Todo list

Matematika pro znalostní inženýrství

Karel Klouda, Štěpán Starosta, Daniel Vašata[1]

KAM FIT ČVUT

LS 2016/2017

MANUAL BUILD

Obsah

[1] karel.klouda@fit.cvut.cz, stepan.starosta@fit.cvut.cz, daniel.vasata@fit.cvut.cz

Seznam obrázků

Předmluva

Zde bude předmluva.

neco jako navod k pouzivani terminologie ruzna, budeme uvadet anglicka priklady cviceni

Kapitola 1

Úvod

1.1 Regresní analýza

Představme si následující situaci: snažíme se prodat nemovitost (řekněme v Praze) a netušíme, za kolik ji máme potenciálním zájemcům nabídnout. Chceme, aby cena nebyla příliš nízká a my neprodělali, ale ani příliš vysoká, abychom vůbec dostali nějaké nabídky. Potáhne-li se prodej měsíce, můžeme prodělat mnoho peněz např. tím, že nemovitost v tomto čase nepronajímáme.

Můžeme postupovat tak, že se zeptáme (= zaplatíme) „odborníka“ z realitní kanceláře, ale jelikož už jsme se s několika realitními makléři setkali a poznali, co jsou zač, tato možnost pro nás není myslitelná. Další způsob je, že budeme brouzdat po realitních serverech a hledat podobné nemovitosti a cenu nastřelíme podle nabytého dojmu. Chceme-li být ale sofistikovanější a mít větší jistotu, že náš odhad není úplně mimo, můžeme postupovat následovně.

Napíšeme si skript, který stáhne data z realitních serverů a uloží je v nějaké strukturované podobě (ideálně tabulce či více tabulkách). Pro jednoduchost uvažujme, že budeme znát u každé nabídky toto:

- Y = cenu, za kterou se prodává,
- X_1 = užitnou plochu,
- X_2 = počet místností,
- X_3 = vzdálenost od nejbližší zastávky metra.

Jako Y jsme si označili veličinu, kterou pro naši nemovitost neznáme a kterou se snažíme *predikovat*, jako X_i jsme si naopak označili veličiny, které pro naši nemovitost umíme zjistit a o kterých věříme, že cenu Y ovlivňují. Cílem regresní analýzy je zjistit a zformulovat (matematicky, jak jinak) tento vliv.

Nyní přejdeme do trochu formálnějšího jazyka a řádně zformulujeme výše naznačené: snažíme se predikovat hodnotu nějaké spojité[1] *vysvětlované proměnné* Y s tím, že před-

[1] O spojité vysvětlované proměnné mluvíme, pokud neočekáváme, že se její hodnoty budou vybírat z malé konečné množiny. Např. cena nemovitosti může nabývat hodnot od několika desítek tisíc po stovky milionů korun, a tak se považuje za spojitou. Kdybychom se ale např. snažili pouze určit, zda se nemovitost prodá (což označíme např. jako $Y = 1$) či neprodá ($Y = 0$) do měsíce, bude mít Y pouze dvě hodnoty a již nebude mít smysl ji považovat za spojitou.

pokládáme závislost této proměnné na jiných veličinách $X_1, X_2, \ldots, X_p$, tzv. *příznacích*. Jelikož nedoufáme, že tato závislost je perfektní v tom smyslu, že pro stejné hodnoty příznaků X_1 až X_p dostaneme vždy stejnou hodnotu vysvětlované proměnné, modelujeme tuto závislost následovně:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon, \tag{1.1}$$

kde ε je náhodná veličina a $f : \mathbb{R}^p \to \mathbb{R}$ je (neznámá[2]) reálná funkce p reálných proměnných.

Přidáním náhodné veličiny ε do modelu (1.1) vyjadřujeme jeho očekávanou nedokonalost. Jinými slovy: věříme, že příznaky X_1 až X_p ovlivňují hodnotu proměnné Y, ale současně nevěříme, že by již neexistovaly jiné vlivy (např. jiné příznaky), které v našem modelu nepostihujeme. Tato neúplná znalost se projeví v tom, že Y se nerovná přesně hodnotě $f(X_1, \ldots, X_p)$ a jelikož vzniklou chybu $Y - f(X_1, \ldots, X_p)$ neznáme a nevíme moc o jejím chování (jinak bychom ji do modelu zahrnuli), považujeme ji za náhodnou veličinu.

Vraťme se nyní k našemu hypotetickému příkladu s prodejem nemovitostí, kdy máme pouze tři (tj. $p = 3$) příznaky a náš model vypadá takto:

$$Y = f(X_1, X_2, X_3) + \varepsilon. \tag{1.2}$$

Do náhodné veličiny ε „schováváme“ vlivy příznaků, které neznáme nebo nezahrnujeme do našeho modelu (např. stáří budovy, počet koupelen, počet oken, ...) ale vlastně i např. chyby, nekonzistence dat a jiné podivnosti v měření příznaků, které v modelu zahrnujeme (např. někdo do užitné plochy zahrnuje balkón či sklep a někdo ne a mi to v datech nemáme vyznačené, nebo bereme vzdálenost od metra jako 500 metrů, ale z dat nevidíme, že bychom cestou museli přeplavat Vltavu).

Nyní se dostáváme k samotné formulaci problému regresní analýzy. Vágně řečeno: naším cílem je najít odhad funkce f tak, aby chyba modelu byla co nejmenší. Odhad funkce f budeme značit $\widehat{f}$, jak je ve statistice a příbuzných oborech zvykem: když někde uvidíte něco se stříškou, skoro jistě je to odhad něčeho. Abychom toto mohli řešit, musíme nejdříve specifikovat, co se myslí chybou modelu a v jakém smyslu má být nejmenší. Jako chybu nejčastěji bereme rozdíl mezi skutečnou hodnotou vysvětlované proměnné Y a její predikovanou hodnotou $f(\boldsymbol{X}) = f(X_1, \ldots, X_p)$, kde $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ je (sloupcový) vektor veličin $X_1, \ldots, X_p$, tj. $Y - f(\boldsymbol{X})$. Velikost (míru) chyby potom značíme $L(Y, f(\boldsymbol{X}))$, kde $L : \mathbb{R}^2 \to \mathbb{R}$ je nějaká nezáporná funkce, kterou obvykle volíme jako

$$L(Y, f(\boldsymbol{X})) = (Y - f(\boldsymbol{X}))^2.$$

Hledáme tedy reálnou funkci p proměnných $\widehat{f} : \mathbb{R}^p \to \mathbb{R}$ takovou, aby minimalizovala míru chyby $L(Y, f(\boldsymbol{X}))$. To ale stále není moc dobře zformulovaná úloha: víme např., že $(Y - f(\boldsymbol{X}))^2 = \varepsilon^2$ je náhodná veličina, a ta se těžko minimalizuje, neb minimalizování náhodné veličiny nedává smysl. Co ovšem umíme rozumně vyjádřit, je *obvyklá hodnota chyby*, neboli její střední hodnota $\mathbb{E}\, L(Y, f(\boldsymbol{X}))$. Abychom ale mohli vyjádřit střední hodnotu, potřebujeme znát pravděpodobnostní rozdělení veličin[3] $Y, X_1, X_2, \ldots, X_p$. Ty můžeme za-

[2] A podle některých názorů nejen neznámá, ale i neexistující funkce. Je to velice zajímavá filozofická debata, kterou ale přeskočíme.

[3] Veličina Y je jistě také náhodnou, veličiny X_i ale už za náhodné považovat nemusíme a můžeme je chápat jako jakési parametry. Toto rozlišení není ale nyní moc důležité.

chytit[4] např. do hustoty pravděpodobností, kterou si označme $h_{Y,\boldsymbol{X}}(y, \boldsymbol{x})$ (je to nezáporná funkce $p+1$ proměnných). Jak víme, střední hodnota je pak rovna následujícímu integrálu

$$\mathbb{E}\, L(Y, f(\boldsymbol{X})) = \int_{\mathbb{R}^{p+1}} L(y, f(\boldsymbol{x}))h_{Y,\boldsymbol{X}}(y, \boldsymbol{x})\,\mathrm{d}y\,\mathrm{d}\boldsymbol{x}. \tag{1.3}$$

Mohlo by se zdát, že máme vyhráno: pro danou množinu příznaků a námi zvolenou míru chyby najdeme mezi reálnými funkcemi p proměnných tu, která minimalizuje hodnotu integrálu (1.3) a to bude ta naše nejlepší hledaná funkce. Bylo by to sice pěkné, ale v reálu je tento přístup nepoužitelný hned z několika důvodů[5]: ten zásadní je, že netušíme, jak vypadá hustota pravděpodobnosti $h_{Y,\boldsymbol{X}}$. Např. u našeho modelu (1.2) bychom potřebovali znát pro všechny čtveřice čísel (y, x_1, x_2, x_3) pravděpodobnost, že náhodně vybraná nemovitost ze všech možných (i hypotetických) nemovitostí má cenu y a má právě x_1 metrů čtverečních, x_2 místností a leží x_3 metrů od metra. Takové informace prakticky nikdy nemáme a výpočet (1.3) je tedy nemožný.

Dalším problémem by bylo i prohledávání prostoru všech funkcí $f : \mathbb{R}^p \to \mathbb{R}$. I když existují postupy, jak minimalizovat nějaký výraz, kde proměnná je funkce (tímto problémem se zabývá oblast matematiky, která se nazývá *variační počet*), i ty se potřebují omezit například pouze na funkce spojité apod.

