X_n \sim Bern(p)$, kur $Bern(p)$ ir Bernulli sadalījums, meklējamajam parametram p izvēlēsimies vienmērīgo sadalījumu $f(p) = 1$. Saskaņā ar Beijesa teorēmu, aposteriorais sadalījums ir formā

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s(1-p)^{n-s} = p^{s+1-1}(1-p)^{n-s+1-1},$$

kur $s = \sum_{t=1}^{n} x_t$. Gadījuma lielumam ir Beta sadalījums ar parametriem α un β, ja tā blīvuma funkcija ir formā

$$f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

Varam ievērot, ka aposteriorais p sadalījums ir Beta sadalījums ar parametriem $s+1$ un $n-s+1$, tas ir

$$f(p|x^n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} p^{(s+1)-1} p^{(n-s+1)-1}.$$

To pārrakstām kā

$$p|x^n \sim Beta(s+1, n-s+1).$$

Varam ievērot, ka esam atraduši normalizējošo konstanti c_n, neaprēķinot integrāli $\int \mathcal{L}_n(p)f(p)dp$. Vidējā vērtība Beta$(\alpha,\beta)$ sadalījumam ir $\alpha/(\alpha+\beta)$, tāpēc Beijesa novērtējums p ir

$$\hat{p} = \frac{s+1}{n+2}.$$

Novērtējumu pārrakstām kā

$$\hat{p} = \lambda_n \hat{p_1} + (1 - \lambda_n)\hat{p_2},$$

kur $\hat{p_1} = \frac{s}{n}$ ir maksimālās ticamības novērtējums, $\hat{p_2}$ ir apriorā sadalījuma vidējā vērtība, un $\lambda_n = \frac{n}{n+2} \approx 1$. n jeb novērojumu skaitam pieaugot, apriorais sadalījums novērtējumu ietekmēs aizvien mazāk. 95 % ticamības intervālu varam iegūt, atrodot a un b tādus, ka $\int_a^b f(p|x^n)dp = 0.95$.

Pieņemsim, vienmērīgā sadalījuma vietā izvēlēsimies $p \sim \text{Beta}(\alpha, \beta)$ kā aprioro sadalījumu. Veicot pārrēķinu, varam iegūt, ka $p|x^n \sim \text{Beta}(\alpha + s, \beta + n - s)$. Vienmērīgas apriorais sadalījums ir īpašs gadījums, kad $\alpha = \beta = 1$. Aposteriorā vidējā vērtība ir

$$\hat{p} = \frac{\alpha + s}{\alpha + \beta + n}.$$

Kad apriorais un aposteriorais sadalījums pieder pie vienas saimes, saka, ka tie ir saistīti (*conjugate*).

Piemērs 2. Pieņemsim, ka $X_1, ..., X_n \sim \mathcal{N}(\theta, \sigma^2)$, kā arī, ka σ ir zināma. Izvēlēsimies $\mathcal{N}(a, b^2)$ kā aprioro θ sadalījumu. Varam iegūt, ka aposteriorais θ sadalījums ir formā

$$\theta|X^n \sim \mathcal{N}(\hat{\theta}, \tau^2),$$

kur

$$\hat{\theta} = w\hat{X} + (1 - w)a,$$

$$w = \frac{\frac{1}{se^2}}{\frac{1}{se^2} + \frac{1}{b^2}}, \qquad \frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2},$$

un $se = \sigma/\sqrt{n}$ ir maksimālās ticamības novērtējuma $\hat{X}$ standartnovirze. Varam ievērot, ka arī šajā piemēra apriorais un aposteriorais ir konjugāti. Ja $n \to \infty$, tad $w \to 1$ un $\tau/se \to 1$, tātad pie lieliem n aposteriorais sadalījums būs $\mathcal{N}(\hat{\theta}, se^2)$. To pašu var novērot, ja n ir fiksēts, bet $b \to \infty$, ko varam interpretēt kā gadījumu, kad apriorais sadalījums kļūst "plakans" (*flat*).

Abos piemēros varam novērot, ka, pieaugot n, apriorā sadalījuma ietekme strauji mazinās.

Var pierādīt [14], ka, ja $\hat{\theta}_n$ ir parametra maksimālās ticamības novērtējums vienādojuma atrisinājums, $\theta_n^* = \int \theta f(\theta|x_n)d\theta$ ir parametra Beijesa novērtējums un θ_0 ir īstā

parametra vērtība, tad pie noteiktiem nosacījumiem

$$\sqrt{n}(\hat{\theta}_n - \theta_n^*) \xrightarrow{p} 0,$$

un

$$\sqrt{n}(\theta_n^* - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1/I(\theta_0)),$$

kur $I(\theta_0)$ ir θ_0 Fišera informācija.

Varam izteikt minējumu, ka frekventistu un beijesiešu metodes asimptotiski uzvedas līdzīgi, kas ir arguments par labu frekvenču tipa procedūrām.

1.2. Hipotēžu pārbaude ar Beijesa metodēm

Lai pārbaudītu hipotēzes, nepieciešams izvēlēties aprioro sadalījumu gan H_0, gan parametriem θ un tad aprēķināt $\mathbb{P}(H_0|X^n)$. Apskatīsim

$$H_0 : \theta = \theta_0 \qquad \text{pret} \qquad H_1 : \theta \neq \theta_0.$$

Izvēlēsimies aprioro sadalījumu $\mathbb{P}(H_0) = \mathbb{P}(H_1) = \frac{1}{2}$. Ja spēkā H_1, tad jāizvēlas apriorais sadalījums θ, ko apzīmēsim ar $f(\theta)$. Izmantojot Beijesa teorēmu,

$$\begin{aligned}
\mathbb{P}(H_0|X^n = x^n) &= \frac{f(x^n|H_0)\mathbb{P}(H_0)}{f(x^n|H_0)\mathbb{P}(H_0) + f(x^n|H_1)\mathbb{P}(H_1)} \\
&= \frac{\frac{1}{2}f(x^n|\theta_0)}{\frac{1}{2}f(x^n|\theta_0) + \frac{1}{2}f(x^n|H_1)} \\
&= \frac{f(x^n|\theta_0)}{f(x^n|\theta_0) + \int f(x^n|\theta)f(\theta)d\theta} \\
&= \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \int \mathcal{L}(\theta_0)f(\theta)d\theta}.
\end{aligned}$$

Ja jānovērtē parametri, apriorā sadalījuma izvēle spēlē salīdzinoši mazu lomu, taču hipotēžu pārbaudē tā ir ļoti svarīga. Turklāt integrālim $\int \mathcal{L}(\theta_0)f(\theta)d\theta$ jākonverģē. Tā kā $0 \leq \int \mathcal{L}(\theta)f(\theta)d\theta \leq \mathcal{L}(\hat{\theta})$, varam iegūt sekojošu nevienādību H_0 ticamībai, neatkarīgi no $f(\theta)$ izvēles

$$\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \mathcal{L}(\hat{\theta})} \leq \mathbb{P}(H_0|X^n = x^n) \leq 1.$$

1.3. Markova ķēžu Montekarlo simulācijas

Tā kā praksē izmantoto parametru skaits var sniegties tūkstošos, integrāļu aprēķināšanai tiek simulētas parametru vērtības no aposteriorā sadalījuma, izmantojot Markova ķēžu Montekarlo tipa procedūras. Aprakstīsim Metropolisa-Heistingsa *(Metropolis-Hastings)* algoritmu, kas ir viens no vienkāršākajiem Markova ķēžu Montekarlo paņēmieniem. Nodaļā izmantota Vasermana grāmatas [12] 24. nodaļa.

Integrāli $I = \int_a^b h(x)dx$ var sadalīt divās daļās

$$I = \int_a^b h(x)dx = \int_a^b w(x)f(x)dx,$$

kur $w(x) = h(x)/f(x)$ un $f(x)$ ir kāda zināma sadalījuma blīvuma funkcija.

Definīcija 1. Procesu $\{X_n : n \in T\}$ sauc par Markova ķēdi, ja

$$\mathbb{P}(X_n = x|X_0, ..., X_{n-1}) = \mathbb{P}(X_n = x|X_{n-1}$$

visiem n.

Definīcija 2. Markova ķēdes sadalījumu π sauc par stacionāru, ja

$$\sum_{x \in S} \pi(x)P(x, y) = \pi(y), \qquad y \in S,$$

kur S ir stāvokļu telpa un $P(x, y)$ ir varbūtība pāriet no stāvokļa x uz stāvokli y.

Teorēma 1. *[12, 408. lpp.] Nereducējama, ergodiskai Markova ķēdei ir unikāls stacionārs sadalījums π. Ja g ir ierobežota funkcija, tad ar varbūtību 1*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g) \equiv \sum_j g(j)\pi_j.$$

Veicot simulācijas $X_1, X_2, ..., X_n$ no Markova ķēdes ar stacionāro sadalījumu f, no 1. teorēmas seko, ka

$$\hat{I} \equiv n^{-1} \sum_{i=1}^{n} w(X_i) \xrightarrow{p} \mathbb{E}_f(w(X)) = I.$$

Pieņemsim, ka $q(y|x)$ ir sadalījuma funkcija, ar ko protam iegūt gadījuma lieluma $(Y|X)$ realizācijas. Patvaļīgi izvēlamies X_0, ģenerējam ķēdi līdz X_i. Lai iegūtu X_{i+1}, veicam sekojošus soļus.

I Ģenerē $Y \sim q(y|X_i)$.

II Aprēķina $r \equiv r(X_i, Y)$, kur

$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}.$$

III Iegūst

$$X_{i+1} = Y_i \quad \text{ar varbūtību } r;$$

$$X_{i+1} = X_i \quad \text{ar varbūtību } 1 - r.$$

Par $q(y|x)$ bieži izvēlas $\mathcal{N}(x, b^2)$ ar $b > 0$. Šādā gadījumā q ir simetrisks $q(y|x) = q(x|y)$, un r var vienkāršot kā

$$r = \min\left\{\frac{f(Y)}{f(X_i)}, 1\right\}.$$

Piemērs 3. Košī sadalījumam blīvuma funkcija ir

$$f(x) = \frac{1}{\pi}\frac{1}{1 + x^2}.$$

Mērķis ir simulēt Markova ķēdi ar $f(x)$ kā stacionāro sadalījumu. Izvēlamies $q(y|x)$ $\mathcal{N}(x, b^2)$ sadalītu. Tad

$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}, 1\right\} = \min\left\{\frac{1 + x^2}{1 + y^2}, 1\right\}.$$

Tātad ģenerējam $Y \sim \mathcal{N}(X_i, b^2)$ un

$$X_{i+1} = Y \quad \text{ar varbūtību } r(X_i, Y);$$

$$X_{i+1} = X_i \quad \text{ar varbūtību } 1 - r(X_i, Y).$$

Ja, ģenerējot Markova ķēdi, iegūto simulāciju sadalījums ātri sāk līdzināties f, saka, ka ķēde "labi jaucas" (*mixing well*).

Sniegsim nelielu ieskatu, kāpēc Metropolisa-Heistingsa algoritma Markova ķēdes sadalījums ir f.

Definīcija 3. Sadalījums π Markova ķēdei ir pamatīgi balansēts (*detailed balance*), ja

$$p_{ij}\pi_i = p_{ji}\pi_j.$$

Teorēma 2. *[12, 391. lpp.] Ja π ir pamatīgi balansēts, tad tas ir ķēdes stacionārais sadalījums.*

Ieviesīsim jaunus apzīmējumus. Ar $p(x, y)$ apzīmēsim varbūtību ķēdei pāriet no stāvokļa x uz y, ar $f(x)$ apzīmēsim π. f ir stacionārais sadalījums, ja $f(x) = \int f(y)p(y, x)dy$, un f ir pamatīgi balansēts, ja

$$f(x)p(x, y) = f(y)p(y, x).$$

Ja f ir pamatīgi balansēts, tad f ir ķēdes stacionārais sadalījums, jo

$$\int f(y)p(y, x)dy = \int f(x)p(x, y)dy = f(x) \int p(x, y)dy = f(x).$$

Apgalvojums 3. *[12, 414. lpp.] f ir pamatīgi balansēts.*

Pierādījums. Izvēlēsimies divus punktus x un y. Vai nu

$$f(x)q(y|x) < f(y)q(x|y) \qquad \text{vai arī} \qquad f(x)q(y|x) > f(y)q(x|y),$$

ekvivalence nepārtrauktiem sadalījumiem ir ar 0 varbūtību. Pieņemsim, ka $f(x)q(y|x) > f(y)q(x|y)$. Tad

$$r(x, y) = \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}$$

un $r(y, x) = 1$. $p(x, y)$ ir varbūtība ķēdei pāriet no x uz y. Lai tā notiktu jārealizējas diviem notikumiem: a) q jāģenerē y; b) y jātiek izvēlētam kā nākamajam ķēdes stāvoklim. Seko, ka

$$p(x, y) = q(y|x)r(x, y) = q(y|x)\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)} = \frac{f(y)}{f(x)}q(x|y).$$

Tātad

$$f(x)p(x, y) = f(y)q(x|y).$$

Līdzīgi $p(y, x)$ ir varbūtība ķēdei pāriet no y uz x; x jātiek ģenerētam un izvēlētam. Tas notiek ar varbūtību $p(y, x) = q(x|y)r(y, x) = q(x|y)$. Tādējādi

$$f(y)p(y, x) = f(y)q(x|y),$$

no kā seko, ka f ir pamatīgi balansēts. □

1.4. Robina-Ritova paradokss

Beijesieši uzsver viņu metožu universālu pielietojamību, taču pastāv daudzi paradoksi, kas parāda būtiskus trūkumus. Šajā nodaļā tiks aprakstīts viens no tiem, sīkāks izklāsts pieejams [15]. Daļa no piemērā izvērstās informācijas balstīta uz tīmeklī pieejamu neformālu diskusiju starp Leriju Vasermanu, Džeimsu Robinsu *(James Robins)* un Kristoferu Simsu *(Christopher Sims)* [16]. Simss, kas 2011. gadā kopā ar Tomasu Sārdžentu (*Thomas Sargent*) saņēma Nobela prēmiju ekonomikā, ir dedzīgs Beijesa metožu aizstāvis.

Pieņemsim, ka Pasaules Veselības organizācija, lai plānotu budžetu, vēlas noskaidrot, cik lielai kādas valsts iedzīvotāju daļai nākošgad būs insults. Pirms gada tikai izvēlēti 5000 iedzīvotāji $(X_1, ..., X_{5000})$, katram no tiem noskaidrojot 300 faktorus: augumu, vecumu, iepriekšējas saslimšanas, utt. $(X_i \in [0,1]^{300})$.

