Osnovna teorija i primjene generaliziranih aditivnih modela

Varjačić, Katja

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Katja Varjačić

OSNOVNA TEORIJA I PRIMJENE GENERALIZIRANIH ADITIVNIH MODELA

Diplomski rad

Voditelj rada:
doc. dr. sc. Hrvoje Planinić

Zagreb, srpanj 2024.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Dedi Iveku

Sadržaj

Uvod

Tradicionalni linearni modeli često pretpostavljaju jednostavne linearne odnose između zavisnih i nezavisnih varijabli. Iako su takvi modeli korisni za mnoge vrste podataka, njihova ograničenost može biti neprikladna za složenije podatke gdje linearni odnosi nisu dovoljni za adekvatno modeliranje. Generalizirani linearni modeli (GLM-ovi) proširuju mogućnosti linearnih modela dopuštajući zavisnoj varijabli da slijedi razne distribucije iz eksponencijalne familije, čime omogućuju veću fleksibilnost u modeliranju.

Međutim, generalizirani aditivni modeli (GAM-ovi) idu korak dalje. Oni uvode aditivnost, što znači da se linearni prediktor dobiva kao zbroj nelinearnih funkcija pojedinačnih prediktora. Jedna od ključnih prednosti GAM-ova je njihova sposobnost da se prilagode podacima bez potrebe za eksplicitnim navođenjem oblika funkcije koja opisuje odnos između zavisnih i nezavisnih varijabli.

Generalizirani aditivni modeli (GAM-ovi) su prvi put predstavljeni od strane Trevora Hastiea i Roberta Tibshiranija 1986. godine [2]. Njihova originalna metoda uključivala je procjenu GAM-ova pomoću tzv. *"backfitting"* algoritma, pri čemu se koristio bilo koji *"smoother"* za procjenu nelinearnih funkcija, no njezin nedostatak je ležao u činjenici da je bilo teško procijeniti stupanj glatkoće. Kasnije, Simon Wood je 2006. godine u svojoj knjizi "Generalized Additive Models: An Introduction with R" predstavio daljnje napretke u metodologiji i primjeni GAM-ova, čineći ih pristupačnijima i korisnijima za širu znanstvenu zajednicu. Ova knjiga je postala referentni materijal za mnoge istraživače i praktičare koji koriste GAM-ove u svojim radovima.

GAM-ovi su relativno nova metoda koja je brzo našla primjenu u različitim područjima zbog svoje sposobnosti da efikasno modeliraju nelinearne odnose. Danas se koriste u mnogim disciplinama, uključujući ekologiju, biostatistiku, ekonomiju i društvene znanosti, gdje pružaju alate za bolju analizu i razumijevanje složenih podataka.

U prvom poglavlju pruža se temeljni pregled generaliziranih linearnih modela (GLM-ova). Ovdje uspoređujemo GLM-ove s linearnim modelima te objašnjavamo ključne koncepte kao što su distribucija odziva, funkcija veze i procjena parametara metodom iterativnih najmanjih kvadrata.

Drugo poglavlje posvećeno je generaliziranim aditivnim modelima (GAM-ovima). Počinjemo s univarijatnim zaglađivanjem, prelazimo na aditivne modele, a zatim uvodimo

općenite GAM-ove. Također, obrađujemo kriterije za odabir glatkoće te alternativni bayesovski pristup.

U trećem poglavlju detaljno objašnjavamo splajnove, koji su ključni za modeliranje nelinearnih funkcija prediktora. Prikazujemo modeliranje pomoću baznih funkcija te uvodimo različite baze (B-splajnovi, kardinalni, ciklični, ...). Posebnu pažnju posvećujemo penaliziranim regresijskim splajnovima zbog njihovih dobrih računalnih svojstava.

Četvrto poglavlje obrađuje dva praktična primjera. Prvi primjer modelira broj smrti kroz vrijeme u gradu Chicagu koristeći prediktorne varijable o kvaliteti zraka. Drugi primjer također se bavi vremenskim nizovima, gdje modeliramo prosječnu temperaturu u Engleskoj kroz nekoliko stoljeća. U ovom slučaju primjenjujemo generalizirane aditivne mješovite modele (GAMM-ove), koji su ovdje posebno pogodni jer mogu modelirati koreliranost reziduala, što je česta pojava kod vremenskih nizova.

Poglavlje 1

Generalizirani linearni modeli

1.1 Usporedba s linearnim modelima

Linearni modeli (LM) dugo su bili temelj statističke analize, pružajući moćne alate za razumijevanje odnosa između varijabli. U svojoj srži, LM-ovi pretpostavljaju linearni odnos između nezavisnih varijabli i zavisne varijable, za koju se očekuje da slijedi normalnu (Gaussovu) distribuciju. Jednostavnost i interpretativnost linearnih modela čine ih popularnim izborom za mnoge primjene. Međutim, restriktivne pretpostavke LM-a ograničavaju njihovu primjenjivost u scenarijima gdje podaci prirodno ne odgovaraju tim pretpostavkama. Ovdje dolaze u igru generalizirani linearni modeli (GLM), proširujući sposobnosti linearnih modela kako bi se prilagodili širem rasponu struktura i odnosa podataka.

Distribucija odziva

Jedna od najznačajnijih razlika između LM-a i GLM-a je u tretmanu distribucije varijable odziva. LM pretpostavljaju da varijabla odziva, Y, slijedi normalnu distribuciju s konstantnom varijancom. Ova pretpostavka nije samo ograničavajuća, već i nerealna za mnoge vrste podataka. Na primjer, podaci o brojanju inherentno slijede Poissonovu distribuciju, koja se ne može točno modelirati s normalnom distribucijom zbog svoje diskretne prirode i odnosa varijance i srednje vrijednosti.

GLM-ovi adresiraju ovo ograničenje dopuštajući da varijabla odziva slijedi bilo koju distribuciju iz eksponencijalne familije, koja uključuje Poissonovu, binomnu, gama i inverznu Gaussovu distribuciju, među ostalima. Ova fleksibilnost omogućava modeliranje širokog spektra tipova podataka, od brojanja do proporcija, do vremena između događaja.

Funkcija veze

Još jedna ključna razlika između LM-ova i GLM-ova je uvođenje funkcije veze u GLM-ovima. U LM-ovima, očekivana vrijednost odziva, $E[Y]$, modelira se izravno kao

linearna kombinacija prediktora. Međutim, ovaj linearni odnos možda nije uvijek prikladan, posebno kada se modeliraju vjerojatnosti ili brojanja. Funkcija veze u GLM-ovima omogućava modeliranje nelinearnog odnosa između linearnih prediktora i srednje vrijednosti distribucije odziva.

Funkcija veze, $g(\cdot)$, transformira očekivanu vrijednost varijable odziva, $E[Y]$, u linearni prediktor, η, tako da $\eta = g(E[Y])$. Ova transformacija osigurava da su modelirane količine prikladne za distribuciju varijable odziva. Na primjer, logit funkcija veze korištena s binomnom distribucijom osigurava da su predviđene vjerojatnosti ograničene između 0 i 1.

Funkcija varijance

Funkcija varijance je još jedan element koji razlikuje GLM-ove od LM-ova. U LM-ovima, pretpostavlja se da je varijanca pogrešaka konstantna kroz cijeli raspon vrijednosti nezavisnih varijabli. Ova pretpostavka homoskedastičnosti često je prekršena u stvarnim podacima. Za razliku od toga, GLM uključuje funkciju varijance koja omogućava da varijanca odziva ovisi o njegovoj srednjoj vrijednosti. To znači da kako se mijenja srednja vrijednost odziva, tako se može mijenjati i varijanca, pružajući točniji prikaz podataka. Na primjer, u podacima o brojanju modeliranim s Poissonovom distribucijom, varijanca je jednaka srednjoj vrijednosti, prirodno prilagođavajući se heteroskedastičnosti koja se opaža u takvim podacima.

Širina primjene

Spomenute razlike između LM-ova i GLM-ova - distribucija odziva, funkcija veze i funkcija varijance - znatno proširuju raspon primjene za GLM-ove. Dok su LM-ovi dobro prilagođeni podacima koji pažljivo slijede njihove osnovne pretpostavke, GLM-ovi proširuju mogućnosti modeliranja na podatke koji su binarni, temeljeni na brojanju, na proporcijama i više, bez gubitka interpretativnosti koji je svojstven regresijskim modelima.

U sažetku, generalizirani linearni modeli predstavljaju fleksibilno i moćno proširenje linearnih modela, sposobni za smještaj širokog spektra tipova podataka i odnosa. Opuštanjem strogih pretpostavki linearnih modela i uvođenjem koncepta kao što su funkcija veze i funkcija varijance, GLM-ovi pružaju robustan okvir za statističku analizu u mnogim područjima studija.

Pregled linearnih i generaliziranih linearnih modela

Razumijevanje temeljnih razlika između LM-ova i GLM-ova može se dodatno pojasniti kroz njihove matematičke formulacije. Osnovna ideja linearnih modela leži u pretpostavci da postoji linearna veza između očekivanja odziva i kovarijata, tj.:

$$\mathbb{E}(Y_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \tag{1.1}$$

gdje je p broj kovarijata (konstanta može biti uključena), uz pretpostavku normalnosti:

$$Y_i \sim N\left(\sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right). \tag{1.2}$$

Za generalizirane linearne modele, pretpostavka se proširuje ne samo da bi se uključila fleksibilnost u obliku veze između očekivanog odziva i kovarijata, nego i da omogući odzivu da prati distribuciju koja nije nužno normalna, već pripada široj klasi eksponencijalnih familija distribucija. To je izraženo preko:

$$\mu_i \equiv \mathbb{E}\left(Y_i\right) = g^{-1}\left(\sum_{j=1}^{p} \beta_j x_{ij}\right), \tag{1.3}$$

$$g\left(\mu_i\right) = \mathbf{X}_i\boldsymbol{\beta} \equiv \boldsymbol{\eta} \tag{1.4}$$

pri čemu je:

- g funkcija veze,
- $\sum_{j=1}^{p} \beta_j x_{ij} = \mathbf{X}_i\boldsymbol{\beta} \equiv \boldsymbol{\eta}$ tzv. linearni prediktor,
- Y_i ima unaprijed određenu distribuciju iz tzv. eksponencijalne familije.

1.2 Eksponencijalna familija distribucija

Odzivna varijabla u GLM-u može imati bilo koju distribuciju iz eksponencijalne familije. Distribucija pripada eksponencijalnoj familiji distribucija ako se njezina vjerojatnosna funkcija gustoće [1] može zapisati kao

$$f_\theta(y) = \exp\left[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)\right], \quad y \in \mathbb{R},$$

gdje su b, a i c proizvoljne funkcije, ϕ proizvoljni "parametar disperzije" ili "skaliranja", a θ je poznat kao "prirodni parametar" distribucije. Najpoznatije i najčešće korištene distribucije iz eksponencijalne familije uključuju normalnu (Gaussovu), binomnu, Poissonovu

[1] Ovdje govorimo ili o gustoći neprekdine slučajne varijable ili o diskretnoj funkciji gustoće kod diskretnih razdioba.

i Gama distribuciju. Pokažimo, na primjer, da je normalna distribucija zaista član eksponencijalne familije. Naime vrijedi sljedeće

$$\begin{aligned} f_\mu(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right] \\ &= \exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right], \end{aligned}$$

što je eksponencijalni oblik, uz $\theta = \mu$, $b(\theta) = \theta^2/2 \equiv \mu^2/2$, $a(\phi) = \phi = \sigma^2$ i $c(\phi, y) = -y^2/(2\phi) - \log(\sqrt{\phi 2\pi}) \equiv -y^2/\left(2\sigma^2\right) - \log(\sigma\sqrt{2\pi})$.

Očekivanje i varijanca izražene preko a, b i ϕ

Logaritam maksimalne vjerodostojnosti θ, uz dani y, jednostavno je $\log\{f_\theta(y)\}$ promatrano kao funkcija od θ. To jest

$$l(\theta) = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi),$$

i stoga

$$\frac{\partial l}{\partial \theta} = \{y - b'(\theta)\}/a(\phi).$$

Tretiranje l kao slučajne varijable, zamjenom opažanja y slučajnom varijablom Y, omogućuje procjenu očekivane vrijednosti $\partial l/\partial\theta$:

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = \{\mathbb{E}(Y) - b'(\theta)\}/a(\phi).$$

Korištenjem općeg rezultata $\mathbb{E}(\partial l/\partial\theta) = 0$ (vidi dokaz u Dodatku A) i preuređivanjem dobivamo

$$\mathbb{E}(Y) = b'(\theta) \tag{1.5}$$

tj., očekivanje bilo koje slučajne varijable eksponencijalne familije dano je prvom derivacijom b s obzirom na θ, gdje b ovisi o određenoj distribuciji. Ova jednadžba ključna je u povezivanju parametara GLM modela, $\boldsymbol{\beta}$, s prirodnim parametrima eksponencijalne familije. U GLM-u, $\boldsymbol{\beta}$ određuje srednju vrijednost varijable odziva i, preko (1.5), određuje prirodni parametar za svako opažanje odziva.

Ponovnim diferenciranjem log-vjerodostojnosti dobivamo

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi)$$

i uključivanjem ovoga u opći rezultat (vidi dokaz u Dodatku A),

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -\mathbb{E}\left\{\left(\frac{\partial l}{\partial \theta}\right)^2\right\},$$

dobivamo

$$b''(\theta)/a(\phi) = \mathbb{E}\left[\{Y - b'(\theta)\}^2\right]/a(\phi)^2,$$

što se preoblikuje u drugi korisni opći rezultat:

$$\text{var}(Y) = b''(\theta)a(\phi). \tag{1.6}$$

Funkcija a bi u principu mogla biti bilo koja funkcija od ϕ, i kada radimo s GLM-ovima, nema poteškoća u rukovanju bilo kojim oblikom a, ako je ϕ poznat. Međutim, situacija se komplicira kada je ϕ nepoznat. Problem se olakšava ako je moguće definirati $a(\phi) = \phi/\omega$, gdje je ω poznata konstanta, pri čemu će najčešće biti jednaka 1. Ovaj ograničeni oblik zapravo pokriva sve slučajeve od praktičnog interesa. Stoga sada imamo

$$\text{var}(Y) = b''(\theta)\phi/\omega.$$

Uz pretpostavku da je $\text{Var}(Y) > 0$ za sve θ, ϕ, imamo da je b' strogo rastuća pa dakle i invertibilna, stoga možemo, zbog (1.5), umjesto parametra θ koristiti parametar očekivanja μ, uz $\theta = \theta(\mu) = (b')^{-1}(\mu)$. Tada imamo vezu između varijance i očekivanja

$$\text{Var}(Y) = \phi b''(\theta(\mu))/\omega =: \phi V(\mu), \tag{1.7}$$

pri čemu funkciju $V(\mu) := b''(\theta(\mu))/\omega$ zovemo funkcijom varijance.

1.3 Procjena parametara

Prisjetimo se da GLM modelira vektor n nezavisnih odzivnih varijabli, $\mathbf{Y}$, gdje je $\boldsymbol{\mu} \equiv \mathbb{E}(\mathbf{Y})$, putem

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} \text{ i } Y_i \sim f_{\theta_i}(y_i),$$

gdje $f_{\theta_i}(y_i)$ označava distribuciju eksponencijalne familije, s prirodnim parametrom θ_i, koji je određen s μ_i i stoga konačno s $\boldsymbol{\beta}$. Uz dani vektor $\mathbf{y}$, jedno opažanje od $\mathbf{Y}$, moguće je procijeniti $\boldsymbol{\beta}$ metodom maksimalne vjerodostojnosti. Budući da su Y_i međusobno nezavisne, vjerodostojnost parametra $\boldsymbol{\beta}$ je

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i)$$

i stoga je log-vjerodostojnost

$$l(\boldsymbol{\beta}) = \log \prod_{i=1}^{n} f_{\theta_i}(y_i) = \sum_{i=1}^{n} \log \{f_{\theta_i}(y_i)\} = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b_i(\theta_i)}{a(\phi)} + c_i(\phi, y_i) \right),$$

gdje ovisnost desne strane o $\boldsymbol{\beta}$ proizlazi iz ovisnosti θ_i o $\boldsymbol{\beta}$. Primijetimo da se funkcije a, b i c mogu mijenjati s i. Na primjer, to omogućuje različite (ali poznate s konstantom) varijance za normalne odzive. S druge strane, ϕ se pretpostavlja isti za sve i. Kao što je spomenuto u prethodnom odjeljku, u praksi je dovoljno razmatrati samo slučajeve u kojima možemo pisati $a_i(\phi) = \phi/\omega_i$, gdje je ω_i poznata konstanta (obično 1) i u tom slučaju je

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\omega_i (y_i \theta_i - b_i(\theta_i))}{\phi} + c_i(\phi, y_i) \right)$$

Newtonova metoda maksimizacije log-vjerodostojnosti

Za maksimiziranje funkcije log-vjerodostojnosti $l(\boldsymbol{\beta})$ u kontekstu Generaliziranih linearnih modela (GLM-ova), gdje izravna rješenja za $\boldsymbol{\beta}$ često nisu dostupna, iterativna Newton-Raphson metoda nudi praktičan i učinkovit pristup. Počevši od početne procjene, metoda precizira $\boldsymbol{\beta}$ iskorištavajući lokalnu zakrivljenost i gradijent funkcije log-vjerodostojnosti.

Suština ovog pristupa leži u primjeni Taylorova razvoja funkcije $l(\boldsymbol{\beta})$ oko trenutačne procjene $\boldsymbol{\beta}^{(\text{stari})}$, omogućavajući kvadratnu aproksimaciju log-vjerodostojnosti. Razvoj je izražen kao:

$$l(\boldsymbol{\beta}) \approx l\left(\boldsymbol{\beta}^{(\text{stari})}\right) + \left(\nabla l(\boldsymbol{\beta}^{(\text{stari})})\right)^T \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(\text{stari})}\right) + \frac{1}{2} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(\text{stari})}\right)^T \mathbf{H}(\boldsymbol{\beta}^{(\text{stari})}) \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(\text{stari})}\right).$$

Iz ove aproksimacije postupak za izračun sljedeće procjene $\boldsymbol{\beta}$ izvodi se postavljanjem derivacije $l(\boldsymbol{\beta})$ u odnosu na $\boldsymbol{\beta}$ na nulu, rezultirajući s:

$$\boldsymbol{\beta}^{(\text{novi})} = \boldsymbol{\beta}^{(\text{stari})} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(\text{stari})}) \nabla l(\boldsymbol{\beta}^{(\text{stari})}),$$

gdje $\nabla l(\boldsymbol{\beta}^{(\text{stari})})$ označava gradijent log-vjerodostojnosti evaluiran u $\boldsymbol{\beta}^{(\text{stari})}$, koji predstavlja smjer u kojem funkcija log-vjerodostojnosti raste, a $\mathbf{H}(\boldsymbol{\beta}^{(\text{stari})})$ je Hesseova matrica, koja odražava zakrivljenost l. Ova formula osigurava da se svaka iteracija parametra $\boldsymbol{\beta}$ usmjerava prema optimalnom povećanju $l(\boldsymbol{\beta})$, uzimajući u obzir i prilagođavajući se zakrivljenosti koju predstavlja $\mathbf{H}(\boldsymbol{\beta})$.

Iteriranjem ovih koraka do konvergencije, metoda Newton-Raphson sustavno pronalazi vrijednosti $\boldsymbol{\beta}$ koje maksimiziraju log-vjerodostojnost, time optimalno prilagođavajući GLM opaženim podacima.

Detaljan izračun gradijenta i Hesseove matrice

Gradijent funkcije log-vjerodostojnosti, $\nabla l(\boldsymbol{\beta})$, ukazuje na smjer u kojem log-vjerodostojnost najbrže raste. Za Generalizirane linearne modele (GLM), izračunava se na sljedeći način: Za svaki koeficijent β_j, gradijent se računa kao parcijalna derivacija $l(\boldsymbol{\beta})$:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i' (\theta_i) \frac{\partial \theta_i}{\partial \beta_j} \right)$$

Primjenjujući lančano pravilo, nalazimo $\frac{\partial \theta_i}{\partial \beta_j}$:

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

S obzirom da je $\eta_i = g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$, $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$, i $\frac{d\mu_i}{d\eta_i} = \frac{1}{g'(\mu_i)}$, te diferenciranjem (1.5), $\frac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i} = \frac{1}{b_i''(\theta_i)}$ dobivamo:

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{x_{ij}}{g'(\mu_i) b_i''(\theta_i)}$$

Uvrštavanjem ovoga u izraz za gradijent dobivamo:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - b_i'(\theta_i)) x_{ij}}{g'(\mu_i) b_i''(\theta_i)/\omega_i}$$

što se dalje pojednostavljuje uvrštavanjem $\mathbb{E}(Y) = b'(\theta)$ i $V(\mu) = b''(\theta)/\omega$:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{g'(\mu_i) V(\mu_i)}.$$

Nadalje, ponovnim diferenciranjem log-vjerodostojnosti dobivamo:

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} &= -\frac{1}{\phi} \sum_{i=1}^{n} \left(\frac{x_{ik} x_{ij}}{g'(\mu_i)^2 V(\mu_i)} + \frac{(y_i - \mu_i) V'(\mu_i) x_{ik} x_{ij}}{g'(\mu_i)^2 V(\mu_i)^2} + \frac{(y_i - \mu_i) x_{ij} g''(\mu_i) x_{ik}}{g'(\mu_i)^3 V(\mu_i)} \right) \\ &= -\frac{1}{\phi} \sum_{i=1}^{n} \frac{x_{ik} x_{ij} \alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)}, \end{aligned}$$

gdje je $\alpha(\mu_i) = 1 + (y_i - \mu_i) \left(\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)} \right)$.

Izvod IRLS metode

Sada, definirajući $\mathbf{W}$ kao dijagonalnu matricu težina

$$\mathbf{W} = \text{diag}\,(w_i)\,, \qquad \text{gdje je } w_i = \frac{\alpha\,(\mu_i)}{g'\,(\mu_i)^2\,V\,(\mu_i)},$$

Hesseova matrica log-vjerodostojnosti postaje $\mathbf{H} = -\mathbf{X}^\top\mathbf{W}\mathbf{X}/\phi$.
Nadalje definiranjem $\mathbf{G} = \text{diag}\,\{g'\,(\mu_i)\,/\alpha\,(\mu_i)\}$, gradijentni vektor log-vjerodostojnosti može se zapisati kao $\nabla l = \mathbf{X}^\top\mathbf{W}\mathbf{G}(\mathbf{y}-\boldsymbol{\mu})/\phi$. Tada jedna Newtonova iteracija ima oblik

$$\begin{aligned}
\boldsymbol{\beta}^{(\text{novi})} &= \boldsymbol{\beta}^{(\text{stari})} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(\text{stari})})\nabla l(\boldsymbol{\beta}^{(\text{stari})}) \\
&= \boldsymbol{\beta}^{(\text{stari})} + (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) \\
&= (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\left\{\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) + \mathbf{X}\boldsymbol{\beta}^{(\text{stari})}\right\} \\
&= (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{z}
\end{aligned}$$

gdje je $z_i = g'\,(\mu_i)\,(y_i-\mu_i)\,/\alpha\,(\mu_i) + \eta_i$. Bitno je primijetiti da su jednadžbe iteracija zapravo procjene najmanjih kvadrata parametra $\boldsymbol{\beta}$ koje proizlaze iz minimiziranja težinske funkcije najmanjih kvadrata

$$\sum_{i=1}^{n} w_i\,(z_i - \mathbf{X}_i\boldsymbol{\beta})^2\,.$$

Dakle GLM-ovi se mogu procjenjivati iterativnom metodom najmanjih kvadrata s težinama (eng., *"Iteratively Re-weighted Least Square"*, kraće IRLS), koji glasi:

1. Inicijaliziraj $\hat{\mu}_i = y_i + \delta_i$ i $\hat{\eta}_i = g\,(\hat{\mu}_i)$, gdje je δ_i obično nula, ali može biti mala konstanta koja osigurava da je $\hat{\eta}_i$ konačan. Iteriraj sljedeća dva koraka do konvergencije.

2. Izračunajte pseudopodatke $z_i = g'\,(\hat{\mu}_i)\,(y_i-\hat{\mu}_i)\,/\alpha\,(\hat{\mu}_i) + \hat{\eta}_i$, i iterativne težine $w_i = \alpha\,(\hat{\mu}_i)\,/\left\{g'\,(\hat{\mu}_i)^2\,V\,(\hat{\mu}_i)\right\}$.

3. Pronađi $\hat{\boldsymbol{\beta}}$, minimizator ciljne funkcije težinskih najmanjih kvadrata

$$\sum_{i=1}^{n} w_i\,(z_i - \mathbf{X}_i\boldsymbol{\beta})^2$$

   zatim ažuriraj $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ i $\hat{\mu}_i = g^{-1}\,(\hat{\eta}_i)$.

Proces se ponavlja dok promjena devijance, koju ćemo objasniti u sljedećem odjeljku, nije dovoljno mala.

1.4 Provjera modela

Devijanca

Pri radu s GLM-ovima u praksi, korisno je imati mjeru koja se može tumačiti na sličan način kao što je to suma kvadrata reziduala u običnom linearnom modeliranju. Ta mjera je devijanca modela, a definira se kao

$$\begin{aligned} D &= 2\left\{l(\hat{\boldsymbol{\beta}}_{\max}) - l(\hat{\boldsymbol{\beta}})\right\}\phi \\ &= \sum_{i=1}^{n} 2\omega_i \left\{y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right\} \end{aligned} \tag{1.8}$$

gdje $l(\hat{\boldsymbol{\beta}}_{\max})$ označava maksimiziranu vjerodostojnost zasićenog modela: model s jednim parametrom po podatku. $l(\hat{\boldsymbol{\beta}}_{\max})$ je najviša vrijednost koju logaritam vjerodostojnosti može imati, uzimajući u obzir podatke. U slučaju distribucije eksponencijalne familije izračunava se jednostavnim postavljanjem $\hat{\boldsymbol{\mu}} = \mathbf{y}$ i evaluacijom vjerodostojnosti. $\tilde{\boldsymbol{\theta}}$ i $\hat{\boldsymbol{\theta}}$ označavaju procjenu maksimalne vjerodostojnosti kanonskih parametara, redom, zasićenog i promatranog modela. Primijetimo kako devijanca ne ovisi o ϕ. Uz devijancu vežemo i skaliranu devijancu,

$$D^* = D/\phi$$

koja ovisi o parametru disperzije. Za Binomne i Poissonove distribucije, gdje je $\phi = 1$, devijanca i skalirana devijanca su jednake.