1.1.1 Regresní analýza v reálné situaci

Jak jsme řekli výše, těžko můžeme doufat, že budeme znát hustotu pravděpodobnosti $h_{Y,\boldsymbol{X}}$. Co obvykle známe, je pouze nějaký (snad) náhodný výběr z možných hodnot veličin Y a $\boldsymbol{X}$. Např. v našem příkladě modelování cen nemovitostí budeme znát cenu a parametry N bytů, které jsme někde získali. Nemáme tedy všechny možné kombinace hodnot Y a $\boldsymbol{X}$, ale jen nějaký jejich výběr $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_N, \boldsymbol{x}_N)$. Budeme proto hledat $\widehat{f}$ takovou, aby se minimalizovala průměrná chyba pouze na těchto N měřeních; tuto chybu pro dané f spočítáme takto:

$$\frac{1}{N}\sum_{k=1}^{N} L(y_k, f(\boldsymbol{x}_k)). \tag{1.4}$$

Kde je problém teď? Problém je ten, že množina všech reálných funkcí p proměnných je příliš bohatá a není problém najít takovou funkci $\widehat{f}$, že suma (1.4) bude nulová. Uvažujme opět náš příklad s nemovitostmi a řekněme, že máme nasbíraná data pro $N = 100$ nemovitostí. Jak si ukážeme v další kapitole, skoro vždy existuje polynom jedné proměnné (např. x_1) a stupně 100 takový, že pokud jej dosadíme za funkci f, je suma (1.4) nulová (vizte také cvičení 1.1).

Znamená to, že jsme našli perfektní model ceny nemovitostí v Praze? Nikoli, znamená to, že náš model perfektně modeluje 100 náhodně vybraných měření. V takové situaci, kdy modelujeme data a nikoli skutečnost, mluvíme o tzv. *přeučení modelu*.

Jak se přeučení vyhnout? Jedna z cest je, že si nějakým způsobem omezíme množinu, ve které hledáme funkci $\widehat{f}$. Například se můžeme omezit pouze na polynomy stupně nejvýše 4, na funkce obsahující pouze lineární kombinace goniometrických funkcí příznaků atp.

[4] Dále předpokládáme, čistě pro jednoduchost, že proměnné X_i jsou taky spojité a má smysle mluvit o hustotě a integrování.

[5] Žádný z těchto důvodů není ten, že se tam vyskytuje vzoreček s integrálem a přes to nejede vlak!

Vraťme se k příkladu s cenami nemovitostí v Praze. Řekněme, že máme důvod věřit, že v modelu (1.2) je dostačující hledat $\widehat{f}$ mezi polynomy tří proměnných stupně jedna, neboli že existují čísla w_0, w_1, w_2 a w_3 tak, že hledaná funkce $\widehat{f}$ má tvar

$$\widehat{f}(x_1, x_2, x_3) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3. \tag{1.5}$$

Mějme opět $N = 100$ měření, neboli 100 čtveřic čísel $(y_i, x_{i,1}, x_{i,2}, x_{i,3})$ pro $i = 1, 2, \ldots, 100$. Zvolme (jako obvykle) míru chyby $L(Y, f(\boldsymbol{X})) = (Y - f(\boldsymbol{X}))^2$. Hledání funkce $\widehat{f}$ tvaru (1.5) se mění na úkol najít hodnotu čísel w_0 až w_3 tak, že získáme minimální hodnotu výrazu

$$\frac{1}{100} \sum_{i=1}^{100} (y_i - w_0 + w_1 x_{i,1} + w_2 x_{i,2} + w_3 x_{i,3})^2,$$

který odpovídá průměrné chybě na našich sto měřeních. Jedná se vlastně o hledání minima funkce čtyř proměnných, a jak uvidíme v další kapitole, konkrétně v tomto případě umíme řešení relativně snadno najít.

Šťouravý čtenář by mohl namítnout, že jsme sice razantním omezením výběru funkce $\widehat{f}$ zabránili přeučení, ale nabízí se otázka, zda jsme to nepřehnali natolik, že náš model již není schopen dobře aproximovat ideální[6] volbu funkce f? Odpověď šťouravému čtenáři je vyhýbavá: možná jsme to přehnali; najít takovou volbu komplexnosti modelu, která brání přílišnému přizpůsobení se našim konkrétním datům (přeučení), ale současně nezabraňuje našemu modelu být dostatečně blízko ideální volbě f, je jedním z klíčových úkolů znalostního inženýra. Více o tomto problému budeme mluvit v části **??** věnované *bias-variance tradeoff*[7].

1.1.2 Nezničitelná chyba

Shrňme si, co jsme se snažili doposud vysvětlit: hledáme model ve tvaru (1.1), který nám umožní predikovat hodnotu Y na základě znalosti příznaků X_1 až X_p. Proměnnou tohoto modelu je funkce f, jejíž optimální volba je funkce minimalizující střední chybu modelu danou integrálem (1.3). Tento integrál bohužel ale umíme spočítat jen za nereálného předpokladu, že perfektně známe chování všech veličin (tj. známe odpovídající hustotu pravděpodobnosti). Tento problém se řeší tak, že se namísto minimalizace (1.3) hledá $\widehat{f}$ minimalizující diskrétní aproximaci (1.4). To ale vede k nebezpečí přeučení, tedy přílišnému přizpůsobení se nasbíranému vzorku dat. Obranou proti přeučení může být výrazné omezení možností, jak volit $\widehat{f}$, ale ani to se nesmí přehnat, abychom se stále s naší funkcí mohli přiblížit k optimálnímu řešení.

Je jasné, že z velké části bude zdrojem chyby našeho modelu odchylka našeho odhadu $\widehat{f}$ od ideální volby f. I kdybychom ale měli štěstí, a náš odhad by se s optimální volbou f shodoval, neznamená to, že dostáváme bezchybný model poskytující dokonalé predikce.

Stále nám tam totiž zbývá náhodná veličina $\varepsilon = Y - f(\boldsymbol{X})$. Jistě můžeme předpokládat, že její střední hodnota je nula: kdyby totiž $\mathbb{E}\,\varepsilon = c$, bude vezmeme namísto $f(\boldsymbol{X})$ funkci $f(\boldsymbol{X}) - c$ a pak již příslušná střední hodnota $Y - f(\boldsymbol{X}) - c$ bude rovna nule. Připomeňme, že díky předpokladu $\mathbb{E}\,\varepsilon = 0$ platí pro rozptyl ε, že

$$\operatorname{var} \varepsilon = \mathbb{E}\,\varepsilon^2 - (\mathbb{E}\,\varepsilon)^2 = \mathbb{E}\,\varepsilon^2.$$

[6] Ideální ve smyslu „minimalizující hodnotu (1.3).“

[7] Je to tak zažitý pojem, že jej nebudeme překládat

Pro střední hodnotu kvadrátu chyby tedy platí i pro optimální volbu funkce f, že

$$\mathbb{E}(Y - f(\boldsymbol{X}))^2 = \mathbb{E}\,\varepsilon^2 = \operatorname{var}\varepsilon.$$

Střední hodnota kvadrátu chyby tedy nikdy nebude nižší, než je rozptyl náhodné veličiny ε a ten nebude nikdy nulový[8], pokud není ε (skoro jistě) nulová funkce a náš model dokonale přesný (tj. není to ani tak model, jako zákon).

1.2 Cvičení

Cvičení 1.1: Najděte polynom (jedné reálné proměnné) co nejnižšího stupně, který prochází body $(0,0), (1,1)$ a $(2,3)$.

[8] Pomíjíme možnost, že rozptyl nebude existovat, resp. nebude spočítatelný, vizte např. Cauchyho rozdělení.

Kapitola 2

Lineární regrese a metoda nejmenších čtverců

V důsledku předchozích úvah tedy předpokládáme, že platí

$$Y = f(\boldsymbol{X}) + \varepsilon = w_0 + \sum_{i=1}^{p} X_i w_i + \varepsilon = \boldsymbol{X}^T \boldsymbol{w} + \varepsilon,$$

kde $\boldsymbol{X} = (1, X_1, \ldots, X_p)$ je vektor příznaků s přidaným konstantním příznakem $X_0 = 1$, $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)$ je vektor vah (koeficientů) lineární závislosti a ε je náhodná veličina reprezentující nedeterministickou část Y. Řekněme, že máme $N = 100$ měření, neboli 100 čtveřic čísel $(y_i, x_{i,1}, x_{i,2}, x_{i,3})$ pro $i = 1, 2, \ldots, 100$. Tato měření můžeme zapsat do vektoru $\boldsymbol{y} \in \mathbb{R}^N$ a matice $\mathbf{X} \in \mathbb{R}^{N,p+1}$, která pro N měření a p příznaků budeme značit takto:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad \text{a} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{pmatrix}. \tag{2.1}$$

Poznamenejme, že v dalším textu budeme i,j-tý prvek $x_{i,j}$ matice $\mathbf{X}$ značit také x_{ij}. Pro náš příklad tak máme sto složkový vektor $\boldsymbol{y}$ s cenami nemovitostí a matici $\mathbf{X}$ o rozměrech 100×3. Hledání funkce f se díky předpokladu, že funkce má tvar (1.5), mění na úkol najít hodnotu čísel $w_0, \ldots, w_3$. Vezměme si za míru chyby obvyklou volbu

2.1 Matematická formulace problému lineární regrese

Základním modelem lineární regrese je model

$$\mathbb{E}(Y | X_1 = x_1, \ldots, X_p = x_p) = w_0 + \sum_{i=1}^{p} X_i w_i = \boldsymbol{x}^T \boldsymbol{w}. \tag{2.2}$$

Přičemž v takovémto zadání je jedno, zda uvažujeme X jako náhodné nebo ne. Vektor $\boldsymbol{X} = (1, X_1, \ldots, X_p)$ je vektor příznaků s přidaným konstantním příznakem $X_0 = 1$ a $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)^T$ je vektor neznámých vah (koeficientů) lineární závislosti.