Daļa no 5000 tika atlasīti novērošanai ar Bernulli gadījuma lielumu R_i ($R_i = 1$: novērots; $R_i = 0$: nenovērots), turklāt varbūtība tikt novērota nav visiem vienāda, to nosaka zināma funkcija $\pi(X) \equiv \mathbb{E}[R|X]$. Gada laikā no novērotajiem daļai bija insults $(Y = 1)$, pārējiem nē $(Y = 0)$.

Definējam arī $\theta(X) \equiv \mathbb{E}[Y|X]$, kas nosaka insulta varbūtību atkarībā no veselības stāvokļa. Šī funkcija nav zināma. Turklāt $\pi(X)$ izvēlēta tā, ka $Cov\{\theta(X), \pi(X)\} \neq 0$.

Formālāk: doti n (5000) *iid* novērojumi

$$(X_1, R_1, RY_1), ..., (X_n, R_n, RY_n),$$

kas pieņem vērtības

$$X_i \in [0,1]^d, \qquad Y_i \in \{0,1\}, \qquad R_i \in \{0,1\} \qquad d = 300.$$

Ņemot vērā, ka

$$\theta(X) = \mathbb{E}[Y|X] \qquad \pi(X) = \mathbb{E}[R|X] \qquad Cov\{\theta(X), \pi(X)\} \neq 0,$$

mūsu mērķis ir novērtēt

$$\psi \equiv \mathbb{E}[Y] = \mathbb{E}\{\mathbb{E}[Y|X]\} = \mathbb{E}\{\mathbb{E}[Y|X, R = 1]\} = \int_{[0,1]^d} \theta(x)dx.$$

Ievērojam arī

$$\theta(x) = \mathbb{E}[Y|X = x] =$$

$$= 1 \cdot \mathbb{P}(Y = 1|X = x) + 0 \cdot \mathbb{P}(Y = 0|X = x) = \mathbb{P}(Y = 1|X = x),$$

$$\mathbb{P}(Y = 0|X = x) = 1 - \theta(x).$$

Līdzīgi

$$\pi(x) = \mathbb{P}(R = 1|X = x) \qquad 1 - \pi(x) = \mathbb{P}(R = 0|X = x).$$

Klasiskā pieeja problēmai ir vienkārša, izmanto Horvica-Tompsona novērtējumu

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i)}.$$

$\hat{\psi}$ ir nenovirzīts un būtisks, taču tam ir trīs defekti: 1) vērtība var pārsniegt 1; 2) netiek ņemta vērā informācija vektora X dimensijās, 3) tas ir neefektīvs (*inefficient*). Neefektīvs novērtējums nozīmē, ka tas nav labākais iespējamais savā klasē.

Novērtējumu var aizstāt ar lokāli semiparametrisko efektīvo regresijas novērtējumu *(locally semiparametric efficient regression estimator)* [17].

$$\hat{\psi} = \int \text{expit} \left(\sum_{m=1}^{k} \hat{\eta}_m \phi_m(x) + \frac{\hat{\omega}}{\pi(x)} \right) dx,$$

kur $\text{expit}(a) = \frac{e^a}{(1+e^a)}$,$\phi_m(x)$ ir bāzes funkcijas, bet $\hat{\eta}_1, ..., \hat{\eta}_k, \hat{\omega}$ ir maksimālās ticamības novērtējumi modelī

$$\log \left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} \right) = \sum_{m=1}^{k} \eta_m \phi_m(x) + \frac{\omega}{\pi(x)}.$$

Atgriežoties pie paradoksa, beijesiešu pieeja ir sekojoša.

Izvēlas aprioro sadalījumu $W(\theta)$.

Ticamības funkcija vienam novērojumam (x, r, ry) ir formā

$$f_X(x) f_{R|X}(r|x) f_{Y|X}(y|x)^r$$

Savukārt n novērojumu $(X_i, R_i, R_i Y_i)$ ticamības funkcija ir

$$\prod_{i=1}^{n} f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} =$$

$$= \prod_i \pi(X_i)^{R_i} (1 - \pi(X_i))^{1-R_i} \theta(X_i)^{Y_i R_i} (1 - \theta(X_i))^{(1-Y_i)R_i}.$$

Taču tā kā ticamības funkcija tiek noteikta ar precizitāti līdz konstantei, tad tās forma saīsinās līdz

$$\mathcal{L}(\theta) \propto \prod_i \theta(X_i)^{Y_i R_i} (1 - \theta(X_i))^{(1-Y_i)R_i}.$$

Kombinējot $W(\theta)$ un $\mathcal{L}(\theta)$, iegūst aposterioro sadalījumu W_n. Tā kā ψ ir funkcija no θ ($\psi = \int_{[0,1]^d} \theta(x)dx$), θ aposteriorais sadalījums noteiks ψ vērtību. Taču ir pierādīta sekojoša teorēma.

Teorēma 4. *[15] Jebkurš novērtējums, kas nav funkcija no $\pi(\cdot)$, nevar būt vienmērīgi būtisks.*

Vispārīgi runājot, tas nozīmē, ka aposteriorais sadalījums nekoncentrēsies ap patieso ψ vērtību.

Viens no beijesiešu piedāvātajiem risinājumiem ir aprioro sadalījumu $W(\theta)$ padarīt atkarīgu no π, tad teorēmas nosacījums neizpildīsies.

Varam šādu apsvērumu ieviest sākotnējā modelī. Pieņemsim, ka ekspertam, pasaules līmeņa kardiologam ir izveidojies priekšstats par θ, un viņš vēlas to izmantot, analizējot datus ar Beijesa metodi. Lai arī PVO var paskaidrot, kādi apsvērumi par θ tika ņemti vērā izvēloties π, kardiologam šī informācija nemainīs uzskatu par θ, jo ir visaugstākās klases speciālists insulta jautājumos. Šādā gadījumā θ apriorais sadalījums būs neatkarīgs no π, un beijesiešu risinājums kļūst nederīgs.

Turklāt $W(\theta)$ atkarība no π ir nepieciešams, bet ne pietiekams nosacījums, apriorā sadalījuma funkcija ir rūpīgi jāizvēlas, visticamāk vadoties pēc frekventistu apsvērumiem.

Cits piedāvātais risinājums ir padarīt $\theta(x)$ gludu (*smooth*) jeb vairākas reizes diferencējamu funkciju, tādējādi katrs datu punkts sniegs informāciju arī par tā apkārtni. Taču tas līdz tikai tad, ja dimensiju skaits ir neliels, bet, ja $d >> n$, arī šis risinājums nederēs.

1.5. Neparametriskā Beijesa statistika

Līdzīgi kā neparametriskā statistika papildina klasiskās metodes, neparametriskā Beijesa procedūras novērš parametrisko metožu trūkumus. Parametriskās Beijesa metodes uzliek pārāk stingrus nosacījumus datus ģenerējošajiem mehānismiem, bet vāji pamatoti nosacījumi var radīt novirzes novērtējumos. Neparametriskās Beijesa metodes piedāvā plašākas iespējas aprioro sadalījumu konstrukcijā, jo spēj modelēt arī gadījumu, kad parametru ir bezgalīgi daudz [3].

Piemēram, klasiskajā, lineārajā Gausa-Markova modelī

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

pieņem, ka kļūdas ε_i ir nekorelēti gadījuma lielumi ar konstantu dispersiju un vidējo vērtību 0. Ar parametriskā Beijesa metodēm nevar neko iesākt, kamēr nav konstruēta ticamības funkcija, savukārt neparametriķis var izvēlēties aprioro sadalījumu kā sadalījumu pa visām blīvuma funkcijām ar vidējo vērtību 0.

Jebkurš modelis, kurš nav pilnībā uzdots, var tikt papildināts ar bezgalīgi dimensionāliem parametriem, kuriem piešķir sadalījumu, tādējādi iegūstot pilnībā uzdotu modeli.

Jebkura neparametriskā Beijesa metode galvenokārt balstās uz vienu no trijiem sadalījumu funkciju ģenerējošajiem mehānismiem: Dirihlē procesa, Polijas koka (*Polya tree*) un Gausa procesa. Pirmo aprakstīsim nākošajā apakšnodaļā, bet pēdējam tiks veltīta visa trešā nodaļa.

1.6. Dirihlē process kā apriorais sadalījums

Beijesa neparametriskajā statistikā viena no lielākajām problēmām ir apriorā sadalījuma izvēle, jo jāizvēlas piemērots varbūtību mērs sadalījuma funkciju telpā. Informācija par šādām telpām pārsvarā nav pieejama. Aprioro sadalījumu izvēlas balstoties uz sarežģītību, pieejamajiem datorresursiem un algoritmiem, kā arī labām frekventistu statistikas īpašībām. Nepieciešams arī, lai sadalījums "noklātu" parametru telpu jeb tam ir pietiekami liels topoloģiskais atbalsts *(support)*. Lai pētītu frekventistu tipa īpašības, tiek pieņemts, ka eksistē noteikta parametra (hiperparametra) patiesā vērtība, kas nosaka ģenerēto datu sadalījumu.

Mēs vēlamies novērtēt blīvuma funkciju, kura var būt jebkādā formā, uz reālās skaitļu ass, ja doti *iid* dati. Klasiskajā statistikā šo problēmu risina ar empīrisko sadalījuma funkciju, taču beijesiešiem jāapraksta varbūtību mērs, kas nav viegls uzdevums un jāattīsta simulāciju algoritmi. Klasiskajās Beijesa metodēs populārs ir Dirihlē apriorais sadalījums, kas neparametriskajā statistikā kalpo par pamatu bezgalīgi dimensionālai blīvuma funkcijai.

Definīcija 4. Parametri $\Theta \equiv \{\theta_1, \theta_2, ..., \theta_m\}$ ir sadalīti pēc Dirihlē sadalījuma ($\Theta \sim$

Dirichlet$(\alpha_1, \alpha_2, ..., \alpha_m)$), ja

$$f(\theta_1, \theta_2, ..., \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1},$$

kur $\Gamma(n) = (n-1)!$ ir gamma funkcija.

Dirihlē apriorajam sadalījumam piemīt noderīga īpašība.

Teorēma 5. *[2, 212. lpp.] Ja* $\Theta \sim$ *Dirichlet*$(\alpha_1, \alpha_2, ..., \alpha_m)$ *un ja* $\mathbb{P}(X = j | \alpha_1, \alpha_2, ..., \alpha_m) = \alpha_j$ *gandrīz droši* $j = 1, ..., k$, *tad* $\alpha_1, \alpha_2, ..., \alpha_m$ *aposteriorais sadalījums, ja* $X = j$ *ir* *Dirichlet*$(\alpha_1^{(j)}, \alpha_2^{(j)}, ..., \alpha_m^{(j)})$, *kur*

$$\alpha_i^{(j)} = \alpha_i \qquad i \neq j,$$

$$\alpha_i^{(j)} = \alpha_i + 1 \qquad i = j.$$

Fergusona ideja bija, ka jebkuri dati, patvaļīgi sadalīti pa grupām, veido Dirihlē sadalījumu, piemēram, ja ir 3 grupas un 7 dati $\boldsymbol{X}$, kas sagrupēti $\{4, 2, 1\}$ (4 dati pirmajā grupā, utt.), tad to varam interpretēt kā Dirichlet$(4, 2, 1)$. Pieņemsim, ka tiek pārkārtoti dati un izveidojas $\{3, 3, 1\}$. Tad mums ir jauns sadalījums Dirichlet$(3, 3, 1)$. Ja pieļaujam iespēju, ka Dirihlē parametru skaits varētu būt bezgalīgs, tad katru reizi datus pārkārtojot tieks iegūts jauns sadalījums. Ja pārkārtošanu atkārtos bezgalīgi ilgi, tad to var saukt par Dirihlē procesu.

Aprakstīsim algoritmu, ar ko iespējams simulēt no Dirihlē procesa apriorā sadalījuma.

"Lauztās nūjas" (stick-breaking) metode.

Dž. Seturamans (*J. Sethuraman*) [18] piedāvā šādu Dirihlē procesa ($\mathcal{DP}$) apriorā sadalījuma praktisku konstrukcijas metodi nezināmam sadalījumam P:

I Izvēlas "bāzes sadalījumu" P_0, parasti $\mathcal{N}(0, 1)$.

II Nosaka intensitāti α, parasti $\alpha = 1$.

III Simulē $V_i \sim \text{Beta}(1, \alpha)$, $\theta_h \sim P_0$, $i = 1, ..., \infty$.

IV Aprēķina $\pi_h = V_h \prod_{l<h}(1 - V_l)$.