U određenim okolnostima, skalirana devijanca može biti aproksimirana sa χ^2 distribucijom, tj. približno vrijedi

$$D^* \sim \chi^2_{n-p}, \tag{1.9}$$

za dovoljno veliki broj podataka, na primjer u problemima s diskretnim podacima gdje su brojke velike. Općenito, međutim, χ^2 aproksimacije za devijancu nisu vrlo dobre čak i kada $n \to \infty$. Daljnja istraživanja o asimptotskoj distribuciji D^* tek trebaju biti provedena.

Usporedba modela

Promotrimo sad testiranje

$$\mathrm{H}_0 : \mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}_0\boldsymbol{\beta}_0$$

naspram

$$\mathrm{H}_1 : \mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1$$

gdje je $\boldsymbol{\mu}$ očekivanje odziva, $\mathbf{Y}$, čiji su elementi nezavisne slučajne varijable iz iste eksponencijalne familije distribucija, i gdje je $\mathbf{X}_0 \subset \mathbf{X}_1$. Ako imamo opažanje, $\mathbf{y}$, vektora odziva,

tada se može izvesti test omjera vjerodostojnosti. Neka su $l(\hat{\boldsymbol{\beta}}_0)$ i $l(\hat{\boldsymbol{\beta}}_1)$ maksimizirane log-vjerodostojnosti dva modela. Ako je H_0 istinita onda za dovoljno veliki broj podataka, približno vrijedi

$$2\left\{l\left(\hat{\boldsymbol{\beta}}_1\right) - l\left(\hat{\boldsymbol{\beta}}_0\right)\right\} \sim \chi^2_{p_1-p_0}, \tag{1.10}$$

gdje je p_i broj (identificiranih) parametara, $\boldsymbol{\beta}_i$, u modelu i. Ako nulta hipoteza nije istinita, tada će model 1 imati znatno veću vjerodostojnost u odnosu na model 0, tako da bi dvostruka razlika u log-vjerodostojnostima bila prevelika da bi bila u skladu s odgovarajućom χ^2 distribucijom.

S obzirom na definiciju devijance, lako se vidi da statistiku dvostrukog omjera log-vjerodostojnosti možemo izraziti kao $D_0^* - D_1^*$. Tada, pod pretpostavkom nulte hipoteze H_0, za dovoljno veliki broj podataka,

$$D_0^* - D_1^* \sim \chi^2_{p_1-p_0},$$

gdje je D_i^* devijanca modela i s p_i identificirajućih parametara. Međutim, ovo je korisno samo ako je parametar disperzije, ϕ, poznat i time se D^* može precizno izračunati. Stoga se rezultat može izravno koristiti s Poissonovim i binomnim modelima, ali ne i s normalnim, gama ili inverznim Gaussovim distribucijama, gdje parametar disperzije nije poznat. Što učiniti u ovim posljednjim slučajevima raspravit ćemo u sljedećem odjeljku.

Usporedba modela s nepoznatim ϕ

Pod H_0 imamo aproksimativne rezultate

$$D_0^* - D_1^* \sim \chi^2_{p_1-p_0} \text{ i } D_1^* \sim \chi^2_{n-p},$$

i, ako se $D_0^* - D_1^*$ i D_1^* tretiraju kao asimptotski nezavisni, to implicira

$$F = \frac{\left(D_0^* - D_1^*\right)/\left(p_1 - p_0\right)}{D_1^*/\left(n - p_1\right)} \sim F_{p_1-p_0,n-p_1}$$

za dovoljno veliki broj podataka (rezultat koji je egzaktan u posebnom slučaju klasičnog linearnog modela). Prednost F statistike je u tome što se može izračunati bez poznavanja ϕ. To je moguće jer se ϕ eliminira iz razlomka, čime se, pod pretpostavkom H_0, dobiva aproksimativan rezultat

$$F = \frac{\left(D_0 - D_1\right)/\left(p_1 - p_0\right)}{D_1/\left(n - p_1\right)} \sim F_{p_1-p_0,n-p_1}.$$

Ovaj pristup omogućava testiranje hipoteza bez prethodnog znanja o parametru ϕ. Međutim, valja uzeti u obzir da se temelji na pretpostavci o distribuciji D_1^*, koja može biti upitna.

AIC

Akaikeov informacijski kriterij (AIC) pristup je odabiru modela u kojem se modeli odabiru kako bi se minimizirala procjena očekivane Kullback-Leiblerove divergencije između prilagođenog modela i "pravog modela". Kriterij je

$$\text{AIC} = -2l + 2p,$$

gdje je l maksimizirana log-vjerodostojnost modela, a p broj procijenjenih parametara modela. Odabire se model s najnižim AIC-om.

Reziduali

Provjera modela iznimno je bitan dio statističkog modeliranja. U kontekstu običnih linearnih modela, ovaj postupak se temelji na analizi reziduala modela, koji uključuju sve podatke neobjašnjene prediktivnim dijelom modela. Analiza reziduala primarni je način provjere modela i u slučaju generaliziranih linearnih modela (GLM-ova), gdje je standardizacija reziduala posebno važna, ali i zahtjevnija. Naime, može pomoći u dijagnosticiranju problema s modelom, poput neispravne specifikacije ili kršenja pretpostavki modela. Za GLM-ove, ključni razlog zašto se ne analiziraju samo neobrađeni reziduali,

$$\hat{\epsilon}_i = y_i - \hat{\mu}_i,$$

jest poteškoća u utvrđivanju ispravnosti pretpostavljene veze između srednje vrijednosti i varijance na temelju tih reziduala. Na primjer, upotreba Poissonova modela podrazumijeva da bi varijanca reziduala trebala rasti proporcionalno s veličinom prilagođenih vrijednosti, $\hat{\mu}_i$. Međutim, procjena s grafa neobrađenih reziduala u odnosu na prilagođene vrijednosti - je li varijabilnost reziduala proporcionalna srednjoj vrijednosti, ili pak kvadratnom korijenu ili kvadratu srednje vrijednosti - gotovo je nemoguća golim okom. Stoga se reziduali GLM-a obično standardiziraju kako bi, ako su pretpostavke modela ispravne, imali približno jednaku varijancu i što više nalikovali rezidualima iz običnog linearnog modela.

Pearsonovi reziduali

Najizravniji način za standardizaciju reziduala je da ih prilagodimo tako što ćemo ih podijeliti s veličinom proporcionalnom njihovoj standardnoj devijaciji prema prilagođenom modelu. To rezultira Pearsonovim rezidualima

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

gdje bi, u idealnom slučaju kada je model precizan, ovi reziduali trebali pokazati malu ili nikakvu sistematsku varijaciju, s prosječnom vrijednošću nula i konzistentnom varijancom

ϕ. Očekuje se da takvi reziduali neće pokazivati ovisnost ni o prilagođenim vrijednostima ni o drugim varijablama, neovisno o tome jesu li one dio modela. Ime "Pearsonovi reziduali" dolazi iz činjenice da je suma kvadrata Pearsonovih reziduala jednaka Pearsonovoj X^2 statistici koja također testira prilagođenost modela:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Zapravo, Pearsonovi reziduali proizlaze iz reziduala koji se dobiju tijekom iterativnog postupka najmanjih kvadrata s težinama (IRLS), podijeljeni s korijenom odgovarajućih konvergiranih težina.

Reziduali devijance

U praksi, distribucija Pearsonovih reziduala može biti prilično asimetrična oko nule, što znači da njihovo ponašanje nije toliko slično rezidualima klasičnog linearnog modela kako bi se možda očekivalo. Reziduali devijance (*"deviance residuals"*) često su bolji izbor u tom pogledu. Ovi reziduali proizlaze iz uvida da devijanca u kontekstu GLM-ova služi sličnoj svrsi kao što je to slučaj sa sumom kvadrata reziduala u klasičnim linearnim modelima, odnosno u tom slučaju devijanca je upravo suma kvadrata reziduala. To jest, reziduali su korijeni komponenata devijacije s odgovarajućim predznakom. Dakle, ako označimo s d_i komponentu devijacije koju doprinosi i-ti podatak (tj. i-ti član u zbroju u (1.8)), imamo

$$D = \sum_{i=1}^{n} d_i$$

i, analogno klasičnom linearnom modelu, možemo definirati

$$\hat{\epsilon}_i^d = \operatorname{sign}(y_i - \hat{\mu}_i)\ \sqrt{d_i}.$$

Primijetimo, zbroj kvadrata ovih "reziduala devijance" daje samu devijancu, tj.

$$D = \sum_{i=1}^{n} (\hat{\epsilon}_i^d)^2.$$

Sada, kada bi se devijanca izračunala za model gdje su svi parametri poznati, tada bi, pod pretpostavkom da je pretpostavljeni model točan, (1.9) postalo $D^* \sim \chi^2_n$. To bi moglo sugerirati da za svaki pojedinačni podatak vrijedi $d_i/\phi \sim \chi^2_1$, implicirajući tako $\epsilon_i^d \sim N(0, \phi)$. Naravno, (1.9) se ne može razumno primijeniti na pojedinačni podatak, ali ipak sugerira da bismo mogli očekivati da se reziduali devijance ponašaju poput slučajnih varijabli $N(0, \phi)$ za dobro prilagođeni model, posebno u slučajevima za koje se očekuje da je (1.9) razumna aproksimacija.

Poglavlje 2

Osnovna teorija GAM-ova

2.1 Uvod

Kao što smo vidjeli, GLM-ovi proširuju korisnost linearne regresije omogućavajući modeliranje podataka kroz razne distribucije, izlazeći izvan okvira isključivo normalno distribuiranih odgovora. Ipak, GLM-ovi zadržavaju pretpostavku linearnog odnosa između prediktora i odzivne varijable. Ovo ograničenje postaje očito kada se analiziraju podaci iz stvarnog svijeta, koji često pokazuju nelinearne veze i kompleksne obrasce koje linearna pretpostavka ne može adekvatno objasniti.

Kao odgovor na to ograničenje, Generalizirani aditivni modeli (GAM-ovi) unaprjeđuju pristup modeliranju uvođenjem aditivne strukture koja omogućava uključivanje nelinearnih veza. Nadograđujući GLM-ove, GAM-ovi zamjenjuju linearni prediktor zbrojem glatkih funkcija prediktora, umjesto obične linearne kombinacije. Ova prilagodba omogućava GAM-ovima da obuhvate širi spektar odnosa, uključujući nelinearne ili promjenjive učinke preko vrijednosti prediktora, sve to bez potrebe za prethodnom definicijom točnog oblika veze.

Formalni matematički prikaz GAM-ova ima strukturu poput

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \tag{2.1}$$

gdje $\mu_i \equiv \mathbb{E}(Y_i)$ označava očekivanu vrijednost odzivne varijable Y_i, koja slijedi distribuciju iz eksponencijalne familije $\mathrm{EF}(\mu_i, \phi)$, karakteriziranu srednjom vrijednošću μ_i i parametrom disperzije ϕ. Izraz $\mathbf{A}_i$ je red matrice modela za bilo koje strogo parametarske komponente modela, s $\boldsymbol{\theta}$ kao odgovarajućim vektorom parametara, a f_j su glatke funkcije preostalih kovarijata, x_k.

Ova formulacija ilustrira sposobnost GAM-ova za fleksibilno određivanje ovisnosti odziva o kovarijatima kroz glatke funkcije, izbjegavajući potrebu za kompleksnim parametarskim odnosima. Osnovna privlačnost GAM-ova leži u njihovoj ravnoteži između

fleksibilnosti i mogućnosti interpretacije. Omogućujući modeliranje složenih, nelinearnih učinaka uz očuvanje strukture aditivnog modela, GAM-ovi osiguravaju interpretativnost utjecaja pojedinačnih prediktora. Ova značajka omogućava detaljnu analizu utjecaja svakog prediktora na odzivnu varijablu, nudeći ključne uvide u ponašanje modeliranih odnosa. Međutim, fleksibilnost i praktičnost dolaze pod cijenu dva nova teorijska problema. Potrebno je predstaviti glatke funkcije na neki način i odabrati koliko ih je potrebno zagladiti.

Ovo poglavlje pokazuje kako se GAM-ovi mogu prikazati proširenjem baznih funkcija za svaku glatku komponentu, uz pridruženu penalizaciju koja regulira glatkoću tih funkcija. Procjena se može izvršiti metodama penalizirane regresije, a odgovarajući stupanj glatkoće za f_j može se procijeniti iz podataka koristeći unakrsnu validaciju ili maksimizaciju marginalne vjerodostojnosti. Da bi se sačuvala osnovna jednostavnost pristupa bez opterećenja brojnim tehničkim pojedinostima, najkompleksniji model razmatran ovdje bit će jednostavan GAM s dvije univarijatne glatke komponente. Nadalje, metode koje će biti predstavljene neće biti one najprikladnije za opću praktičnu upotrebu, već metode koje omogućuju da se osnovni okvir jednostavno objasni.

2.2 Univarijatno zaglađivanje

Reprezentacija i procjena komponentnih funkcija u modelu najbolje se uvodi razmatranjem modela koji sadrži jednu funkciju jednog prediktora,

$$y_i = f(x_i) + \epsilon_i, \tag{2.2}$$

gdje je y_i varijabla odziva, x_i prediktor, f glatka funkcija, a ϵ_i nezavisne slučajne varijable s normalnom distribucijom $N\left(0, \sigma^2\right)$.

Reprezentacija funkcije pomoću baznih funkcija

Za procjenu funkcije f, koristeći metode razrađene za linearne i generalizirane linearne modele, potrebno je f predstaviti na način da (2.2) postane linearni model. To se može učiniti odabirom baze, definirajući prostor funkcija kojeg je f (ili bliska aproksimacija) element. Odabir baze svodi se na odabir nekih baznih funkcija, koje ćemo tretirati kao potpuno poznate: ako je $b_j(x)$ j-ta takva bazna funkcija, tada se pretpostavlja da f ima reprezentaciju

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j, \tag{2.3}$$

za neke vrijednosti nepoznatih parametara, β_j. Supstitucijom (2.3) u (2.2) jasno se dobiva linearni model.

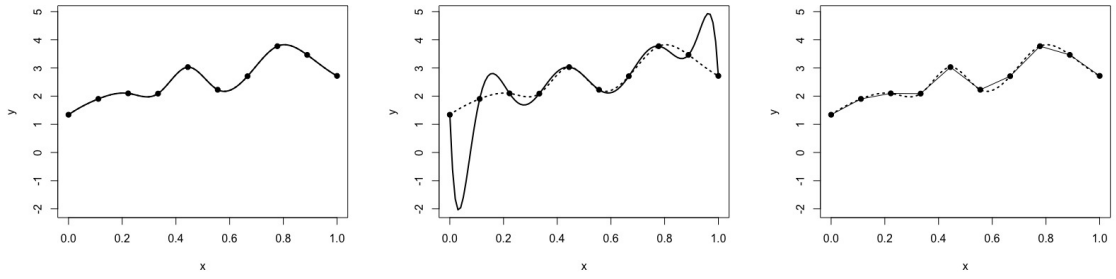
Polinomijalna baza

Prvi, vrlo jednostavni odabir baze bi bio prostor polinoma. Recimo ako vjerujemo da je f polinom četvrtog reda, prostor polinoma reda 4 ili niže bi sadržavao f. U tom slučaju (2.3) postaje

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5.$$

Taylorov teorem sugerira da će polinomijalne baze biti korisne u situacijama gdje je važno analizirati ponašanje funkcije f u neposrednoj blizini određene točke. Međutim, kada je cilj sagledati funkciju f kroz cijelu njezinu domenu, korištenje polinomijalnih baza postaje problematično.

Izazovi postaju najvidljiviji prilikom razmatranja interpolacije što je ilustrirano na slici 2.1. Na središnjem grafu sa slike pokušava se aproksimirati funkcija prikazana na lijevom grafu, korištenjem polinomijalne interpolacije označenih točaka. Polinom na nekim mjestima pokazuje snažne oscilacije u pokušaju da udovolji zahtjevima za interpolacijom podataka i za očuvanjem neprekidnosti svih derivacija u odnosu na x. Odustajanjem od zahtjeva za neprekidnosti derivacija i primjenom po dijelovima linearne interpolacije, postiže se mnogo bolja aproksimacija, kako je prikazano na desnom grafu.



Slika 2.1: Rekonstrukcija glatke funkcije: polinomijalna i po dijelovima linearna interpolacija.

Po dijelovima linearna baza

Jasno je da ima smisla koristiti baze koje su učinkovite u aproksimaciji poznatih funkcija kako bismo prikazali nepoznate funkcije. Slično, baze koje su uspješne u interpolaciji egzaktnih observacija funkcije također su dobra osnova za usko povezan zadatak zaglađivanja zašumljenih observacija funkcije. U narednim poglavljima vidjet ćemo da se po dijelovima linearne baze mogu unaprijediti korištenjem splajn baza koje osiguravaju neprekidnost derivacija do određenog reda, ali zbog svoje praktičnosti, u ovom poglavlju koristit ćemo po dijelovima linearnu bazu.

Baza za po dijelovima linearnu funkciju jedne varijable x u potpunosti je određena mjestima gdje dolazi do diskontinuiteta derivacije, odnosno mjestima gdje se linearni segmenti spajaju. Neka su ti čvorovi označeni s $\left\{x_j^* : j = 1, \ldots, k\right\}$ i pretpostavimo da je $x_j^* > x_{j-1}^*$. Tada za $j = 2, \ldots, k-1$

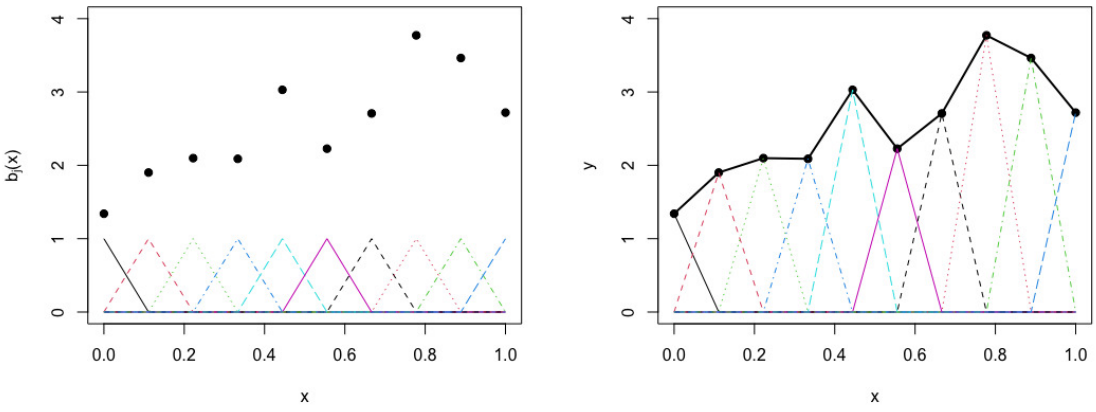
$$b_j(x) = \begin{cases} \left(x - x_{j-1}^*\right) / \left(x_j^* - x_{j-1}^*\right) & x_{j-1}^* < x \leq x_j^* \\ \left(x_{j+1}^* - x\right) / \left(x_{j+1}^* - x_j^*\right) & x_j^* < x < x_{j+1}^* \\ 0 & \text{inače} \end{cases} \tag{2.4}$$

dok

$$b_1(x) = \begin{cases} \left(x_2^* - x\right) / \left(x_2^* - x_1^*\right) & x < x_2^* \\ 0 & \text{inače} \end{cases}$$

i

$$b_k(x) = \begin{cases} \left(x - x_{k-1}^*\right) / \left(x_k^* - x_{k-1}^*\right) & x > x_{k-1}^* \\ 0 & \text{inače} \end{cases}$$



Slika 2.2: Lijevi graf prikazuje primjer baze šatorskih funkcija za interpolaciju podataka prikazanih crnim točkama. Svaka od obojenih linija predstavlja pojedinu šatorsku funkciju koja doseže svoj maksimum od 1 na x-osi kod jedne od točaka. Desni graf ilustrira postupak u kojem se svaka bazna funkcija množi odgovarajućim koeficijentom, a zatim se rezultati zbrajaju kako bi se formirala konačna interpolacija, prikazana debljom crnom linijom.

Dakle, $b_j(x)$ je nula posvuda, osim u intervalu između čvorova neposredno s jedne i druge strane x_j^*. Funkcija $b_j(x)$ linearno raste od 0 u x_{j-1}^* do 1 u x_j^*, a zatim linearno

opada do 0 u x^*_{j+1}. Za bazne funkcije poput ove, koje su ne nula samo na nekim konačnim intervalima, kaže se da imaju kompaktni nosač. Zbog svog oblika, funkcije b_j često se nazivaju šatorskim funkcijama (eng., *"tent functions"*). Primjer možemo vidjeti na slici 2.2.

Korištenjem ove baze za prikazivanje $f(x)$, (2.2) sada postaje linearni model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ gdje je $X_{ij} = b_j(x_i)$.

Kontrola glatkoće penalizacijom vijugavosti

Jedna od mogućnosti za odabir stupnja zaglađivanja je fiksiranje dimenzije baze na malo većoj veličini od one za koju se vjeruje da bi mogla biti razumno potrebna, ali kontrolirajući glatkoću modela dodavanjem kazne za "vijugavost" u cilj prilagodbe najmanjim kvadratima. Na primjer, umjesto prilagodbe modela minimizacijom

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

model bi se mogao prilagoditi minimizacijom

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=2}^{k-1} \left\{ f\left(x^*_{j-1}\right) - 2f\left(x^*_j\right) + f\left(x^*_{j+1}\right) \right\}^2,$$

gdje pridodana suma mjeri vijugavost kao zbroj kvadriranih drugih razlika funkcije u čvorovima (što grubo aproksimira integrirani kvadrirani penalizacijski izraz druge derivacije koji se koristi u zaglađivanju kubičnim splajnovima: vidi odjeljak 3.3). Kada je funkcija f jako vijugava, penalizacija će imati visoke vrijednosti, a kada je funkcija 'glatka', bit će niska. Ako je f pravac, tada je penalizacija zapravo nula. Dakle, penalizacija ima nul-prostor funkcija koje nisu penalizirane: u ovom slučaju pravci. Dimenzija mu je 2, budući da je baza za pravce dvodimenzionalna.

Parametar zaglađivanja, λ, kontrolira ravnotežu između glatkoće procijenjene funkcije f i vjernosti podacima. Kada λ teži prema beskonačnosti, procjena za f postaje pravac, dok $\lambda = 0$ rezultira nepenaliziranom procjenom po dijelovima linearne regresije.

Za bazu "šatorskih funkcija", lako je uočiti da su koeficijenti funkcije f zapravo vrijednosti funkcije u čvorovima, tj. $\beta_j = f(x^*_j)$. To olakšava izražavanje penalizacije u kvadratnoj formi, $\boldsymbol{\beta}^\top \mathbf{S}\boldsymbol{\beta}$, preko koeficijenata baze (iako je za to zapravo potrebna samo linearost funkcije f u koeficijentima baze). Prvo je važno primijetiti da je

$$\begin{bmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \beta_3 - 2\beta_4 + \beta_5 \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & . & . & \cdot \\ 0 & 1 & -2 & 1 & 0 & . & \cdot \\ 0 & 0 & 1 & -2 & 1 & 0 & . \\ . & . & . & . & . & . & \cdot \\ . & . & . & . & . & . & \cdot \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \end{bmatrix}$$

pa zapisivanje desne strane kao $\mathbf{D}\boldsymbol{\beta}$, definiranjem matrice $\mathbf{D}$ dimenzija $(k-2)\times k$, penalizacija postaje

$$\sum_{j=2}^{k-1}\left(\beta_{j-1}-2\beta_j+\beta_{j+1}\right)^2=\boldsymbol{\beta}^\top\mathbf{D}^\top\mathbf{D}\boldsymbol{\beta}=\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}$$

gdje je $\mathbf{S}=\mathbf{D}^\top\mathbf{D}$.

Stoga je problem prilagodbe penalizirane regresije minimizacija izraza

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} \tag{2.5}$$

u odnosu na $\boldsymbol{\beta}$. Problem procjene stupnja glatkoće modela sada je problem procjene parametra zaglađivanja λ. No, prije nego što se pozabavimo procjenom λ, razmotrimo procjenu $\boldsymbol{\beta}$ uz dani λ.

Pokažimo sada da je formalni izraz za minimizator (2.5), penalizirani procjenitelj najmanjih kvadrata (eng., *"penalized least squares estimator"*) za $\boldsymbol{\beta}$ koji glasi

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{2.6}$$

Iz toga će nam slijediti i da je odgovarajuća *"hat matrix"* jednaka

$$\mathbf{H}=\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top, \tag{2.7}$$

odnosno vrijedi

$$\hat{\boldsymbol{\mu}}=\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{H}\mathbf{y}.$$

Označimo sada funkciju koju trebamo minimizirati s $J(\boldsymbol{\beta})$

$$J(\boldsymbol{\beta})=\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}.$$

Dalje proširujemo i pišemo kao

$$J(\boldsymbol{\beta})=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}$$

$$J(\boldsymbol{\beta})=\mathbf{y}^\top\mathbf{y}-\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+\lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}.$$

Da bismo pronašli minimizator, potrebno je izračunati derivaciju $J(\boldsymbol{\beta})$ u odnosu na $\boldsymbol{\beta}$ i postaviti je na nulu. Derivacija je dana formulom

$$\frac{\partial J}{\partial\boldsymbol{\beta}}=-2\mathbf{X}^\top\mathbf{y}+2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+2\lambda\mathbf{S}\boldsymbol{\beta}.$$

Postavljanjem na nulu dobivamo

$$-2\mathbf{X}^\top\mathbf{y}+2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+2\lambda\mathbf{S}\boldsymbol{\beta}=0$$

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{S}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}$$

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}.$$

Nadalje, matrica $\mathbf{X}^\top\mathbf{X}$ je uvijek pozitivno semidefinitna ako je $\mathbf{X}$ realna. Dodavanjem $\lambda\mathbf{S}$, gdje je $\mathbf{S}$ pozitivno semidefinitna i $\lambda > 0$, osigurava se da je zbroj $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}$ također pozitivno semidefinitan. U našem slučaju $\mathbf{S} = \mathbf{D}^\top\mathbf{D}$. Za proizvoljni vektor $\mathbf{v}$, $\mathbf{v}^\top\mathbf{S}\mathbf{v} = \mathbf{v}^\top(\mathbf{D}^\top\mathbf{D})\mathbf{v} = (\mathbf{D}\mathbf{v})^\top(\mathbf{D}\mathbf{v})$, što je suma kvadrata elemenata $\mathbf{D}\mathbf{v}$, a to je uvijek nenegativno. Prema tome, $\mathbf{v}^\top\mathbf{S}\mathbf{v} \geq 0$ za bilo koji $\mathbf{v}$, pa je $\mathbf{S}$ zaista pozitivno semidefinitna. Nadalje ako $\mathbf{X}^\top\mathbf{X}$ ima puni rang, ili ako $\mathbf{S}$ dovoljno kompenzira bilo kakav nedostatak ranga u $\mathbf{X}^\top\mathbf{X}$, tada je $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}$ doista pozitivno definitna pa i invertibilna. Sada rješavamo za $\boldsymbol{\beta}$ kako bismo dobili

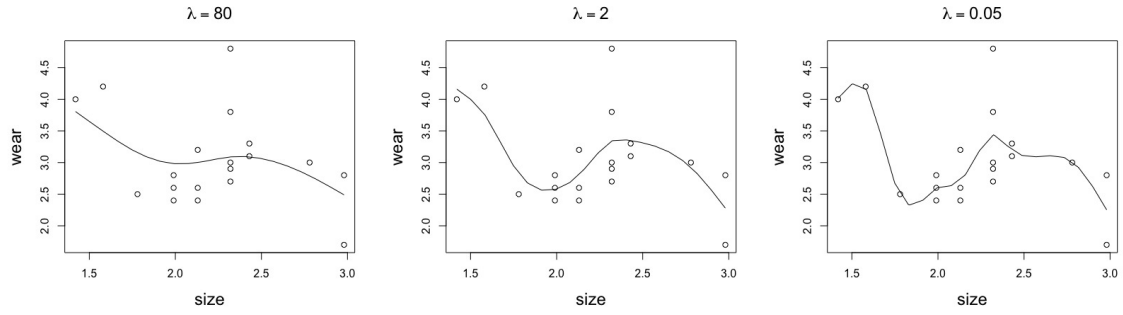
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Pokažimo još samo da je to zaista minimum, što će vrijediti ako je Hessijan pozitivno definitan. No, Hessijan od $J(\boldsymbol{\beta})$, što je matrica drugih derivacija, je dan s:

$$\nabla^2 J(\boldsymbol{\beta}) = 2\mathbf{X}^\top\mathbf{X} + 2\lambda\mathbf{S},$$

a to smo pokazali da je pozitivno definitno. Time je dokaz završen.