Mějme nyní náhodný výběr z výše uvedeného náhodného modelu provedeného v různých bodech $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. To znamená, že máme n párů typu $(Y_i, \boldsymbol{x}_i)$, kde Y_i je náhodná veličina ukazující výsledek vysvětlované proměnné v i-tém bodě $\boldsymbol{x}_i = (1, x_{i;1}, \ldots, x_{i;p})^T$ konkrétních hodnot příznaků.

Uspořádáme-li veličiny $Y_1, \ldots, Y_n$ do náhodného vektoru $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ a hodnoty bodů $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ do matice $\mathbf{X} \in \mathbb{R}^{N,p+1}$, v jejíž řádcích jsou zapsané složky $\boldsymbol{x}_i$:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{a} \quad \mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{1;0} & x_{1;1} & x_{1;2} & \cdots & x_{1;p} \\ x_{2;0} & x_{2;1} & x_{2;2} & \cdots & x_{2;p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n;0} & x_{n;1} & x_{n;2} & \cdots & x_{n;p} \end{pmatrix},$$

kde $x_{i;0} = 1$ pro každé $i = 1, \ldots, n$, můžeme psát

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}. \tag{2.3}$$

Tento vztah zapsaný po řádcích dává

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{w} + \varepsilon_i$$

pro každé $i = 1, \ldots, n$. Náhodný vektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ reprezentuje všechny možné neznámé náhodné příspěvky k vysvětlovaným veličinám $Y_1, \ldots, Y_n$, které nepocházejí od hodnot příznaků $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Poznamenejme, že v dalším textu budeme i, j-tý prvek matice $\mathbf{X}$ značit také x_{ij}.

Jediný předpoklad na $\boldsymbol{\varepsilon}$ je v tuto chvíli tedy $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$. postupně budeme přidávat i další předpoklady, které nám umožní určovat charakteristiky získaného odhadu $\hat{\boldsymbol{w}}$.

nepoužijeme \wide-hat?

Jak už bylo zmíněno v předchozí sekci, rozumný odhad $\hat{\boldsymbol{w}}$ neznámé hodnoty $\boldsymbol{w}$ můžeme získat jako argument minima residuálního součtu čtverců $\mathrm{RSS}(\boldsymbol{w})$ definovaného vztahem

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{p} x_{i,j} w_j \right)^2 = (\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{w}). \tag{2.4}$$

Tedy

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^{p+1}} \mathrm{RSS}(\boldsymbol{w}). \tag{2.5}$$

Kritický bod vzhledem ke složkám $\boldsymbol{w}$ získáme parciálním zderivováním $\mathrm{RSS}(\boldsymbol{w})$ podle složek $\boldsymbol{w}$ a položením rovno 0. Získáme vztah

$$-2\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{w}) = 0.$$

Zderivováním podruhé získáme Hessián jako

$$\mathbf{H} = \left(\frac{\partial^2 RSS(\boldsymbol{w})}{\partial w_i \partial w_j} \right) = 2\mathbf{X}^T\mathbf{X}$$

což je matice nezávislá na $\boldsymbol{w}$. Navíc $\boldsymbol{c}^T\mathbf{W}\boldsymbol{c} \geq 0$ pro každé $\boldsymbol{c} \in \mathbb{R}^n$ a tedy jakýkoliv kritický bod bude lokálním minimem. Poznamenejme, že tento bod bude ostrým lokálním minimem, právě když matice $\mathbf{X}^T\mathbf{X}$ bude mít plnou hodnost a bude tak pozitivně definitní, viz věta A.1.

To znamená, že nutnou podmínkou minima je platnost vztahu

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{w}} = \mathbf{X}^T\boldsymbol{Y}, \tag{2.6}$$

který bývá nazýván *normální rovnice*.

Předpokládejme nyní, že hodnost $h(\mathbf{X})$ matice $\mathbf{X}$ je $p+1$. Z toho již nutně plyne, že hodnost matice $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{p+1,p+1}$ je také $p+1$ a tedy, že je tato matice regulární. Celkově tak získáváme

$$\hat{\boldsymbol{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}. \tag{2.7}$$

Při porovnávání s jinými metodami budeme takto získaný odhad značit $\widehat{\boldsymbol{w}}_{\mathrm{OLS}}$, kde index OLS je z anglického *ordinary least squares*, tedy obyčejné nejmenší čtverce.

2.2 Cvičení

Cvičení 2.1:

je treba doladit znaceni, dopsat text + napovedu; co si o tom zadani myslite?
DV: Mě to přijde dobrý. Akorát bych tu otázku zformuloval jako: najděte hodnoty koeficientů A_n, B_n minimalizující výraz...

$$\begin{aligned} Y =& w_0 + w_1 \cos\left(\frac{2\pi x}{N}\right) + w_2 \sin\left(\frac{2\pi x}{N}\right) + w_3 \cos\left(\frac{4\pi x}{N}\right) + w_4 \sin\left(\frac{4\pi x}{N}\right) + \ldots + \\ &+ w_{2N-1} \cos\left(\frac{2N\pi x}{N}\right) + w_{2N} \sin\left(\frac{2N\pi x}{N}\right) \end{aligned}$$

Přeznačme koeficienty w_n na koeficienty A_n a B_n tak, aby platilo

$$Y = \sum_{n=0}^{N} A_n \cos\left(\frac{2n\pi x}{N}\right) + \sum_{n=0}^{N} B_n \sin\left(\frac{2n\pi x}{N}\right).$$

Nelezněte minimum funkce

$$\sum_{k=1}^{N}\left(y_k - \sum_{n=0}^{N} A_n \cos\left(\frac{2n\pi x_k}{N}\right) + \sum_{n=0}^{N} B_n \sin\left(\frac{2n\pi x_k}{N}\right)\right)^2$$

Kapitola 3

Pravděpodobnostní pohled na lineární regresi

Základním modelem lineární regrese je model

$$\mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = w_0 + \sum_{i=1}^{p} x_i w_i = \boldsymbol{x}^T \boldsymbol{w}. \tag{3.1}$$

Přičemž v takovémto zadání je jedno, zda uvažujeme X jako náhodné nebo ne. Vektor $\boldsymbol{X} = (1, X_1, \ldots, X_p)$ je vektor příznaků s přidaným konstantním příznakem $X_0 = 1$ a $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)$ je vektor neznámých vah (koeficientů) lineární závislosti.

Mějme nyní náhodný výběr z výše uvedeného náhodného modelu provedeného v různých bodech $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. To znamená, že máme N párů typu $(Y_i, \boldsymbol{x}_i)$, kde Y_i je náhodná veličina ukazující výsledek vysvětlované proměnné v i-tém bodě $\boldsymbol{x}_i = (1, x_{i;1}, \ldots, x_{i;p})^T$ konkrétních hodnot příznaků.

Uspořádáme-li veličiny $Y_1, \ldots, Y_n$ do náhodného vektoru $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ a hodnoty bodů $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ do matice $\mathbf{X} \in \mathbb{R}^{N,p+1}$, v jejíž řádcích jsou zapsané složky $\boldsymbol{x}_i$, můžeme psát

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}. \tag{3.2}$$

Tento vztah zapsaný po řádcích dává

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{w} + \varepsilon_i$$

pro každé $i = 1, \ldots, n$. Náhodný vektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ reprezentuje všechny možné neznámé náhodné příspěvky k vysvětlovaným veličinám $Y_1, \ldots, Y_n$, které nepocházejí od hodnot příznaků $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Poznamenejme, že v dalším textu budeme i, j-tý prvek matice $\mathbf{X}$ značit také x_{ij}.

Jediný předpoklad na $\boldsymbol{\varepsilon}$ je v tuto chvíli tedy $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$. postupně budeme přidávat i další předpoklady, které nám umožní určovat charakteristiky získaného odhadu $\hat{\boldsymbol{w}}$. Jak už bylo zmíněno v předchozí sekci, rozumný odhad $\hat{\boldsymbol{w}}$ neznámé hodnoty $\boldsymbol{w}$ můžeme získat jako argument minima residuálního součtu čtverců $\mathrm{RSS}(\boldsymbol{w})$ definovaného vztahem

$$\mathrm{RSS}(\boldsymbol{w}) = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{w})^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{w}) = \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{w}\|^2. \tag{3.3}$$

Výsledný odhad je určen platností tzv. normálních rovnic

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{w}} = \mathbf{X}^T\boldsymbol{Y}. \tag{3.4}$$

V případě, že má matice $\mathbf{X}$ plnou hodnost je určen jednoznačně a platí

$$\hat{\boldsymbol{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{Y}. \tag{3.5}$$

3.1 Vlastnosti odhadu vektorů vah

Předpokládejme tedy, že data pocházejí z modelu (3.2) a předpokládejme nyní pouze $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$. Z toho plyne $\mathbb{E}\,\boldsymbol{Y} = \mathbf{X}\boldsymbol{w}$. Snadno dokážeme následující tvrzení.