V $P \sim \mathcal{DP}(\alpha P_0)$ ekvivalents

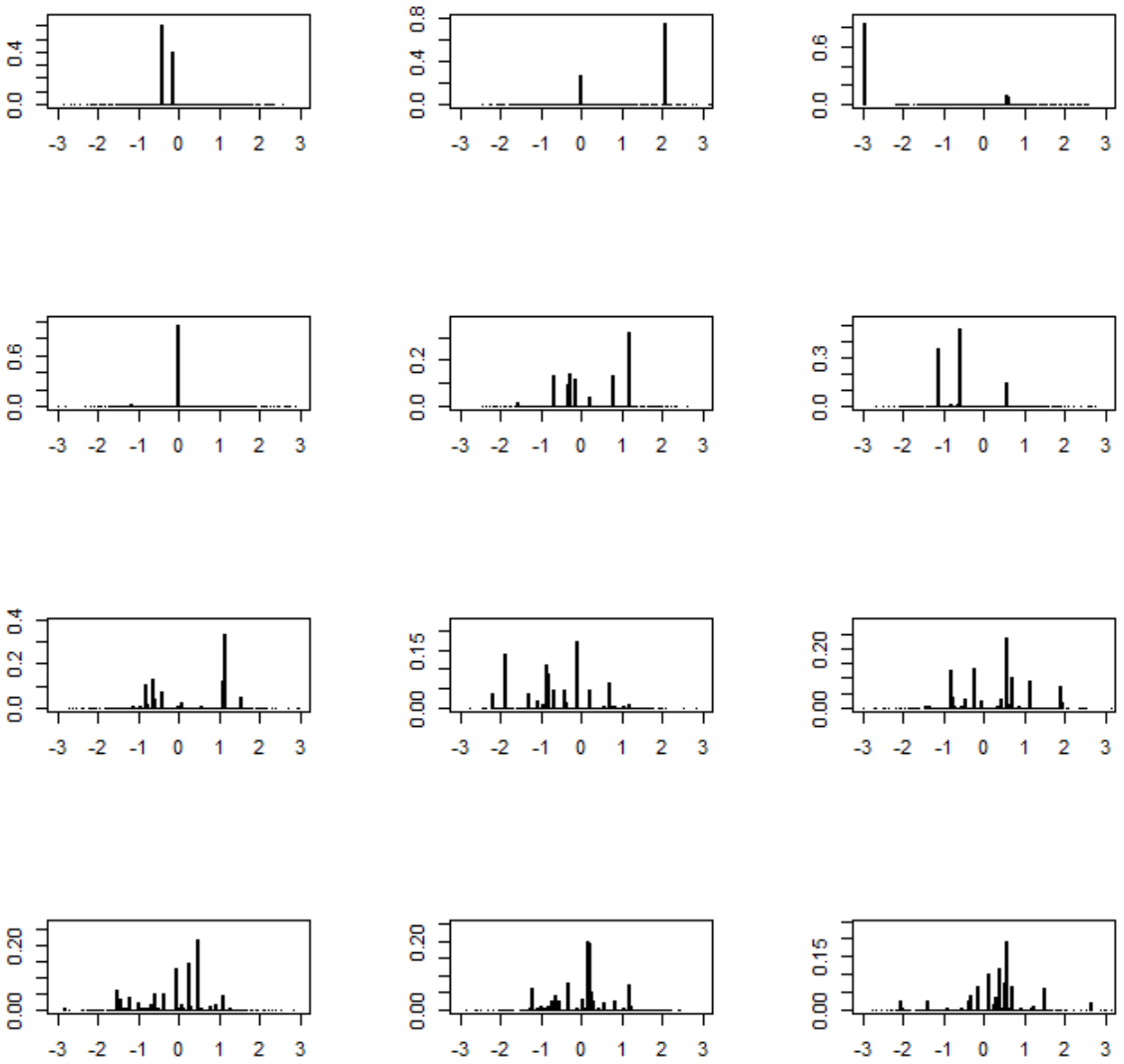
$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h},$$

kur δ_θ ir punkta θ varbūtības masa.

Simulācijas no $\mathcal{DP}(\alpha P_0)$ ar $P_0 = \mathcal{N}(0,1)$ var apskatīt 1.1. attēlā.

Jo lielāka intensitāte, jo simulētie sadalījumi būs līdzīgāki P_0. Ja intensitāte ir zema, gandrīz visa sadalījuma masa būs uz dažiem punktiem jeb atomiem.

Vizuāli varam to iztēloties, kā nūjas ar garumu 1 laušanu. Pirmajā solī ģenerējam lielumu v_1 no $V_1 \sim \text{Beta}(1,\alpha)$ un izdarām atzīmi v_1 ($0 \leq v_i \leq 1$) attālumā no nūjas gala un tur to pārlaužam. Nolauztā daļa būs jaunā sadalījuma P varbūtības masa punktā θ_1 jeb $\mathbb{P}(X = \theta_1) = v_1$, $X \sim P$. Otrajā solī jaunais v_2 no $V_2 \sim \text{Beta}(1,\alpha)$ būs jāreizina ar atlikušo nūjas garumu, lai noteiktu jauno laušanas vietu. Tad atkal piešķir nolauzto masu punktā θ_2. Teorētiski tā turpina bezgalīgi ilgi, bet praktiski pēc 50 soļiem nūja ir pārāk īsa, lai tālākie soļi ko ietekmētu. Nodaļas izklāsts balstīts uz [19].



1.1. att.: Simulēts Dirihlē procesa sadalījums ar intensitāti $\alpha \in \{0,5; 1; 5; 10\}$. Ar katru intensitāti attēlotas trīs sadalījuma realizācijas.

2. Maiņas punktu noteikšana ar Beijesa tipa metodi

2.1. Berija-Hartigana metode

Maiņas punktu noteikšanas problemātika statistikā galvenokārt parādās divos veidos. Pirmā ir situācija, kad pirms padziļinātas apkopoto datu analīzes nepieciešams pārbaudīt pieņēmumu, ka datu ģenerējošais process ir viens un tas pats visiem mērījumiem. Ja pieņēmums tiek noraidīts, datu kopa jāsadala vairākās atsevišķi pētāmās daļās.

Otra situācija ir, kad novērojumi pienāk sekvenciāli, viens pēc otra, un katrs jauns mērījums jāsalīdzina ar iepriekšējiem, lai noteiktu vai nav notikušas izmaiņas. Šāda problemātika sastopama, piemēram, kvalitātes kontroles sistēmās. Darbā tiks apskatītas metodes pirmā veida situācijās.

Klasiskās, frekventistu statistikas pieejā tiek atrasti specifiski punkti, kuros novērojamas datu struktūras izmaiņas. Piemēram, ja pieņemts, ka ir m maiņas punkti, Bai un Perona (*J. Bai*, *P. Perron*) dinamiskās programmēšanas metode [20] aprēķina ticamākās atrašanās vietas, minimizējot atlikumu kvadrātu summu katrā datu kopas dalījumā, pēc tam labākais maiņas punktu skaits tiek atrasts, izmantojot BIC kritēriju.

Beijesa metodes, tai skaitā arī Berija-Hartigana (*D. Barry and J. A. Hartigan*) [21] (BH) pieeja, piešķir varbūtību sadalījumu, tas ir, maiņas punkts ar aprēķināmu varbūtību iespējams jebkur datu kopā, atstājot pētnieka ziņā lēmumu, vai varbūtība ir pietiekoši liela.

Formālāk, pieņemsim, ka doti neatkarīgu gadījuma lielumu virkne $\boldsymbol{X} \equiv X_1, X_2, ..., X_n$, $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$. Varbūtību, ka punktā i ir maiņas punkts, apzīmēsim ar p_i. Pieņemsim arī, ka maiņas punkti virkni sadala b blokos $i+1, ..., j$ (ērtībai apzīmēsim ar ij), un piešķirsim katra bloka, kas sākas ar indeksu $i+1$ un beidzas ar j, vidējai vērtībai μ_{ij}

aprioro sadalījumu $\mathcal{N}(\mu_0, \sigma_0^2/(j-i))$, tādējādi īsu bloku vidējai vērtībai tiks pieļauta salīdzinoši liela dispersija, un ar metodi varēs atrast arī tādus maiņas punktus, starp kuriem ir maz datu.

Erdmans un Emersons [22] savā `R` bibliotēkā `bcp`, lai pielietotu BH metodi, pielieto sekojošu Markova ķēžu Montekarlo algoritmu.

Definē partīciju (*partition*) $\rho = (U_1, U_2, ..., U_n)$, $U_i = 1$ nozīmē, ka punktā $i+1$ ir maiņas punkts. Partīcija tiek inicializēta ar vērtībām $U_i = 0, U_n = 1$. Katrā ķēdes solī katram i, aprēķina b, bloku skaitu pie pašreizējās partīcijas ar pieņēmumu, ka $U_i = 0$. Tad aprēķina maiņas punkta varbūtību p_i no

$$\frac{p_i}{1-p_i} = \frac{\mathbb{P}(U_i = 1 | \boldsymbol{X}, U_j, j \neq i)}{\mathbb{P}(U_i = 0 | \boldsymbol{X}, U_j, j \neq i)} \tag{2.1}$$

$$= \frac{[\int_0^\gamma p^b (1-p)^{n-b-1} dp] \left[\int_0^\lambda \frac{w^{b/2}}{(W_1 + B_1 w)^{(n-1)/2}} dw\right]}{[\int_0^\gamma p^{b-1} (1-p)^{n-b} dp] \left[\int_0^\lambda \frac{w^{(b-1)/2}}{(W_0 + B_0 w)^{(n-1)/2}} dw\right]}. \tag{2.2}$$

Kad iegūta varbūtība, ģenerē jauno U_i un pāriet pie nākamā virknes punkta. Kad visas virknes maiņas punktu varbūtības p_i ir noskaidrotas, aprēķina μ_{ij} pie pašreizējās partīcijas, tad procedūru atkārto sākot no ρ pirmā elementa. Pēc pietiekoši liela soļu skaita iegūst simulācijas no diviem sadalījumiem, aposteriorā maiņas punktu varbūtību, kā arī vidējo vērtību sadalījumu, no kuriem iegūst ticamākos novērtējumus.

Apskatīsim tuvāk (2.2) vienādojuma skaitītāju (analoģiski var interpretēt arī saucēju). $p^b(1-p)^{n-b-1}$ ir varbūtība, ka starp $n-1$ punktiem, kas ir zināmi, ir b maiņas punkti, integrālis tiek ņemts pa visām iespējamajām p vērtībām. Simulācijās ir noskaidrots, ka ar $\gamma = 0.2$ metode strādā vislabāk, tas ir, tiek pieņemts, ka nav vairāk par 1 maiņas punktu uz 5 novērojumiem.

$\frac{w^{b/2}}{(W_1 + B_1 w)^{(n-1)/2}}$, savukārt ir $\boldsymbol{X}$ ticamības funkcija, ja partīcija ir zināma, kur

$$W_1 = \sum_{ij \in \rho} \sum_{l=i+1}^{j} (X_l - \bar{X}_{ij})^2$$

jeb kopējā kvadrātu summa blokos un

$$B_1 = \sum_{ij \in \rho} (j-i)(\bar{X}_{ij} - \bar{X})^2$$

jeb kopējā kvadrātu summa pa blokiem. Parametrs $w = \frac{\sigma^2}{(\sigma_0^2 + \sigma^2)}$ parasti tiek izvēlēts 0.2, balstoties uz pieņēmumu, ka izmaiņas virknē, kad tās notiek, nav niecīgas. (2.2) specifiskā forma izriet no pieņēmuma par $\boldsymbol{X}$ normalitāti un vienmērīgi sadalītiem μ_0, σ^2 un w.

BH autori norāda, ka viņu metode varētu tikt pielāgota arī atkarīgiem datiem: "...pieņēmums par neatkarību varētu tikt vājināts, jo viss, kas nepieciešams, ir, lai dati dažādos blokos ir savstarpēji neatkarīgi, ja parametri un partīcija ir doti"[21, 310. lpp.]. Erdmans un Emersons diemžēl šādu iespēju neapskata.

2.2. Salīdzinājums ar EL metodi

Lai noskaidrotu BH metodes priekšrocības un trūkumus, salīdzināsim to ar EL metodi, ko savā darbā aprakstījis A. Vaselāns [10]. Tā ir klasiskās statistikas neparametriska metode, kas strādā arī tad, ja dati ir vāji atkarīgi. Ja doti dati $X_1, X_2, ..., X_n$, no tiem tiek atlasītas 2 apakšizlases ar garumu N, $\boldsymbol{X}_1 \equiv X_i, X_{i+1}, ...X_{i+N-1}$, $\boldsymbol{X}_2 \equiv X_{j+1}, X_{j+2}, ...X_{j+N-1}$, $i \in \{1, 2, ..., n-2N\}$, $j \in \{N+1, N+2, ..., n-N\}$, un tad tiek veikta hipotēžu pārbaude