Promotrimo sada još kako je promjenom vrijednosti parametra zaglađivanja, λ, moguće dobiti različite modele s različitim stupnjevima glatkoće. Slika 2.3 to ilustrira, ali i nameće pitanje, koja vrijednost λ je "najbolja"?



Slika 2.3: Penalizirane prilagodbe po dijelovima linearnih funkcija na podatke iz skupa podataka engine iz paketa gamair, koristeći tri različite vrijednosti parametra zaglađivanja, λ. Primijetimo kako penalizacija omogućuje vrlo glatke procjene, unatoč korištenju po dijelovima linearne baze.

Odabir parametra zaglađivanja, metodom unakrsne validacije

Ako je λ prevelik, podaci će biti prekomjerno zaglađeni, a ako je premali, podaci će biti nedovoljno zaglađeni: u oba slučaja to znači da procjena $\hat{f}$ neće biti blizu stvarne funkcije f. Idealno bi bilo odabrati λ tako da $\hat{f}$ bude što bliže funkciji f. Prikladan kriterij bi mogao biti odabir λ koja minimizira

$$M = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i - f_i\right)^2$$

gdje je notacija $\hat{f}_i \equiv \hat{f}(x_i)$ i $f_i \equiv f(x_i)$ usvojena radi sažetosti. Budući da je f nepoznata, M se ne može koristiti izravno, ali je moguće izvesti procjenu $\mathbb{E}(M) + \sigma^2$, što je očekivana kvadratna pogreška u predviđanju nove varijable. Definirajmo $\hat{f}^{[-i]}$ kao model prilagođen svim podacima osim (x_i, y_i). Rezultat unakrsne validacije metodom izostavljanja jednog po jednog podatka (eng., *"leave-one-out cross validation"*, kraće LOOCV ili *"ordinary cross validation"*, kraće OCV), $\mathcal{V}_o$, računa se na sljedeći način:

$$\begin{aligned}
\mathcal{V}_o &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - y_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i - \epsilon_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i\right)^2 - 2\left(\hat{f}_i^{[-i]} - f_i\right)\epsilon_i + \epsilon_i^2.
\end{aligned}$$

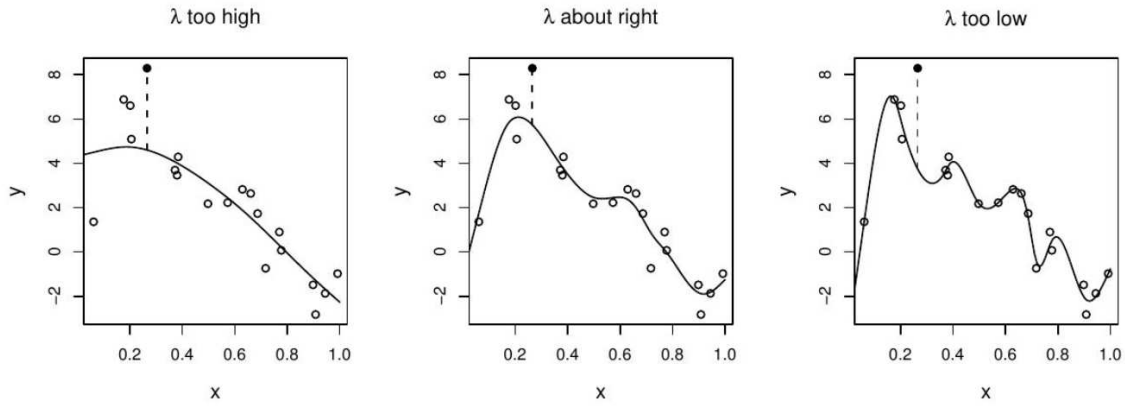
Budući da je $\mathbb{E}(\epsilon_i) = 0$ te su ϵ_i i $\hat{f}_i^{[-i]}$ nezavisni, drugi izraz u zbroju nestaje kada se uzmu očekivanja:

$$\mathbb{E}(\mathcal{V}_o) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i\right)^2\right) + \sigma^2.$$

Intuitivno, za veliki broj podataka $\hat{f}^{[-i]} \approx \hat{f}$ pa $\mathbb{E}(\mathcal{V}_o) \approx \mathbb{E}(M) + \sigma^2$ također za dovoljno velik broj podataka. Stoga je odabir λ kako bi se minimizirao $\mathcal{V}_o$ razuman pristup ako nam je cilj minimizirati M.

Unakrsna validacija je sama po sebi smislena, čak i bez opravdanja srednje kvadratnom greškom. Ako se modeli ocjenjuju samo prema njihovoj sposobnosti prilagodbe podacima iz kojih su procijenjeni, tada se uvijek odabiru složeniji modeli umjesto jednostavnijih. Odabir modela kako bi se maksimizirala sposobnost predviđanja podataka kojima model nije prilagođen, ne pati od ovog problema, kako ilustrira slika 2.4.

Računalno je zahtjevno izračunati $\mathcal{V}_o$ izostavljanjem jednog po jednog podatka, ponovnim prilagođavanjem modela za svaki od n rezultirajućih skupova podataka, no pokazat ćemo da to zapravo nije potrebno. Naime, pokazali smo da se procjene $\hat{y}$ za model



Slika 2.4: Ilustracija principa unakrsne validacije. Peti podatak (crna točka) izostavljen je iz prilagodbe, a neprekidna krivulja prikazuje penalizirani regresijski splajn prilagođen preostalim podacima. Kada je parametar zaglađivanja prevelik, splajn se loše prilagođava mnogim podacima i ne postiže bolje rezultate ni s izostavljenom točkom. Kada je λ premali, splajn prilagođava buku kao i signal, a posljedična dodatna varijabilnost uzrokuje da loše predviđa izostavljeni podatak. Za srednje vrijednosti λ, splajn prilično dobro pristaje osnovnom signalu, ali zaglađuje kroz šum: stoga je izostavljeni podatak razumno dobro predviđen. Unakrsna validacija redom izostavlja svaki podatak iz skupa podataka i razmatra prosječnu sposobnost modela prilagođenih preostalim podacima za predviđanje izostavljenih podataka. Slika je preuzeta iz [5], poglavlje 4.

dobiven minimizacijom penaliziranih najmanjih kvadrata mogu zapisati kao $\hat{\mathbf{y}} = \mathbf{H} * \mathbf{y}$, pri čemu je $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top$. Izbacivanjem $i-$tog podatka, gubimo redak matrice $\mathbf{X}$, no ako umjesto podatka (x_i, y_i) stavimo $(x_i, \hat{f}_i^{[-i]})$ matrica $\mathbf{X}$ ostaje nepromijenjena, pa tako i matrica $\mathbf{H}$.

Dakle, promatrajući i-tu predikciju, imamo da je

$$\hat{f}_i^{[-i]} = \mathbf{H}_i\mathbf{y}^* = \mathbf{H}_i\mathbf{y} - H_{ii}y_i + H_{ii}\hat{\mu}_i^{[-i]} = \hat{f}_i - H_{ii}y_i + H_{ii}\hat{f}_i^{[-i]}, \tag{2.8}$$

gdje je $\mathbf{y}^* = \mathbf{y} - \mathbf{e_i}(y_i - \hat{f}_i^{[-i]})$, pri čemu je $\mathbf{e_i}$ vektor s n nula i jedinicom na i-tom mjestu, a $\hat{f}_i$ predstavlja procjenu izvedenu iz kompletnog vektora $\mathbf{y}$. Dodavanjem y_i na obje strane i

uz malo premještanja dobivamo:

$$\begin{aligned} \hat{f}_i^{[-i]} + y_i &= \hat{f}_i - H_{ii} y_i + H_{ii} \hat{f}_i^{[-i]} + y_i \\ y_i - \hat{f}_i &= y_i - \hat{f}_i^{[-i]} - H_{ii} \left(y_i - \hat{f}_i^{[-i]} \right) \\ y_i - \hat{f}_i &= \left(y_i - \hat{f}_i^{[-i]} \right) (1 - H_{ii}) \\ y_i - \hat{f}_i^{[-i]} &= \frac{\left(y_i - \hat{f}_i \right)}{(1 - H_{ii})}. \end{aligned}$$

Stoga LOOCV izraz sada postaje:

$$\mathcal{V}_o = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i - \hat{f}_i \right)^2}{(1 - H_{ii})^2}$$

U praksi H_{ii} su često zamijenjeni njihovim prosjekom $\mathrm{tr}(\mathbf{H})/n$ što rezultira s generaliziranom unakrsnom validacijom (eng., *"generalized cross validation"*, kraće GCV)

$$\mathcal{V}_g = \frac{n \sum_{i=1}^{n} \left(y_i - \hat{f}_i \right)^2}{[n - \mathrm{tr}(\mathbf{H})]^2},$$

koja ima računalne prednosti u odnosu na običnu unakrsnu validaciju (OCV).

2.3 Aditivni modeli

Sada pretpostavimo da su za varijablu odziva y dostupne dvije prediktorne varijable, x i v, i da je prikladna jednostavna aditivna struktura modela

$$y_i = \alpha + f_1(x_i) + f_2(v_i) + \epsilon_i. \tag{2.9}$$

α je slobodan parametar, f_j su glatke funkcije, a ϵ_i su nezavisne $N\left(0, \sigma^2\right)$ slučajne varijable.

Postoje dvije važne stvari koje treba primijetiti o ovom modelu. Prvo, pretpostavka aditivne strukture je prilično jaka: $f_1(x) + f_2(v)$ je prilično restriktivan poseban slučaj od opće glatke funkcije dvije varijable $f(x, v)$. Drugo, činjenica da model sada sadrži više od jedne funkcije uvodi problem identifikacije (eng., *"identifiability problem"*): f_1 i f_2 svaka je procjenjiva samo do na aditivnu konstantu. Da bismo to vidjeli, primijetimo da se bilo koja konstanta može istovremeno dodati f_1 i oduzeti od f_2, a da se ne promijene predikcije modela. Stoga je potrebno nametnuti ograničenja u model koja osiguravaju jedinstvenost, prije procjene modela.

Pod uvjetom da se riješi problem identifikacije, aditivni model se može predstaviti pomoću penaliziranih regresijskih splajnova, procijenjenih pomoću penaliziranih najmanjih kvadrata, a stupanj zaglađivanja odabran unakrsnom validacijom ili REML-om (*"Restricted Maximum Likelihood"*), na isti način kao i kod jednostavnog univarijatnog modela. Ovdje ćemo obraditi unakrsnu validaciju, a odabir stupnja zaglađivanja REML metodom možete pogledati u [5].

Reprezentacija aditivnog modela

Svaka glatka funkcija u (2.9) može se predstaviti korištenjem penalizirane po dijelovima linerane baze. Konkretno,

$$f_1(x) = \sum_{j=1}^{k_1} b_j(x)\delta_j,$$

gdje su δ_j nepoznati koeficijenti, dok su $b_j(x)$ bazne funkcije oblika (2.4) definirane korištenjem niza od k_1 čvorova x_j^*, ravnomjerno raspoređenih u rasponu varijable x. Analogno,

$$f_2(v) = \sum_{j=1}^{k_2} \mathcal{B}_j(v)\gamma_j,$$

gdje su γ_j nepoznati koeficijenti, a $\mathcal{B}_j(v)$ bazne funkcije, definirane korištenjem niza od k_2 čvorova, v_j^*, ravnomjerno raspoređenih u rasponu varijable v.

Definirajući n-dimenzionalni vektor $\mathbf{f}_1 = [f_1(x_1), \ldots, f_1(x_n)]^\top$, imamo $\mathbf{f}_1 = \mathbf{X}_1\boldsymbol{\delta}$ gdje je $b_j(x_i)$ element i, j matrice $\mathbf{X}_1$. Slično, $\mathbf{f}_2 = \mathbf{X}_2\boldsymbol{\gamma}$, gdje je $\mathcal{B}_j(v_i)$ element i, j matrice $\mathbf{X}_2$.

Penalizacija oblika (2.2) također je povezana sa svakom funkcijom: $\boldsymbol{\delta}^\top\mathbf{D}_1^\top\mathbf{D}_1\boldsymbol{\delta} = \boldsymbol{\delta}^\top\overline{\mathbf{S}}_1\boldsymbol{\delta}$ za f_1 i $\boldsymbol{\gamma}^\top\mathbf{D}_2^\top\mathbf{D}_2\boldsymbol{\gamma} = \boldsymbol{\gamma}^\top\overline{\mathbf{S}}_2\boldsymbol{\gamma}$ za f_2.

Kako bismo riješili problem identifikacije, uvodimo praktično rješenje kroz linearna ograničenja. Iako teoretski različita ograničenja mogu riješiti problem identifikacije, većina ih rezultira nepotrebno širokim intervalima pouzdanosti za ograničene funkcije. Uzeći to u obzir, najbolja ograničenja su ograničenja sume nula (eng., *"sum-to-zero constraints"*), poput:

$$\sum_{i=1}^{n} f_1(x_i) = 0, \text{ ili ekvivalentno } \mathbf{1}^\top\mathbf{f}_1 = 0$$

gdje je $\mathbf{1}$ vektor n jedinica.

Da bismo primijenili ograničenje, primijetimo da zahtijevamo da je $\mathbf{1}^\top\mathbf{f}_1 = \mathbf{1}^\top\mathbf{X}_1\boldsymbol{\delta} = 0$ za svaki $\boldsymbol{\delta}$, što implicira da je $\mathbf{1}^\top\mathbf{X}_1 = \mathbf{0}$. To se postiže normalizacijom $\mathbf{X}_1$, odnosno oduzimanjem srednje vrijednosti svakog stupca od njegovih vrijednosti:

$$\tilde{\mathbf{X}}_1 = \mathbf{X}_1 - \mathbf{1}\mathbf{1}^\top\mathbf{X}_1/n$$

Nakon ovog prilagodbe, $\tilde{\mathbf{f}}_1 = \tilde{\mathbf{X}}_1\boldsymbol{\delta}$ postaje naša nova reprezentacija funkcije. Primijetimo kako ovo ograničenje mijenja samo vertikalni pomak $\mathbf{f}_1$, na način da mu je srednja vrijednost nula, bez utjecaja na njezin oblik ili penalizaciju, to se lako provjeri na sljedeći način:

$$\tilde{\mathbf{f}}_1 = \tilde{\mathbf{X}}_1\boldsymbol{\delta} = \mathbf{X}_1\boldsymbol{\delta} - \mathbf{1}\mathbf{1}^\top\mathbf{X}_1\boldsymbol{\delta}/n = \mathbf{X}_1\boldsymbol{\delta} - \mathbf{1}c = \mathbf{f}_1 - c$$

uz definiciju skalara $c = \mathbf{1}^\top\mathbf{X}_1\boldsymbol{\delta}/n$.

Konačno, valja napomenuti da centriranje stupaca smanjuje rang $\tilde{\mathbf{X}}_1$ na $k_1 - 1$, tako da se samo $k_1 - 1$ elemenata vektora $\boldsymbol{\delta}$ s k_1 elemenata može jedinstveno procijeniti. Jednostavno ograničenje rješava ovaj problem: jedan element $\boldsymbol{\delta}$ postavlja se na nulu, a odgovarajući stupci matrica $\tilde{\mathbf{X}}_1$ i $\mathbf{D}$ se brišu.

Nakon što smo postavili ograničene baze za f_j, sada je jednostavno izraziti (2.9) kao

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

gdje je $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2)$ i $\boldsymbol{\beta}^\top = \left(\alpha, \boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top\right)$. Radi jednostavnije notacije korisno je izraziti penalizaciju kao kvadratne forme u punom vektoru koeficijenata $\boldsymbol{\beta}$, što je jednostavno učiniti dodavanjem nula u $\overline{\mathbf{S}}_j$, po potrebi. Na primjer,

$$\boldsymbol{\beta}^\top\mathbf{S}_1\boldsymbol{\beta} = \left(\alpha, \boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top\right)\begin{bmatrix} 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{S}}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \alpha \\ \boldsymbol{\delta} \\ \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{\delta}^\top\overline{\mathbf{S}}_1\boldsymbol{\delta}.$$

Procjena aditivnih modela penaliziranim najmanjim kvadratima

Koeficijenti procjene $\hat{\boldsymbol{\beta}}$ modela (2.9) dobivaju se minimizacijom cilja penaliziranih najmanjih kvadrata

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1\boldsymbol{\beta}^\top\mathbf{S}_1\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^\top\mathbf{S}_2\boldsymbol{\beta}.$$

Za sada, pretpostavimo da su parametri zaglađivanja λ_1 i λ_2 dani. Slično kao u slučaju univarijatnog modela, imamo

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda_1\mathbf{S}_1 + \lambda_2\mathbf{S}_2\right)^{-1}\mathbf{X}^\top\mathbf{y},$$

no ovi su izrazi suboptimalni s obzirom na računalnu stabilnost i bolje je preformulirati cilj kao

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1\boldsymbol{\beta}^\top\mathbf{S}_1\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^\top\mathbf{S}_2\boldsymbol{\beta} = \left\|\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix}\boldsymbol{\beta}\right\|^2, \tag{2.10}$$

gdje je

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \sqrt{\lambda_1}\mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{\lambda_2}\mathbf{D}_2 \end{bmatrix}.$$

Desna strana izraza (2.10) jednostavno je nepenalizirani cilj najmanjih kvadrata za proširenu verziju modela i odgovarajućih odziva. Stoga se model može procijeniti standardnom linearnom regresijom koristeći stabilne metode zasnovane na ortogonalnim matricama.

2.4 Generalizirani aditivni modeli

Generalizirani aditivni modeli (GAM-ovi) proizlaze iz aditivnih modela, kao što generalizirani linearni modeli proizlaze iz linearnih modela. Drugim riječima, linearni prediktor sada predviđa neku poznatu glatku funkciju očekivane vrijednosti odziva, a on može slijediti bilo koju distribuciju eksponencijalne familije. Standardni generalizirani aditivni model ima općeniti oblik

$$g(\mu_i) = \mathbf{A}_i\gamma + \sum_j f_j\left(x_{ji}\right), \quad y_i \sim \mathrm{EF}(\mu_i, \phi) \tag{2.11}$$

gdje je $\mathbf{A}_i$ i-ti red parametarske matrice modela (sadrži strogo parametarske kovarijate) s pripadajućim parametrima γ, f_j je glatka funkcija (moguće vektorska) kovarijata x_j, a $\mathrm{EF}(\mu_i, \phi)$ označava distribuciju eksponencijalne familije s očekivanom vrijednošću μ_i i parametrom skaliranja ϕ. Pretpostavlja se da su y_i neovisni uz dane μ_i.

Odabiremo zaglađujuće baze i penalizacije za svaku funkciju f_j, što implicira matrice modela $\mathbf{X}^{[j]}$ i penalizacije $\mathbf{S}^{[j]}$. Ako je $b_{jk}(x)$ k-ta bazna funkcija za f_j, tada je $X^{[j]}_{ik} = b_{jk}(x_{ji})$. Nadalje, moramo primijeniti identifikacijsko ograničenje (eng., *"identifiability constraint"*) na bilo koju glatku funkciju koja uključuje $\mathbf{1}$ u ljusci $\mathbf{X}^{[j]}$; inače će glatke komponente biti pomiješane sa slobodnim članom uključenim u $\mathbf{A}$.

Identifikacijska ograničenja, u obliku $\sum_i f_j\left(x_{ji}\right) = 0$, uključujemo u bazu reparametrizacijom (dobar primjer toga može se vidjeti u odjeljku 5.4.1 u [5]). Neka $\mathcal{X}^{[j]}$ i $\mathcal{S}_j$ označavaju redom matricu modela i matricu penalizacije za f_j nakon ove reparemetrizacije. Zatim možemo kombinirati $\mathbf{A}$ i $\mathcal{X}^{[j]}$ spajajući ih po stupcima, kako bismo stvorili cjelovitu matricu modela

$$\mathbf{X} = \left(\mathbf{A} : \mathcal{X}^{[1]} : \mathcal{X}^{[2]} : \cdots\right).$$

Odgovarajući vektor koeficijenata modela $\boldsymbol{\beta}$ sadrži γ i vektore koeficijenata pojedinih glatkih komponenti koji su složeni jedan za drugim. Ukupna penalizacija modela tada može biti napisana kao

$$\sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta},$$

gdje je λ_j parametar zaglađivanja, a $\mathbf{S}_j$ je jednostavno $\mathcal{S}_j$ ugrađen kao dijagonalni blok u matricu koja inače sadrži samo nule, tako da je $\lambda_j\boldsymbol{\beta}^\top\mathbf{S}_j\boldsymbol{\beta}$ penalizacija za f_j.

Procjena modela

Stoga se naš model pretvorio u preparametrizirani GLM

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}, \quad y_i \sim \mathrm{EF}(\mu_i, \phi)$$

koji ćemo, za razliku od aditivnog modela kojeg smo procijenili pomoću penaliziranih najmanjih kvadrata, procijeniti maksimizacijom penalizirane vjerodostojnosti

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi}\sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}, \tag{2.12}$$

pri čemu skaliramo s parametrom $\frac{1}{2\phi}$ kako se kasnije ne bi eksplicitno pojavljivao u izrazu kojeg minimiziramo. U praksi ta maksimizaicija je postignuta penaliziranom iterativnom metodom najmanjih kvadrata (eng., *"Penalized Iteratively Re-weighted Least Squares"*, kraće PIRLS).

Izvod metode je isti kao i u odjeljku 1.3 uz uključenje penalizacije. Označimo s $p(\boldsymbol{\beta}) := \sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}$. Izračunajmo prvo gradijent $\nabla_\beta p(\boldsymbol{\beta})$:

$$\nabla_\beta p(\boldsymbol{\beta}) = \sum_j \lambda_j \nabla_\beta \left(\boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}\right) = \sum_j 2\lambda_j \mathbf{S}_j \boldsymbol{\beta}.$$

Nadalje, izračunajmo i Hessijan $\nabla^2_\beta p(\boldsymbol{\beta})$:

$$\nabla^2_\beta p(\boldsymbol{\beta}) = \sum_j 2\lambda_j \nabla_\beta \left(\mathbf{S}_j \boldsymbol{\beta}\right) = \sum_j 2\lambda_j \mathbf{S}_j.$$

Dakle gradijentni vektor log-vjerodostojnosti može se zapisati kao $\nabla l_p = \mathbf{X}^\top \mathbf{W}\mathbf{G}(\mathbf{y}-\boldsymbol{\mu})/\phi - \sum_j \lambda_j \mathbf{S}_j \boldsymbol{\beta}/\phi$, a Hesseova matrica log-vjerodostojnosti postaje $\mathbf{H} = -\mathbf{X}^\top \mathbf{W}\mathbf{X}/\phi - \sum_j \lambda_j \mathbf{S}_j/\phi$, uz oznake kao i u odjeljku 1.3.

Sada jedna Newtonova iteracija ima oblik

$$\begin{aligned}
\boldsymbol{\beta}^{(\text{novi})} &= \boldsymbol{\beta}^{(\text{stari})} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(\text{stari})})\nabla l(\boldsymbol{\beta}^{(\text{stari})}) \\
&= \boldsymbol{\beta}^{(\text{stari})} + \phi(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1} \left\{ \mathbf{X}^\top \mathbf{W}\mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) - \sum_j \lambda_j \mathbf{S}_j \boldsymbol{\beta}^{(\text{stari})} \right\} /\phi \\
&= (\mathbf{X}^\top \mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1} \mathbf{X}^\top \mathbf{W} \left\{ \mathbf{G}(\mathbf{y}-\boldsymbol{\mu}) + \mathbf{X}\boldsymbol{\beta}^{(\text{stari})} \right\} \\
&= (\mathbf{X}^\top \mathbf{W}\mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{z}
\end{aligned}$$

Primijetimo da su jednadžbe iteracija zapravo procjene najmanjih kvadrata parametra $\boldsymbol{\beta}$ koje proizlaze iz minimiziranja težinske funkcije penaliziranih najmanjih kvadrata

$$\sum_{i=1}^{n} w_i \left(z_i - \mathbf{X}_i \boldsymbol{\beta}\right)^2 + \sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}.$$

Dakle PIRLS algoritam je sljedeći:

1. Inicijaliziraj $\hat{\mu}_i = y_i + \delta_i$ i $\hat{\eta}_i = g(\hat{\mu}_i)$, gdje je δ_i obično nula, ali može biti mala konstanta koja osigurava da je $\hat{\eta}_i$ konačan. Iteriraj sljedeće korake do konvergencije:

2. Izračunaj pseudopodatke $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i$, i iterativne težine $w_i = \alpha(\hat{\mu}_i) / \left\{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)\right\}$.