Věta 3.1: Odhad $\widehat{\boldsymbol{w}}_{\mathrm{OLS}}$ získaný metodou nejmenších čtverců je za předpokladu $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ nestranný, tj. $\mathbb{E}\,\widehat{\boldsymbol{w}}_{\mathrm{OLS}} = \boldsymbol{w}$.

Důkaz.

$$\mathbb{E}\,\widehat{\boldsymbol{w}}_{\mathrm{OLS}} = \mathbb{E}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{Y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\mathbb{E}\,\boldsymbol{Y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{w} = \boldsymbol{w}$$

□

Předpokládejme navíc $\operatorname{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$, tj. že chyby ε_i v jednotlivých bodech $\boldsymbol{x}_i$ jsou nekorelované veličiny se stejným rozptylem σ^2.

Věta 3.2: Za předpokladu $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ a $\operatorname{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$ platí $\operatorname{var}\widehat{\boldsymbol{w}}_{\mathrm{OLS}} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Důkaz. Protože $\mathbf{X}^T\mathbf{X}$ je symetrická, je také $(\mathbf{X}^T\mathbf{X})^{-1}$ symetrická a platí

$$\begin{aligned}\operatorname{var}\widehat{\boldsymbol{w}}_{\mathrm{OLS}} &= \operatorname{var}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{Y}\right) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\operatorname{var}\boldsymbol{Y})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.\end{aligned}$$

□

Dále se pojďme zabývat residuálním součtem čtverců. Po dosazení vztahu (3.5) do RSS a označení $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ dostaneme

$$\mathrm{RSS}(\hat{\boldsymbol{w}}) = \|\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{w}}\|^2 = \|\boldsymbol{Y} - \mathbf{P}\boldsymbol{Y}\|^2 = \|(\mathbf{I}-\mathbf{P})\boldsymbol{Y}\|^2,$$

kde $\mathbf{I}$ je jednotková matice z $\mathbb{R}^{n,n}$. Matice $\mathbf{P}$ je projekce do lineárního obalu sloupců matice $\mathbf{X}$. Dále platí

$$\mathbf{P}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}$$

a

$$(\mathbf{I}-\mathbf{P})^2 = \mathbf{I}^2 - \mathbf{I}\mathbf{P} - \mathbf{P}\mathbf{I} + \mathbf{P}^2 = \mathbf{I} - \mathbf{P}.$$

Obě matice $\mathbf{P}$ i $\mathbf{I}-\mathbf{P}$ jsou tedy idempotentní. Protože $\mathbf{X}^T\mathbf{X}$ je zjevně symetrická, je symetrická i její inverze $(\mathbf{X}^T\mathbf{X})^{-1}$ a tudíž i matice $\mathbf{P}$ a $\mathbf{I}-\mathbf{P}$. Z idempotence $\mathbf{I}-\mathbf{P}$ plyne

$$\mathrm{RSS}(\hat{\boldsymbol{w}}) = \|(\mathbf{I}-\mathbf{P})\boldsymbol{Y}\|^2 = \boldsymbol{Y}^T(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\boldsymbol{Y} = \boldsymbol{Y}^T(\mathbf{I}-\mathbf{P})\boldsymbol{Y}.$$

Nakonec si všimneme, že

$$(\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}.$$

Nyní už můžeme dokázat následující větu.

Věta 3.3: Nechť $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ a $\operatorname{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$. Potom $s^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{w}})}{n-p-1}$ je nestranný odhad σ^2.

Důkaz. Pro střední hodnotu podle předchozích vyjádření dostáváme

$$\begin{aligned}\mathbb{E}\,\mathrm{RSS}(\hat{\boldsymbol{w}}) &= \mathbb{E}\,\boldsymbol{Y}^T(\mathbf{I}-\mathbf{P})\boldsymbol{Y} = \mathbb{E}\operatorname{Tr}\boldsymbol{Y}^T(\mathbf{I}-\mathbf{P})\boldsymbol{Y} = \mathbb{E}\operatorname{Tr}(\mathbf{I}-\mathbf{P})\boldsymbol{Y}\boldsymbol{Y}^T \\ &= \operatorname{Tr}(\mathbf{I}-\mathbf{P})\,\mathbb{E}\,\boldsymbol{Y}\boldsymbol{Y}^T = \operatorname{Tr}(\mathbf{I}-\mathbf{P})(\operatorname{var}\boldsymbol{Y}+\mathbb{E}\,\boldsymbol{Y}(\mathbb{E}\,\boldsymbol{Y})^T) = \operatorname{Tr}(\mathbf{I}-\mathbf{P})\sigma^2\mathbf{I}+\operatorname{Tr}(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{w}(\mathbf{X}\boldsymbol{w})^T \\ &= \sigma^2\operatorname{Tr}(\mathbf{I}-\mathbf{P}) = \sigma^2(n-p-1),\end{aligned}$$

protože $(\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{0}$, $\mathbf{I}$ je jednotková matice $n \times n$ a

$$\operatorname{Tr}\mathbf{P} = \operatorname{Tr}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \operatorname{Tr}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \operatorname{Tr}\mathbf{I}_{p+1},$$

kde $\mathbf{I}_{p+1}$ je jednotková matice $(p+1)\times(p+1)$. □

3.2 Testy hypotéz a intervaly spolehlivosti

Stále uvažujme model (3.2) ale zesilme předpoklad o rozdělení $\boldsymbol{\varepsilon}$ na $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, tj. vektor náhodných odchylek $\boldsymbol{\varepsilon}$ má vícerozměrné normální rozdělení se střední hodnotou $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ a varianční maticí $\operatorname{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$ (viz dodatek B.4). Z věty B.15 tak dostáváme

$$\boldsymbol{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{w}, \sigma^2\mathbf{I}).$$

Následující věta představuje rozšíření vět 3.1 a 3.2.

Věta 3.4: Nechť $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Potom $\hat{\boldsymbol{w}} \sim \mathrm{N}\big(\boldsymbol{w}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\big)$.

Důkaz. Analogickým postupem jako v důkazu věty 3.2 plyne z vyjádření (3.5) a věty B.15. □

Z předpokladu normálního rozdělení $\boldsymbol{\varepsilon}$ lze odvodit rozdělení residuálního součtu čtverců $\mathrm{RSS}(\hat{\boldsymbol{w}})$.

Věta 3.5: Nechť $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Potom $\frac{\mathrm{RSS}(\hat{\boldsymbol{w}})}{\sigma^2} \sim \chi^2_{n-p-1}$.

Důkaz. Plyne z vyjádření $\mathrm{RSS}(\hat{\boldsymbol{w}}) = \boldsymbol{Y}^T(\mathbf{I}-\mathbf{P})\boldsymbol{Y}$ a věty B.16. □

Kombinací předchozích dvou vět dostaneme tvrzení, které je základem pro testování hypotéz o hodnotách jednotlivých složek vektoru vah $\boldsymbol{w} = (w_1, \ldots, w_{p+1})^T$ a také pro tvorbu pásu spolehlivosti pro regresní přímku.

Věta 3.6: Nechť $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Potom pro každé $\boldsymbol{c} = (c_1, \ldots, c_{p+1}) \in \mathbb{R}^{p+1}$ platí

$$T_{\boldsymbol{c}} = \frac{\boldsymbol{c}^T(\hat{\boldsymbol{w}} - \boldsymbol{w})}{\sqrt{s^2\boldsymbol{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{c}}} \sim \mathrm{t}_{n-p-1},$$

kde $s^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{w}})}{n-p-1}$.

Důkaz. Z věty 3.4 a definice B.10 vícerozměrného normálního rozdělení dostaneme

$$\boldsymbol{c}^T\hat{\boldsymbol{w}} \sim \mathrm{N}\Big(\boldsymbol{c}^T\boldsymbol{w}, \boldsymbol{c}^T\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{c}\Big)$$

a po provedení standardizace (v souladu s větou B.1)

$$\frac{\boldsymbol{c}^T(\hat{\boldsymbol{w}}-\boldsymbol{w})}{\sqrt{\sigma^2\boldsymbol{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{c}}} \sim \mathrm{N}(0,1).$$

Podle věty 3.5 má $\frac{\mathrm{RSS}(\hat{\boldsymbol{w}})}{\sigma^2}$ rozdělení χ^2_{n-p-1}, přičemž lze dokázat (viz [**?**, Theorem 3.5]), že $\hat{\boldsymbol{w}}$ a $\mathrm{RSS}(\hat{\boldsymbol{w}})$ jsou nezávislé veličiny. Z věty B.3 plyne, že

$$\frac{\frac{\boldsymbol{c}^T(\hat{\boldsymbol{w}}-\boldsymbol{w})}{\sqrt{\sigma^2\boldsymbol{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{c}}}}{\sqrt{\frac{\mathrm{RSS}(\hat{\boldsymbol{w}})}{(n-p-1)\sigma^2}}} = \frac{\boldsymbol{c}^T(\hat{\boldsymbol{w}}-\boldsymbol{w})}{\sqrt{s^2\boldsymbol{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{c}}} = T_{\boldsymbol{c}}$$

má Studentovo t rozdělení s $n-p-1$ stupni volnosti. □

Důsledek 3.7: Nechť $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Potom pro každé $i = 1, \ldots, p+1$ platí

$$T_i = \frac{\hat{w}_i - w_i}{\sqrt{s^2 v_{ii}}} \sim \mathrm{t}_{n-p-1},$$

kde $s^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{w}})}{n-p-1}$ a v_{ii} je i-tá diagonální složka matice $(\mathbf{X}^T\mathbf{X})^{-1}$.