$$H_0 : \mu_2 - \mu_1 = 0$$

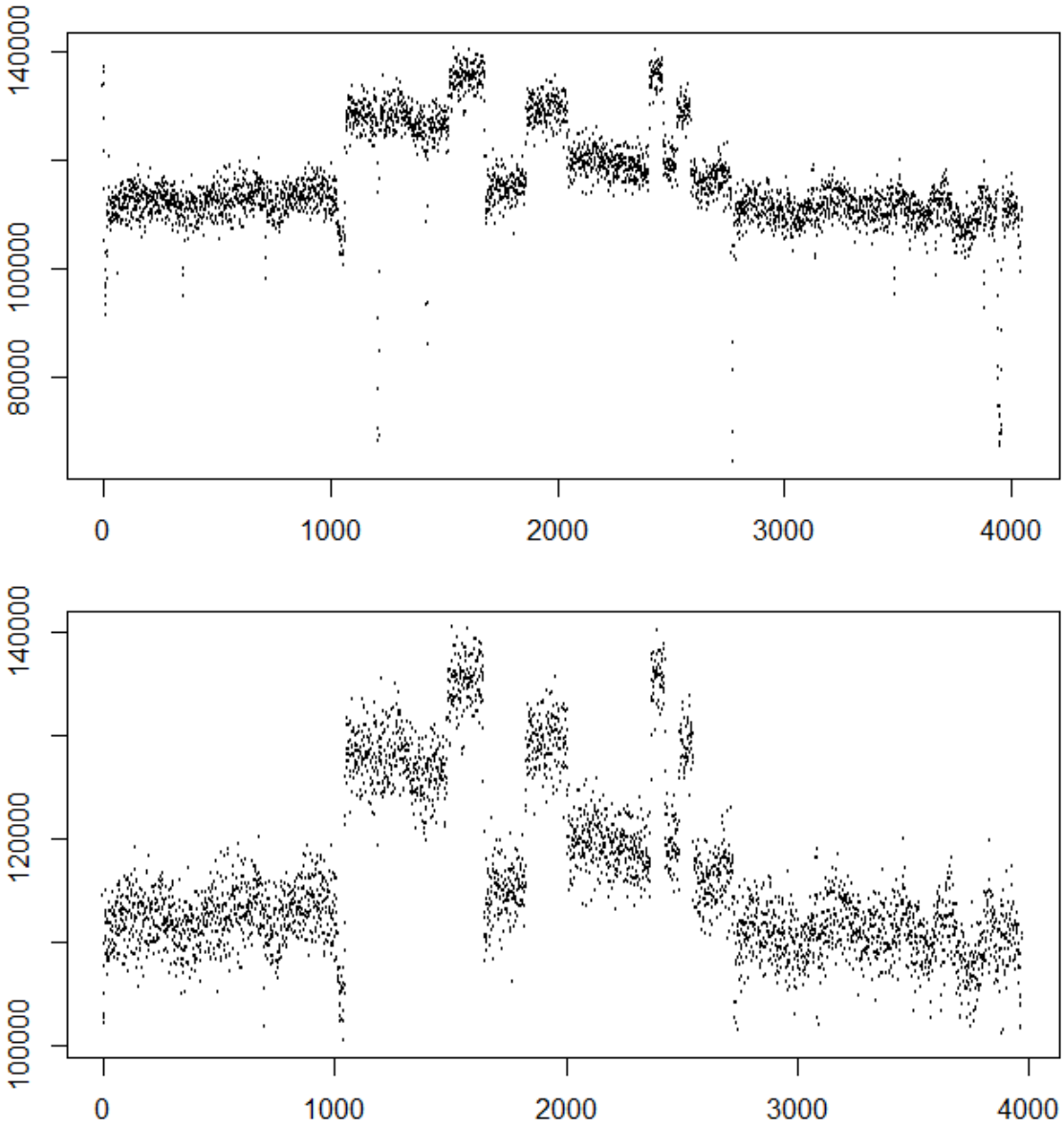
pret

$$H_1 : \mu_2 - \mu_1 \neq 0,$$

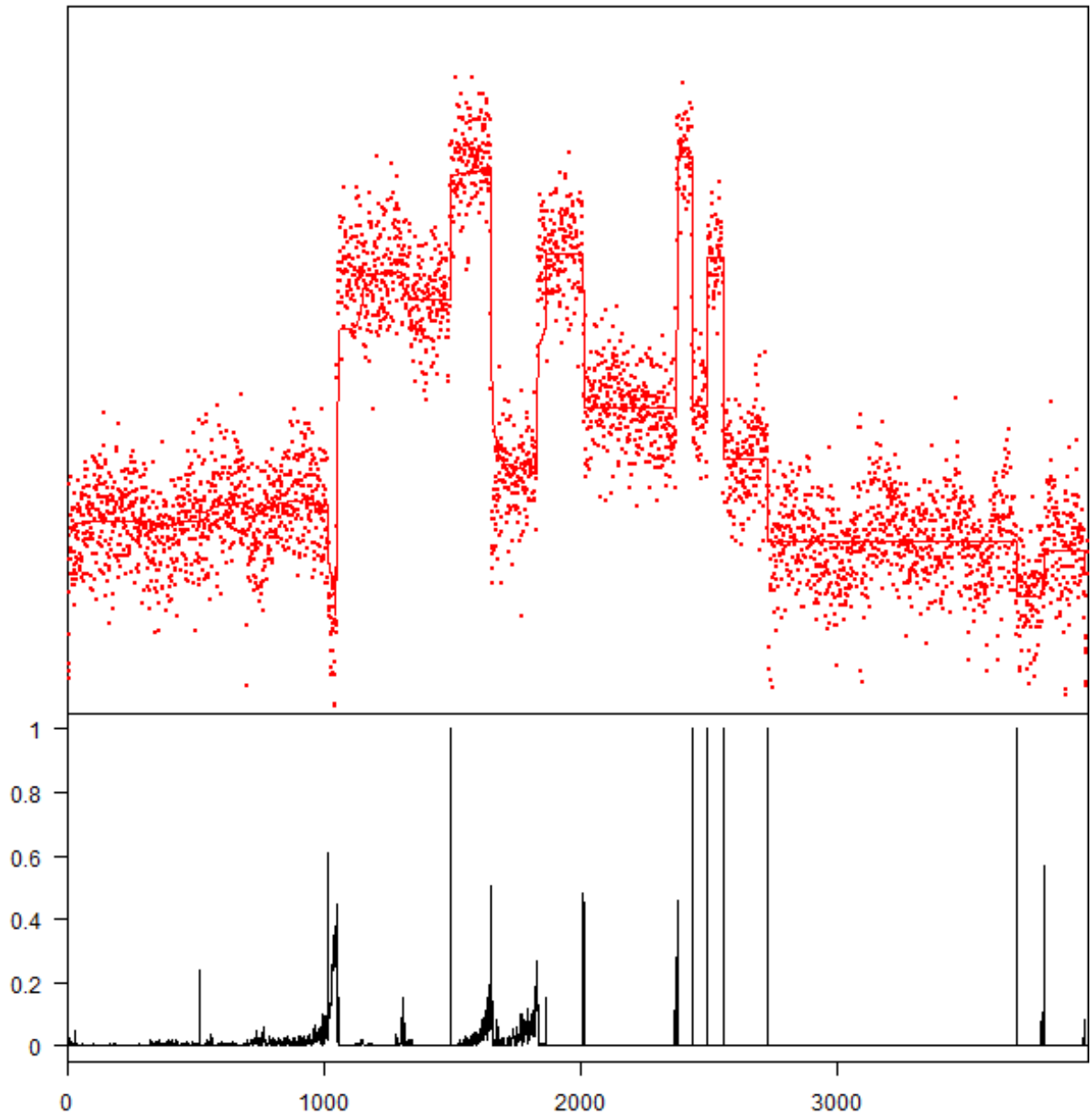
kur μ_1 un μ_2 ir attiecīgo apakšizlašu vidējās vērtības, izmantojot empīriskās ticamības vidējo vērtību testu no `EL` bibliotēkas [23]. Hipotēzi pārbauda visām iespējamajām izlasēm, un iegūtās p-vērtības kalpo kā indikators, vai punkts starp apakšizlasēm ir maiņas punkts. Datu atkarības izraisītie efekti tiek novērsti, apakšizlasēs dalot datus blokos. Bloki var daļēji pārklāties, vai nepārklāties nemaz, mūsu pielietojumā tie nekad nepārklāsies.

Apskatīsim "well-log" datus 2.1. attēlā no [24] 5. nodaļas, kas zinātniskajā literatūrā ir bieži apskatīti. Tie ir mērījumi, kas ņemti naftas ieguves vietu urbšanas laikā un raksturo zemes slāņa, kurā atrodas urbja gals, fizikālās īpašības. Ja novērojamas izmaiņas, tiek uzskatīts, ka slāņa tips ir mainījies. Dati ir ar izlēcējiem, taču precīzākai analīzei tos izņemsim, līdzīgi kā tika darīts Dž. Liu (*J. Liu*) doktora disertācijā [25, 137. lpp]. Tiek uzskatīts, ka mērījumi ir atkarīgi un ka maiņas punktu skaits ir 13.

Kā redzam 2.2. attēlā, BH metode spēj atrast gandrīz visus maiņas punktus, ja pieņemam, ka aposteriorā varbūtība ir pietiekoši liela (izvēlēsimies to lielāku par 0.4). Tik labus rezultātus gan nevar iegūt ar rekomendētajām vērtībām $\gamma = 0.2$ un $\lambda = 0.2$, bet gan ar $\gamma = \lambda = 10^{-5}$. Tas ir, apriori jāpieņem gan maiņas punktu retums, gan arī augsta bloka vidējās vērtības dispersija.

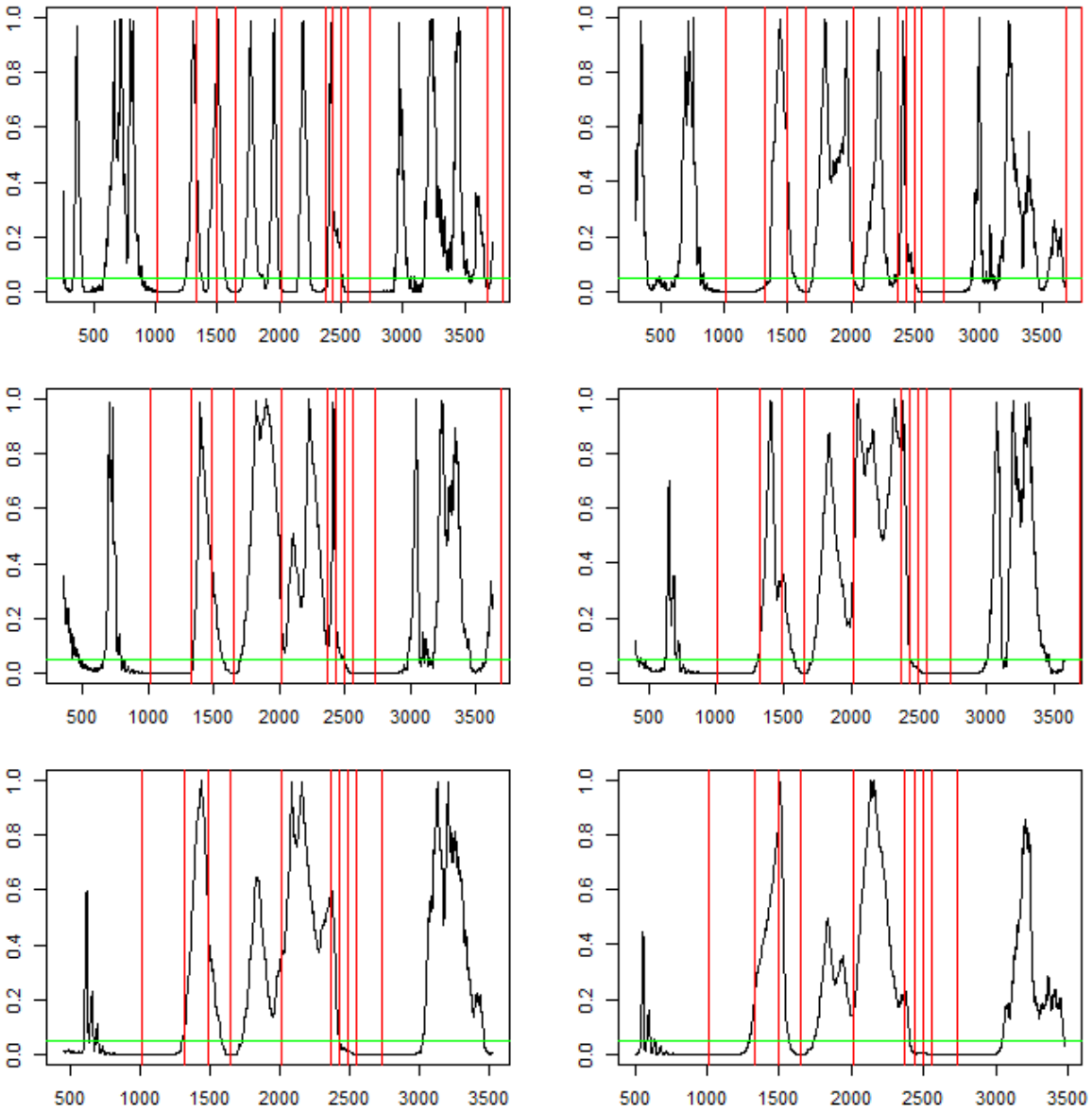


2.1. att.: 'well-log' dati ar izlēcējiem (augšā) un bez tiem (apakšā).



2.2. att.: "well-log" datu izpēte ar BH metodi, attēla augšdaļā dati un to ticamākā aposteriorās vidējās vērtība, zem tā ticamākās maiņaspunktu atrašanās vietas un to varbūtība. $\gamma = \lambda = 10^{-5}$.

EL metode savukārt uzrāda sliktākus rezultātus, skat. 2.3. attēlu. Ar visiem logu garumiem $N \in \{250, 300, 350, 400, 450, 500\}$ tiek fiksēti maiņas punkti tur, kur tie nav, un otrādi. Turklāt pie $N = 500$ tiek "pazaudētas" aptuveni ceturtā daļa sākotnējo vērtību.



2.3. att.: "well-log" datu izpēte ar EL metodi pie 6 dažādiem logu garumiem N. Sākot no augšējā, kreisā stūra, attēlotas testa p vērtības pie $N \in \{250, 300, 350, 400, 450, 500\}$. Ar sarkanām līnijām iezīmēti BH metodes atrastie maiņas punkti ar aposterioro varbūtību $p_i > 0.4$. Ar zaļu līniju attēlota EL testu kritiskā varbūtība 0.05.

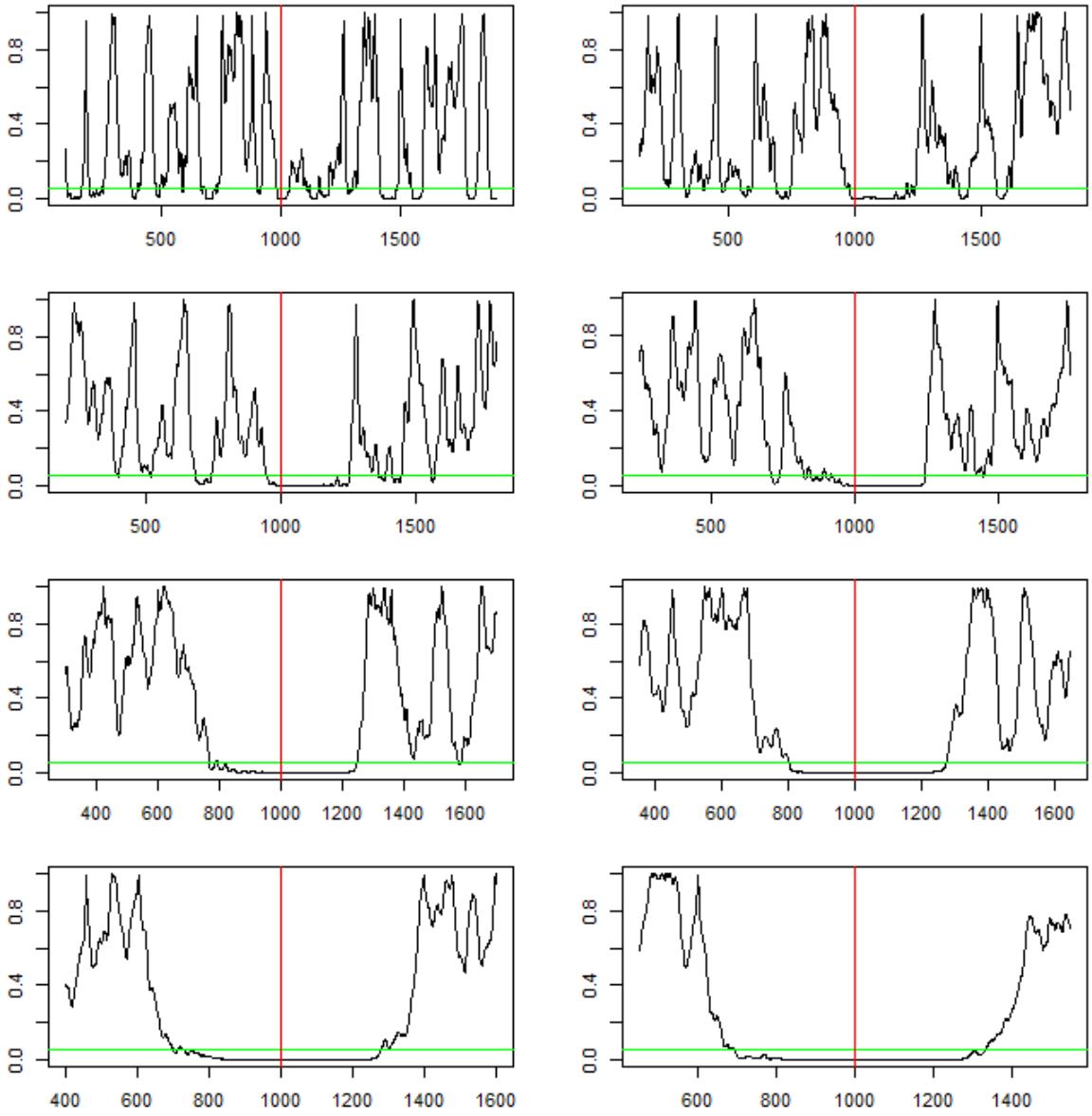
Taču iespējams arī konstruēt piemēru, kurā EL metode strādā labāk nekā BH. Ģenerēsim divus autoregresīvus AR(1; 0.9) procesus ($n = 1000$) ar vidējām vērtībām 0 un 5 un ievietosim tos vienā laikrindā vienu aiz otra. Kā varam novērot 2.4. attēlā, ar EL metode

maiņas punktu var atrast precīzi pie dažādiem logu garumiem. Logu garumam pieaugot, kritiskās p vērtības "centrējas" uz patieso punktu vidū. Varam secināt, ka, izmantojot EL metodi, ļoti ieteicams apskatīt dažādus logu garumus.