3. Pronađi $\hat{\boldsymbol{\beta}}$ minimizirajući najmanje kvadrate s težinama

$$\sum_{i=1}^{n} w_i (z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}$$

   i zatim ažuriraj $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ i $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

Prisjetimo se, (iz odjeljka 1.3) da je $V(\mu)$ funkcija varijance određena distribucijom eksponencijalne familije, dok je $\alpha(\mu_i) = [1 + (y_i - \mu_i)\{V'(\mu_i)/V(\mu_i) + g''(\mu_i)/g'(\mu_i)\}]$. Alternativno, možemo koristiti pristup 'Fisher scoring' u kojem se Hessian log-vjerodostojnosti zamjenjuje njegovim očekivanjem, što odgovara postavljanju $\alpha(\mu_i) = 1$.

2.5 Alternativni bayesovski pristup

U ovom ćemo odjeljku predstaviti alternativni pristup procjeni koeficijenata modela. Dosadašnji frekvencionistički pristup pretpostavljao je da postoji jedan točan skup koeficijenata β koji čine naš model približno točnim. Sada, međutim, pretpostavljamo da su koeficijenti β slučajni te da slijede određenu apriornu distribuciju (eng. *"prior"*).

Prisjetimo se, u odjeljku 2.2, u svrhu izbjegavanja overfittinga metodom najmanjih kvadrata uveli smo penalizaciju vijugavosti, za koju smo pokazali da ju možemo zapisati u obliku $\lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$, pri čemu je $\mathbf{S} = \mathbf{D}^\top\mathbf{D}$. No sada taj problem umjesto eksplicitnim uključivanjem penalizacije tijekom prilagodbe modela, kako bi se nametnulo uvjerenje da je prava funkcija vjerojatnije glatka nego vijugava, rješavamo izborom apriorne distribucije $\pi(\boldsymbol{\beta})$ koja stavlja veću masu na one koeficijente β za koje je $\lambda\beta^\top\mathbf{S}\beta$ što manji. Stoga ima smisla uzeti apriornu distribuciju

$$\pi(\boldsymbol{\beta} \mid \lambda) \propto \exp\left(-\lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}/2\sigma^2\right),$$

gdje je σ^2 uključen radi kasnije praktičnosti, a inače se može apsorbirati i u λ.

Možemo primijetiti da to ima oblik višedimenzionalne normalne *"improper"* razdiobe s očekivanjem 0 i kovarijacijskom matricom $\sigma^2/\lambda \cdot \mathbf{S}^-$, tj.

$$\boldsymbol{\beta} \sim N\left(\mathbf{0}, \sigma^2/\lambda \cdot \mathbf{S}^-\right),$$

gdje je $\mathbf{S}^-$ pseudoinverz (budući da $\mathbf{S}$ nije punog ranga zbog dimenzije nul-prostora penalizacije). Ovdje *"improper"* razdioba, znači da funkcija gustoće nije prava (tj., integral nije jednak 1), a to je upravo zato jer matrica $\mathbf{S}$ nije punog ranga pa kovarijacijska matrica višedimenzionalne normalne razdiobe ne može biti pozitivno definitna.

MAP procjenitelj

Uz dane bazne funkcije, prisjetimo se, model

$$y_i = f(x_i) + \epsilon_i, \qquad \epsilon_i \sim N\left(0, \sigma^2\right)$$

možemo izraziti kao $\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2\right)$. Dakle, sada imamo apriornu distribuciju $\pi(\boldsymbol{\beta})$ i vjerodostojnost $\pi(\mathbf{y} \mid \boldsymbol{\beta})$ pa možemo izračunati aposteriorni mod, tj. tako zvani MAP ("Maximum a posteriori probability") procjenitelj (vidi dodatak B.1). Uvrštavanjem izraza za multivarijatnu normalnu funkciju gustoće, logaritmiranjem, množenjem s -1 i zanemarivanjem nerelevantnih konstanti dobivamo sljedeće:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{MAP}}(\mathbf{y}) &= \arg\max_{\beta} \pi(\mathbf{y} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}) \\
&= \arg\max_{\beta} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \cdot e^{-\frac{\lambda}{2}\boldsymbol{\beta}^\top \mathbf{S}\boldsymbol{\beta}} \\
&= \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/\sigma^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta},
\end{aligned}$$

pri čemu smo mogli aposrbirati λ u σ^2 tako da se ne pojavljuje eksplicitno u izrazu kojeg minimiziramo. Vidimo da je izraz s desne strane koje minimiziramo upravo jednak 2.5, odnosno $\hat{\boldsymbol{\beta}}_{\text{MAP}}(\mathbf{y})$ jednak 2.6:

$$\hat{\boldsymbol{\beta}}_{\text{MAP}}(\mathbf{y}) = \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Analogija s linearnim mješovitim modelom

Definirajmo sada linearni mješoviti model (LMM) kako bismo pokazali analogiju s gam modelima, odnosno analogiju slučajnih učinaka i glatkih funkcija kod gamova. Dakle, opći linearni mješoviti model možemo prikladno zapisati kao

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N\left(\mathbf{0}, \boldsymbol{\psi}_\theta\right), \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \boldsymbol{\Lambda}_\theta\right) \tag{2.13}$$

gdje je $\boldsymbol{\psi}_\theta$ pozitivno definitna kovarijacijska matrica za slučajne efekte $\mathbf{b}$, a $\mathbf{Z}$ je matrica fiksnih koeficijenata koja opisuje kako zavisna varijabla $\mathbf{y}$ ovisi o slučajnim učincima. Na kraju, $\boldsymbol{\Lambda}_\theta$ je pozitivno definitna matrica koja obično ima jednostavnu strukturu, često

samo $\mathbf{I}\sigma^2$, ili je ponekad kovarijacijska matrica jednostavnog autoregresivnog rezidualnog modela, što daje trakastu (vrpčastu, eng. "banded") $\mathbf{\Lambda}_\theta^{-1}$ i stoga omogućava učinkovito računanje. Primijetimo da model kaže da je $\mathbf{y}$ linearna kombinacija normalnih slučajnih varijabli, što implicira da ima multivarijatnu normalnu distribuciju:

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\psi}_\theta\mathbf{Z}^\top + \mathbf{\Lambda}_\theta\right).$$

Može se pokazati (vidi npr. poglavlje 2.4. u [5]) da su aposteriorna distribucija od b i procijenjen vektor slučajnih učinaka, odnosno MAP procjenitelj $\hat{\mathbf{b}}$ jednaki:

$$\mathbf{b} \mid \mathbf{y}, \hat{\boldsymbol{\beta}} \sim N\left(\hat{\mathbf{b}}, \left(\mathbf{Z}^\top\mathbf{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1}\right)^{-1}\right)$$

$$\hat{\mathbf{b}} = \left(\mathbf{Z}^\top\mathbf{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1}\right)^{-1}\mathbf{Z}^\top\mathbf{\Lambda}_\theta^{-1}\tilde{\mathbf{y}}.$$

Primijetimo sada kako na naš model možemo gledati kao linearni mješoviti model uz ekvivalenciju $\beta = b$, $X = Z$ (uz zanemarivanje β u 2.13) te $\Psi_\theta = \sigma^2/\lambda \cdot S^{-1}$ i $\Lambda_\theta = \sigma^2 I_n$:

$$\mathbf{y} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N\left(\mathbf{0}, \sigma^2/\lambda \cdot \mathbf{S}^-\right), \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I_n}\right)$$

Sada koristeći rezultate linearnih mješovitih modela dobivamo aposteriornu distribuciju $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \mid \mathbf{y} \sim N\left(\hat{\boldsymbol{\beta}}, \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\sigma^2\right),$$

te isto tako uvrštavanjem možemo dobiti i $\hat{\boldsymbol{\beta}}$, za koji vidimo da je to upravo isti onaj procjenitelj koji smo dobili gornjim računom.

Reparametrizacija modela u LMM oblik

Pokažimo sada kako možemo glatke funkcije gam modela eksplicitno napisati u obliku linearnog mješovitog modela, što nam je korisno kako bismo mogli koristiti već razvijene metode i softverske pakete za procjenu standardnih mješovitih modela.

Modificirat ćemo parametre $\boldsymbol{\beta}$ na sljedeći način: koeficijente β_1 i β_2 ostavimo istima (oni odgovaraju linearnom dijelu naše glatke funkcije), a $\beta_3, \ldots, \beta_k$ zamijenimo vektorom $\mathbf{D}\boldsymbol{\beta}$ (pri čemu je $\mathbf{D}$ $(k-2)\times k$ matrica t.d. $\mathbf{D}^\top\mathbf{D} = \mathbf{S}$). Označimo s $\boldsymbol{\beta}^* := (\beta_1, \beta_2)$, $\mathbf{b} := \mathbf{D}\boldsymbol{\beta}$ i $\boldsymbol{\beta}' := (\boldsymbol{\beta}^*, \mathbf{b})$.

Reparametrizaciju ćemo sada provesti tako da model napišemo u terminima parametara $\boldsymbol{\beta}' = (\boldsymbol{\beta}^*, \mathbf{b}) = \mathbf{D}_+\boldsymbol{\beta}$, gdje je

$$\mathbf{D}_+ = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \\ & \mathbf{D} \end{bmatrix}.$$

Sada imamo $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{D}_+^{-1}\boldsymbol{\beta}'$ i penal postaje $\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} = (\mathbf{D}\boldsymbol{\beta})^\top\mathbf{D}\boldsymbol{\beta} = \sum_{i=3}^k \beta_i'^2 = \mathbf{b}^\top\mathbf{b}$. Primijetimo da penal ne ovisi o $\boldsymbol{\beta}^*$, a to, uz gornji bayesovski pristup, znači da ih možemo

tretirati kao fiksne učinke. Nadalje, parametre $\mathbf{b}$ penaliziramo s $\mathbf{b}^\top\mathbf{b}$, na što, kao i gore, možemo gledati kao da imaju normalnu apriornu, sada pravu, razdiobu s kovarijacijskom matricom $\mathbf{I}\sigma^2/\lambda$ (primijetimo, uz reparametrizaciju matrica $\mathbf{S}$ je sada postala identiteta), tj. $\mathbf{b} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2/\lambda\right)$.

Da bismo potpuno razjasnili vezu sa standardnim mješovitim modelom, neka $\mathbf{X}^*$ sada označava prva 2 stupca $\mathbf{X}\mathbf{D}_+^{-1}$, dok je $\mathbf{Z}$ matrica preostalih stupaca. Tada glatki model postaje

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2/\lambda\right), \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2\right).$$

Konačno, promatranje našeg GAM-a bayesovskim pristupom je donekle neizbježna posljedica ekvivalencije slučajnih učinaka i splajnova - ako smatramo da postoji neka prava glatka funkcija (*"smoother"*) koju želimo procijeniti, moramo zauzeti Bayesov pristup prema našim slučajnim učincima (splajnovima), jer ne mislimo da se prava glatka funkcija mijenja svaki put kada prikupljamo podatke, kao što bismo obično očekivali za frekvencionistički slučajni učinak.

2.6 Kriteriji za odabir glatkoće

Do sada smo razmatrali procjenu β uz dane zaglađujuće parametre λ, no njih također moramo procijeniti, što je najizazovniji dio procjene modela. Općenito se koriste dvije klase metoda: metode predikcijske greške, poput GCV-a i AIC-a, ili metode bazirane na vjerodostojnosti.

Korišteni kriteriji temeljeni na predikcijskoj grešci su obična unakrsna validacija (OCV), koju smo obradili u slučaju univarijatnog modela u odjeljku 2.2, te generalizirana unakrsna validacija (GCV) kada je parametar skaliranja nepoznat ili nepristrani procjenitelj rizika (eng., *"Un-Biased Risk Estimator"*, kraće UBRE) kada je poznat. Alternativno, za odabir glatkoće može se koristiti metoda ograničene maksimalne vjerodostojnosti (eng., *"restricted maximum likelihood)"*, kraće REML), promatrajući glatke komponente kao slučajne efekte (eng., *"random effects"*), koju ćemo obraditi u nastavku.

Za obje klase metoda postoje zatim dvije glavne alternativne računalne strategije: ili se kriterij odabira glatkoće definira za sam model i optimizira izravno, ili se kriterij odabira glatkoće definira za radni model u iteraciji prilagodbe PIRLS-a te se primjenjuje na taj radni model pri svakom koraku PIRLS-a. Iako ova druga strategija ne garantira konvergenciju, može biti iznimno brza, posebno kod velikih skupova podataka i modela. Za detalje pogledajte [5].

REML metoda

Bayesovski pogled na glatkoću kroz apriornu distribuciju također olakšava pristup procjeni parametara zaglađivanja maksimiziranjem marginalne vjerodostojnosti (zajedničke

gustoće podataka i koeficijenata β, s integriranim koeficijentima - vidi dodatak B.2), odnosno pronalaženjem njegove MAP procjene pod uniformnom apriornom distribucijom, poznat kao empirijski Bayes (eng., *"empirical Bayes"*) (za razliku od potpunog Bayesovskog pristupa u kojem se postavlja apriorna distribucija na λ i zatim dobiva odgovarajuća aposteriorna gustoća). Ovaj pristup je inače pogodan za odabir hiperparametara. Dakle tražimo parametar λ koji maksimizira sljedeći izraz:

$$\pi(\mathbf{y} \mid \lambda) = \int \pi(\mathbf{y} \mid \boldsymbol{\beta}, \lambda)\pi(\boldsymbol{\beta} \mid \lambda)d\boldsymbol{\beta}.$$

Intuitivno integral možemo interpretirati kao prosječnu vjerodostojnost slučajnih uzoraka uzetih iz apriorne distribucije $\pi(\boldsymbol{\beta} \mid \lambda)$. Biramo λ tako da slučajni uzorci iz apriorne distribucije imaju odgovarajuću razinu glatkoće kako bi se dovoljno približili podacima i imali razumno visoku vjerodostojnost. Osim u slučaju Gaussove vjerodostojnosti, integral je nerješiv egzaktno, no može se dobro aproksimirati.

Napomenimo da ova bayesovska marginalna vjerodostojnost ima potpuno isti oblik kao REML kriterij za generalizirani linearni mješoviti model (vidi odjeljak 3.4. u [5]), budući da slučajni koeficijenti u tom slučaju također imaju Gaussovu distribuciju. Stoga se ova metoda procjene zaglađujućih parametara naziva REML metoda i ona je unaprijed postavljena metoda u standardnom paketu za procjenu gam-ova.

Poglavlje 3

Splajnovi

3.1 Uvod u splajnove

Po dijelovima linearna glatka funkcija iz prethodnog poglavlja nudi potpuno prihvatljiv način prikaza glatkih funkcija u aditivnim modelima, ali postoji prostor za znatno poboljšanje. Posebno, ako prikažemo elemente glatkog modela koristeći baze *splajnova*, tada je moguće znatno smanjiti pogrešku aproksimacije funkcije. Poglavlje započinjemo razmatranjem jednodimenzionalnog zaglađivanja splajnovima, a zatim prelazimo na računalno efikasne penalizirane regresijske splajnove smanjenog ranga. Osim toga, kratko ćemo raspraviti tzv. *"thin-plate"* regresijske splajnove koji nude općenitije rješenje za procjenu glatkih funkcija s više varijabli.

Za fiksne "čvorove" (eng., *"knots"*), $\tau_0 := a < \tau_1 < \tau_2 < \cdots < \tau_K < b =: \tau_{K+1}$, splajn reda $d \in \mathbb{N}$ (s čvorovima $\tau_1, ..., \tau_K$) je svaka funkcija $g : [a, b] \to \mathbb{R}$ takva da vrijedi

1. g je polinom reda d na $C_k := [\tau_{k-1}, \tau_k], \forall k = 1, \ldots, K + 1$
2. $g^{(k)}$ je neprekidna na $\langle a, b \rangle, \forall k = 0, 1, \ldots, d - 1$

Uočimo, splajnovi reda d s K čvora čine vektorski prostor dimenzije

$$(K + 1) * (d + 1) - d * K = K + d + 1.$$

Naime uz K čvorova imamo $K + 1$ intervala, u svakom polinom reda d (što je $d + 1$ parametara) i d uvjeta za svaki čvor, tj $d * K$ ukupno.

Prikaz pomoću baznih funkcija

Pretpostavimo da je nepoznata funkcija f predstavljena splajnom s fiksnim nizom K čvorova i fiksnim stupnjem d. Budući da te funkcije čine vektorski prostor V, moguće je izraziti f

kao

$$f(X) = \sum_{k=1}^{K+d+1} \beta_k B_k(X) \tag{3.1}$$

gdje su B_k skup baznih funkcija koje definiraju V, a β_k su pripadajući koeficijenti splajna.

Ovaj način prikaza ima prednost što se procjena funkcije f svodi na procjenu koeficijenata β_k. Preciznije, izraz je linearan u odnosu na vektor koeficijenata $\beta = (\beta_1, \ldots, \beta_{K+d+1})$. Stoga se procjena funkcije f može smatrati optimizacijskim problemom koji je linearan u odnosu na transformirane varijable $B_1(X), \ldots, B_{K+d+1}(X)$, omogućujući korištenje već dobro uspostavljenih tehnika procjene za upotrebu splajnova u širokom rasponu (generaliziranih) modela multivarijatne regresije. Važno je primijetiti da modeliranje splajnovima reducira problem procjene funkcija f na procjenu malog skupa realnih koeficijenata.

Velika fleksibilnost modeliranja splajnovima dolazi uz cijenu potrebe za brojnim podesivim parametarima. Dva od tih, izbor baznih funkcija B i stupanj d osnovnih polinoma, pokazalo se da imaju malo utjecaja. Zapravo, prilagodbe splajnovima pokazuju iznenađujuću otpornost na promjene stupnja d. Kubični polinomi $d = 3$ su uobičajeni standard jer rezultiraju krivuljama koje izgledaju savršeno glatko ljudskom oku. Prema definiciji, izbor između dvije skupine baznih funkcija B i B^* ne utječe na predikcije dobivene prilagodbom, stoga se sve svodi na pitanja praktičnosti upotrebe.

3.2 Popularne baze splajnova

Budući da je prostor splajnova određenog reda i niza čvorova vektorski prostor, postoji mnogo ekvivalentnih baza za njihovo prikazivanje (kao što je to slučaj i s običnim polinomima), pri čemu se različite baze splajnova razlikuju s obzirom na njihova numerička svojstva. U ovom odjeljku predstavit ćemo neke od najpopularnijih baza splajnova, naime tzv. *"truncated power series"* bazu, bazu B-splajnova i bazu kardinalnih splajnova.

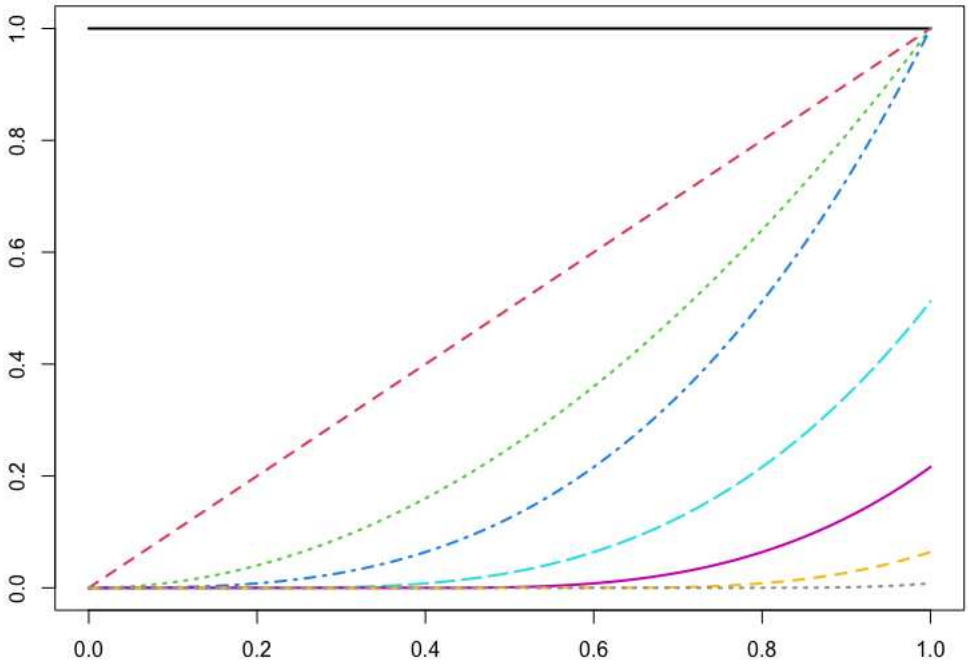
Truncated power series

"Truncated power series" baza stupnja d, s čvorovima $\tau_1, \ldots, \tau_K$ definirana je pomoću baznih funkcija:

$$B_1(x) = 1, B_2(x) = x, \ldots, B_{d+1}(x) = x^d$$
$$B_{d+2}(x) = (x - \tau_1)_+^d, \ldots, B_{K+d+1} = (x - \tau_K)_+^d$$

Prednost navedenih baznih funkcija je njihova jednostavna interpretacija: Počevši s "osnovnim" polinomom stupnja d definiranim na intervalu $[a, b]$ (prvi redak jednadžbe), odstupanja od osnovnog polinoma postupno se dodaju splajn funkciji desno od svakog od K čvorova (drugi redak).

Jedna od karakteristika "truncated power series" baze je to što nosači funkcija nisu lokalni, pri čemu su neke od funkcija B_k definirane preko cijelog raspona podataka $[a, b]$. To može dovesti do visokih korelacija između nekih baznih splajnova, što implicira numeričke nestabilnosti u procjeni splajnova. Neka x predstavlja neka opažanja u intervalu $[0, 1]$. "Truncated power series" baza trećeg stupnja s pet jednako razmaknutih čvorova duž raspona x je ilustrirana na slici 3.1.



Slika 3.1: "Truncated power" bazne funkcije splajna trećeg stupnja (d=3) s pet ekvidistantnih čvorova (K=5).

B-splajnovi

Iako je "truncated power series" baza konceptualno jednostavna, numerički nije previše privlačna zbog potenciranja velikih brojeva koje može dovesti do ozbiljnih problema s zaokruživanjem. Stoga u praksi, za predstavljanje kubičnih splajnova (i splajnova višeg ili nižeg reda), često koristimo B-splajn bazu, koja omogućuje učinkovite izračune čak i

pri velikom broju čvorova K. Ova baza je posebno privlačna zbog svoje stroge lokalnosti baznih funkcija, koje su različite od nule samo na intervalima između $d+2$ susjedna čvora, gdje je d red baze (npr., $d = 3$ za kubični splajn).

Da bismo definirali B-splajn bazu prvo trebamo proširiti niz od K unutarnjih čvorova, $x_1, \ldots, x_K$, s $d+1$ vanjskih čvorova sa svake strane. Pri tom dobivamo niz od ukupno $K + 2d + 2$ čvora. Interval na kojem će se splajn procijeniti leži unutar $[\tau_{d+1}, \tau_{K+d+2}]$ (tako da su položaji prva i posljednja d čvora u suštini proizvoljna no uobičajeno ih je staviti sve iste i jednake τ_{d+1} i τ_{K+d+2} redom).

Definirajmo sada prošireni niz čvorova $\tau_1, \ldots, \tau_{K+2d+2}$:

- $\tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_d \leqslant \tau_{d+1} < x_1$
- $\tau_{d+1+i} = x_i, i = 1, \ldots, K$
- $x_K < \tau_{K+d+2} \leqslant \tau_{K+d+3} \leqslant \ldots \leqslant \tau_{K+2d+2}$

Splajn reda d se dakle može reprezentirati kao:

$$f(x) = \sum_{i=1}^{K+d+1} B_i^d(x)\beta_i$$

pri čemu su B-splajn bazne funkcije rekurzivno definirane na sljedeći način:

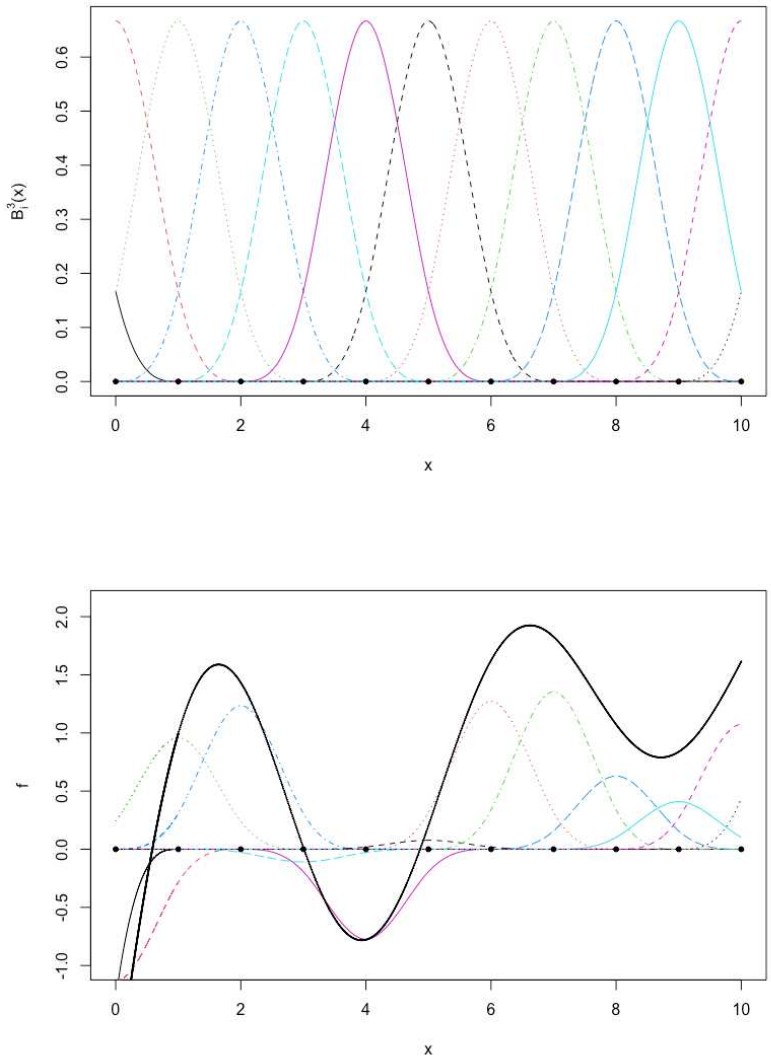
$$B_i^d(x) = \frac{x - \tau_i}{\tau_{i+d} - \tau_i} B_i^{d-1}(x) + \frac{\tau_{i+d+1} - x}{\tau_{i+d+1} - \tau_{i+1}} B_{i+1}^{d-1}(x) \quad i = 1, \ldots, K + d + 1$$

$$B_i^0(x) = \begin{cases} 1 & \text{za } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{inače} \end{cases} \qquad i = 1, \ldots, K + 2d + 1$$

Svojstvo lokalnog nosača B-splajn baze rezultira visokom numeričkom stabilnosti, kao i efikasnim algoritmom za njenu konstrukciju. B-splajn bazne funkcije, reda 3, prikazane su na prvom grafu na slici 3.2. Drugi graf iste slike ilustrira reprezentaciju neke glatke funkcije f koristeći prikazane bazne funkcije.