Důkaz. Plyne z předchozí věty po dosazení $\boldsymbol{c} = (c_1, \ldots, c_{p+1})^T$, kde $c_i = 1$ a $c_j = 0$ pro každé $j \neq i$. □

Předchozí věta umožňuje testovat hypotézy o hodnotách jednotlivých složek vektoru vah $\boldsymbol{w}$. Např. pokud chceme testovat hypotézu $H_0 : w_1 = 0$ proti alternativě $H_A : w_1 \neq 0$, použijeme testovací statistiku

$$T_1 = \frac{\hat{w}_1 - 0}{\sqrt{s^2 v_{ii}}},$$

která má při platnosti H_0 Studentovo t rozdělení s $n-p-1$ stupni volnosti. Test na hladině α tedy provedeme porovnáním T_1 s kritickou hodnotou $t^{\alpha/2}_{n-p-1}$. Je-li $|T_1| \geq t^{\alpha/2}_{n-p-1}$, zamítneme H_0 a je-li $|T_1| < t^{\alpha/2}_{n-p-1}$, nezamítneme H_0.

Poladit ty nerovnosti.

Další možností je testovat hypotézy o hodnotách několika složek vektoru vah najednou. K tomu se hodí následující věta.

Věta 3.8: Nechť $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Potom pro libovolnou matici $\mathbf{A} \in \mathbb{R}^{q,p+1}$ s hodností q platí

$$F_{\mathbf{A}} = \frac{\Big(\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})\Big)^T\Big(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\Big)^{-1}\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})}{qs^2} \sim \mathrm{F}_{q,n-p-1}.$$

Důkaz. Podle věty 3.4 platí $\hat{\boldsymbol{w}} \sim \mathrm{N}\Big(\boldsymbol{w}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\Big)$. Tudíž podle věty B.15 dostáváme $\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w}) \sim \mathrm{N}\Big(\mathbf{0}, \sigma^2\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\Big)$.

Ověřme nyní předpoklady věty B.16 pro $\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})$ místo $\boldsymbol{X}$ a $\left(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\right)^{-1}/\sigma^2$ místo $\mathbf{A}$. Matice $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T$ je zjevně symetrická a positivně definitní. Tedy její inverze $\left(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\right)^{-1}$ existuje a je také symetrická a positivně definitní. Dále platí

$$\frac{\left(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\right)^{-1}}{\sigma^2}\sigma^2\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T=\mathbf{I}_q,$$

kde $\mathbf{I}_q$ je jednotková matice $q\times q$ - tedy idempotentní matice různá od $\mathbf{0}$. Protože $\operatorname{Tr}\mathbf{I}_q=q$, dostáváme, že

$$\left(\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})\right)^T\frac{\left(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\right)^{-1}}{\sigma^2}\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})\sim\chi^2_q$$

přičemž tato veličina je nezávislá na veličině $\mathrm{RSS}(\hat{\boldsymbol{w}})$, což plyne z věty 3.5 v knize [**?**]. Podle věty B.4 tedy platí

$$\frac{\left(\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})\right)^T\left(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\right)^{-1}\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{w})}{\sigma^2 q}\frac{\sigma^2(n-p-1)}{\mathrm{RSS}(\hat{\boldsymbol{w}})}=F_{\mathbf{A}}\sim\mathrm{F}_{q,n-p-1}.$$

□

Chceme-li tedy například testovat hypotézu $H_0: w_0=0$ a $w_2=1$ proti alternativě $H_A: w_0\neq 0$ nebo $w_2\neq 1$, zvolíme matici $\mathbf{A}\in\mathbb{R}^{2,p+1}$ jako

$$\mathbf{A}=\begin{pmatrix}1&0&0&0&\cdots\\0&0&1&0&\cdots\end{pmatrix},$$

vektor $\boldsymbol{c}$ jako $\boldsymbol{c}=(0,1)^T$ a použijeme testovací statistiku

$$F_{\mathbf{A}}=\frac{\left(\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{c})\right)^T\left(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\right)^{-1}\mathbf{A}(\hat{\boldsymbol{w}}-\boldsymbol{c})}{qs^2},$$

která má při platnosti H_0 rozdělení F s q a $n-p-1$ stupni volnosti. Test na hladině α tedy provedeme porovnáním $F_{\mathbf{A}}$ s kritickou hodnotou $F^{\alpha}_{q,n-p-1}$. Pokud $F_{\mathbf{A}}\geq F^{\alpha}_{q,n-p-1}$, zamítneme H_0 a pokud $F_{\mathbf{A}}<F^{\alpha}_{q,n-p-1}$, nezamítneme H_0.

3.2.1 Pásy spolehlivosti a predikční interval

Větu 3.6 můžeme kromě testování hypotéz o hodnotách vektoru vah použít také ke konstrukci pásu spolehlivosti kolem regresní přímky, což nyní provedeme. Mějme libovolný bod $\boldsymbol{x}_0$ z prostoru příznaků $\mathbb{R}^{p+1}$. Vysvětlovaná proměnná Y_0 v tomto bodě je podle uvažovaného modelu (3.1) určena vztahem

$$Y_0=\boldsymbol{x}_0^T\boldsymbol{w}+\varepsilon_0,$$

kde $\mathbb{E}\,\varepsilon_0=0$ a $\operatorname{var}\varepsilon_0=\sigma^2$.

Výraz $\mathbb{E}\,Y_0=\boldsymbol{x}_0^T\boldsymbol{w}$ nazýváme hodnotou regresní přímky v bodě $\boldsymbol{x}_0$. Tuto hodnotu přímočaře odhadujeme pomocí $\hat{Y}_0=\boldsymbol{x}_0^T\hat{\boldsymbol{w}}$, což je zjevně nestranný bodový odhad $\mathbb{E}\,Y_0$. V důsledku věty 3.1 totiž platí

$$\mathbb{E}\,\hat{Y}_0=\mathbb{E}\,\boldsymbol{x}_0^T\hat{\boldsymbol{w}}=\boldsymbol{x}_0^T\,\mathbb{E}\,\hat{\boldsymbol{w}}=\boldsymbol{x}_0^T\boldsymbol{w}=\mathbb{E}\,Y_0.$$

Obrázek 3.1: Pás spolehlivosti kolem regresní přímky a pás spolehlivosti pro regresní přímku.

Předpokládejme nyní $\varepsilon_0 \sim \mathrm{N}(0, \sigma^2)$. Z věty 3.6 plyne

$$\frac{\boldsymbol{x}_0^T(\hat{\boldsymbol{w}} - \boldsymbol{w})}{\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0}} = \frac{\hat{Y}_0 - \mathbb{E}\, Y_0}{\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0}} \sim \mathrm{t}_{n-p-1}.$$

Dostáváme tak

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-t_{n-p-1}^{\alpha/2} \le \frac{\hat{Y}_0 - \mathbb{E}\, Y_0}{\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0}} \le t_{n-p-1}^{\alpha/2} \right) \\ &= \mathbb{P}\left(\hat{Y}_0 - t_{n-p-1}^{\alpha/2}\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0} \le \mathbb{E}\, Y_0 \le \hat{Y}_0 + t_{n-p-1}^{\alpha/2}\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0} \right). \end{aligned}$$

Interval

$$\left(\hat{Y}_0 - t_{n-p-1}^{\alpha/2}\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0}, \hat{Y}_0 + t_{n-p-1}^{\alpha/2}\sqrt{s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0} \right)$$

je tedy $(1-\alpha)\%$ interval spolehlivosti pro hodnotu regresní přímky $\mathbb{E}\, Y_0$ v bodě $\boldsymbol{x}_0$. Zaneseme-li tento interval pro každé $\boldsymbol{x}_0$ ve zvoleném rozsahu do grafu regresní přímky, dostaneme takzvaný *pás kolem regresní přímky*. Příklad je zobrazen na obrázku 3.1. Je třeba si uvědomit, že tento interval spolehlivosti platí pro každý bod $\boldsymbol{x}_0$ zvlášť.

Chceme-li získat pás spolehlivosti, který platí ve všech bodech současně, tzv. *pás spolehlivosti pro regresní přímku*, musíme využít sofistikovanějšího přístupu pomocí Scheffého S-metody (viz část 5.1.1 v [**?**]). Výsledkem je pás spolehlivosti určený v bodě $\boldsymbol{x}_0$ intervalem

$$\left(\hat{Y}_0 - \sqrt{(p+1)F_{p+1,n-p-1}^{\alpha} s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0}, \hat{Y}_0 + \sqrt{(p+1)F_{p+1,n-p-1}^{\alpha} s^2 \boldsymbol{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0} \right), \tag{3.6}$$

který platí pro všechny body $\boldsymbol{x}_0$ současně. To znamená, že celá regresní přímka je obsažena v pásu určeném tímto intervalem s pravděpodobností $1-\alpha$. Pás spolehlivosti pro regresní přímku je vždy širší než pás spolehlivosti kolem regresní přímky. Příklad vyobrazení konkrétního pásu spolehlivosti kolem regresní přímky je na obrázku 3.1.