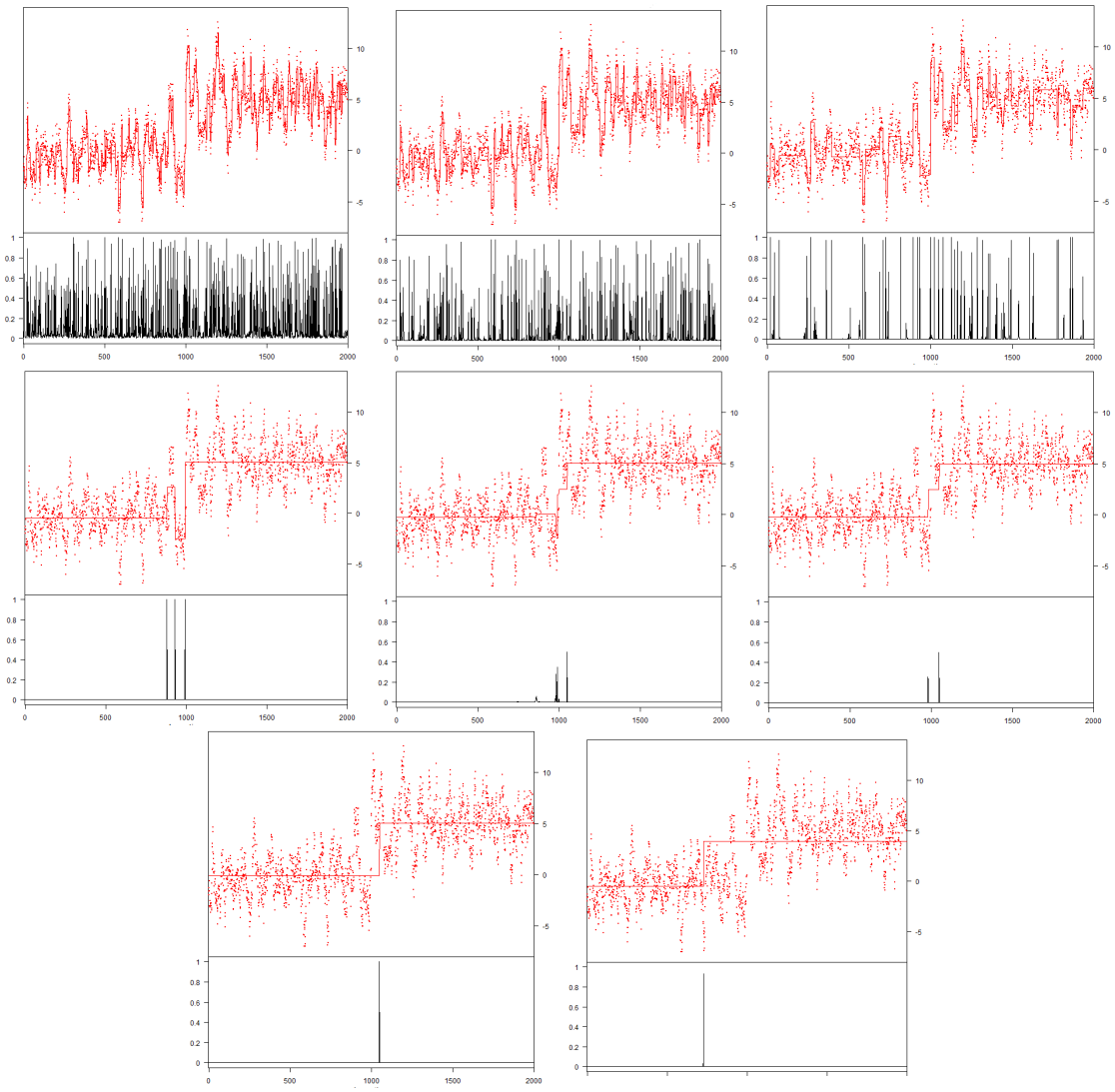
Savukārt 2.5. attēlā varam redzēt, kā parametri γ un λ spēcīgi ietekmē metodes veikumu. Varam uzskatīt, ka maiņas punkts tiek atrasts pie $\lambda = \gamma = 10^{-64}$, taču, nezinot patiesību, patieso atrašanās vietu noteikt būtu sarežģīti, un to lielā mērā ietekmētu pētnieka personiskas vērtējums. Šeit ļoti labi parādās Beijesa metožu subjektivitāte, par ko tās tiek kritizētas.

Salīdzinot abas metodes, varam novērot optimizācijas teorijā populārā principa "Brīvpusdienu nav"(*No Free Lunch*) patiesumu [26]. Vispārīgi runājot, teorēma apgalvo, ka, viena metode var būt labāka pār otru tikai tad, ja tā ir specifiski piemērota pētāmajai problēmai. Ar BH var atrast maiņas punktus akas dziļurbuma datos, bet atkarīgi dati ir pārāk svārstīgi, pat pielāgojot parametrus. Savukārt ar EL nav iespējams atpazīt maiņas punktus, starp kuriem ir pārāk maz datu. Neviena no metodēm nav izteikti pārāka.

Pastāv vairākas neparametriskās Beijesa maiņas punktu noteikšans metodes, kas iespējams varētu uzrādīt labākus rezultātus. A. Mira un S. Petrone (*A. Mira*, *S. Petrone*) apskata punktu meklēšanu ar Dirihlē procesu mikstūru (*mixture*) palīdzību [27], kas strādā arī ar atkarīgiem datiem. R. Gārnets, M. Osborns un S. Robertss (*R. Garnett*, *M. A. Osbourne*, *S. J. Roberts*) aplūko maiņaspunktu problemātiku, izmantojot Gausa procesus [28]. Viņu metode strādā ar atkarīgiem datiem, turklāt spēj arī izmaiņas prognozēt. Diemžēl neviena no šīm metodēm pašlaik nav publiski pieejama pielietojamā veidā, un šeit varētu būt plašs lauks tālākam darbam.



2.4. att.: Ģenerēti divi AR(1; 0.9) procesi ar vidējām vērtībām 0 un 5, tie apstrādāti ar EL metodi, $N \in \{100, 150, 200, 250, 300, 350, 400, 450\}$, bloki izlasēs nepārklājas. Ar sarkanu līniju patiesā maiņas punkta atrašanā vieta, ar zaļu testu kritiskā vērtība 0.05.



2.5. att.: Ģenerēti divi AR(1; 0.9) procesi ar vidējām vērtībām 0 un 5, tie apstrādāti ar BH metodi ar parametriem $\lambda = \gamma \in \{0.2, 10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}, 10^{-32}, 10^{-64}, 10^{-128}\}$. Maiņas punkta noteikšana šķiet lielā mērā subjektīva.

3. Gausa process kā apriorais sadalījums funkciju telpā

3.1. Lineārā regresija ar bāzes funkcijām

Pieņemsim, ka doti dati $\mathcal{D}$, kas sastāv no pāriem $(t_1, \boldsymbol{x}_1), (t_2, \boldsymbol{x}_2), ..., (t_n, \boldsymbol{x}_n)$ [1], kur $\boldsymbol{x}_i$ ir d-dimensionāli. Mūsu mērķis ir atrast tādu funkciju $y(\boldsymbol{x})$, ka

$$t_i = y(\boldsymbol{x}_i) + \varepsilon_i,$$

kur ε_i ir gadījuma lielumi ar dispersiju σ_v^2 un vidējo vērtību 0. Šīs nodaļas teorētiskais materiāls, galvenokārt, balstīts uz [29] un [30].

Viena no vienkāršākajām metodēm ir piešķirt svarus katrai $\boldsymbol{x}$ komponentei, proti,

$$t_i = \sum_{j=1}^{d} w_j x_j + \varepsilon_i.$$

Klasiskajā statistikā labākos w_j mēs varētu atrast izmantojot, piemēram, mazāko kvadrātu metodi, taču no Beijesu metožu skatpunkta jāņem vērā, ka $\boldsymbol{w} \equiv \{w_1, w_2, ..., w_d\}$ ir gadījuma lielums ar aprioro sadalījumu.

$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ datus izskaidros labi tikai tad, ja pareizs būs pieņēmums par linearitāti. Taču iespējams, ka, piemēram, polinomiālā telpā ar bāzes funkcijām $\{1, x, x^2, ...\}$, t_i varētu izteikt precīzāk. Tāpēc apskatīsim vispārīgu Beijesa tipa lineāru regresiju, tas ir, lineāru regresiju ar m bāzes funkcijām $\{\phi_i(\boldsymbol{x})\}$. Tad regresijas funkcija būs formā $y(\boldsymbol{x}) = \sum_{i=1}^{m} w_i\phi_i(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$. Ja mēs $\boldsymbol{w}$ aprioro sadalījumu izvēlamies kā daudzdimensionālu normālo, tad meklējamo regresijas funkciju var iegūt divos ekvivalentos veidos:

[1] Literatūrā parasti datu apzīmēšanai izmanto y un x, taču šeit mēs izmantojam t (no *target*), lai uzsvērtu, ka y ir funkcija no $\boldsymbol{x}$.

vai nu izmantojot svaru $\boldsymbol{w}$ īpašības, vai arī aplūkojot modeli kā Gausa procesu. Problēmas ilustrēšanai noderīgas abas pieejas, tāpēc apskatīsim tās.

Pieņemot, ka $\boldsymbol{w}$ ir daudzdimensionāli normāli sadalīts $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_w)$, kur Σ_w ir kovariācijas matrica ar elementiem $\{\mathrm{Cov}[w_i, w_j]\}_{ij}$, tā blīvuma funkcija būs formā

$$f(\boldsymbol{w}) = \frac{1}{(2\pi)^{m/2}|\Sigma_w|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T\Sigma_w^{-1}\boldsymbol{w}\right).$$

Ja t_i ir ģenerēti ar meklējamo funkciju, pieskaitot Gausa troksni ar dispersiju σ_v^2, $\boldsymbol{w}$ ticamība ir

$$f(t_1, ..., t_n|\boldsymbol{w}) = \frac{1}{(2\pi\sigma_v^2)^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{(t_i - y(\boldsymbol{x}_i; \boldsymbol{w}))^2}{2\sigma_v^2}\right).$$

Tā kā gan apriorais sadalījums, gan ticamība ir normālā sadalījuma blīvuma funkcijas, tāds būs arī aposteriorais sadalījums.

Apgalvojums 6. *[31, 241. lpp.] Svaru $\boldsymbol{w}$ aposteriorā sadalījuma vidējā vērtība $\bar{\boldsymbol{w}}$ ir minimums kvadrātiskai formai*

$$E = \frac{1}{2\sigma_v^2}\sum_i (t_i - \boldsymbol{w}^T\phi(\boldsymbol{w}_i))^2 + \frac{1}{2}\boldsymbol{w}^T\Sigma_w^{-1}\boldsymbol{w}.$$

Ieviesīsim apzīmējumus $\beta = 1/\sigma_v^2$, $\boldsymbol{t} = (t_1, ..., t_n)$ un $n \times m$ matricu

$$\Phi = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_1) & \cdots & \phi_m(\boldsymbol{x}_1) \\ \phi_1(\boldsymbol{x}_2) & \phi_2(\boldsymbol{x}_2) & \cdots & \phi_m(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \phi_2(\boldsymbol{x}_n) & \cdots & \phi_m(\boldsymbol{x}_n) \end{pmatrix}.$$

Tad varam pārrakstīt kvadrātisko formu E kā

$$\frac{1}{2}\boldsymbol{w}^T(\Sigma_w^{-1} + \beta\Phi^T\Phi)\boldsymbol{w} - \beta\boldsymbol{w}^T\Phi^T\boldsymbol{t} + \frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t},$$

un $\bar{\boldsymbol{w}}$ ir atrisinājums vienādojumam

$$(\Sigma_w^{-1} + \beta\Phi^T\Phi)\bar{\boldsymbol{w}} = \beta\Phi^T\boldsymbol{t}.$$

Apzīmēsim $A = \Sigma_w^{-1} + \beta\Phi^T\Phi$. Tad $\bar{\boldsymbol{w}} = \beta\Phi^T\boldsymbol{t}$ un svaru $\boldsymbol{w}$ aposteriorā sadalījuma kovariācijas matrica ir A^{-1}.

Prognoze patvaļīgam $\boldsymbol{x}_*$ svaru telpā būs vidējā vērtība

$$\mu_{st}(\boldsymbol{x}_*) = \boldsymbol{\phi}(\boldsymbol{x}_*)^T \bar{\boldsymbol{w}} = \beta \boldsymbol{\phi}(\boldsymbol{x}_*)^T \Phi^T \boldsymbol{t} \tag{3.1}$$

ar dispersiju

$$\sigma_y^2(\boldsymbol{x}_*) = \boldsymbol{\phi}(\boldsymbol{x}_*)^T A^{-1} \boldsymbol{\phi}(\boldsymbol{x}_*). \tag{3.2}$$

Lai iegūtu prognozes dispersiju var $t(\boldsymbol{x}_*)$, jāpieskaita trokšņa dispersija σ_v^2.