Prirodni kubični i kardinalni splajnovi

Polinomijalni splajnovi kao što su kubični ili B-splajnovi mogu biti nepredvidivi na rubovima podataka, a ekstrapolacija može biti opasna. Polinomi koji se prilagođavaju izvan granica čvorova ponašaju se još ekstremnije nego odgovarajući globalni polinomi u tom području. Da bi se riješio ovaj problem, uvodimo prirodne splajnove. To su kubični splajnovi koji imaju dodatna ograničenja da su linearni izvan granica čvorova. To se postiže zahtjevom da funkcija splajna f zadovoljava $f'' = f''' = 0$ na $(-\infty, \tau_1], [\tau_K, +\infty)$. Time se

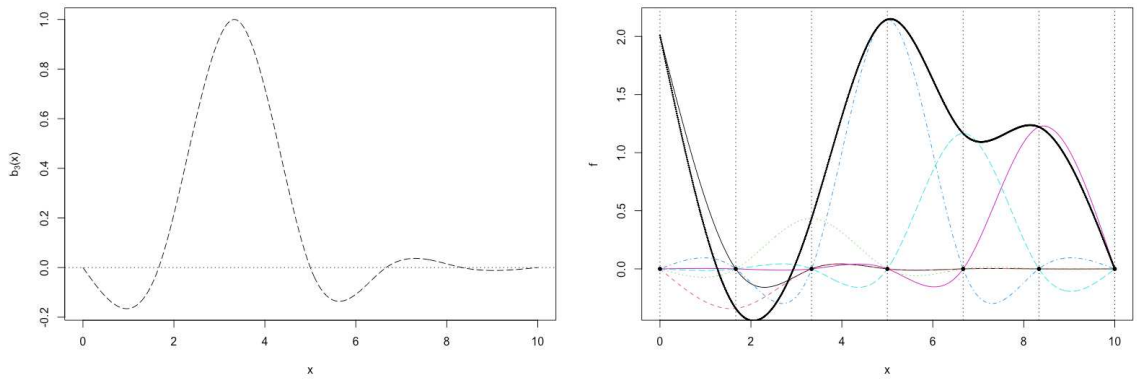


Slika 3.2: Ilustracija reprezentacije glatke funkcije f koristeći B-splajn bazu. Tanke krivulje prikazuju B-splajn bazne funkcije pomnožene s pripadajućim koeficijentima, koje u sumi daju sam splajn, prikazan podebljanom crnom linijom. Primijetimo kako je svaka bazna funkcije ne nula samo na 4 intervala, tj između 5 susjednih čvorova prikazanih podebljanim točkama.

oslobađaju četiri stupnja slobode (po dva ograničenja u oba granična područja), što rezultira s ukupno K stupnja slobode za prirodni kubični splajn.

Jedna baza za prirodne kubične splajnove je kardinalna baza splajnova (eng., *"cardinal spline basis"*). K baznih funkcija kardinalnih splajnova (svaki stupnja $d = 3$) definirane su njihovim vrijednostima u čvorovima $\tau_1, \ldots, \tau_K$. Preciznije, definirane su tako da k-ta bazna funkcija zadovoljava $B_k(\tau_k) = 1$ i $B_k(\tau_j) = 0$, za $\tau_j \neq \tau_k$. Posljedično, koeficijenti β_k imaju jednostavnu interpretaciju: svaki koeficijent je jednak vrijednosti funkcije splajna f u čvoru τ_k. Za učinkovitu konstrukciju baze kardinalnih splajnova upućujemo na [5], poglavlje 4.

Kada prilagođavamo model, procjenjujemo koeficijent za svaku od tih baznih funkcija. Konačni splajn dobiva se kao težinski zbroj baznih funkcija, pri čemu se procijenjeni koeficijenti koriste kao težine. To je ilustrirano na desnom grafu slike 3.3. Važno je primijetiti kako postoji veliki diskontinuitet u vrijednosti splajna na krajevima podataka. Ako bi x predstavljao nešto poput dana u godini ili mjeseca, takav diskontinuitet bi mogao biti problematičan. Ovdje na scenu stupa baza cikličnog kubičnog splajna.

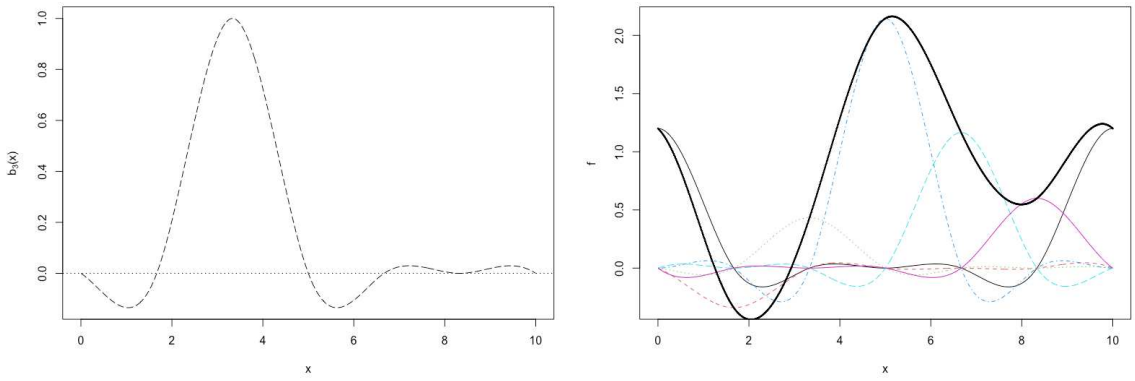


Slika 3.3: Na lijevom grafu prikazana je jedna bazna funkcija kardinalnih splajnova, $b_2(x)$, koja poprima vrijednost jedan u jednom specifičnom čvoru splajna, a nula u svim ostalima. Na desnom grafu vidimo kako se te bazne funkcije kombiniraju s ciljem formiranja glatke krivulje. Različite obojane tanje krivulje prikazuju bazne funkcije, $b_j(x)$, kubičnog regresijskog splajna, svaka pomnožena s odgovarajućim koeficijentom β_j. Zbroj ovih skaliranih baznih funkcija rezultira glatkom krivuljom prikazanom debelom neprekinutom linijom. Vertikalne tanke linije označavaju položaje čvorova.

Ciklični kubični splajn

Često je prikladno da model glatke funkcije bude "cikličan", što znači da funkcija ima istu vrijednost i prvih nekoliko derivacija na svojim granicama. Na primjer, u većini primjena ne bi bilo prikladno da glatka funkcija dana u godini doživi diskontinuitet na kraju godine.

Kubični regresijski splajn iz prethodnog odjeljka može se modificirati kako bismo dobili takvu glatku funkciju. Ciklični splajn i njegove bazne funkcije prikazane su na slici 3.4. Slična ograničenja mogu se primijeniti i na druge vrste splajnova, kao što su ciklični B-splajnovi.



Slika 3.4: Na lijevom grafu prikazana je jedna bazna funkcija cikličnog kubičnog splajna, $b_3(x)$, koja poprima vrijednost jedan u jednom specifičnom čvoru splajna, a nula u svim ostalima. Primijetimo kako se vrijednosti bazne funkcije i prve dvije derivacije podudaraju u točkama $x = 0$ i $x = 1$. Budući da se sve bazne funkcije glatko spajaju na krajevima intervala x, isto vrijedi i za prilagođeni ciklični kubični splajn, prikazan debelom crnom krivuljom na desnom grafu.

3.3 Penalizirani splajnovi

Do sada predstavljeni splajnovi često se nazivaju regresijski splajnovi. Osim izbora baze splajna (B-splajn, *"truncated power series"*, kardinalni splajnovi itd.), potrebno je odabrati broj čvorova i položaje čvorova. Očito je da ti podesivi parametri mogu značajno utjecati na procijenjeni oblik funkcije splajna. Primjerice, veliki broj čvorova implicira veliku fleksibilnost, ali može dovesti i do prekomjernog prilagođavanja podacima. S druge strane, mali broj čvorova može rezultirati "prezaglađenom" procjenom sklonom pristranosti zbog nedovoljnog prilagođavanja. Popularan pristup olakšanju izbora položaja čvorova u modeliranju splajnovima je upotreba penaliziranih splajnova.

Na uzorku nezavisnih i jednako distribuiranih podataka $(x_1, y_1), \ldots (x_n, y_n)$, penalizirani splajn je rješenje problema

$$\hat{\beta} = \text{argmax}_\beta \left[l_\beta (x_1, y_1, \ldots, x_n, y_n) - \lambda \cdot J_\beta \right],$$

gdje l_β označava log-vjerodostojnost kada odzive y_i procjenjujemo splajnom f_β, dobivenim kada za koeficijente baze uzmemo baš β, a J_β je penalizacija vijugavosti koja nameće glatkoću (tj. kažnjava vijugavost u f_β). Općenito, penalizirani splajnovi temelje se na ideji da se nepoznata funkcija f modelira splajnom s velikim brojem čvorova, što omogućava visok stupanj fleksibilnosti. S druge strane, gruba procjena splajna koja ima visoku vrijednost l_β i bliska je vrijednostima podataka rezultira velikom vrijednošću J_β. Maksimizacija ove funkcije stoga implicira kompromis između glatkoće i prilagođavanja modela koji kontrolira podesiv parametar (eng., *"tuning parameter"*) $\lambda \geq 0$. Poseban slučaj je problem penaliziranih najmanjih kvadrata

$$\hat{\beta} = \text{argmin}_\beta \left[\sum_{i=1}^{n} \left(y_i - f_\beta (x_i) \right)^2 + \lambda \cdot \int \left(f''(x) \right)^2 dx \right], \tag{3.2}$$

pri čemu je prvi izraz, tako zvana funkcija gubitka zapravo suma kvadrata reziduala (eng., *"residual sum of squares"*, kraće RSS),

3.4 Smoothing splajn

Dakle, sada promatramo sljedeći problem: među svim funkcijama $f(x)$ koje su dva puta neprekidno diferencijabilne nađi onu koja minimizira penaliziranu sumu kvadrata:

$$\sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 + \lambda \int \left(f''(x) \right)^2 dx \tag{3.3}$$

gdje je λ nenegativan podesiv parametar.

Prije pronalaska funkcije koja minimizira (3.3), tzv. *"smoothing spline"* objasnimo intuiciju iza odabira penalizacije. Prva derivacija $f'(t)$ mjeri nagib funkcije u točki t, a druga derivacija odgovara količini promjene nagiba. Dakle, općenito govoreći, druga derivacija funkcije mjeri njezinu vijugavost: ona je velika po apsolutnoj vrijednosti ako je $f(t)$ vrlo vijugava blizu t, a inače je bliska nuli. (Druga derivacija pravca je nula, savršeno je gladak.) Integral možemo shvatiti kao zbrajanje preko raspona t. Drugim riječima, $\int f''(t)^2 dt$ jednostavno je mjera ukupne promjene funkcije $f'(t)$ duž njenog cijelog raspona. Ako je f vrlo glatka, tada će $f'(t)$ biti blizu konstante, a $\int f''(t)^2 dt$ će zauzeti malu vrijednost. Suprotno tome, ako je f skakutava i promjenjiva, tada će $f'(t)$ značajno varirati, a $\int f''(t)^2 dt$ će zauzeti veliku vrijednost. Stoga, u (3.3), izraz $\lambda \int f''(t)^2 dt$ potiče f da bude glatka.

Odnos između prilagodbe modela i glatkoće kontrolira se zaglađujućim parametrom (eng. *"smoothing parameter"*), λ. Velike vrijednosti λ induciraju glađe krivulje, dok manje vrijednosti proizvode vijugave krivulje. U jednoj krajnosti kako $\lambda \to 0$, penalizacija postaje nevažna, pa će rješenje biti jako vijugavo i točno će interpolirati dana opažanja,

tj. rješenje će težiti prema interpolirajućoj funkciji koja je dvaput derivabilna. Na drugom kraju, kako $\lambda \to \infty$ penalizacija dominira, prisiljavajući $f''(x) = 0$ posvuda, i tako rješenje postaje savršeno glatko - pravac koji najbolje opisuje podatke. Zapravo, u ovom slučaju, rješenje će biti pravac dobiven metodom najmanjih kvadrata, budući da funkcija gubitka u (3.3) znači minimiziranje sume kvadrata reziduala. Za srednju vrijednost λ, f će se prilagoditi opažanjima, ali će zadržati glatkoću. Vidimo da λ kontrolira ravnotežu između pristranosti i varijance za smoothing splajn.

Zanimljivo je da izraz (3.3) ima eksplicitnog, jedinstvenog minimizatora i taj minimizator je prirodni kubični splajn s čvorovima u svakoj točki podataka, jedinstvenim vrijednostima x_i. Pokažimo to u sljedećem teoremu.

Teorem 3.4.1. *Pretpostavimo da je $n \geq 2$ te $x_1 < \cdots < x_n$ (dakle, $x_i \neq x_j, \forall i \neq j$). Funkcija f koja minimizira*

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{x_1}^{x_n} (f''(x))^2 \, dx$$

među svim dva puta neprekidno diferencijabilnim funkcijama, je prirodni kubični splajn s čvorovima $x_i, i = 1, \ldots, n, \forall \lambda > 0$.

Dokaz. Neka je f proizvoljna dva puta neprekidno diferencijabilna funkcija. Postoji jedinstveni prirodni kubični splajn g s čvorovima u svakoj točki x_i takav da vrijedi $g(x_i) = f(x_i), \forall i = 1 \ldots n$. Naime, kako je prostor prirodnih kubičnih splajnova n dimenzionalan, možemo odrediti koeficijente tako da g interpolira točke $\{x_i, f(x_i)\}$. Tvrdnja će slijediti ako pokažemo

$$\int_{x_1}^{x_n} f''(x)^2 dx \geq \int_{x_1}^{x_n} g''(x)^2 dx.$$

Definirajmo $h(x) = f(x) - g(x)$. Sada koristeći supstituciju $f(x) = h(x) + g(x)$ dobivamo

$$\begin{aligned} \int_{x_1}^{x_n} f''(x)^2 dx &= \int_{x_1}^{x_n} \{g''(x) + h''(x)\}^2 \, dx \\ &= \int_{x_1}^{x_n} g''(x)^2 dx + 2 \int_{x_1}^{x_n} g''(x) h''(x) dx + \int_{x_1}^{x_n} h''(x)^2 dx. \end{aligned}$$

Zatim parcijalno intergrirajući drugi izraz u drugom redu dobivamo

$$\begin{aligned}\int_{x_1}^{x_n} g''(x)h''(x)dx &= \left[\begin{array}{ll} du = h'' & v = g'' \\ u = h' & dv = g^{(3)} \end{array}\right] \\ &= g''(x_n)\,h'(x_n) - g''(x_1)\,h'(x_1) - \int_{x_1}^{x_n} g'''(x)h'(x)dx \\ &= -\int_{x_1}^{x_n} g'''(x)h'(x)dx = -\sum_{i=1}^{n-1} g'''(x_i^+)\int_{x_i}^{x_{i+1}} h'(x)dx \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+)\{h(x_{i+1}) - h(x_i)\} = 0\end{aligned}$$

pri čemu jednakost 2. i 3. reda slijedi iz činjenice da $g''(x_1) = g''(x_n) = 0$. Jednakost u 3. redu vrijedi zbog činjenice da je $g(x)$ napravljen od dijelova kubičnih polinoma tako da je $g'''(x)$ konstanta unutar svakog intervala $[x_i, x_{i+1}]$; x_i^+ označava jedan element takvog intervala. Iz definicije prirodnog kubičnog splajna g vrijedi $g(x_i) = f(x_i)$, to jest $h(x_i) = 0, \forall i = 1 \ldots n$, iz čega slijedi da je zadnja jednakost jednaka nuli.

Dakle pokazali smo sljedeće

$$\int_{x_1}^{x_n} f''(x)^2 dx = \int_{x_1}^{x_n} g''(x)^2 dx + \int_{x_1}^{x_n} h''(x)^2 dx \geq \int_{x_1}^{x_n} g''(x)^2 dx$$

pri čemu vrijedi jednakost ako i samo ako $h''(x) = 0$ za $x_1 < x < x_n$. Međutim, $h(x_1) = h(x_n) = 0$, stoga zapravo imamo jednakost ako i samo ako $h(x) = 0$ na $[x_1, x_n]$.

□

Dakle, umjesto da se odabere unaprijed, baza kubičnih splajnova prirodno proizlazi iz izraza kojeg želimo minimizirati definiranog u 3.3, gdje su prilagodba modela i glatkoća precizno određeni na način koji ne ovisi o bazi. Smoothing splajnovi čine se kao gotovo idealni zaglađivači. Jedini značajni problem je činjenica da imaju onoliko slobodnih parametara koliko ima podataka koji se trebaju zagladiti. To je nepotrebno, s obzirom na to da će u praksi parametar λ gotovo uvijek biti dovoljno velik da rezultirajući splajn bude znatno glađi nego što bi to sugeriralo n stupnjeva slobode. O alternativi će biti riječ u sljedećem odjeljku.

Prisjetimo se sada, prirodni kubični splajnovi mogu se prikazati pomoću baze kardinalnih splajnova, kao

$$f_\beta(x) = \sum_{i=1}^{K} b_i(x)\beta_i$$

pretpostavljajući da imamo K čvorova. Može se dodatno pokazati da se penalizacija J_β može izraziti kao $\beta^\top S\beta$ s odgovarajuće definiranom penalizacijskom matricom S. U ovom

slučaju:

$$J_\beta = \int_{x_1}^{x_k} f''(x)^2 dx = \boldsymbol{\beta}^\top S \boldsymbol{\beta}.$$

Stoga je rješenje (3.3) dano penaliziranom procjenom najmanjih kvadrata

$$\hat{\beta} = \left(B^\top B + \lambda\Omega\right)^{-1} B^\top y,$$

gdje je B matrica dimenzija $n \times n$ koja sadrži bazne funkcije prirodnih splajnova evaluirane u podatacima. Vektor y sadrži vrijednosti odziva $y_1, \ldots, y_n$. Umjesto specificiranja baze prirodnih splajnova za f, dalje je moguće raditi s neograničenom bazom B-splajnova, jer penalizacija u (3.2) automatski nameće ograničenja linearnosti u čvorovima $x_{(1)}$ i $x_{(n)}$.

Penalizirani regresijski splajnovi niskog ranga

Ako je n velik i interval $[a, b]$ gusto pokriven opaženim podacima, obično nije potrebno postaviti čvorove u svaki $x_i, i = 1, \ldots, n$. Umjesto toga, smoothing splajn može se aproksimirati penaliziranim regresijskim splajnom koji koristi reducirani skup čvorova. To je dobar kompromis između očuvanja dobrih svojstava splajnova i računalne efikasnosti. U najosnovnijem obliku, to uključuje kreiranje baze splajnova (i pripadne penalizacije) za znatno manji skup podataka od onog koji će se analizirati, a zatim korištenjem te baze (plus penalizacije) modelira se izvorni skup podataka. Vrijednosti kovarijata u manjem skupu podataka trebale bi biti raspoređene tako da dobro pokrivaju distribuciju vrijednosti kovarijata u izvornom skupu podataka.

P-splajnovi

Jedna vrlo popularna klasa penaliziranih regresijskih splajnova su P-splajnovi, koji se temelje na bazi kubičnih B-splajnova i na 'velikom' skupu ekvidistantnih čvorova. P-splajnovi koriste penalizaciju razlike (eng., *"difference penalty"*) koja se izravno primjenjuje na parametre β_i. Primjer toga je penalizacija kvadrata udaljenosti između vrijednosti susjednih koeficijenata β_i i u tom slučaju ona iznosi:

$$\mathcal{P} = \sum_{i=1}^{k-1} (\beta_{i+1} - \beta_i)^2 = \boldsymbol{\beta}^\top \mathbf{P}^\top \mathbf{P} \boldsymbol{\beta}$$

gdje je

$$\mathbf{P} = \begin{bmatrix} -1 & 1 & 0 & . & . \\ 0 & -1 & 1 & 0 & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \text{ tako da je } \begin{bmatrix} \beta_2 - \beta_1 \\ \beta_3 - \beta_2 \\ . \\ . \end{bmatrix} = \mathbf{P}\boldsymbol{\beta},$$

i stoga

$$\mathcal{P} = \boldsymbol{\beta}^\top \mathbf{P}^\top \mathbf{P} \boldsymbol{\beta} = \boldsymbol{\beta}^\top \begin{bmatrix} 1 & -1 & 0 & . & . \\ -1 & 2 & -1 & . & . \\ 0 & -1 & 2 & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \boldsymbol{\beta}.$$

Jedna praktična osobina P-splajnova je da su numerički stabilni te ih je vrlo lako definirati i implementirati. Posebno, jednostavnije je postaviti matricu razlika P nego matricu S. Osim toga, omogućuju veliku fleksibilnost, budući da se bilo koji red penalizacije može kombinirati s bilo kojim redom B-splajn baze. Međutim, njihova mana je u tome što su, u usporedbi s uobičajenim penalizacijama za splajnove (*"derivative penalties"*), diskretne penalizacije manje interpretabilne u smislu svojstava prilagođene glatke krivulje.

3.5 "Thin plate" splajnovi

Smoothing splajnovi su poseban slučaj šire klase tzv. *"thin plate"* splajnova, koji omogućavaju proširenje kriterija u 3.3 na višedimenzionalne x_i.

Dosad obrađene baze korisne su u praksi, ali su podložne nekim kritikama. Primjerice, potrebno je odabrati lokacije čvorova: to uvodi dodatnu razinu subjektivnosti u proces prilagođavanja modela te su korisne samo u slučaju jedne prediktorne varijable. U ovom odjeljku razvija se pristup koji djelomično rješava ove probleme, proizvodeći baze bez čvorova za glatke funkcije bilo kojeg broja prediktornih varijabli, koje su u određenom smislu "optimalne".

"Thin plate" regresijski splajnovi vrlo su elegantno i općenito rješenje problema procjene glatke funkcije više prediktorskih varijabli iz zašumljenih observacija funkcije u određenim vrijednostima tih prediktora. Razmotrimo problem procjene glatke funkcije $g(\mathbf{x})$ iz n observacija $(\mathbf{x}_i, y_i)$ takvih da

$$y_i = g(\mathbf{x}_i) + \epsilon_i$$

gdje je ϵ_i slučajna greška, a $\mathbf{x}$ je p-dimenzionalni vektor. "Thin-plate" splajn procjenjuje funkciju g pronalazeći funkciju $\hat{f}$ koja minimizira

$$\|\mathbf{y} - \mathbf{f}\|^2 + \lambda J_{mp}(f),$$

gdje je $\mathbf{y}$ vektor y_i, $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)]^\top$, a m red derivacije takac da $2m > p$. $J_{mp}(f)$ je penalizacijski funkcional koji ovdje preuzima ulogu penalizacije iz 3.2 u slučaju više kovarijata. Definiramo ga kao

$$J_{mp} = \int_{\Re^p} \sum_{\nu_1 + \ldots + \nu_p = m} \frac{m!}{\nu_1! \ldots \nu_p!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} \ldots \partial x_p^{\nu_p}} \right)^2 dx_1 \ldots dx_p.$$

Općeniti oblik penalizacije može biti pomalo zastrašujući, pa ju pogledajmo na primjeru funkcije dvaju prediktora mjerenjem zakrivljenosti korištenjem drugih derivacija:

$$J_{22} = \iint \left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 dx_1 dx_2.$$

"Thin plate" splajn, $\hat{f}$, je nešto poput idealnog zaglađivača; zadržava sva dobra svojstva kao i već spomenuti smoothing splajnovi, uz dodatna svojstva da se mogu nositi s bilo kojim brojem prediktornih varijabli i omogućuju određenu fleksibilnost u odabiru reda derivacije koji se koristi u mjerenju zakrivljenosti funkcije. No, slično problemu koji smo ranije raspravljali kod zaglađujućih splajnova, i ovi zaglađivači nailaze na problem: računalne troškove. Imaju onoliko nepoznatih parametara koliko i podataka, što može biti neefikasno u praksi.

Kako bi se to riješilo, postoji alternativa u obliku aproksimacije niskog ranga koja oponaša glatkoću *"thin plate"* splajna, ali bez prekomjerne računalne zahtjevnosti, s bazom bilo kojeg niskog ranga. Osim toga, čuvaju karakteristiku rotacijske invarijance (izotropije) punog *"thin plate"* splajna, što znači da daju iste predikcije odzivne varijable bez obzira na rotaciju ili refleksiju prediktornih varijabli. Za više detalja pogledajte odjeljak 5.5 u [5], te zgodnu vizualizaciju možete vidjeti u [3].

Poglavlje 4

Praktični primjeri

4.1 Zagađenje zraka u Chicagu

U ovom odjeljku bavit ćemo se jednim praktičnim primjerom, a to je modeliranje ovisnosti zdravlja o zagađenju zraka na primjeru podataka iz grada Chicaga. Konkretno, ovisnost broja smrtnih slučajeva u danu, `death`, o razini ozona, `o3median`, razini sumporovog dioksida, `so2median`, prosječnoj dnevnoj temperaturi, `tmpd`, i razini štetnih čestica u zraku, `pm10median` (kao što su proizvedene ispušnim plinovima dizela, na primjer). Osim ovih varijabli kvalitete zraka, osnovna stopa smrtnosti varira s vremenom (posebno tijekom godine), iz razloga koji imaju malo ili nemaju uopće veze s kvalitetom zraka. Podaci su dani u tablici podataka `chicago` iz paketa `gamair`. Grafički prikaz možemo vidjeti na slici 4.1.