Na závěr se vraťme k náhodné veličině Y_0 v bodě $\boldsymbol{x}_0$. Její střední hodnotu bodově odhadujeme pomocí $\hat{Y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{w}}$ a intervalově pomocí intervalu (3.6). Zabývejme se nyní přímo hodnotou náhodné veličiny Y_0. Číselným[1] odhadem této hodnoty je samozřejmě

[1] Nepíšeme bodovým odhadem, protože tento termín je vyhrazen pro číselné odhady neznámých ale pevných parametrů rozdělení. Náhodná veličina jako taková nemá, žádnou pevnou hodnotu a nelze ji tedy v jednoduchém slova smyslu chápat jako parametr rozdělení.

opět $\hat{Y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{w}}$. Sestrojme nyní interval předpokládaného výskytu, tj. interval, který pokrývá Y_0 s předem danou pravděpodobností $1-\alpha$. Předpokládejme, že v bodě $\boldsymbol{x}_0$ opět platí

$$Y_0 = \boldsymbol{x}_0^T \boldsymbol{w} + \varepsilon_0,$$

kde $\varepsilon_0 \sim \mathrm{N}(0, \sigma^2)$ a pro trénovací data platí

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Navíc předpokládejme, že trénovací data jsou nezávislá na Y_0, což znamená, že ε_0 a $\boldsymbol{\varepsilon}$ jsou nezávislé.

Díky předpokladu o rozdělení ε_0 platí $Y_0 \sim \mathrm{N}(\boldsymbol{x}_0^T\boldsymbol{w}, \sigma^2)$. Z věty 3.4 víme $\hat{\boldsymbol{w}} \sim \mathrm{N}\left(\boldsymbol{w}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$ a tedy $\hat{Y}_0 \sim \mathrm{N}\left(\boldsymbol{x}_0^T\boldsymbol{w}, \sigma^2\boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0\right)$. Podle důsledku **??** platí

$$\hat{Y}_0 - Y_0 \sim \mathrm{N}\left(0, \sigma^2(\boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0 + 1)\right).$$

Analogickým postupem jako v důkazu věty 3.6 dostáváme

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{s^2(\boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0 + 1)}} \sim \mathrm{t}_{n-p-1}.$$

Nyní již stejným postupem jako u intervalu spolehlivosti pro $\mathbb{E}\, Y_0$ zjistíme, že interval pokrývající Y_0 s pravděpodobností $1-\alpha$ je

$$\left(\hat{Y}_0 - t_{n-p-1}^{\alpha/2}\sqrt{s^2(\boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0 + 1)}, \hat{Y}_0 + t_{n-p-1}^{\alpha/2}\sqrt{s^2(\boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0 + 1)}\right).$$

3.3 Vztah vychýlení a rozptylu

Uvažujme nyní bod $\boldsymbol{x}_0$ v prostoru příznaků $\mathbb{R}^{p+1}$ a vysvětlovanou proměnnou Y_0 v tomto bodě. Dále mějme nějaký odhad $\hat{Y}_0$ veličiny Y_0, který je založen na vstupních datech nazývaných *trénovací data* představovaných N dvojicemi $(Y_i, \boldsymbol{x}_i)$, kde Y_i je náhodná veličina ukazující výsledek vysvětlované proměnné v i-tém bodě $\boldsymbol{x}_i = (1, x_{i;1}, \ldots, x_{i;p})^T$ konkrétních hodnot příznaků. Opět označme $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Předpokládejme, že $\operatorname{var} Y_0 = \sigma^2$ a $\operatorname{var}\boldsymbol{Y} = \sigma^2\mathbf{I}$. Navíc budeme předpokládat, že Y_0 a $\boldsymbol{Y}$ jsou nezávislé.

Chybu tohoto odhadu $\hat{Y}_0$ veličiny Y_0 v bodě $\boldsymbol{x}_0$ můžeme zkoumat pomocí střední chyby odhadu (*expected prediction error*) v bodě $\boldsymbol{x}_0$, která je definovaná vztahem

$$\mathbb{E}\,\mathrm{PE}(Y_0, \hat{Y}_0) = \mathbb{E}(Y_0 - \hat{Y}_0)^2.$$

Po přičtení a odečtení $\mathbb{E}\, Y_0$ uvnitř závorky dostaneme

$$\mathbb{E}\,\mathrm{PE}(Y_0, \hat{Y}_0) = \mathbb{E}(Y_0 - \mathbb{E}\, Y_0 + \mathbb{E}\, Y_0 - \hat{Y}_0)^2 = \mathbb{E}(Y_0 - \mathbb{E}\, Y_0)^2 + \mathbb{E}(\hat{Y}_0 - \mathbb{E}\, Y_0)^2,$$

protože nezávislost Y_0 a $\boldsymbol{Y}$ implikuje nezávislost Y_0 a $\hat{Y}_0$, a tedy

$$\begin{aligned}\mathbb{E}\left((Y_0 - \mathbb{E}\, Y_0)(\mathbb{E}\, Y_0 - \hat{Y}_0)\right) &= (\mathbb{E}\, Y_0)^2 - \mathbb{E}(Y_0\hat{Y}_0) - (\mathbb{E}\, Y_0)^2 + \mathbb{E}\, Y_0\,\mathbb{E}\,\hat{Y}_0 \\ &= (\mathbb{E}\, Y_0)^2 - \mathbb{E}\, Y_0\,\mathbb{E}\,\hat{Y}_0 - (\mathbb{E}\, Y_0)^2 + \mathbb{E}\, Y_0\,\mathbb{E}\,\hat{Y}_0 \\ &= 0.\end{aligned}$$

S využitím $\operatorname{var} Y_0 = \sigma^2$ tak máme

$$\mathbb{E}\,\mathrm{PE}(Y_0, \hat{Y}_0) = \sigma^2 + \mathbb{E}(\hat{Y}_0 - \mathbb{E}\,Y_0)^2.$$

Poznamenejme, že druhý člen se někdy označuje jako střední chyba modelu (*expected model error*) v bodě $\boldsymbol{x}_0$, tj. $\mathbb{E}\,\mathrm{ME}(Y_0, \hat{Y}_0) = \mathbb{E}(\hat{Y}_0 - \mathbb{E}\,Y_0)^2$. Tento člen ještě můžeme upravit jako

$$\begin{aligned}\mathbb{E}\,\mathrm{ME}(Y_0, \hat{Y}_0) &= \mathbb{E}(\hat{Y}_0 - \mathbb{E}\,Y_0)^2 = \mathbb{E}(\hat{Y}_0 - \mathbb{E}\,\hat{Y}_0 + \mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0)^2 \\ &= \mathbb{E}(\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0)^2 + \mathbb{E}(\hat{Y}_0 - \mathbb{E}\,\hat{Y}_0)^2 + 2\,\mathbb{E}(\hat{Y}_0 - \mathbb{E}\,\hat{Y}_0)(\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0) \\ &= (\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0)^2 + \mathbb{E}(\hat{Y}_0 - \mathbb{E}\,\hat{Y}_0)^2 + 2(\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,\hat{Y}_0)(\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0) \\ &= (\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0)^2 + \operatorname{var}\hat{Y}_0.\end{aligned}$$

Celkem tak můžeme psát

$$\mathbb{E}\,\mathrm{PE}(Y_0, \hat{Y}_0) = \sigma^2 + (\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0)^2 + \operatorname{var}\hat{Y}_0. \tag{3.7}$$

Protože $\mathbb{E}\,\hat{Y}_0 - \mathbb{E}\,Y_0$ představuje vychýlení odhadu $\hat{Y}_0$ a $\operatorname{var}\hat{Y}_0$ jeho rozptyl, bývá předchozí vztah nazýván vztahem vychýlení a rozptylu (*bias-variance tradeoff*). Vidíme, že vychýlení i rozptyl přispívají k celkové střední chybě odhadu. Typicky platí, že obě tyto hodnoty nelze současně snižovat nastavováním parametrů modelu. Je tedy třeba volit model, který mezi těmito příspěvky vhodně balancuje.

Pokud je $\hat{Y}_0$ nestranným odhadem Y_0, tj. $\mathbb{E}\,\hat{Y}_0 = \mathbb{E}\,Y_0$, je střední chyba odhadu rovna součtu σ^2 a rozptylu $\operatorname{var}\hat{Y}_0$. Nejlepším odhadem je v takovém případě odhad s nejmenším rozptylem. Uvažujme nyní model (3.2), tj. $\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$, spolu s předpoklady $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ a $\operatorname{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$. Z věty 3.1 víme, že $\hat{\boldsymbol{w}}$ je nestranným odhadem $\boldsymbol{w}$. Protože $\hat{Y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{w}}$, platí $\mathbb{E}\,\hat{Y}_0 = \boldsymbol{x}_0^T\,\mathbb{E}\,\hat{\boldsymbol{w}} = \boldsymbol{x}_0^T\boldsymbol{w} = \mathbb{E}\,Y_0$ a tedy odhad $\hat{Y}_0$ je nestranný. Dosadíme-li přesný tvar (3.5) odhadu $\hat{\boldsymbol{w}}$ získáme

$$\hat{Y}_0 = \boldsymbol{x}_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{Y},$$

což znamená, že $\hat{Y}_0$ je lineární v $\boldsymbol{Y}$. Následující věta ukazuje, že tento odhad $\hat{Y}_0$ získaný metodou nejmenších čtverců je nejlepší mezi všemi nestrannými odhady lineárními v $\boldsymbol{Y}$.