3.2. Lineārā regresija funkciju telpā

Iepriekšējā apakšnodaļā apskatījām lineāro regresiju ar m bāzes funkcijām $\phi_i(\boldsymbol{x})$. Ieguvām arī vidējās vērtības un dispersijas vienādojumus patvaļīgam $\boldsymbol{x}_*$. Parādīsim, ka šos pašus vienādojumus varam iegūt arī funkciju telpā.

Definīcija 5. Gausa process ir tāda gadījuma lielumu kopa, ka, izvēloties jebkurus no tiem saskaitāmā skaitā, tie būs daudzdimensionāli normāli sadalīti.

Vienkāršu Gausa procesa piemēru varam iegūt no iepriekš aplūkotās vispārīgās regresijas $y(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{w}$ ar aprioro sadalījumu uz svariem $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_w)$. Aprēķināsim vidējo vērtību un kovariāciju,

$$\mathbb{E}[y(\boldsymbol{x})] = \boldsymbol{\phi}(\boldsymbol{x})^T \mathbb{E}[\boldsymbol{w}] = 0$$

$$\mathbb{E}[y(\boldsymbol{x}) y(\boldsymbol{x}')] = \boldsymbol{\phi}(\boldsymbol{x})^T \mathbb{E}[\boldsymbol{w}^T \boldsymbol{w}] \boldsymbol{\phi}(\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \Sigma_w \boldsymbol{\phi}(\boldsymbol{x}').$$

Tātad $y(\boldsymbol{x})$ un $y(\boldsymbol{x}')$ ir sadalīti normāli ar vidējo vērtību 0 un kovariācijas matricu $\boldsymbol{\phi}(\boldsymbol{x})^T \Sigma_w \boldsymbol{\phi}(\boldsymbol{x}')$.

Pirms pievēršamies lineārajai regresijai funkciju telpā, sekojot C. Viljamsa (*C. K. I. Williams*) [30, 6. lpp.] paraugam, apskatīsim vispārīgi prognozēšanu ar Gausa procesiem. Pieņemsim, ka mums doti $n+1$ gadījuma lielumi $(Z_1, ..., Z_n, Z_*)$, kas sadalīti normāli ar vidējo vērtību $\boldsymbol{0}$ un kovariācijas matricu K_+. Vēlamies prognozēt Z_*. Sadalīsim K_+ $n \times n$ matricā K, $n \times 1$ vektorā k un skalārā vērtībā k_*,

$$K_+ = \begin{pmatrix} K & \boldsymbol{k} \\ \boldsymbol{k}^T & k_* \end{pmatrix}.$$

Teorēma 7. *[32, 117. lpp.] Ja novērotas vērtības $Z_1 = z_1, ..., Z_n = z_n$, tad nosacītais sadalījums Z_* būs normāls ar vidējo vērtību un dispersiju*

$$\mathbb{E}[Z_*] = \boldsymbol{k}^T K^{-1} \boldsymbol{z}, \tag{3.3}$$

$$\text{var}[Z_*] = k_* - \boldsymbol{k}^T K^{-1} \boldsymbol{k}. \tag{3.4}$$

Atgriežoties pie lineārās regresijas, mūs interesē sadalījums $Y(\boldsymbol{x}_*) = Y_*$ (apzīmējam Y ar lielo burtu, lai uzsvērtu, ka tas ir gadījuma lielums), ja doti dati $(t_1, ..., t_n)$, kas satur troksni. Lai to izdarītu, jāatrod $(T_1, ..., T_n, Y_*)$ sadalījums un no tā jāiegūst nosacītā blīvuma funkcija $p(Y_*|\boldsymbol{t})$, ja doti $T_1 = t_1, ..., T_n = t_n$.

$(T_1, ..., T_n, Y_*)$ sadalījumu var iegūt vispirms apskatot $\boldsymbol{Y}_+ = (Y_1, ..., Y_n, Y_*)$. Ja modelējam tos kā lineārajā regresijā, tad $\boldsymbol{Y}_+ = \mathcal{N}(\boldsymbol{0}, \Phi_+ \Sigma_w \Phi_+^T)$, kur Φ_+ ir paplašināta Φ matrica, tai apakšā pievienojot vēl vienu rindu $\boldsymbol{\phi}_* = \boldsymbol{\phi}(\boldsymbol{x}_*) = (\phi_1(\boldsymbol{x}_*), ..., \phi_m(\boldsymbol{x}_*))$. $\Phi_+ \Sigma_w \Phi_+^T$ varam pārrakstīt kā

$$\Phi_+ \Sigma_w \Phi_+^T = \begin{pmatrix} \Phi \Sigma_w \Phi^T & \Phi \Sigma_w \boldsymbol{\phi}_* \\ \boldsymbol{\phi}_*^T \Sigma_w \Phi^T & \boldsymbol{\phi}_*^T \Sigma_w \boldsymbol{\phi}_* \end{pmatrix}.$$

Tā kā T tiek iegūti, pie gadījuma lielumiem Y pieskaitot Gausa troksni, varam secināt, ka $(T_,..., T_n, Y_*) \sim \mathcal{N}(\boldsymbol{0}, \Phi_+ \Sigma_w \Phi_+^T + E_+)$, kur E_+ ir $n \times n$ matrica $\sigma_v^2 I_n$, kurai labajā pusē un apakšā pievienota kolonna un rinda ar 0. Izmantojot vienādojumus (3.3) un (3.4), varam iegūt Y_* prognozi funkciju telpā,

$$\mathbb{E}[Y_*] = \mu_{ft}(\boldsymbol{x}_*) = \Phi_*^T \Sigma_w \Phi^T P^{-1} \boldsymbol{t}, \tag{3.5}$$

$$\text{var}[Y_*] = \phi_*^T \Sigma_w \phi_* - \phi_*^T \Sigma_w \Phi^T P^{-1} \Phi \Sigma_w \phi_*, \tag{3.6}$$

kur $P = (\Phi \Sigma_w \Phi^T + \sigma_v^2 I_n)$.

Apgalvojums 8. *[30, 7. lpp.] Funkciju telpas prognozes vienādojumi* (3.5) *un* (3.6) *ir ekvivalenti ar svaru telpas prognozes vienādojumiem* (3.1) (3.2).

Pierādījums. Ievērojam, ka

$$A \Sigma_w \Phi^T = \Phi^T + \beta \Phi^T \Phi \Sigma_w \Phi^T = \beta \Phi^T (\sigma_v^2 I + \Phi \Sigma_w \Phi^T) = \beta \Phi^T P.$$

Sareizinot vienādojumu ar A^{-1} un P^{-1}, varam iegūt $\Sigma_w \Phi^T P^{-1} = \beta A^{-1} \Phi^T$, ko ievietojot vienādojumā (3.5), iegūstam ekvivalenci.

Dispersijas pierādījumam izmantojam formulu

$$(X+YZ)^{-1} = X^{-1} - X^{-1}Y(I+ZX^{-1}Y)^{-1}ZX^{-1},$$

lai pārveidotu $A^{-1} = (\Sigma_w^{-1} + \beta\Phi^T\Phi)^{-1}$, iegūstot

$$A^{-1} = \Sigma_w - \Sigma_w\Phi^T(\sigma_v^2 I + \Phi\Sigma_w\Phi^T)^{-1}\Phi\Sigma_w,$$

ko, ievietojot (3.6), varam iegūt vēlamo rezultātu. □

Apkopojot nodaļā rakstīto, mēs, aplūkojot regresiju no Gausa Procesa puses, ieguvām tos pašus rezultātus ko salīdzinoši vienkāršajā svaru telpas gadījumā. Taču funkciju telpā iespējas paveras daudz plašākas, jo, lai izmantotu prognozēšanas vienādojumus (3.3) un (3.4), vienīgais nosacījums ir, lai $\boldsymbol{Y}$ būtu Gausa process ar zināmu kovariācijas matricu K.

Lineārajā regresijā matrica ir formā $\Phi\Sigma_w\Phi^T$, taču izrādās, ka tas ir tikai viens specifisks Gausa procesa apriorā sadalījuma gadījums, K var konstruēt dažādi, izvēloties situācijai piemērotu procedūru. Protams, ne jau jebkura funkcija no $(\boldsymbol{x}, \boldsymbol{x}')$ dos derīgu kovariācijas matricu, jo tai jāpiemīt specifiskām īpašībām, tāpēc nākamā apakšnodaļa tiks veltīta to aprakstam.

3.3. Kovariācijas funkcijas

Kovariācijas funkcijas, sauktas arī par čaulām (*kernel*), nosaka kā un kādā mērā datu punkti ietekmē viens otru. Ja mums, piemēram, ir pamats uzskatīt, ka starp diviem datu punktiem x un x' distancei $|x - x'|$ pieaugot, to savstarpējā ietekme mazinās lēni, tad varam izvēlēties čaulu, kas to atspoguļos prognozēs. Šeit parādās viens no Beijesa metožu pamatprincipiem, proti, mēs izmantojam iepriekš zināmu (aprioru) informāciju, pieņēmumus, lai iegūtu attiecīgajai problēmai piemērotu modeli.

Definīcija 6. Ja kovariācijas funkcija ir atkarīga tikai no attāluma $|x - x'|$, tad to sauc par stacionāru.

Lielākā daļa praksē pielietotās čaulas ir stacionāras. Apskatīsim vienu no biežāk pielietotajām, kvadrātiski eksponenciālo (SE) kovariācijas funkciju [2]

$$k_{SE} = \exp\left(-\frac{|x-x'|^2}{2\ell^2}\right),$$

[2]Pielietojumos kovariācijas funkcija tiek reizināta ar funkcijas dispersiju σ_Y^2, taču demonstrācijas nolūkos pieņemsim, ka tā ir 1.

kur parametru ℓ sauc par harakteristiko garuma skalu (*characteristic length-scale*). Var pierādīt [33], ka no ℓ ir atkarīgs, cik bieži Gausa process ar vidējo vērtību 0 šķērso ordinātu asi vienības intervālā. Jo mazāks ℓ, jo svārstīgāka funkcija. SE ir bezgalīgi daudz reižu diferencējama, un tāpēc ļoti gluda.

Izmantojot kovariācijas funkcijas spektrālo formu, var pierādīt [29, 84. lpp.], ka regresija ar SE čaulu ir ekvivalenta regresijai ar bāzes funkcijām, ko aplūkojām iepriekš, tikai šajā gadījumā bāzes funkciju skaits būs bezgalīgs.

Salīdzinājumam aplūkosim arī eksponenciālo kovariācijas funkciju

$$k_{OU} = \exp\left(-\frac{|x - x'|}{\ell}\right),$$

kas pieder Materna (*Matern*) klasei un viendimensionālā gadījumā ir Ornšteina-Ūlenbeka (*Ornstein-Uhlenbeck*) procesa kovariācijas funkcija. OU procesa čaulai atvasinājumi neeksistē, tamdēļ ar to konstruētie Gausa procesi ir saraustīti.

Aprakstīsim algoritmu, ar kuru var praktiski simulēt gan aprioro, gan aposterioro sadalījumu un to izmantosim, lai parādītu kā izskatās Gausa procesi ar k_{SE} un k_{OU} kovariācijas funkcijām.

Simulēšana no Gausa procesa $\boldsymbol{Y} \sim \mathcal{GP}(0, k(x, x'))$ apriorā sadalījuma:

I izvēlas $(x_1, ..., x_n)$, punktus, kuros tiks iegūtas Gausa procesa vērtības;

II izvēlas kovariācijas funkciju $k(x, x')$;

III aprēķina kovariācijas matricu

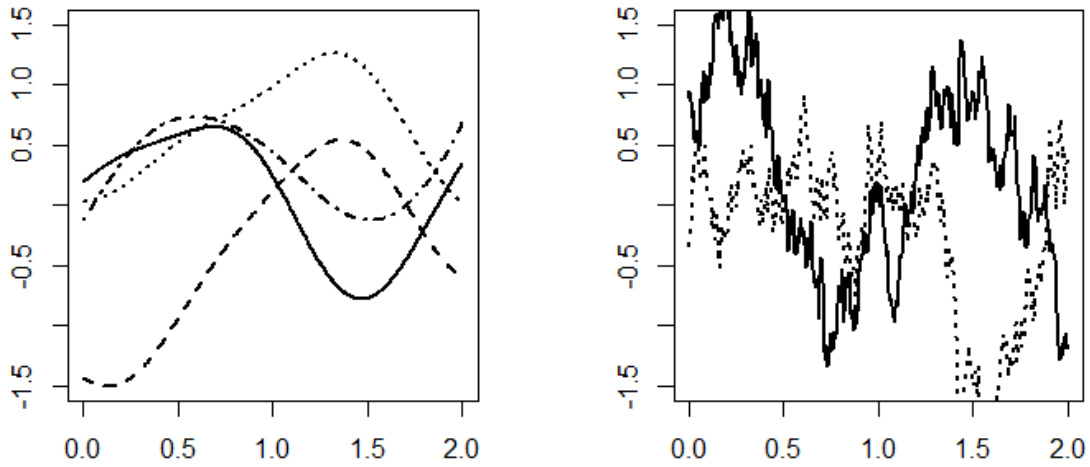
$$K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{pmatrix};$$

IV simulē funkcijas vērtības $\boldsymbol{Y}$ no $\mathcal{N}(\boldsymbol{0}, K)$.

3.1. attēlā apskatāmas trīs realizācijas no Gausa apriorā sadalījuma ar SE un OU kovariācijas funkcijām.

Pieņemsim, ka ir novērotas prognozējamās funkcijas $\boldsymbol{Y}$ vērtības $\boldsymbol{y}$ punktos $\boldsymbol{x}$. Lai aprēķinātu prognozi $\boldsymbol{Y}_*$ punktos $\boldsymbol{x}_*$, varam izmantot sekojošu algoritmu.

Simulēšana no aposteriorā Gausa procesa sadalījuma $\boldsymbol{Y}_* | \boldsymbol{x}_*, \boldsymbol{x}, \boldsymbol{y}$:



3.1. att.: Realizācijas no Gausa procesa apriorā sadalījuma. Pa kreisi četras realizācijas no procesa ar SE kovariācijas funkciju. Pa labi divas no procesa ar OU kovariācijas funkciju. $\ell = 0.5$.