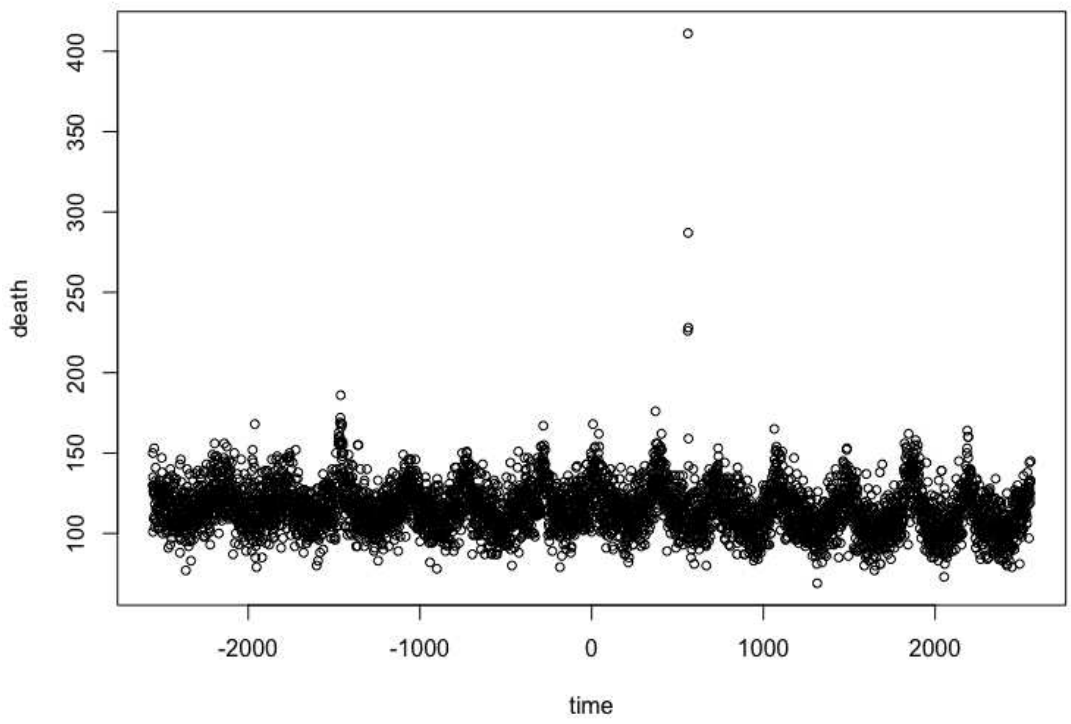
Kako je dnevni broj umrlih osoba diskretna veličina, pretpostavit ćemo da se radi o Poissonovoj slučajnoj varijabli koja ovisi o vremenski promjenjivoj stopi smrtnosti, modificiranoj množenjem efektima vezanim za zagađenje zraka. Tj. radi se o sljedećem modelu:

$$\log\{\mathbb{E}(\texttt{death}_i)\} = f(\texttt{time}_i) + \beta_1 \texttt{pm10median}_i + \beta_2 \texttt{so2median}_i + \beta_3 \texttt{o3median}_i + \beta_4 \texttt{tmpd}_i,$$

pri čemu $\texttt{death}_i$ prati Poissonovu distribuciju, a f je glatka funkcija. Model se lako procijeni pozivom `gam` funkcije iz `mgcv` paketa na sljedeći način:

```
ap0 <- gam(death~s(time,bs="cr",k=200)+pm10median+so2median+
           o3median+tmpd,data=chicago,family=poisson)
```

Ovdje, `s(time,bs="cr",k=200)` znači da procijenjujemo glatku funkciju varijable `time` iz prostora kubičnih regresijskih splajnova dimenzije 200. Nadalje, model provjerimo korištenjem `gam.check` funkcije, koja prima procijenjeni gam objekt, a vraća neke informacije o konvergenciji optimizacije izbora parametra zaglađivanja. Osim toga, provodi dijagnostičke testove o adekvatnosti izbora dimenzije baze i reproducira 4 grafa o rezidualima. Oni su redom: "QQ plot" koji prikazuje reziduale devijance u odnosu na približne

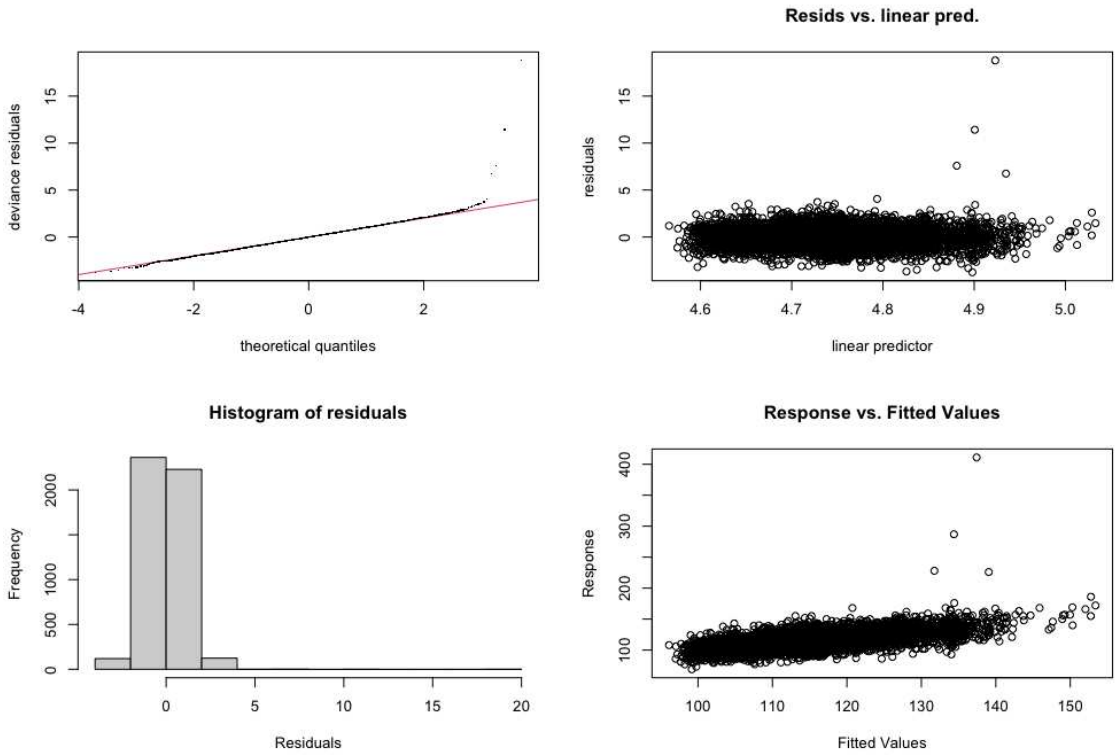


Slika 4.1: Graf broja umrlih kroz vrijeme u Chicagu

teorijske kvantile njihove distribucije, koja ovisi o prilagođenom modelu, graf reziduala u ovisnosti o linearnom prediktoru, histogram reziduala i graf prilagođenih vrijednosti naspram stvarnim vrijednostima. Grafovi provjere su prikazani na slici 4.2 i ukazuju na očite probleme koji su rezultat nekoliko značajnih odskočnika.

Prikažimo sada grafički procijenjene glatke funkcije vremena s i bez parcijalnih reziduala koji naglašavaju veličinu odskočnika. Pri tom ćemo koristiti `plot.gam` funkciju koja uzima procijenjeni gam objekt dobiven pozivom funkcije `gam()` i crta komponentne glatke funkcije koje ga sačinjavaju, na skali linearnog prediktora. Opcionalno crta i grafove za parametarske varijable u modelu. Parcijalne reziduale dobivamo stavljanjem argumenta `residuals = TRUE`. Oni su jednostavno, ranije spomenuti, Pearsonovi reziduali dodani na glatke komponente evaluirane u odgovarajućim vrijednostima kovarijata, tj. reziduali koji bi se dobili izostavljanjem dotičnog člana iz modela, dok su sve ostale procjene fiksne. Na primjer, parcijalni reziduali prikazani na donjem grafu slike 4.3 dani su formulom

$$\hat{\epsilon}_i^{\text{parcijalni}} = f(\texttt{time}_\text{i}) + \hat{\epsilon}_i^p$$



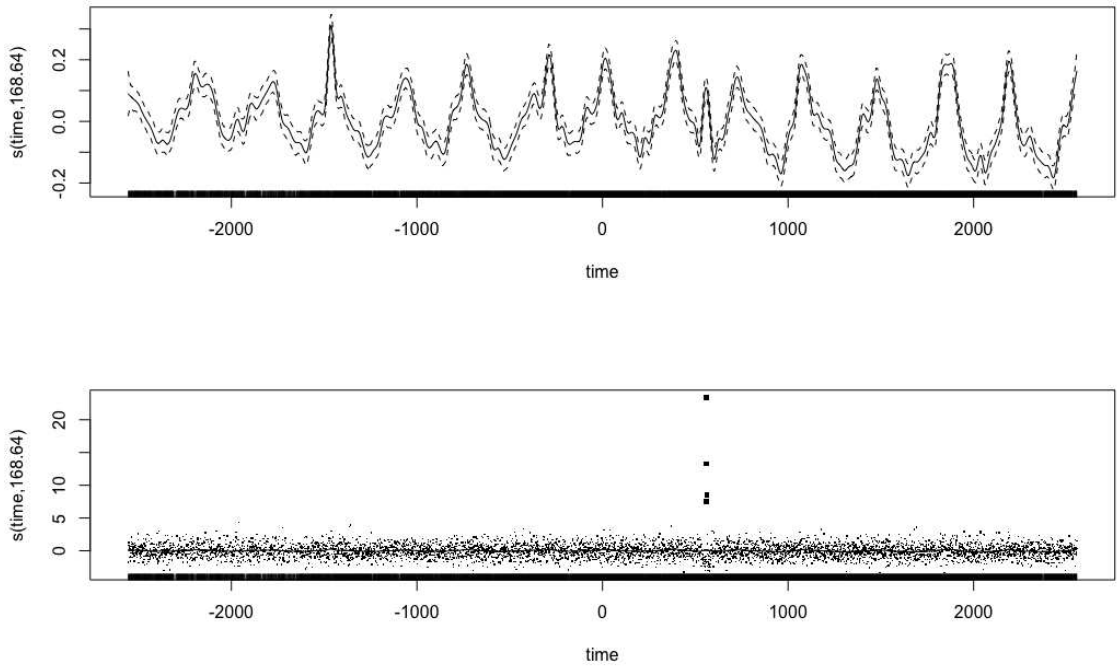
Slika 4.2: Grafovi dobiveni iz gam.check funkcije ap0 gam modela. Za Poissonove podatke s umjereno visokim srednjim vrijednostima, distribucija standardiziranih reziduala bi trebala biti blizu normalne, što nam ukazuje da je QQ-plot očito problematičan. Kao što se vidi na svim grafovima, ima nekoliko očitih odskočnika koji su jako problematični u ovom modelu.

gdje je $\hat{\epsilon}_i^p$ Pearsonov rezidual. Za dobro prilagođen model, parcijalni reziduali bi trebali biti ravnomjerno raspršeni oko krivulje na koju se odnose. Željene grafove dobivamo pozivom sljedećih funkcija, pri čemu parametar `n` određuje broj točaka korištenih za svaki jednodimenzionalni graf, pa je povećan radi dobivanja glatkog prikaza.

```
plot(ap0,n=1000)
plot(ap0,residuals=TRUE,n=1000)
```

Grafovi su prikazani na slici 4.3 na kojoj su jasno vidljiva četiri značajna ekstrema, u neposrednoj blizini jedan drugog. Ispitivanjem podataka zaključujemo da su odskočnici zapravo 4 najveća broja umrlih u danu i da su se dogodila u uzastopnim danima.

```
> chicago$death[3111:3125]
```



Slika 4.3: Procjena glatke funkcije iz modela ap0 prikazana je sa i bez parcijalnih reziduala. Primijetimo 4 velika odskočnika koja su očito u blizini jedno drugome.

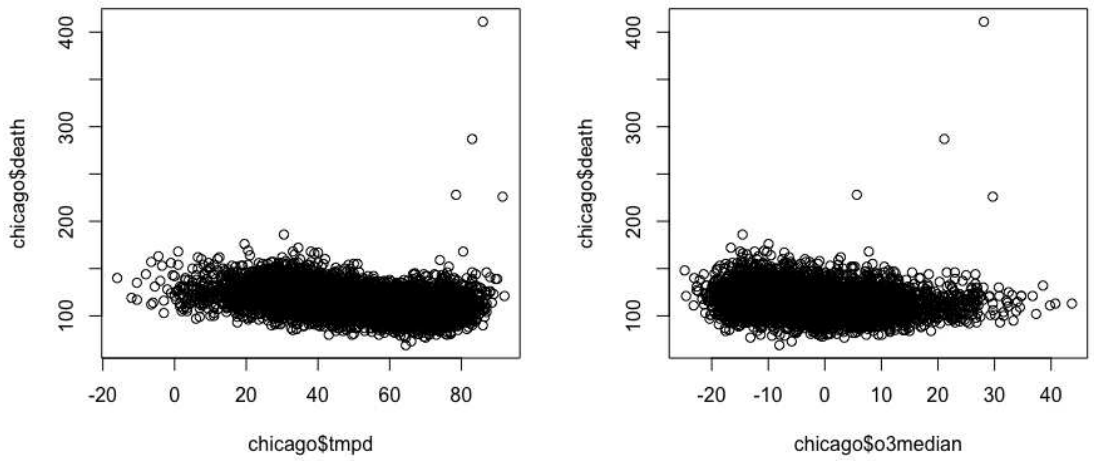
```
[1] 112 97 122 119 116 121 226 411 287 228 159 142 123 102
   94
```

Crtanje ovog dijela podataka također indicira da je ovaj maksimum u dnevnoj stopi smrtnosti asociran s periodom jako visoke temperature i visoke razine ozona kao što možemo vidjeti na slici 4.4. Jedna očita mogućnost je da je model jednostavno nefleksibilan i da je potrebna neka nelinearna ovisnost stope smrtnosti o temperaturi i razini ozona. To sugerira da umjesto uključivanja varijabli koje govore o kvaliteti zraka linearno u model ih uključimo kao neku glatku funkciju, tako da nam model sada postaje:

$$\log\{\mathbb{E}(\texttt{death}_i)\} = f_1(\texttt{time}_i) + f_2(\texttt{pm10median}_i) + f_3(\texttt{so2median}_i) + f_4(\texttt{o3median}_i) + f_5(\texttt{tmpd}_i),$$

gdje su f_i glatke funkcije. Model se sada lagano procijeni sljedećim pozivom `gam()` funkcije:

```
ap1<-gam(death ~ s(time,bs="cr",k=200) + s(pm10median,bs="cr
   ") + s(so2median,bs="cr") + s(o3median,bs="cr") + s(tmpd,
   bs="cr"), data=chicago, family=poisson)
```

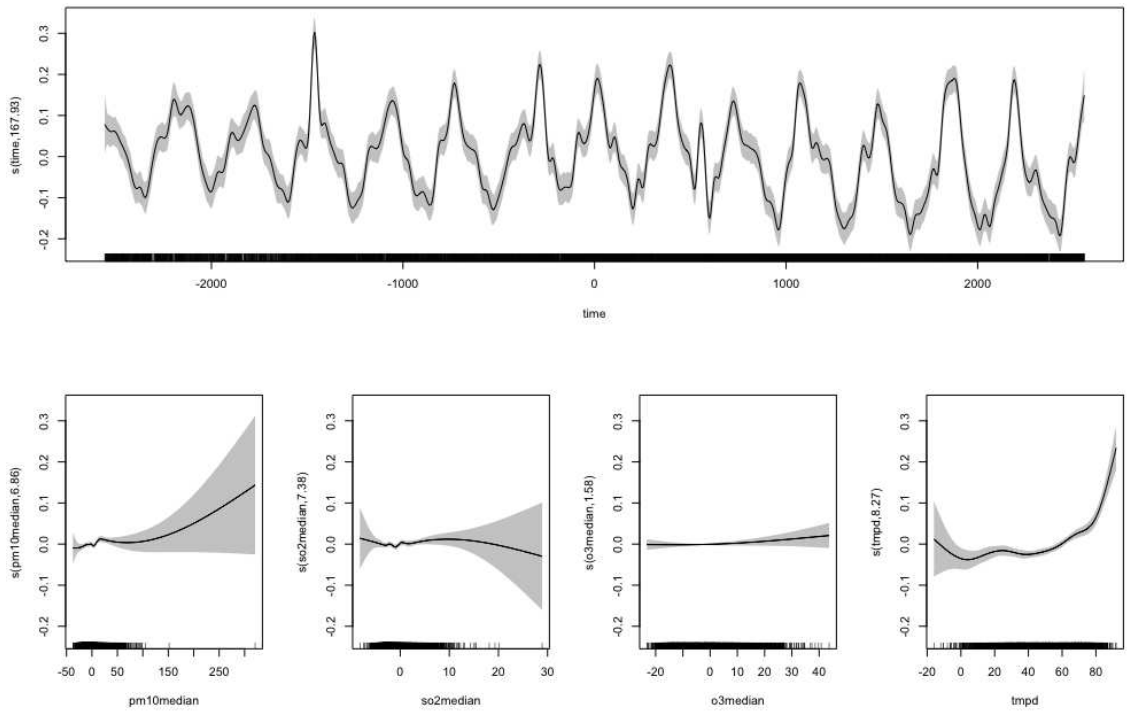


Slika 4.4: Stopa smrtnosti u ovisnosti o temperaturi i razini ozona

no `gam.check()` grafovi su gotovo isti kao i na slici 4.2.

Na slici 4.5 prikazani su grafovi procijenjenih glatkih funkcija za svaku kovarijatu, tzv. parcijalnih funkcija, i iz njih možemo uočiti problem s distribucijom `pm10median` vrijednosti, za koje bi se moglo očekivati da uzrokuju probleme s točkama visoke poluge. Točke visoke poluge su odskočnici u odnosu na nezavisnu varijablu i imaju potencijal uzrokovati velike promjene u procjenama parametara kada se izbace iz modela, tj. da budu utjecajne točke. To vidimo pomoću takozvanog "rug plota", odnosno vrijednosti kovarijata iscrtane uz donji rub grafa. Slični grafovi, sada s parcijalnim rezidualima, ponovno ukazuju na očitu nemogućnost modela u procjeni naša 4 odskočnika kao što se vidi sa slike 4.6.

Detaljnija analiza podataka u blizini naša 4 odskočnika ukazuje na to da je najviša izmjerena temperatura zapravo nekoliko dana prije navedena 4 dana s najvišom stopom smrtnosti, također i najviša razina ozona. Ovo sugerira da bi prosječna temperatura i razine onečišćenja, tijekom nekoliko dana koji prethode određenoj stopi smrtnosti, mogle bolje predvidjeti istu nego temperatura i razine na isti dan. Takav bi model mogao biti razumniji i s biološke, medicinske strane. Naime razine onečišćenja i temperature zabilježene u podacima nisu dovoljno visoke da izazovu neposrednu akutnu bolest i smrtnost, a čini se vjerojatnijim da bi za bilo kakve učinke trebalo neko vrijeme da se manifestiraju, na primjer, pogoršanjem postojećih zdravstvenih stanja. Navedena razmišljanja probat ćemo implementirati tako zvanim *distributed lag modelom*, odnosno model s distribuiranim kašnjenjem. U njemu odziv ovisi o zbroju glatkih funkcija kovarijati s kašnjenjem. Obično glatke funkcije s različitim kašnjenjima glatko variraju s kašnjenjem. Primjerice, ne očekujemo da će odgovor na jučerašnje zagađenje imati potpuno drugačiji oblik od

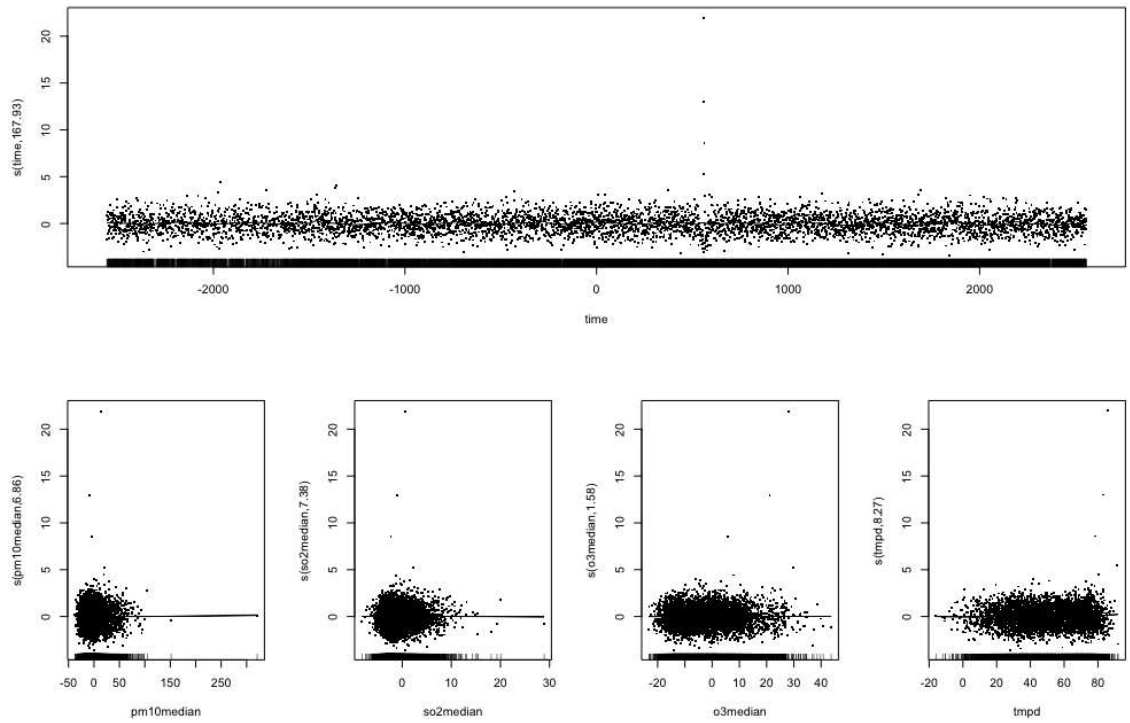


Slika 4.5: Procjena glatkih funkcija iz modela ap1 prikazana bez parcijalnih reziduala. Primjetna je značajna razlika, praznina, u vrijednostima pm10medijana.

odgovora na prekjučerašnje zagađenje.

Model s distribuiranim kašnjenjem

Model s distribuiranim kašnjenjem predstavlja učinak zagađivača kao zbroj glatkih funkcija zagađivača u rasponu fiksnog broja dana kašnjenja. Na primjer, učinak `pm10` na smrt i-tog dana može biti predstavljen sa $\sum_{k=0}^{5} f_k(\texttt{pm10}_{i-k})$, gdje su f_k glatke funkcije koje treba procijeniti. Ovaj učinak je lako uključiti u model, preko komponente formule modela oblika nešto poput `s(pm10) + s(pm10.1) + ... + s(pm10.5)`, gdje `pm10.j` sadrži opažanja pm10 uz j dana kašnjenja. Naravno, možda ne bismo htjeli da su svi ti učinci procijenjeni potpuno zasebno. Jedna mogućnost bi bila prisiliti kovarijate da imaju isti parametar zaglađivanja, što možemo postići stavljajući na istu vrijednost, u svim kovarijatama učinka `pm10`, parametar `id`, nešto poput `s(pm10, id = 1) + s(pm10.1, id = 1) + ... + s(pm10.5, id = 1)`. Međutim, procjene učinaka još uvijek bi mogle biti jako različite između susjednih dana kašnjenja, što možda nije vjerojatno. Druga mogućnost



Slika 4.6: Glatke funkcije komponenti ap1 modela s parcijalnim rezidualima

je zahtjevati da se same glatke funkcije glatko mijenjaju s kašnjenjem. Kao priprema za prilagođavanje modela, korisno je pripremiti skup matrica s recimo 6 stupaca koje sadrže varijable i njihova kašnjenja do 5 dana u zasebnim stupcima. Na primjer

```
lagard <- function(x,n.lag=6) {
  n <- length(x)
  X <- matrix(NA,n,n.lag)
  for (i in 1:n.lag) X[i:n,i] <- x[i:n-i+1]
  X
}
dat <- list(lag=matrix(0:5,nrow(chicago),6,byrow=TRUE))
dat$pm10 <- lagard(chicago$pm10median)
dat$o3 <- lagard(chicago$o3median)
dat$tmp <- lagard(chicago$tmpd)
dat$death <- chicago$death
dat$time <- chicago$time
```

Pretpostavimo sada da je f bivarijatna glatka funkcija, dok je $\texttt{lag}_{ik} = k - 1$ i $\texttt{PM10}_{ik} = \texttt{pm10}_{i-k+1}$. Tada bi odgovarajući model za doprinos varijable `pm10` broju smrti i-tog dana mogao biti:

$$\sum_{k=1}^{6} f\left(\texttt{PM10}_{ik}, \texttt{lag}_{ik}\right).$$

Koristeći "konvenciju zbrajanja" u `mgcv` paketu, takvi izrazi mogu biti specificirani koristeći komponentu formule modela `te(PM10,lag)`, gdje su `PM10` i `lag` matrice s 6 stupaca definirane gore. Model s distribuiranim kašnjenjem možemo pisati kao:

$$\log\left\{\mathbb{E}\left(\texttt{death}_i\right)\right\} = f_1\left(\texttt{time}_i\right) + \sum_{k=1}^{6} f_2(\texttt{PM10}_{ik}, \texttt{lag}_{ik}) + \sum_{k=1}^{6} f_2(\texttt{O3}_{ik}, \texttt{lag}_{ik}) + \sum_{k=1}^{6} f_2(\texttt{TMP}_{ik}, \texttt{lag}_{ik})$$

kojeg procjenjujemo pozivom sljedeće funkcije:

```
ap2 <- bam(death~s(time,bs="cr",k=200)+te(pm10,lag,k=c(10,5)
   )+te(o3,lag,k=c(8,5))+ te(tmp,lag,k=c(8,5)), family=
   poisson,data=dat)
```

Promatrajući podatke možemo uočiti da se nekoliko dana prije dana s izrazito velikim brojem smrti istovremeno opaža i velika temperatura i razina ozona. To nam je motivacija da ubacimo i interakciju te dvije varijable u model. Konačan model bi sada bio:

$$\log\left\{\mathbb{E}\left(\texttt{death}_i\right)\right\} = f_1\left(\texttt{time}_i\right) + \sum_{k=1}^{6} f_2(\texttt{PM10}_{ik}, \texttt{lag}_{ik}) + \sum_{k=1}^{6} f_3(\texttt{O3}_{ik}, \texttt{TMP}_{ik}, \texttt{lag}_{ik})$$

kojeg možemo procijeniti pozivom sljedeće funkcije:

```
ap3 <- bam(death~s(time,bs="cr",k=200)+te(pm10,lag,k=c(10,5)
   )+ te(o3,tmp,lag,k=c(8,8,5)),family=poisson,data=dat)
```

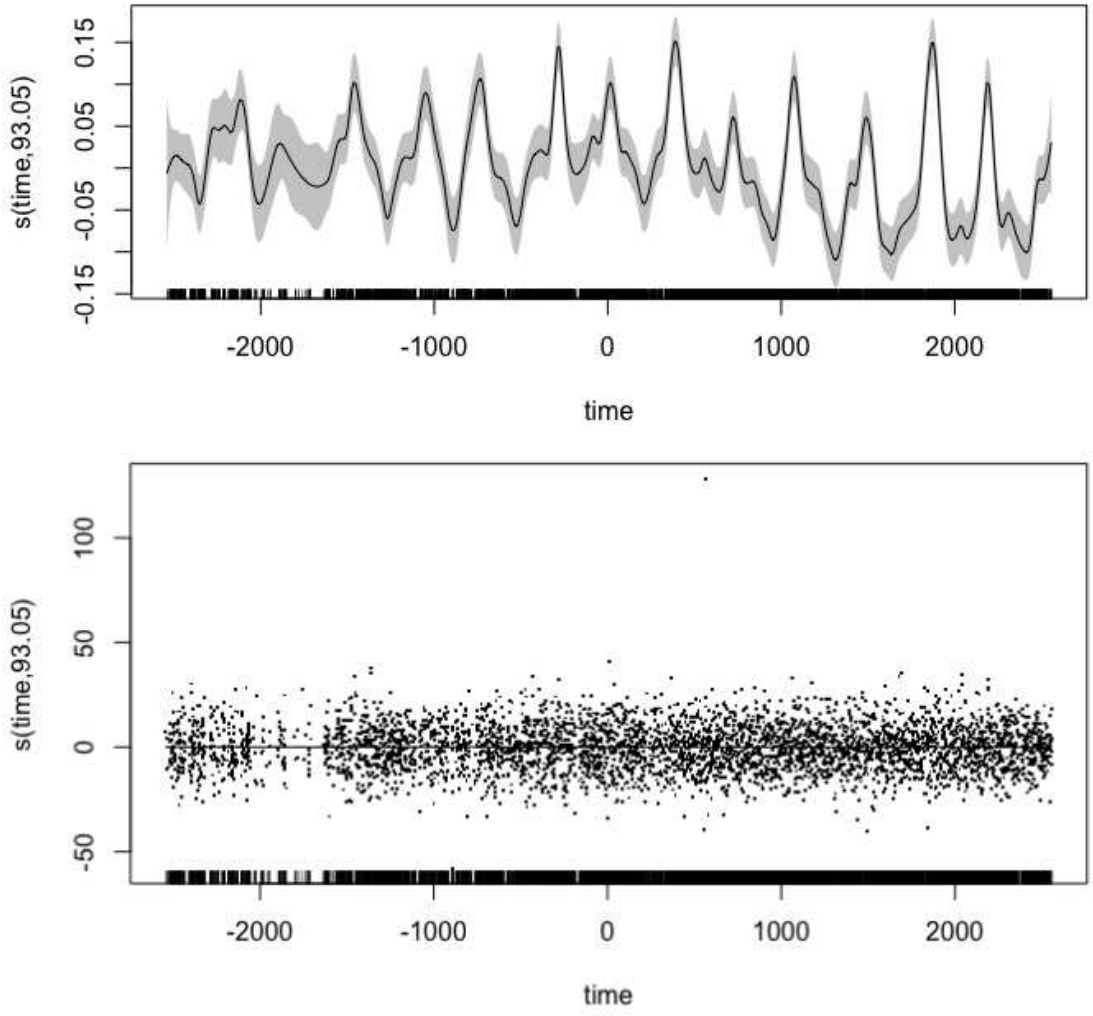
Ovdje koristimo `bam` funkciju umjesto `gam` funkcije jer radimo s velikim skupom podataka, a ona je optimizirana za takve slučajeve, jer troši manje memorije i brža je.