Věta 3.9 (Gauss-Markov)**:** Předpokládejme, že $\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$ a $Y_0 = \boldsymbol{x}_0^T\boldsymbol{w} + \varepsilon_0$, kde $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathbb{E}\,\varepsilon_0 = 0$, $\operatorname{var}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$, $\operatorname{var}\varepsilon_0 = \sigma^2$ a $\boldsymbol{\varepsilon}, \varepsilon_0$ jsou nezávislé. Označme $\hat{Y}_0 = \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}$ odhad Y_0 získaný metodou nejmenších čtverců. Potom pro libovolný lineární v $\boldsymbol{Y}$ odhad $\tilde{Y}_0 = \boldsymbol{d}^T\boldsymbol{Y}$, který je nestranný, tj. $\mathbb{E}\,\tilde{Y}_0 = \boldsymbol{x}_0^T\boldsymbol{w}$ pro každé $\boldsymbol{w} \in \mathbb{R}^{p+1}$, platí

$$\operatorname{var}\hat{Y}_0 \leq \operatorname{var}\tilde{Y}_0.$$

Důkaz. Položme $\mathbf{A} = \frac{\boldsymbol{x}_0\boldsymbol{d}^T}{\|\boldsymbol{x}_0\|^2}$ Platí

$$\begin{aligned}\operatorname{var}\tilde{Y}_0 = \operatorname{var}\boldsymbol{d}^T\boldsymbol{Y} &= \mathbb{E}(\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\boldsymbol{w})^2 = \mathbb{E}(\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} + \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} - \boldsymbol{x}_0^T\boldsymbol{w})^2 \\ &= \mathbb{E}(\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} - \boldsymbol{x}_0^T\boldsymbol{w})^2 + \mathbb{E}(\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})^2 + 2\,\mathbb{E}(\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})(\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} - \boldsymbol{x}_0^T\boldsymbol{w}).\end{aligned}$$

Dokážeme, že třetí člen tohoto součtu je nulový. Z nestrannosti $\tilde{Y}_0$ a $\hat{Y}_0$ dostáváme

$$\begin{aligned}&\mathbb{E}(\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})(\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} - \boldsymbol{x}_0^T\boldsymbol{w}) = \mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}\right) - \mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\boldsymbol{w}\right) \\ &= \mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}\right) - (\mathbb{E}\,\boldsymbol{d}^T\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\boldsymbol{w} = \mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}\right).\end{aligned}$$

Po dosazení $\widehat{\boldsymbol{w}}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ s využitím faktu $\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} = \widehat{\boldsymbol{w}}_{\mathrm{OLS}}{}^T\boldsymbol{x}_0$ získáme

$$\mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}\right) = \mathbb{E}\left(\left(\boldsymbol{d}^T - \boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\boldsymbol{Y}\boldsymbol{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0\right),$$

protože $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ je symetrická matice. Využijeme-li faktu $\mathbb{E}\,\boldsymbol{Y}\boldsymbol{Y}^T = \operatorname{var}\boldsymbol{Y} + \mathbb{E}\,\boldsymbol{Y}\,\mathbb{E}\,\boldsymbol{Y}^T$ a $\mathbb{E}\,\boldsymbol{Y} = \mathbf{X}\boldsymbol{w}$, $\operatorname{var}\boldsymbol{Y} = \sigma^2\mathbf{I}$, dostaneme

$$\begin{aligned}
\mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}\right) &= \left(\boldsymbol{d}^T - \boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbb{E}(\boldsymbol{Y}\boldsymbol{Y}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0 \\
&= \left(\boldsymbol{d}^T - \boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\left(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{w}\boldsymbol{w}^T\mathbf{X}^T\right)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0 \\
&= \left(\boldsymbol{d}^T - \boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{X}\left(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \boldsymbol{w}\boldsymbol{w}^T\right)\boldsymbol{x}_0 \\
&= \left(\boldsymbol{d}^T\mathbf{X} - \boldsymbol{x}_0^T\right)\left(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \boldsymbol{w}\boldsymbol{w}^T\right)\boldsymbol{x}_0
\end{aligned}$$

Nestrannost $\tilde{Y}_0 = \boldsymbol{d}^T\boldsymbol{Y}$ znamená

$$\boldsymbol{x}_0^T\boldsymbol{w} = \mathbb{E}\,\boldsymbol{d}^T\boldsymbol{Y} = \boldsymbol{d}^T\,\mathbb{E}\,\boldsymbol{Y} = \boldsymbol{d}^T\mathbf{X}\boldsymbol{w}$$

pro každé $\boldsymbol{w}$ a tudíž $\boldsymbol{d}^T\mathbf{X} - \boldsymbol{x}_0^T = \mathbf{0}$. To znamená

$$\mathbb{E}\left((\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}}\right) = 0.$$

Jelikož $\mathbb{E}(\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}} - \boldsymbol{x}_0^T\boldsymbol{w})^2 = \operatorname{var}\hat{Y}_0$, dostáváme celkem

$$\operatorname{var}\tilde{Y}_0 = \operatorname{var}\hat{Y}_0 + \mathbb{E}(\boldsymbol{d}^T\boldsymbol{Y} - \boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\mathrm{OLS}})^2 \geq \operatorname{var}\hat{Y}_0$$

a důkaz je hotov. □

3.4 Cvičení

Kapitola 4

Redukce počtu vysvětlujících proměnných

změnit název, takto to nesedí: ne vždy jde o redukci počtu

V této části se budeme věnovat postupům, které mají za úkol snížit počet vysvětlujících proměnných. Důvodem ke snížení je to, že odhad získaný nejmenšími čtverci pomocí rovnice (2.7) nemusí být zcela uspokojivý v následujícím smyslu. Odhad $\widehat{\boldsymbol{w}}_{\mathrm{OLS}}$ má často nízké vychýlení (je dokonce nestranný, vizte větu 3.1) a velký rozptyl. Podle věty 3.9 je odhad $\widehat{\boldsymbol{w}}_{\mathrm{OLS}}$ dokonce v jistém smyslu nejlepší mezi nestrannými odhady.

Motivace pro snížení počtu vysvětlujících proměnných je obětování trochu vychýlení (zvýšíme v absolutní hodnotě), které ovšem povede ke snížení rozptylu. Obě tyto hodnoty jsou přítomny ve střední chybě modelu (3.7) a my doufáme, že snížení rozptylu převýší zvýšení vychýlení a celkově se nám tedy střední chyba modelu sníží.

Druhým důvodem ke snižování počtu vysvětlujících proměnných je interpretace samotného odhadu. Vynecháním proměnných, které mají jen malý vliv, nám může při interpretaci výsledku umožnit soustředit se na důležité faktory.

Snižování počtu proměnných se též říká redukce dimenzionality. Nejprve se budeme věnovat metodám, které vybírají podmnožiny příznaků. Poté se budemě věnovat metodám, které konstruují vychýlené odhady. Na závěr se budeme věnovat metodám, které identifikují korelované příznaky a slučují je do nových souhrnných příznaků. (U těchto metod už se ovšem nejedná o faktické snížení počtu příznaků, ale o jejich transformaci.)

4.1 Výběr podmnožiny (*Subset selection*)

4.1.1 Výběr nejlepší podmnožiny (*Best-subset selection*)

Výběrem nejlepší podmnožiny rozumíme nalezení podmnožiny velikosti k, pro kterou je reziduální suma čtverců nejmenší. Tedy přesněji, hledáme k-prvkovou množinu $I \subset \{1, \dots, p\}$ takovou, že

$$\mathrm{RSS}(\boldsymbol{w}_I) = \sum_{i=1}^{N}(y_i - w_0 - \sum_{j\in I} x_{ij}w_j)^2 \leq \sum_{i=1}^{N}(y_i - w_0 - \sum_{j\in I'} x_{ij}w_j)^2 = \mathrm{RSS}(\boldsymbol{w}_{I'})$$

pro všechny k-prvkové množiny $I' \subset \{1, \dots, p\}$. (Symbolem $\boldsymbol{w}_I$ jsme označili odhad pomocí obyčejných nejmenších čtverců při zúžení množiny příznaků na ty indexované mno-

žinou I.) Účinný přístup k tomuto úkolu je algoritmus *leaps and bounds*[1] uvedený v [?]. (Dobrý popis tohoto algoritmu lze nalézt v [?].) V krátkosti lze říct, že tento algoritmus neprochází celý prostor možných podmnožin I, ale vyhýbá se kandidátům, o kterých ví, že nemohou být hledanou množinou.

4.1.2 Dopředný postupný výběr (*Forward-stepwise selection*)

Do výběru se může dostat příznak, který je nedůležitý a to tak, že je přidán před příznakem, který jej „zahrnuje“.