I aprēķina matricas $K(\boldsymbol{x}_*, \boldsymbol{x})$, $K(\boldsymbol{x}, \boldsymbol{x}_*)$, $K(\boldsymbol{x}, \boldsymbol{x})^{-1}$, $K(\boldsymbol{x}_*, \boldsymbol{x}_*)$;

II simulē funkcijas vērtības Y_* no nosacītā sadalījuma

$$\mathcal{N}\big(K(\boldsymbol{x}_*, \boldsymbol{x})K(\boldsymbol{x}, \boldsymbol{x})^{-1}\boldsymbol{y}, K(\boldsymbol{x}_*, \boldsymbol{x}_*) - K(\boldsymbol{x}_*, \boldsymbol{x})K(\boldsymbol{x}, \boldsymbol{x})^{-1}K(\boldsymbol{x}, \boldsymbol{x}_*)\big).$$

3.2. attēlā apskatāmas realizācijas no aposteriorā Gausa procesa sadalījuma. Ģenerēšanu no aposteriorā sadalījuma varam interpretēt kā tādu funkciju meklēšanu, kuras iet caur novērotajiem punktiem.

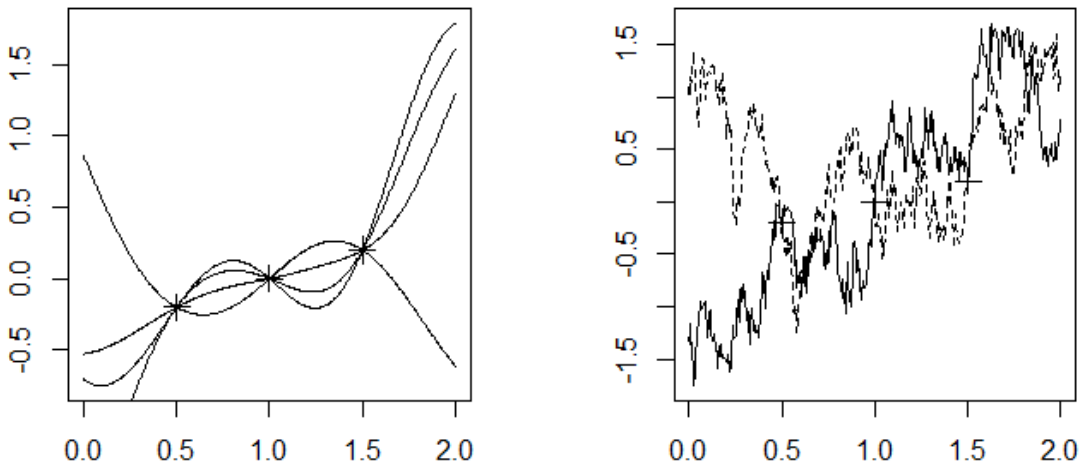
Veicot regresiju praktiskos pielietojumos, parasti nav zināma Gausa procesa dispersija σ_Y^2, nedz arī kovariācijas čaulas parametrs ℓ, tāpat bieži pieņem, ka dati ir trokšņaini, un jānovērtē arī trokšņa dispersija σ_ε^2. Piemēram, SE kovariācijas funkcija būs formā

$$k_{SE} = \sigma_Y^2 \exp\left(-\frac{|x - x'|^2}{2\ell^2}\right) + \sigma_\varepsilon^2 \delta,$$

kur $\delta = 1$, kad $x = x'$, un $\delta = 0$, kad $x \neq x'$.

Lai atrastu parametrus ℓ, σ_Y^2, σ_ε^2, ievērojam, ka, ja $T = Y + \varepsilon$, tad $\boldsymbol{T} \sim \mathcal{N}(\boldsymbol{0}, K + \sigma_\varepsilon^2 I)$ un, izmantojot normālā sadalījuma īpašības, varam iegūt logaritmu no ticamības funkcijas

$$\log \mathcal{L}(\boldsymbol{x}, \ell, \sigma_Y^2, \sigma_\varepsilon^2) = -\frac{1}{2}\log|K| - \frac{1}{2}\boldsymbol{t}^T K^{-1}\boldsymbol{t} - \frac{n}{2}\log 2\pi,$$



3.2. att.: Realizācijas no Gausa procesa aposteriorā sadalījuma. Pa kreisi četras realizācijas no procesa ar SE kovariācijas funkciju. Pa labi divas no procesa ar OU kovariācijas funkciju. $\ell = 0.5$, $\boldsymbol{x} = \{0.5; 1; 1.5\}$, $\boldsymbol{y} = \{-0.2; 0; 0.2\}$.

kuru maksimizējot var atrast parametru ticamākās vērtības.

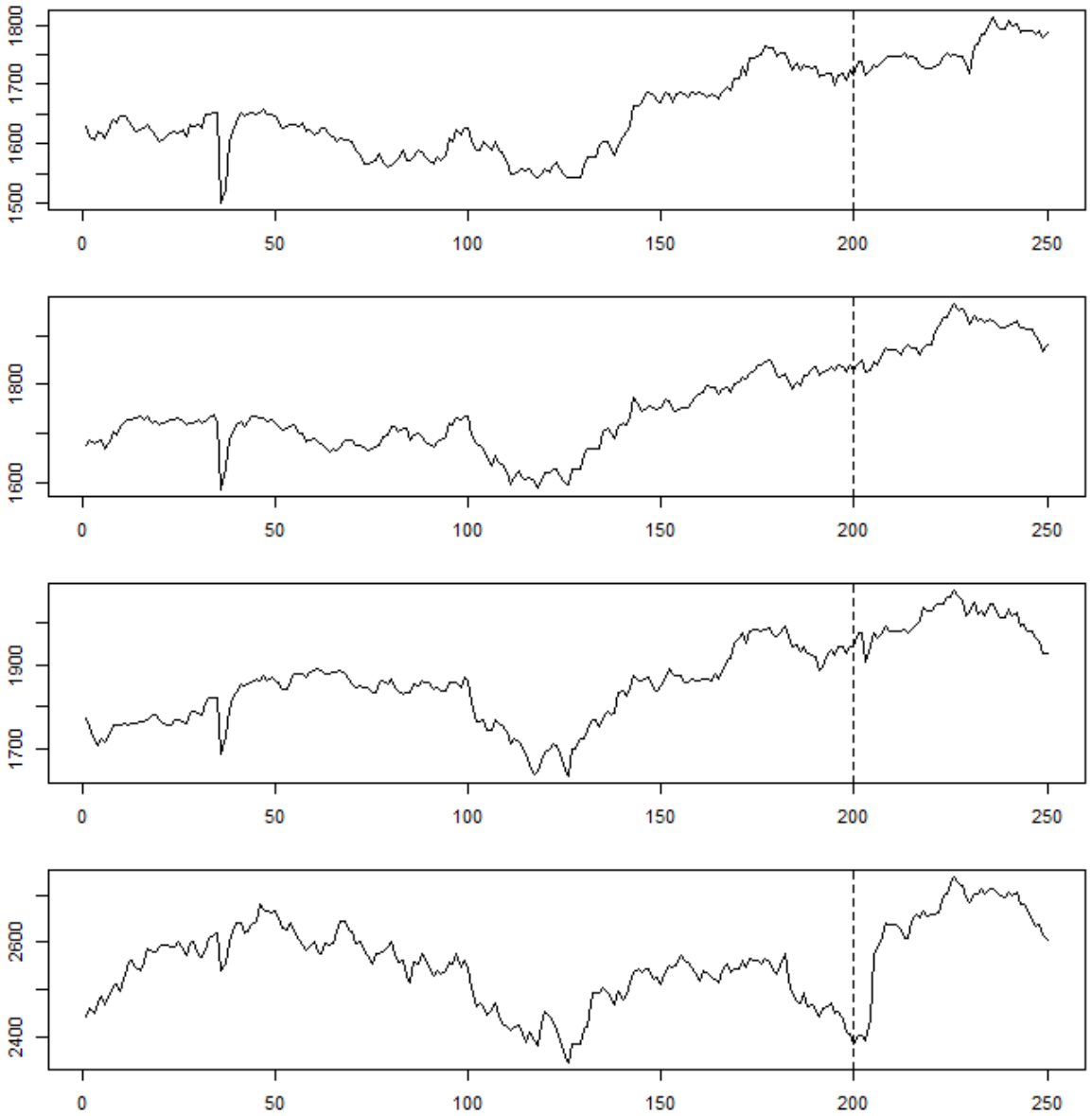
3.4. Laikrindu prognozēšana, salīdzinājums ar ARIMA metodoloģiju

Lai pārbaudītu regresijas ar Gausa procesu prognozēšanas spēju, izmantosim divu veidu kovariācijas funkcijas: kvadrātiski eksponenciālo un kādu no Materna klases funkciju saimes. To izvēli galvenokārt nosaka fakts, ka tās ir pieejamas pašlaik vienīgajā `R` Gausa procesu bibliotēkā `tgp` [34].

Salīdzināsim to ar ARIMA procesiem, pielietojot `forecast` bibliotēku [35]. Lai arī var pierādīt [29, 212. lpp.], ka eksistē ekvivalence starp prognozēšanu ar AR(p) procesu un Gausa procesu ar noteiktu kovariācijas funkciju, ARIMA parasti netiek klasificēta kā Beijesa tipa pieeja.

Kā sava veida pārbaudi, vai ARIMA un Gausa procesa modeļi ir piemēroti biržu indeksu prognozēšanai, salīdzināsim arī ar pēdējās dienas vērtību kā prognozi dienu uz priekšu jeb $\mathbb{E}(x_{t+1}) = x_t$, sauksim to par "naivo" metodi.

Prognozēšanai tiks izmantota Eiropas Savienības biržu indeksi `EuStockMarkets`, kas iekļauta `R` bāzes versijā. Tajā glabājas vēsturiskie dati par DAX (Vācija), SMI (Šveice), CAC (Francija) un FTSE (Lielbritānija) indeksu vērtībām darbadienas beigās no 1991. līdz 1998. gadam. Indeksu pirmās 250 vērtības redzamas 3.3. attēlā.



3.3. att.: 4 Eiropas biržu indeksi no 1991. gada 1. janvāra 250 dienas uz priekšu. Ar pārtrauktu līniju atzīmēta vieta, no kuras veiktas prognozes. No augšas DAX, SMI, CAC, FTSE.

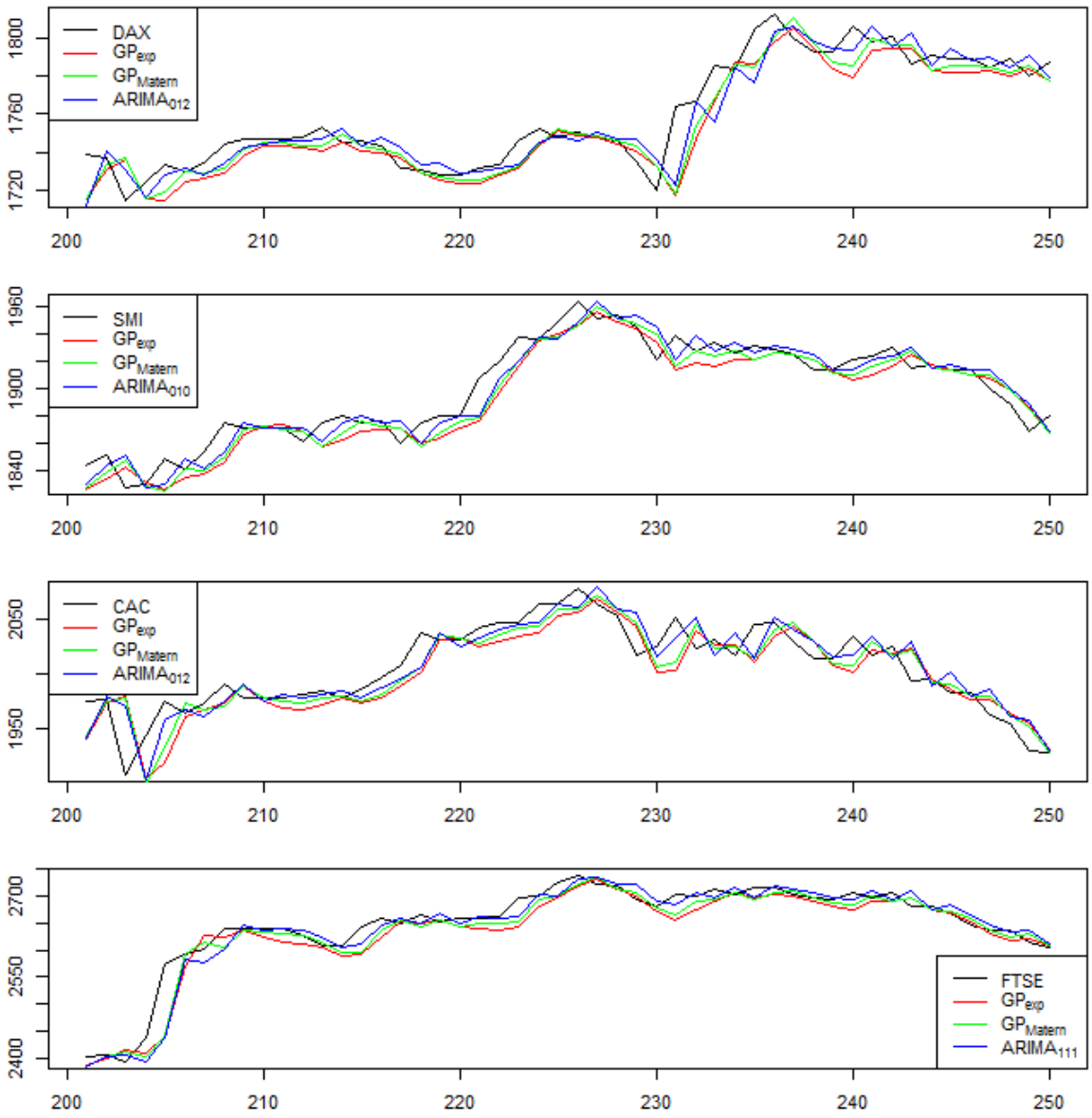
Pirmajā solī pirmie n pārveidotie dati izmantoti parametru noteikšanai, un ar rezultējošo modeli veikta prognoze vienai dienai uz priekšu x^*_{n+1}, tad aprēķināta absolūtā kļūda $|x^*_{n+1} - x_{n+1}|$ un kvadrātiskā kļūda $|x^*_{n+1} - x_{n+1}|^2$. Nākošajā solī parametri aprēķināti no

jauna, izmantojot jau $n+1$ vērtības. Šādi veikta prognozes 50 dienām. To novērtēšanai izmantosim divas statistikas,

$$\mathrm{RMSE} = \sqrt{n^{-1}\sum_i (x_i^* - x_i)^2}, \qquad \mathrm{MAE} = n^{-1}\sum_i |x_i^* - x_i|.$$

Katra indeksu 50 dienu prognozi, izmantojot aprakstītās metodes, varam apskatīt 3.4. attēlā. 3.1. tabulā nolasāmi indeksa prognožu precizitātes novērtējumi. Varam izdarīt divus secinājumus, pirmkārt, ARIMA metodoloģija uzrāda labākus rezultātus nekā Gausa procesi visiem indeksiem, lai arī Gausa process ar Materna klases čaulu cieši seko. Taču, ja aplūkojam arī laika patēriņu 3.2. tabulā, varam ievērot, ka Gausa procesiem nepieciešams daudz vairāk laika, kas ir liels pluss klasiskajām metodēm. Ilgā prognozes rēķināšana skaidrojama ar faktu, ka nepieciešams invertēt $n \times n$ matricu, kas ir procedūra ar $O(n^3)$ kompleksitāti.