Usporedimo sada AIC kriterij za početni model, model s uključenim *"lagovima"* bez interakcije i za konačni model s uključenim *"lagovima"* i interakcijom.

```
> AIC(ap1, ap2, ap3)
          df       AIC
ap1 193.0265 37856.90
ap2 149.1193 32250.73
ap3 183.0127 31971.88
```

Ovo nam opravdava uključivanje interakcije te također pokazuje znatno poboljšanje modela s distribuiranim kašnjenjem u odnosu na početni model.

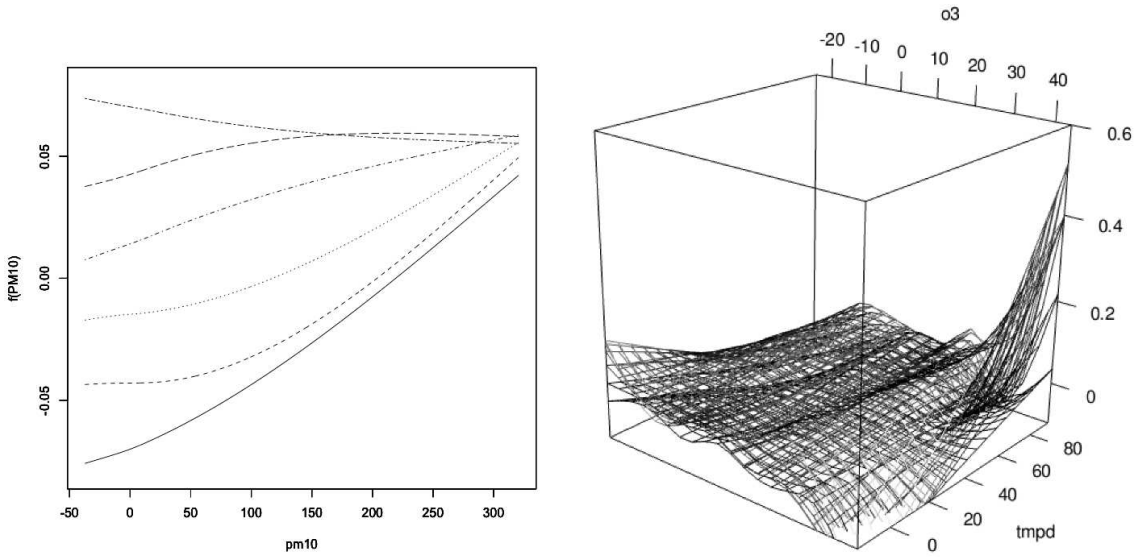
Model s dodatnom komponentom s distribuiranim kašnjenjem, `so2`, daje veliku p-vrijednost za dodatni član, tako da nema razloga da ju uključimo u model. Razlog zašto model sa i bez `so2` ne uspoređujemo pomoću AIC-a ili anove je taj da nedostaje puno podataka o `so2`, pa se ta dva modela procjenjuju s različitim brojem podataka stoga takva usporedba ne bi bila valjana, osim ako ponovno ne prilagodimo jednostavniji model samo na podskup podataka bez nedostajućeg `so2`.



Slika 4.7: Procijenjeni učinci za model s distribuiranim kašnjenjem.

Na slici 4.7 ponovno možemo vidjeti procijenjenu glatku funkciju za osnovnu stopu

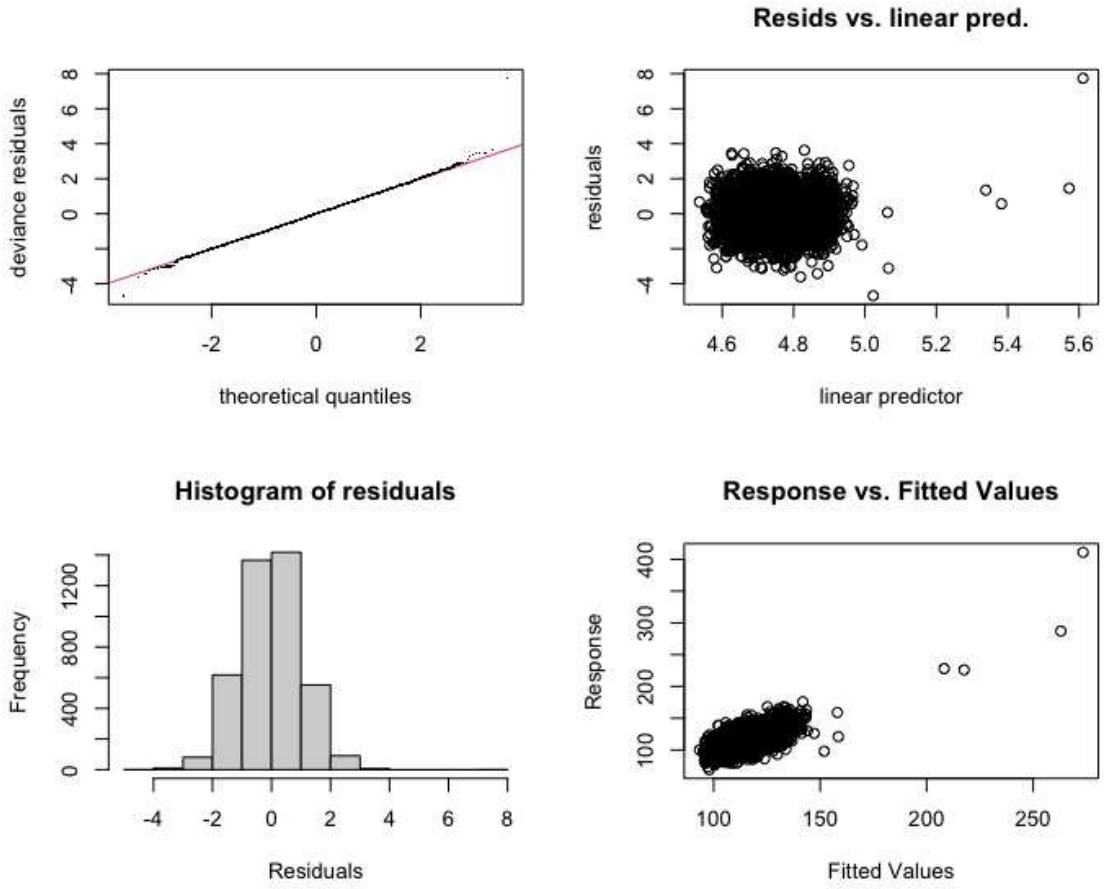
smrtnosti zajedno s pripadajućim grafom parcijalnih reziduala, sada za model s distrubuiranim kašnjenjem. Četiri velika odskočnika koja su se nalazila blizu jedno drugome su uklonjena, no jedan značajan odskočnik je i dalje prisutan.



Slika 4.8: Procijenjeni učinci interakcija za model s distribuiranim kašnjenjem.

Kompaktna vizualizacija modela s distribuiranim kašnjenjem nije uvijek jednostavna, no jedan pokušaj možemo vidjeti na slici 4.8. Na lijevom grafu prikazan je učinak PM10 za svako kašnjenje od 0 do 5 dana. Kretanjem vertikalno od donjeg lijevog dijela grafa, krivulje su poredane po povećanju kašnjenja. Primijetimo kako gotovo da nema učinka kod kašnjenja 4 i 5 (krivulja je gotovo ravna), ali kod kraćih kašnjenja povećanje PM10 je povezano s povećanjem stope smrtnosti. Na desnom grafu vidimo perspektivne prikaze interakcije ozona i temperature po kašnjenju. Kombinacija visokog ozona i visoke temperature, za srednja kašnjenja, najviše povećava rizik smrtnosti.

Promotrimo za kraj grafove dobivene pozivom `gam.check` funkcije, za model s distribuiranim kašnjenjem (`ap3`), prikazane na slici 4.9. Primijetimo da su očiti odskočnici, prisutni na grafu reziduala u ovisnosti o linearnom prediktoru i grafu prilagođenih vrijednosti naspram stvarnih vrijednosti, sada uklonjeni. QQ-plot je sada mnogo bliži pravcu, a histogram reziduala poprimio je oblik normalne distribucije. Na sva četiri grafa su primjetna znatna poboljšanja u odnosu na početni model pa možemo zaključiti da je model zadovoljavajući.



Slika 4.9: Grafovi dobiveni gam.check funkcijom modela s distribuiranim kašnjenjem i interakcijama (ap3).

4.2 Modeliranje sezonskih podataka pomoću GAM-ova

U području analize vremenskih nizova, posebice u klimatologiji i ekološkim istraživanjima, susrećemo skupove podataka koji pokazuju složena ponašanja tijekom vremena, obuhvaćajući ne samo dugoročne trendove već i sezonske oscilacije. Ti podaci su obično prikupljeni tijekom više godina i uključuju mjerenja provedena u redovitim intervalima svake godine. Primarni cilj naše analize je precizno modeliranje sezonskih varijacija i dugoročnih trendova u vremenskim nizovima. Ovo modeliranje nam omogućuje bolje razumijevanje osnovne vremenske dinamike i poboljšava našu sposobnost predviđanja budućih trendova na temelju povijesnih obrazaca.

Generalizirani aditivni modeli (GAM-ovi) nude fleksibilan okvir za suočavanje s tim

izazovima, omogućujući definiranje glatkih, nelinearnih odnosa među varijablama. Ovaj pristup je posebno učinkovit u identificiranju suptilnih obrazaca sezonskih promjena i trendova tijekom vremena. U ovom kontekstu razmatramo dvije glavne komponente u našem modelu: sezonski učinak i dugoročni trend. Primjer koji ćemo obraditi inspiriran je s [4].

U matematičkoj formulaciji našeg modela, temperaturu možemo izraziti jednadžbom:

$$\texttt{temp} = f_{\text{seasonal}}(\texttt{time\_of\_year}) + f_{\text{trend}}(\texttt{year}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2\mathbf{\Lambda})$$

Funkcija $f_{\text{seasonal}}(\texttt{time\_of\_year})$ glatko modelira sezonsku varijaciju gdje `time_of_year` može biti određeni dan ili mjesec, što je ključno za hvatanje ciklične prirode podataka. Funkcija $f_{\text{trend}}(\texttt{year})$ hvata dugoročne promjene, prateći trendove kako godine prolaze. Komponenta ϵ, koja predstavlja grešku, slijedi normalnu distribuciju sa srednjom vrijednosti nula i kovarijacijskom matricom $\sigma^2\mathbf{\Lambda}$, što pomaže u upravljanju potencijalnom autokorelacijom unutar podataka.

Opis skupa podataka

Skup podataka koji se koristi u ovoj analizi je niz temperatura Središnje Engleske (”Central England Temperature”, CET), preuzet s web stranice UK Met Office. Ovaj niz podataka predstavlja jedan od najduljih dostupnih zapisa o temperaturi, koji datira još od 1659. godine. Pruža mjesečne prosječne temperature širom Središnje Engleske, obuhvaćajući širok raspon klimatskih varijacija kroz nekoliko stoljeća. Ovo sveobuhvatno vremensko pokrivanje čini ga neprocjenjivim resursom za proučavanje dugoročnih klimatskih trendova i sezonskih uzoraka.

CET skup podataka strukturiran je sa stupcima koji predstavljaju svaki mjesec u godini, uz dodatni stupac za godišnji prosjek. Evo primjera skupa podataka kako bismo ilustrirali njegov format:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
1659	3	4	6	7	11	13	16	16	13	10	5	2	8.9
1660	0	4	6	9	11	14	15	16	13	10	6	5	9.1
1661	5	5	6	8	11	14	15	15	13	11	8	6	9.8
1662	5	6	6	8	11	15	15	15	13	11	6	3	9.5
1663	1	1	5	7	10	14	15	15	13	10	7	5	8.6
1664	4	5	5	8	11	15	16	16	13	9	6	4	9.3

Tablica 4.1: Uzorak tablice podataka CET koji prikazuje mjesečne i godišnje prosječne temperature od 1659. do 1664. godine.

Za analizu pomoću GAM-a, skup podataka pretvoren je u ”dugi” format, gdje svaki redak predstavlja mjesečno promatranje s detaljima o godini i temperaturi. Dodatno, stvoren

je numerički indeks mjeseca (od 1 do 12) koji će zajedno s godinom služiti kao ključne varijable za modeliranje sezonskih i godišnjih obrazaca. Podaci su potom sortirani kronološki te je uvedena i vremenska varijabla, koja će nam služiti za modeliranje trenda. Ta varijabla prikazuje broj sekundi koji je protekao od 1.1.1970. (taj datum nam je referentna točka), te je skalirana radi numeričke stabilnosti. U sljedećoj tablici prikazano je nekoliko redova modificirane tablice podataka:

Month	Temperature	Year	nMonth	Date	Time
Jan	3	1659	1	1659-01-15	-113.576
Feb	4	1659	2	1659-02-15	-113.545
Mar	6	1659	3	1659-03-15	-113.517
Apr	7	1659	4	1659-04-15	-113.486
May	11	1659	5	1659-05-15	-113.456
Jun	13	1659	6	1659-06-15	-113.425

Tablica 4.2: Uzorak modificirane tablice podataka prikazan kroz mjesečna promatranja od 1659. do 1664. godine

Na sljedećem grafu (4.10) prikazan je vremenski niz godišnjih prosječnih temperatura iz skupa podataka CET, koji se proteže kroz nekoliko stoljeća od 1659. godine do nedavnih godina. Graf otkriva fluktuacije u temperaturi tijekom vremena, ilustrirajući varijabilnost unutar godina i potencijalne dugoročne trendove.

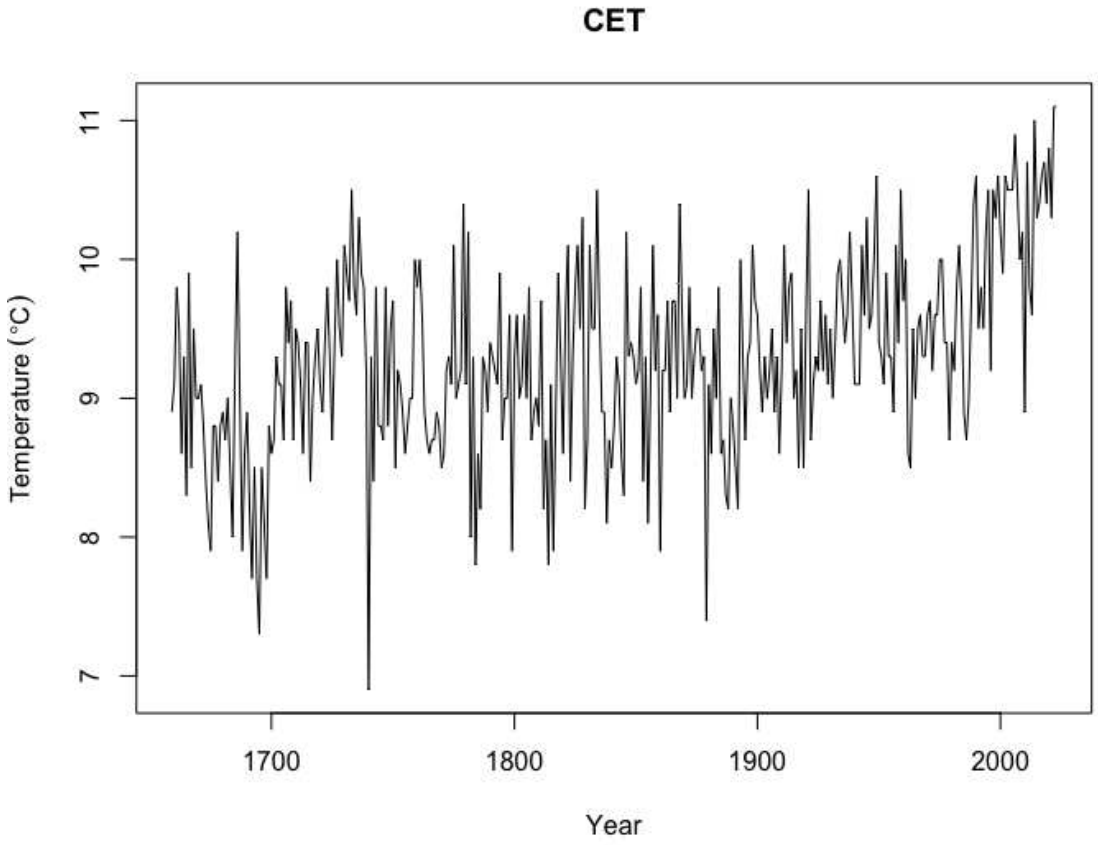
Model s nekoreliranim greškama

Započet ćemo modeliranje s osnovnim modelom korištenjem funkcije `gam` iz paketa `mgcv`. Model je specificiran na sljedeći način:

```
m <- gam(Temperature ~ s(nMonth, bs = "cc", k = 12) +
         s(Time), data = cet)
```

Ovdje `s(nMonth, bs = "cc", k = 12)` predstavlja ciklični kubični splajn za mjesečnu varijablu `nMonth`. Ciklični kubični splajn odabran je kako bi se osiguralo da model uzima u obzir sezonsku komponentu podataka, reflektirajući kontinuiranu prirodu temperature preko granice kraja godine od prosinca do siječnja. Parametar k određuje dimenziju baze splajna. Ovdje je postavljen na maksimalno mogući za `nMonth`, što je 12, broj jedinstvenih vrijednosti. Dodatno, `s(Time)` uključen je kako bi se uhvatio dugoročni trend tijekom vremena.

Pri analizi početne prilagodbe modela pomoću funkcije `gam()` koristit ćemo funkciju `summary(m)` kako bismo ocijenili njegovu uspješnost. Funkcija uzima prilagođeni GAM objekt i generira razne korisne sažetke iz njega, od kojih jedan dio možemo vidjeti ovdje:



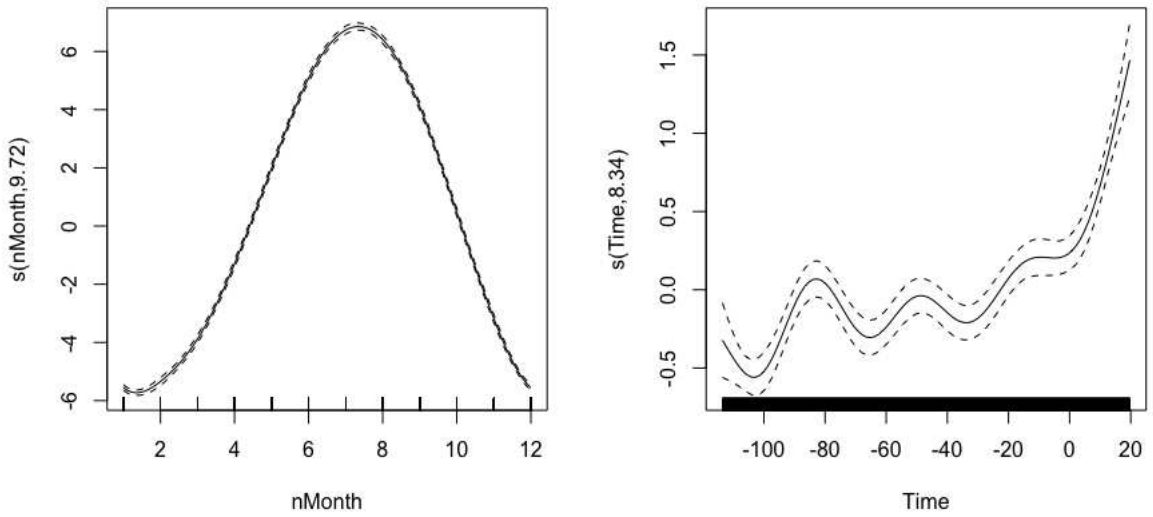
Slika 4.10: Vremenski niz godišnjih prosječnih temperatura iz skupa podataka CET

```
Approximate significance of smooth terms:
            edf Ref.df       F p-value
s(nMonth) 7.812 10.000 4667.07  <2e-16 ***
s(Time)   8.709  8.975   26.92  <2e-16 ***

R-sq.(adj) =  0.917   Deviance explained = 91.7%
```

Izlaz pokazuje da su i sezonska i dugoročna komponenta trenda statistički značajne. Međutim, ovaj početni model ne uzima u obzir inherentne ovisnosti u podacima, što je značajan nedostatak. Prikazivanje parcijalnih utjecaja komponenata modela ilustrira potencijalne probleme koji se mogu pojaviti ako zanemarimo ove ovisnosti.

```
    plot(m, scale = 0)
```



Slika 4.11: Komponente glatkih funkcija naivnog modela koji pretpostavlja nekorelirane greške

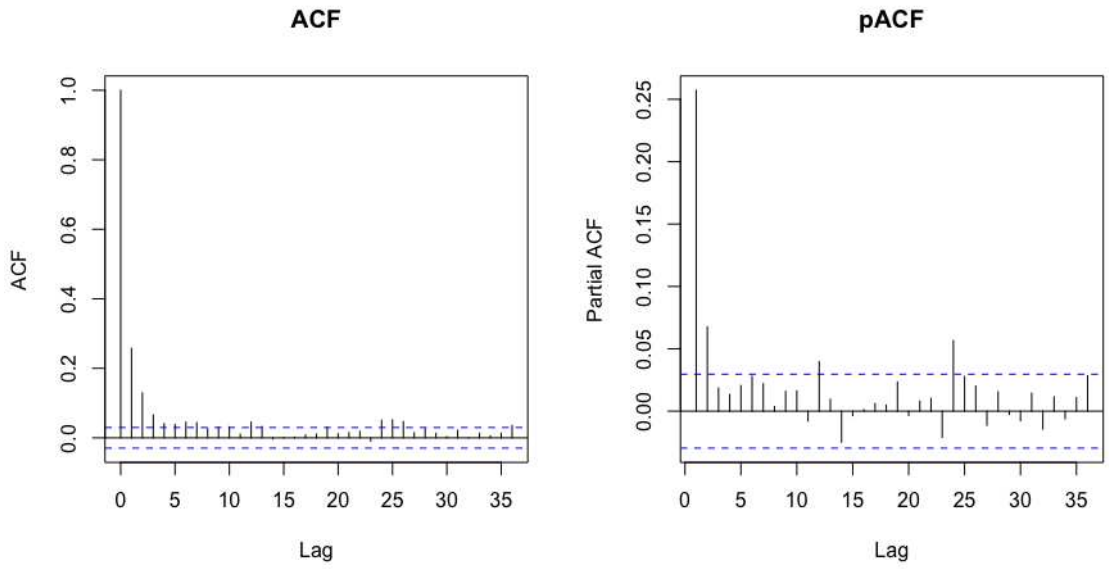
Slika prikazuje dva splajna. Sezonska komponenta, prikazana kao ciklični kubični splajn, pokazuje besprijekoran prijelaz između kraja i početka godine, kako smo i očekivali. S druge strane, komponenta trenda pokazuje pretjeranu fleksibilnost, što se vidi po njegovom valovitom izgledu, što bi potencijalno moglo dovesti do prekomjernog prilagođavanja podacima. Splajnovi su na vrlo različitim skalama (`scale=0`) što ilustrira relativne stupnjeve varijacije u sezonskoj i trendovskoj komponenti. Naime, temperatura se tijekom cjelokupnog razdoblja povećava za otprilike 1,5 stupnjeva, ali unutar jedne godine postoji oko 12 stupnjeva varijacije u temperaturi, u prosjeku. Očito, stvarni podaci variraju oko ovih vrijednosti, to je neobjašnjena varijanca.