Otrkārt, vislabākie rezultāti ir naivajai metodei. Atliek secināt, ka vērtspapīru biržu indeksi nav bijis tas labākais pielietojums aprakstītajām metodēm. Interesanti, ka SMI indeksa gadījumā ARIMA metodoloģijas un naivās metodes prognozes, kas notiek tad, ja $\mathrm{ARIMA}_{0,1,0}$ tiek izvēlēts par piemērotāko.



3.4. att.: DAX, SMI, CAC, FTSE indeksu 50 dienu prognoze pa vienam solim uz priekšu.

3.1. tabula: DAX, SMI, CAC, FTSE indeksu prognožu precizitātes novērtējumi katrai no metodēm. Vislabākos rezultātus dod naivā metode.

DAX				
	GP_{exp}	GP_{Matern}	$ARIMA_{0,1,2}$	Naivā
RMSE	12.20	11.34	11.30	10.80
MAE	9.11	8.02	7.80	7.37
SMI				
	GP_{exp}	GP_{Matern}	$ARIMA_{0,1,0}$	Naivā
RMSE	13.61	12.35	11.80	11.80
MAE	11.13	9.76	9.47	9.47
CAC				
	GP_{exp}	GP_{Matern}	$ARIMA_{0,1,2}$	Naivā
RMSE	22.23	20.39	19.30	19.08
MAE	16.52	15.09	14.52	13.81
FTSE				
	GP_{exp}	GP_{Matern}	$ARIMA_{1,1,1}$	Naivā
RMSE	27.47	25.30	25.21	25.14
MAE	25.14	15.60	15.48	14.99

3.2. tabula: Prognozes laiks vienam solim DAX indeksam katrai no metodēm, ja $n = 249$.

	GP_{exp}	GP_{Matern}	$ARIMA_{0,1,2}$	Naivā
Laiks sekundēs	566.20	384.08	0.10	≈ 0

Secinājumi

Darbā tika sniegts neliels ieskats Beijesa statistikas metodēs, to pielietojumi maiņas punktu noteikšanā, kā arī laikrindu prognozēšanā.

Pirmajā nodaļā tika parādīts, kā Beijesa formula sniedz elegantu, vienotu pieeju dažādu problēmu risināšanā. Taču Robina-Ritova paradokss uzskatāmi parāda, ka uz Beijesa metodēm nevar vienmēr paļauties un ka tās jāpielieto saprātīgi. Iespējams, ka neparametriskā statistika spēs piedāvāt risinājumu šai problēmai.

Otrajā nodaļā salīdzinātās maiņas punkta atrašanas procedūras sniedz ieskatu potenciālā, kas ir tāda statistiķa rīcībā, kurš pārvalda gan klasiskās, gan Beijesa metodes. Proti, viņš var izvēlēties pieeju, kas ir vislabāk piemērota situācijai. Jāsecina, ka Beijesa metodes neaizvietos klasiskās vai otrādi, visdrīzāk nākotnē novērosim, ka tās tiks vienlīdz bieži izmantotas.

Beidzamajā sadaļā demonstrēts, cik plašu metožu klasi aptver Gausa procesi. Mainot kovariācijas funkciju iespējams ar vienu un to pašu algoritmu risināt dažādas problēmas, un tas varētu daudzu pētnieku darbu padarīt vieglāku. Ievērojams defekts gan ir ilgais rēķināšanas laiks, kas visticamāk kavēs procesu atpazīstamību kā efektīvu instrumentu praktiskās problēmās.

Beijesietis D. Lindlijs (*D. Lindley*) [36] 1975. gadā rakstīja, ka 21. gadsimts būs Beijesa gadsimts. Lai arī darbā iegūtie rezultāti nav tik pārliecinoši, lai viņam pilnībā piekristu, jāatzīst, ka nozīmīgākie atklājumi statistikā tik tiešām varētu nākt no Beijesa nometnes, jo nozares ideju, it īpaši neparametrisko, lauks šķiet bagātīgs.

Izmantotā literatūra un avoti

[1] A. Hald. *A history of parametric statistical inference from Bernoulli to Fisher, 1713-1935*. Springer, 2006.

[2] T.S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[3] S. Ghosal and AW Van der Vaart. Fundamentals of nonparametric bayesian inference, 2011.

[4] S. Jackman. Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *American Journal of Political Science*, pages 375–404, 2000.

[5] T. Zheng, M.J. Salganik, and A. Gelman. How many people do you know in prison? *Journal of the American Statistical Association*, 101(474):409–423, 2006.

[6] A. Gelman, J.S. Liebman, V. West, and A. Kiss. A broken system: The persistent patterns of reversals of death sentences in the united states. *Journal of Empirical Legal Studies*, 1(2):209–261, 2004.

[7] C.A. Sims and T. Zha. Bayesian methods for dynamic multivariate models. *International Economic Review*, pages 949–968, 1998.

[8] C.A. Sims. Bayesian methods in applied econometrics, or, why econometrics should always and everywhere be bayesian. *Hotelling lecture, June*, 29:2007, 2007.

[9] D.G. Mayo. *Error and the growth of experimental knowledge*. University of Chicago Press, 1996.

[10] A. Vaselāns. Maiņas punktu noteikšana laikrindu analīzē. Diplomdarbs, Latvijas Universitāte, 2012.

[11] Larry Wasserman. http://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/.

[12] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2003.

[13] A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

[14] J.K. Ghosh and RV Ramamoorthi. *Bayesian nonparametrics*. Springer, 2003.

[15] J.M. Robins, Y. Ritov, et al. Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

[16] Larry Wasserman. http://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/.

[17] D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[18] J. Sethuraman. A constructive definition of dirichlet priors. Technical report, DTIC Document, 1991.

[19] N.L. Hjort, C. Holmes, P. Mueller, and S.G. Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[20] J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2002.

[21] D. Barry and J.A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, pages 309–319, 1993.

[22] C. Erdman and J.W. Emerson. bcp: an r package for performing a bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.

[23] J. Valeinis and E. Cers. Extending the two-sample empirical likelihood (preprint, available on homepage). 2011.

[24] J.J.K. O'Ruanaidh and W.J.F. Gerald. *Numerical Bayesian methods applied to signal processing*, volume 5. Springer-Verlag New York, 1996.

[25] Z. Liu. *Direct simulation methods for multiple changepoint problems*. PhD thesis, Ph. D. thesis, Department of Mathematics and Statistics, Lancaster University, 2007.

[26] Y.C. Ho and D.L. Pepyne. Simple explanation of the no-free-lunch theorem and its implications. *Journal of Optimization Theory and Applications*, 115(3):549–570, 2002.

[27] A. Mira and S. Petrone. Bayesian hierarchical nonparametric inference for change-point problems. *Bayesian Statistics*, 5:693–703, 1996.

[28] R. Garnett, M.A. Osborne, and S.J. Roberts. Sequential bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 345–352. ACM, 2009.

[29] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

[30] C.K.I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. *NATO ASI SERIES D BEHAVIOURAL AND SOCIAL SCIENCES*, 89:599–621, 1998.

[31] J.K. Ghosh, Mohan. Delampady, and Tapas. Samanta. *An introduction to Bayesian analysis*. Springer New York, 2006.

[32] M.L. Eaton and ML Eaton. *Multivariate statistics: a vector space approach*. Wiley New York, 1983.

[33] M.F. Kratz. Level crossings and other level functionals of stationary gaussian processes. *Probability Surveys*, 3:230–288, 2006.

[34] R.B. Gramacy and M.A. Taddy. tgp: Bayesian treed gaussian process models. *R package version*, 2(2), 2008.

[35] R.J. Hyndman and Y. Khandakar. Automatic time series for forecasting: The forecast package for r. Technical report, Monash University, Department of Econometrics and Business Statistics, 2007.

[36] DV Lindley. The future of statistics: A bayesian 21st century. *Advances in Applied Probability*, 7:106–115, 1975.

Pielikums

```
####################Dirihlē procesa simulācijas
intensity = 10
n = 500
theta <- rnorm(n)
V<- rbeta(n, 1, intensity); v_pi<-c(); v_pi[1] = V[1]
for (i in 1:(n-1))
{
prod = 1;
for (j in 1:i) prod = prod * (1-V[j])
v_pi[i+1] <- prod*V[i+1]
}

####################WELLOG
akas_log <- scan("welllog.txt")

####################Izlēcēju izmešana
akas_log_NO <- akas_log
for(j in 1:1000){
for (i in 1: length(akas_log_NO))
{
if (akas_log_NO[i]<100000) {akas_log_NO<-akas_log_NO[-i]; break}
}}
plot(seq(1, length(akas_log_NO)), akas_log_NO)

for(j in 1:20){
for (i in 1: 500)
{
if (akas_log_NO[i]>120000) {akas_log_NO<-akas_log_NO[-i]; break}
}}

for(j in 1:4){
for (i in 1100:1500)
{
if (akas_log_NO[i]<118000) {akas_log_NO<-akas_log_NO[-i]; break}
}}

####################Beijesa maiņas punkts
```

```
library(bcp)
bh_akas <- bcp(akas_log_NO, w0 = 0.00001, p0 = 0.00001)

#################### AR(1; 0.9)
set.seed(6); p = 10^(-128)
nn <- 1000
ar1<-arima.sim(n = nn, list(ar = 0.9))
ar2<-arima.sim(n = nn, list(ar = 0.9))+5
ar_12<-c(ar1,ar2);
bh_ar09<-bcp(ar_12, w0 = p, p0 = p)
plot(bh_ar09)

################### EL maiņas punkts
#J. Valeiņa kods

library(EL)
dati.sim<-ar_12
dati.sim<-akas_log_NO
NN<-1000; # NN<-length(dati.sim);
delta0<-0
el<-0

N<-300; N; #datu apjoms logos N<-100; par(mfrow=c(2,2))
M<-trunc(N^(3/5));M; #bloka garums -  window width no publikacijas
L<-M; L; #bloku skaits, ja non-overlaping
Q<-trunc((N-M)/L)+1; Q; #- bloku skaits
#Q<-8

blocking<-function(X.data, Y.data)
{
X.block<-c()
mu1=mean(X.data)
for(i in 1:Q) X.block[i]<- mean(X.data[((i-1)*L+1):((i-1)*L+M)]-mu1)
#transformetie X bloku datiti
Y.block<-c()
mu2=mean(Y.data)
for(i in 1:Q) Y.block[i]<-mean(Y.data[((i-1)*L+1):((i-1)*L+M)]-mu1-delta0)
EL.means(X.block,Y.block)$p.value
## Robusti EL.Huber / EL.means
```

```
}
pp.values<-c()
for (i in 1:(NN-2*N))
{
F1.block<-dati.sim[i:(i+N-1)];
F2.block<-dati.sim[(i+N):(i+2*N-1)];
pp.values[i]<-blocking(F1.block,F2.block);
}
plot(N:(NN-N-1),pp.values,type="l", xlab = "", ylab = "");
abline(h = 0.05, col = "green")

####################Gausa procesa apriorie/aposteriorie sadalījumi
library(MASS)
kov_kernel <- function(x, y, sigma_f, ell)
{
return(sigma_f^2 * exp( -(x - y)^2 / (2*ell^2)))
}

kovar2 <- function(X1, X2, sigma_f, ell)
{
n1 <-length(X1); n2<-length(X2); A <- matrix(rep(0, n1*n2), nrow = n1)
for (i in 1:n1) {
for (j in 1:n2) {
A[i,j] = kov_kernel(X[i], X[j], sigma_f, ell)
#if (i == j) A[i,j] = A[i,j] #+ sigma_n
};}; return(A)}

X<-1.0
X_star <- seq(0,2, by = 0.005)

K_XstarX <- c()
for (i in 1:length(X_star))
K_XstarX[i] <- kov_kernel(X_star[i], X, 1, 0.5)

K_XX<-1

K_XXstar <- c()
for (i in 1:length(X_star))
K_XXstar[i] <- kov_kernel(X, X_star[i], 1, 0.5)
```

```
K_star_star<-kovar(X_star)

K_pr<- K_star_star - K_XstarX%*%t(K_XXstar)

Y <- mvrnorm(n = 1, mu = rep(0, length(X_star)), Sigma = K_pr)

#####################Prognozēšana ar Gausa procesiem

library(tgp)
predDaxExp<-c()
for(i in 1:50)
{
diffDax<-EuStockMarkets[,1][1:(199+i)]
fit_bgp <- bgp(X = seq(1,199+i), XX = (200+i), Z = diffDax)
predDaxExp[i]<-fit_bgp$ZZ.mean
}

####################Prognozēšana ar Arima procesiem
library(forecast)
predDaxArima<-c()
for(i in 1:50)
{
predArima<-auto.arima(EuStockMarkets[,1][1:(199+i)])
predDaxArima[i]<-forecast(predArima, h = 1)$mean
}
```

Diplomdarbs “Parametriskās un neparametriskās Beijesa metodes un to pielietojumi” izstrādāts LU Fizikas un matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Aleksis Jurševskis

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.

Vadītājs: doc. Dr. math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: doc. Dr. math. Nadežda Siņenko

_____ _____

(paraksts) (datums)

Darbs iesniegts _____

(datums)

_____

(darbu pieņēma)

Diplomdarbs aizstāvēts valsts pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)