Analizirajmo sada također ostatke ovog modela koristeći autokorelacijsku (ACF) i parcijalnu autokorelacijsku funkciju (pACF) (za definiciju vidi [1]). Ukratko, ACF prikazuje ukupnu korelaciju s prošlim "lagovima", dok PACF izolira izravnu korelaciju s svakim specifičnim "lagom".

```
acf(resid(m), lag.max = 36, main = "ACF")
pacf(resid(m), lag.max = 36, main = "pACF")
```

Rezultati su prikazani na slici 4.12

Pri evaluaciji reziduala modela pomoću funkcija autokorelacije nailazimo na očitu prisutnost autokorelacije. To sugerira da, unatoč sposobnosti modela da objasni velik dio



Slika 4.12: ACF i pACF grafovi reziduala iz naivnog modela

varijabilnosti, još uvijek postoji sustavna varijacija koja nije uzeta u obzir, tj. postoji neki oblik zavisnosti između trenutnih i prošlih grešaka u vremenskom nizu. Ovaj obrazac korelacije ukazuje na to da greške nisu slučajne i da mogu biti pod utjecajem faktora koji nisu uzeti u obzir u modelu. Na primjer, autokorelacija može dovesti do pristranih procjena parametara, posebno u varijanci, što utječe na razumijevanje odnosa između varijabli. To rezultira nevaljanim zaključcima izvučenim iz modela te neefikasnim predikcijama.

Obrasci uočeni na ACF i pACF grafu impliciraju da bi za adekvatno modeliranje temporalne korelacije unutar reziduala možda bilo potrebno uključiti autoregresijski proces reda p. Dakle, formula modela uz uključenje, AR(p) modela mogla bi izgledati ovako:

$$\texttt{temp}_i = f_{\text{seasonal}}(\texttt{time\_of\_year}_i) + f_{\text{trend}}(\texttt{year}_i) + e_i, \quad e_i = \sum_{k=1}^{p} \phi_k e_{i-k} + \epsilon_i$$

pri čemu su ϵ_i nezavisne i jednako distribuirane $N\left(0, \sigma^2\right)$ slučajne varijable i ϕ_k za $k = 1, ..., p$ realne konstante.

GAMMs

Prije uvođenja novog modela, u ovom odjeljku opisat ćemo jedno proširenje GAM-ova, generalizirane aditivne mješovite modele (eng., *"generalised additive mixed models"*, kraće GAMM). GAMM je vrsta statističkog modela koji kombinira fleksibilnost generaliziranih

aditivnih modela (GAM-ova) s mogućnosti uključivanja slučajnih učinaka iz mješovitih modela, omogućujući modeliranje složenih struktura podataka koje mogu pokazivati nezavisna opažanja. GAMM-ovi su posebno korisni pri radu s hijerarhijskim ili grupiranim podacima, gdje opažanja unutar iste grupe mogu biti korelirana.

Dok GAM-ovi prilagođavaju glatku krivulju podacima i izvrsni su za hvatanje nelinearnih trendova i obrazaca, rade pod pretpostavkom da su opažanja nezavisna jedna od drugih. Međutim, u mnogim stvarnim scenarijima, kao što su longitudinalna istraživanja ili ugniježđeni podaci, pretpostavka nezavisnosti može biti prekršena zbog prisutnosti grupiranih struktura u podacima. GAMM-ovi rješavaju ovaj problem uključivanjem slučajnih efekata, koji uzimaju u obzir korelaciju unutar grupa podataka. Ti slučajni efekti dodaju se linearnom prediktoru modela, slično kao u modelima mješovitih efekata.

Generalizirani aditivni mješoviti model (GAMM) može se zapisati kao:

$$g(\mu_i) = \mathbf{A}_i\gamma + \sum_j f_j\left(x_{ji}\right) + \mathbf{Z_i b}, \quad y_i \sim \mathrm{EF}(\mu_i, \phi), \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_\theta),$$

gdje su prve dvije komponente iste kao i kod običnih gam-ova, a izraz $\mathbf{Z_i b}$ predstavlja slučajne učinke u modelu. $\mathbf{Z_i}$ je matrica koja specificira dizajn slučajnih učinaka za i-to opažanje, a $\mathbf{b}$ je vektor koeficijenata slučajnih učinaka.

Procjena GAMM-ova

Pokazali smo postojanje dualnosti između glatkih funkcija i slučajnih učinaka. Kao posljedica toga, jednostavni Gaussovi slučajni učinci mogu se procijeniti kao da su glatke funkcije u modelu, koristeći metode procjene koje su već obrađene u ovom poglavlju.

S druge strane, možemo procijeniti glatke funkcije u GAMM-u kao da su slučajni učinci u (generaliziranom) linearnom mješovitom modelu, koristeći opće metode i softver mješovitih modela u tu svrhu. Za ovakav pristup potrebno nam je glatke funkcije postaviti u obliku koji odgovara strukturi slučajnih učinaka koju softver prepoznaje. U tu svrhu slijedimo recept iz odjeljka 2.5 kako bismo svaku glatku funkciju eksplicitno predstavili u obliku $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}$, gdje se $\boldsymbol{\beta}$ tretira kao vektor (nepenaliziranih) fiksnih učinaka, a $\mathbf{b} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_b^2\right)$ tretira kao vektor njd. slučajnih učinaka.

Ovo je način na koji funkcionira funkcija `gamm()` iz paketa `mgcv`, koristeći funkciju `lme()` iz `nlme` paketa za procjenu u slučaju normalne distribucije i funkcije veze identite, a u suprotnom koristeći `gammPQL()` funkciju koja je modifikacija `glmmPQL()` funkcije iz `MASS` paketa.

Modeli s koreliranim greškama

U našem modelu pretpostavljamo normalnu razdiobu odziva te funkciju veze identitetu pa naš model svodimo na linearni mješoviti model i pritom za prilagodbu koristimo funkciju

`lme`. Nadalje htjeli bismo uvrstiti korelaciju reziduala u model, a to ćemo napraviti na način da korelacijsku matricu $\mathbf{\Lambda}_\theta$ iz $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{\Lambda}_\theta)$ modificiramo tako da nije sada jednaka $\mathbf{I}\sigma^2$ već odgovarajuća kovarijacijska matrica nekog jednostavnog autoregresijskog modela.

U praksi korelaciju grešaka uključujemo u funkciju `gamm` definirajući parametar `correlation`. Sada ćemo prilagoditi osnovni model s funkcijom `gamm` i budući da izgleda da je potreban neki AR model niskog reda za reziduale, prilagodit ćemo tri: AR(1), AR(2) i AR(3) unutar jedne godine.

```
## AR(0)
m <- gamm(Temperature ~ s(nMonth, bs = "cc", k = 12) +
        s(Time), data  = cet)
## AR(1)
m1 <- gamm(Temperature ~ s(nMonth, bs = "cc", k = 12) +
        s(Time, k = 20), data = cet,
        correlation = corARMA(form = ~ 1|Year, p = 1))

## AR(2)
m2 <- gamm(Temperature ~ s(nMonth, bs = "cc", k = 12) +
        s(Time, k = 20), data = cet,
        correlation = corARMA(form = ~ 1|Year, p = 2))

## AR(3)
m3 <- gamm(Temperature ~ s(nMonth, bs = "cc", k = 12) +
        s(Time, k = 20), data = cet,
        correlation = corARMA(form = ~ 1|Year, p = 3))
```

Važno je napomenuti što argument korelacije, `correlation`, radi ovdje: `corARMA(form = ~1|Year, p = p)` znači prilagodbu ARMA procesa rezidualima, gdje p označava red za AR dio ARMA modela (q označava red MA, tj, "moving average" dijela, no to je u našem slučaju 0, što je i unaprijed zadana vrijednost), a `form = ~1|Year` znači da je ARMA ugniježđen unutar svake godine (radi brzine izvršavnja prilagodbe).

Usporedimo sada prilagođene modele kako bismo izabrali najbolji. To ćemo napraviti putem općeg testa omjera vjerodostojnosti `anova()` metodom za `lme` objekte. Ovo je valjana usporedba jer su modeli ugniježđeni (možemo prijeći od AR(3) do AR(1) postavljanjem nekih AR koeficijenata na 0).

```
anova(m$lme, m1$lme, m2$lme, m3$lme)

        df      AIC      logLik   Test   L.Ratio p-value
m$lme    5 14946.17  -7468.083
m1$lme   6 14661.27  -7324.633 1 vs 2 286.89853  <.0001
```

```
m2$lme    7 14637.00   -7311.500 2 vs 3   26.26645   <.0001
m3$lme    8 14636.38   -7310.189 3 vs 4    2.62281   0.1053
```

AR(1) model pruža značajno poboljšanje prilagodbe u odnosu na naivni model, a AR(2) dodatno značajno poboljšava prilagodbu. Međutim, prijelaz na AR(3) daje vrlo malo poboljšanja. Procijenjene koeficijente ϕ_1 i ϕ_2 za AR(2) proces možemo vidjeti ovdje:

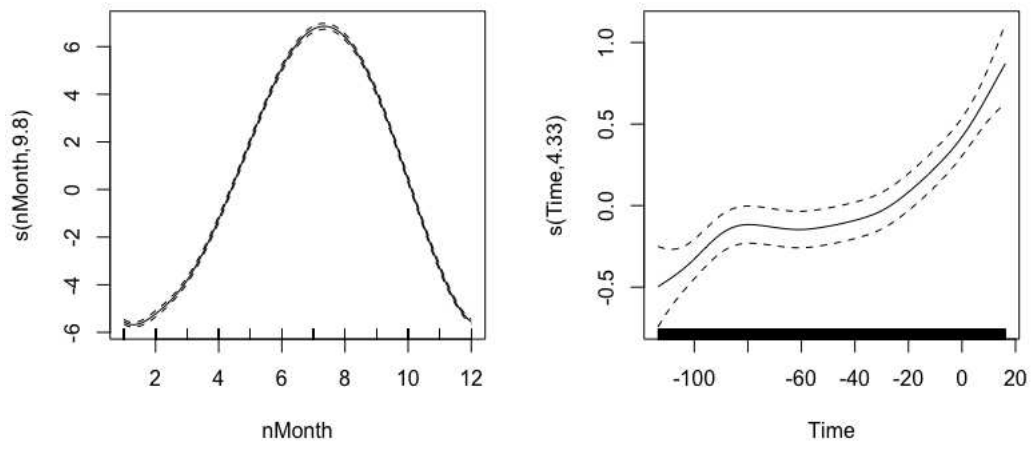
```
intervals(m2$lme, which = "var-cov")$corStruct

           lower      est.     upper
Phi1 0.24473306 0.2674293 0.2867680
Phi2 0.06976492 0.1071881 0.1443099
attr(,"label")
[1] "Correlation structure:"
```

Pouzdani interval za ϕ lako se uočava i pruža vrlo snažne dokaze da je AR(2) model prikladniji u odnosu na početni model ($\phi_1 = \phi_2 = 0$).

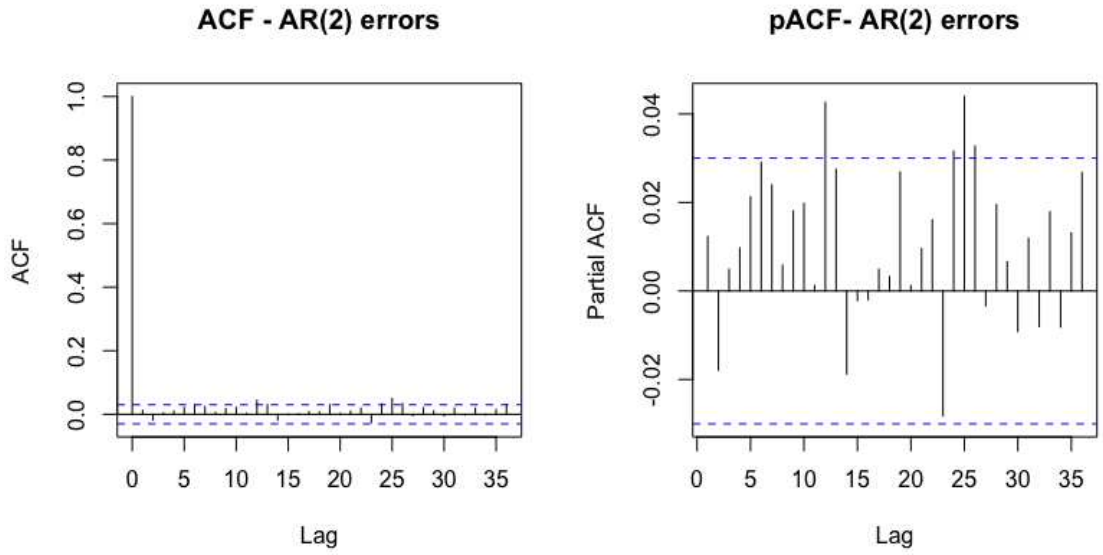
Grafički prikaz parcijalnih utjecaja novog odabranog modela, sa slike 4.13 pokazuje koliko je naivni model s nekoreliranim greškama bio preprilagođen. Sada je trend mnogo glađi i više u skladu s našim očekivanjima.

```
plot(m2$gam, scale = 0)
```



Slika 4.13: Glatke funkcije za najbolje prilagođeni GAM model s AR(2) modelom za reziduale

Gledajući sada ponovno ACF i pACF grafove sa slike 4.14, ali sada za normalizirane reziduale (`resid(m2, type = "normalised")`) ne vidimo značajnu autokorelaciju, što sugerira da je AR(2) model dovoljan i da možemo, do određene mjere, izvući zaključke iz ovog modela. Napomenimo samo zašto je bitno gledati normalizirane reziduale i što su oni točno. Naime, pozivom `resid(m2)` dobimo tako zvane *"raw residuals"*, koji su jednaki razlici opažene i predviđene vrijednosti i pritom ne uzimaju u obzir uključenu korelaciju reziduala, samo fiksne efekte. Stoga bismo dobili iste vrijednosti kao da smo uzeli i `resid(m0)`. S druge strane, normalizirani reziduali su standardizirani reziduali prethodno pomnoženi s inverznim kvadratnim korijenom procijenjene matrice korelacije pogrešaka.



Slika 4.14: ACF i pACF grafovi reziduala iz GAM modela s AR(2) korelacijskom matricom

Na kraju, možemo zaključiti da korištenje GAM-ova za modeliranje sezonskih podataka, uz uključivanje korelacije grešaka, omogućava detaljnu analizu složenih vremenskih nizova poput CET skupa podataka. Ovaj pristup omogućuje precizno hvatanje sezonskih obrazaca i dugoročnih trendova, te pruža bolje razumijevanje klimatskih promjena kroz povijesne podatke.

Nastavak istraživanja mogao bi uključivati daljnje unapređenje modela, kao što su uključivanje dodatnih kovarijata ili ispitivanje drugih struktura korelacije, kako bi se postigla još bolja prilagodba i razumijevanje podataka. Ovim završavamo poglavlje o modeliranju sezonskih podataka pomoću GAM-ova.

Bibliografija

[1] Peter J. Brockwell i Richard A. Davis, *Introduction to Time Series and Forecasting*, Springer Texts in Statistics, Springer International Publishing, 2016, https://doi.org/10.1007/978-3-319-29854-2.

[2] Trevor Hastie i Robert Tibshirani, *Generalized Additive Models*, Statistical Science **1** (1986), br. 3, 297 – 310, https://doi.org/10.1214/ss/1177013604.

[3] Elias K. Pedersen, David L. Miller, Gavin L. Simpson, Finn Lindgren, Rachel N. Thomas i Simon N. Wood, *Hierarchical generalized additive models in ecology: an introduction with mgcv*, PeerJ **7** (2019), e6876, https://peerj.com/articles/6876/.

[4] Gavin Simpson, *Modelling seasonal data with GAMs*, 2014, https://fromthebottomoftheheap.net/2014/05/09/modelling-seasonal-data-with-gam/.

[5] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd., Chapman and Hall/CRC, 2017, https://doi.org/10.1201/9781315370279.

Dodatak A

Maksimalna vjerodostojnost

A.1 Dokaz općih rezultata

Promotrimo funkciju log-vjerodostojnosti $l(y;\theta,\phi) = \log(f(y;\theta,\phi))$ unutar neke eksponencijalne familije. Ona nam je potrebna pri procjeni GLM-a. Pokazat ćemo dva vrlo dobro znana rezultata iz statističke teorije:

$$\mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] = 0 \quad \text{i} \quad \mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = 0. \tag{A.1}$$

Da bismo pokazali prvu od gornjih jednakosti, pretpostavimo da možemo diferencirati

$$\int f(y;\theta,\phi)\mathrm{d}x$$

po θ jednostavno uvođenjem znaka diferenciranja pod integral (to je uvijek moguće unutar eksponencijalnih familija). Kako je gornji integral jednak 1 za sve θ, diferenciranjem ćemo dobiti 0. Dakle

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta}\int f(y;\theta,\phi)\mathrm{d}x = \int \frac{\partial}{\partial \theta} f(y;\theta,\phi)\mathrm{d}x = \int \frac{\partial}{\partial \theta} f(y;\theta,\phi)\frac{f(y;\theta,\phi)}{f(y;\theta,\phi)}\mathrm{d}x \\ &= \int \frac{\partial}{\partial \theta} l(y;\theta,\phi) f(y;\theta,\phi)\mathrm{d}x = \mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] \end{aligned}$$

Slično ako po θ diferenciramo jednakost

$$\int \frac{\partial}{\partial \theta} l(y;\theta,\phi) f(y;\theta,\phi)\mathrm{d}x = 0$$

dobijemo sljedeće:

$$\int \frac{\partial^2 l(y;\theta,\phi)}{\partial \theta^2} f(y;\theta,\phi) + \frac{\partial l(y;\theta,\phi)}{\partial \theta}\frac{\partial f(y;\theta,\phi)}{\partial \theta}\mathrm{d}x = 0,$$

ali

$$\frac{\partial l(y;\theta,\phi)}{\partial \theta} = \frac{1}{f(y;\theta,\phi)} \frac{\partial f(y;\theta,\phi)}{\partial \theta}$$

stoga

$$\int \frac{\partial^2 l(y;\theta,\phi)}{\partial \theta^2} f(y;\theta,\phi) \mathrm{d}x = -\int \left(\frac{\partial l(y;\theta,\phi)}{\partial \theta} \right)^2 f(y;\theta,\phi) \mathrm{d}x$$

što je zapravo

$$\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2} \right] = -\mathbb{E}\left[\left(\frac{\partial l}{\partial \theta} \right)^2 \right] = 0.$$

Dakle, pokazali smo i drugu jednakost u (A.1).

Dodatak B

Bayesovska statistika

B.1 MAP procjena

Pretpostavimo da želimo procijeniti neopaženi populacijski parametar θ na temelju opažanja x. Neka je f uzorak distribucije x, tako da je $f(x \mid \theta)$ vjerojatnost da smo opazili upravo x uz dani parametar θ. Tada je funkcija:

$$\theta \mapsto f(x \mid \theta)$$

poznata kao funkcija vjerodostojnosti, a procjena:

$$\hat{\theta}_{\mathrm{MLE}}(x) = \arg\max_{\theta} f(x \mid \theta)$$

je procjena maksimalne vjerodostojnosti parametra θ.

Sada pretpostavimo da postoji apriorna distribucija g za θ. To nam omogućava da tretiramo θ kao slučajnu varijablu, kao u Bayesovskoj statistici. Možemo izračunati aposteriornu distribuciju θ koristeći Bayesov teorem:

$$\theta \mapsto f(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\int_{\Theta} f(x \mid \vartheta) g(\vartheta) d\vartheta}$$

gdje je g funkcija gustoće za θ, a Θ je domena funkcije g.

Metoda maksimalne aposteriorne procjene (*"MAP estimate"*) zatim procjenjuje θ kao mod aposteriorne distribucije ove slučajne varijable:

$$\begin{aligned}
\hat{\theta}_{\mathrm{MAP}}(x) &= \arg\max_{\theta} f(\theta \mid x) \\
&= \arg\max_{\theta} \frac{f(x \mid \theta) g(\theta)}{\int_{\Theta} f(x \mid \vartheta) g(\vartheta) d\vartheta} \\
&= \arg\max_{\theta} f(x \mid \theta) g(\theta)
\end{aligned}$$

Nazivnik aposteriorne distribucije (tzv. marginalna vjerodostojnost) uvijek je pozitivan i ne ovisi o θ, te stoga ne igra nikakvu ulogu u maksimizaciji. Primijetimo da se MAP procjena θ podudara s procjenom maksimalne vjerodostojnosti (ML) kada je apriorna distribucija g uniformna (tj. kada je g konstantna funkcija).

B.2 Marginalna vjerodostojnost

Marginalna vjerodostojnost je vjerodostojnost koja je integrirana duž prostora parametara. U Bayesovskoj statistici ona predstavlja vjerojatnost generiranja opaženog uzorka za sve moguće vrijednosti parametara.

Formalno, za skup neovisnih i jednako distribuiranih podataka $\mathbf{X} = (x_1, \ldots, x_n)$, gdje $x_i \sim \pi(x \mid \theta)$ prema nekoj vjerojatnosnoj distribuciji parametriziranoj s θ, pri čemu je θ sama slučajna varijabla opisana distribucijom, tj. $\theta \sim \pi(\theta \mid \alpha)$, marginalna vjerodostojnost $\pi(\mathbf{X} \mid \alpha)$ se dobije marginaliziranjem (integriranjem) θ:

$$\pi(\mathbf{X} \mid \alpha) = \int_\theta \pi(\mathbf{X} \mid \theta)\pi(\theta \mid \alpha)\mathrm{d}\theta,$$

pri čemu je $\pi(\theta \mid \alpha)$ apriorna gustoća, a $\pi(\mathbf{X} \mid \theta)$ vjerodostojnost.

Sažetak

Tradicionalni linearni modeli često pretpostavljaju jednostavne linearne odnose između zavisnih i nezavisnih varijabli, što može biti ograničavajuće za složenije podatke. Generalizirani linearni modeli (GLM-ovi) ublažavaju ovu pretpostavku, dopuštajući da očekivana vrijednost odziva ovisi o glatkoj monotonoj funkciji linearnog prediktora te omogućujući da odziv slijedi bilo koju distribuciju iz eksponencijalne familije (npr. normalnu, Poissonovu, binomnu, gamma, itd.). Generalizirani aditivni model (GAM) je proširenje GLM-a u kojem linearni prediktor ovisi linearno o glatkim funkcijama prediktora. Parametarski oblik ovih funkcija nije unaprijed poznat, kao ni stupanj glatkoće prikladan za svaku od njih, nego se procjenuje iz podataka.

U ovom radu prikazali smo kako odabir baze za glatke funkcije zajedno s odgovarajućim mjerama kazne za njihovu "vijugavost" omogućuje preoblikovanje procjene generaliziranog aditivnog modela u problem procjene parametara zaglađivanja i koeficijenata modela za problem maksimizacije penalizirane vjerodostojnosti. U praksi se problem maksimizacije penalizirane vjerodostojnosti rješava penaliziranom iterativnom metodom najmanjih kvadrata (PIRLS), dok se parametri zaglađivanja mogu procijeniti koristeći zaseban kriterij poput unakrsne validacije ili REML metode. Naša rasprava o splajnovima predstavila je širok spektar zaglađivača koji koriste linearno proširenje baze i kvadratnu penalizaciju, stoga odgovaraju ovom okviru. Osim toga, istražili smo dualnost između glatkih funkcija i slučajnih efekata, osobito iz bayesovske perspektive te uključili generalizirane aditivne mješovite modele (GAMM-ove).

Na kraju smo metode razvijene u početnim poglavljima primijenili na dva praktična primjera, oba iz područja vremenskih nizova. U prvom primjeru modelirali smo broj smrti kroz vrijeme u gradu Chicagu. Primijenili smo GAM-ove uvodeći glatke funkcije kovarijata, što su u ovom slučaju varijable koje opisuju kakvoću zraka. Model smo dodatno unaprijedili uvođenjem "distributed lag" modela s uključenim interakcijama kovarijata. U drugom primjeru promatrali smo prosječnu temperaturu u Engleskoj kroz nekoliko stoljeća s posebnim naglaskom na modeliranje trenda i sezonalnosti. Korištenjem GAMM-ova uspješno smo uključili korelaciju reziduala, što je ključan aspekt pri radu s vremenskim nizovima.

Summary

Traditional linear models often assume simple linear relationships between dependent and independent variables, which can be limiting for more complex data. Generalized linear models (GLMs) relax this assumption by allowing the expected value of the response to depend on a smooth monotonic function of the linear predictor and by permitting the response to follow any distribution from the exponential family (e.g., normal, Poisson, binomial, gamma, etc.). A generalized additive model (GAM) is an extension of the GLM where the linear predictor depends linearly on smooth functions of the predictors. The parametric form of these functions and the degree of smoothness appropriate for each are not known a priori but are estimated from the data.

In this thesis, we illustrated how selecting a basis for the smooth functions along with appropriate penalty measures for their "wiggliness" allows the estimation of a generalized additive model to be reframed as a problem of estimating smoothing parameters and model coefficients for a penalized likelihood maximization problem. In practice, maximizing penalized likelihood is typically addressed through the penalized iterative reweighted least squares (PIRLS) approach. Meanwhile, the estimation of smoothing parameters can be performed using alternative criteria like cross-validation or the REML method. Our discussion on splines introduced various smoothers that utilize linear basis expansion and quadratic penalization, which align well within this framework. Additionally, we explored the duality between smooth functions and random effects, particularly from a Bayesian perspective, and included generalized additive mixed models (GAMMs).

Finally, we applied the methods developed in the initial chapters to two practical examples, both from the field of time series analysis. In the first example, we modeled the number of deaths over time in the city of Chicago. We applied GAMs by introducing smooth functions of covariates, which in this case are variables describing air quality. The model was further enhanced by introducing a "distributed lag" model with included covariate interactions. In the second example, we observed the average temperature in England over several centuries, with a particular focus on modeling trend and seasonality. By using GAMMs, we successfully included residual correlation, which is a crucial aspect when working with time series data.

Životopis

Rođena sam 1. lipnja 1999. godine u Zagrebu. Nakon završene osnovne škole Stenjevec, upisujem prirodoslovno-matematičku XV. gimnaziju u Zagrebu (MIOC), gdje sam maturirala 2018. godine. Tijekom srednjoškolskog obrazovanja rado sam pohađala natjecanja iz matematike i logike. Svoj matematički put nastavljam upisom na Matematički odsjek Prirodoslovno-matematičkog fakulteta u Zagrebu. 2021. godine kao prvostupnik matematike upisujem diplomski studij Matematička statistika. Tijekom studiranja dodijeljene su mi dvije nagrade za najuspješnije studente na Matematičkom odsjeku. Svoje studentske dane upotpunila sam držanjem dodatne matematike u MIOC-u i volontiranjem u udruzi Mladi nadareni matematičari Marin Getaldić pripremajući učenike za matematička natjecanja.