4.1.3 Zpětný postupný výběr (*Backward-stepwise selection*)

jinak *Recursive feature elimination (RFE)*

Pokud je vyloučen příznak, který je důležitý, neexistuje možnost nápravy.

4.2 Smrskávací modely (*Shrinkage models*)

4.2.1 Hřebenová regrese (*Ridge regression*)

Hřebenová regrese byla navržena v roce 1970 v [?] jako řešení problému singularity matice $\mathbf{X}^T\mathbf{X}$. Kromě vyřešení tohoto problému má hřebenová regrese i další zajímavé vlastnosti, které si popíšeme níže.

Hřebenová regrese má parametr $\lambda \in \mathbb{R}$ splňující $\lambda \geq 0$. Odhad pomocí hřebenové regrese budeme značit $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ a, podobně jako v případě obyčejných nejmenších čtverců, se jedná o řešení dané optimalizační úlohy. Konkrétně položíme $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ takto:

$$\begin{aligned}\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{p+1}} \left\{ \text{RSS}(\boldsymbol{w}) + \lambda \sum_{j=1}^{p} w_j^2 \right\} \\ &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{p+1}} \left\{ \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{p} x_{i,j} w_j \right)^2 + \lambda \sum_{j=1}^{p} w_j^2 \right\}.\end{aligned} \tag{4.1}$$

Z definice odhadu pomocí nejmenších čtverců (2.5) plyne $\widehat{\boldsymbol{w}}_{\text{ridge}(0)} = \widehat{\boldsymbol{w}}_{\text{OLS}}$. Abychom zjistili, jaký je vliv nenulového parametru λ, uvažme podobnou úlohu. Mějme parametr τ splňující $\tau \geq 0$ a řešme následující problém argumentu vázaného extrému:

$$\begin{aligned}\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)} &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{p+1}} \{\text{RSS}(\boldsymbol{w})\} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{p+1}} \left\{ \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{p} x_{i,j} w_j \right)^2 \right\} \\ &\text{za podmínky } \sum_{j=1}^{p} w_j^2 \leq \tau.\end{aligned} \tag{4.2}$$

Souvislost obou úloh je shrnuta následující větou.

Věta 4.1: Úloha (4.1) je ekvivalentní úloze (4.2) v následujícím smyslu: pokud máme řešení $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ úlohy (4.1) pro nějaké $\lambda \geq 0$, pak existuje $\tau \geq 0$, že řešení úlohy (4.2) je

[1] Výraz „in/by leaps and bounds“ je idiom, který lze do češtiny volně přeložit jako „mílovými kroky“.

Obrázek 4.1: Ukázka souvislosti úlohy hřebenové regrese s hledáním argumentu vázaného extrému. Modře jsou zobrazeny vrstevnice minimalizované funkce RSS($\boldsymbol{w}$), červeně je zobrazena vazba, tedy kružnice v počátku s poloměrem τ.

rovno $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$; a naopak, pokud máme řešení $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$ úlohy (4.2) pro nějaké $\tau \geq 0$, pak existuje $\lambda \geq 0$, že řešení úlohy (4.1) je rovno $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$.

Důkaz. Mějme řešení $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ úlohy (4.1) pro nějaké $\lambda \geq 0$. Položme $\tau = \sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}\right)_j^2$. Vyřešme úlohu (4.2) s tímto τ. Dostaneme tak nějaké řešení $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$.

Předpokládejme, že $\sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}\right)_j^2 < t$. Máme

$$\text{RSS}(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}) + \lambda \underbrace{\sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}\right)_j^2}_{=t} > \text{RSS}(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}) + \lambda \sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}\right)_j^2 .$$

Protože $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ je také splňuje podmínku úlohy (4.2), tak platí $\text{RSS}(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}) \geq \text{RSS}(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)})$. Dostáváme $\text{RSS}(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}) + \lambda \sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}\right)_j^2 \geq \text{RSS}(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}) + \lambda \sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}\right)_j^2$ a tedy celkem

$$\text{RSS}(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}) + \lambda \sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}\right)_j^2 > \text{RSS}(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}) + \lambda \sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}\right)_j^2 .$$

Tedy úloha (4.1) by měla řešení $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$. Jelikož $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)} \neq \widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$, dostáváme spor. Platí tedy $\sum_{j=1}^{p} \left(\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}\right)_j^2 = t$. Jelikož na první souřadnici není žádná podmínka, platí $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} = \widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$. Tím je první část tvrzení dokázána.

Předpokládejme nyní, že máme řešení $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$ úlohy (4.2) pro nějaké $\tau \geq 0$. Pokud platí $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)} < \tau$, pak nastavíme $\lambda = 0$ a řešíme úlohu (4.1). Protože funkce RSS má jedno lokální minimum, pak dostaneme $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} = \widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$. Pokud platí $\widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)} = \tau$, pak lze úlohu (4.2) chápat jako vázaný extrém s jednou rovnostní podmínkou. Sestavíme Langrangeovu funkci:

$$L(\mathbf{w}, \mu) = \text{RSS}(\mathbf{w}) + \mu \left(\sum_{j=1}^{p} (\mathbf{w})_j^2 - \tau \right).$$

Taková úloha bude mít právě jedno řešení, které bude lokálním minimem. Tomuto řešení bude odpovídat nějaká hodnota μ_0 Langrangeova multiplikátoru μ. Jednoduchým ověřením lze ověřit, že pro $\lambda = \mu_0$ bude mít úloha (4.1) právě toto řešení, a tedy opět $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} = \widehat{\boldsymbol{w}}'_{\text{ridge}(\tau)}$. □

Z předchozího důkazu plyne, že parametr λ buď nemá vliv a při řešení úlohy (4.1) najdeme globální minimum funkce RSS, nebo najdeme minimum na hranici koule o poloměru τ se středem v počátku (nebereme-li v potaz posun w_0). Ukázka druhého případu je na obrázku 4.1. Platí, že čím větší λ, tím menší τ a odhad $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ je blíže a blíže k

0. To platí opět s výjímkou posunu w_0; Nechceme, aby byl posun posunován směrem k 0, neboť se jedná o odhad absolutního členu neznámé funkce.

Kvůli tomuto chování se parametru λ také říká penalizační parametr. Než přistoupíme k řešení úlohy (4.1), tak si vystředíme příznaky a ukážeme si, že pro vystředěné příznaky platí $\widehat{w_0} = \frac{\sum y_i}{N}$.

Definujme matici příznaků po vystředění (centralizaci) bez prvního sloupce jedniček takto:

$$\mathbf{X}_c = \begin{pmatrix} x_{1,1} - \overline{x_1} & x_{1,2} - \overline{x_2} & \cdots & x_{1,p} - \overline{x_p} \\ x_{2,1} - \overline{x_1} & x_{2,2} - \overline{x_2} & \cdots & x_{2,p} - \overline{x_p} \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ x_{N,1} - \overline{x_1} & x_{N,2} - \overline{x_2} & \cdots & x_{N,p} - \overline{x_p} \end{pmatrix} \in \mathbb{R}^{N,p},$$

kde $\overline{x_i} = \frac{\sum_{j=1}^{N} x_{j,i}}{N}$ (tedy aritmetický průměr hodnot i-tého příznaku). Dále označme matici $\mathbf{X}'_c = (\mathbf{1}, \mathbf{X}_c) \in \mathbb{R}^{N,p+1}$, kde $\mathbf{1}$ je sloupec samých jedniček. Řešení úlohy (4.1) s vystředěnými příznaky pak odpovídá hledání minima funkce $g : \mathbb{R}^{p+1} \to \mathbb{R}$ dané

$$g(\boldsymbol{w}) = (\boldsymbol{Y} - \mathbf{X}'_c)^T(\boldsymbol{Y} - \mathbf{X}'_c) + \lambda \boldsymbol{w}^T \mathbf{I}' \boldsymbol{w}, \qquad \text{kde} \quad \mathbf{I}' = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Minimalizaci provedeme obdobně jako v části 2.1. To je umožněno především tím, že ve vzorci (4.1) jsou u posledního členu druhé mocniny. Pro gradient a Hessián funkce f platí

$$\nabla f(\boldsymbol{w}) = -2\mathbf{X}_c'^T(\boldsymbol{Y} - \mathbf{X}'_c\boldsymbol{w}) + 2\lambda\mathbf{I}'\boldsymbol{w} \quad \text{a} \quad \nabla^2 f(\boldsymbol{w}) = 2\mathbf{X}_c'^T\mathbf{X}'_c + 2\lambda\mathbf{I}'. \tag{4.3}$$

Hledané řešení tedy splňuje $-\mathbf{X}_c'^T(\boldsymbol{Y} - \mathbf{X}'_c\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}) + \lambda\mathbf{I}'\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} = 0$. Jednoduchými úpravami odvodíme

$$\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} = \left((\mathbf{X}_c'^T\mathbf{X}'_c - \lambda\mathbf{I}'\right)^{-1}\mathbf{X}_c'^T\boldsymbol{Y},$$

samozřejmě za předpokladu, že lze matici $\mathbf{X}_c'^T\mathbf{X}'_c + \lambda\mathbf{I}'$ invertovat. Pro kladné λ toto lze, protože matice $\mathbf{X}_c'^T\mathbf{X}'_c + \lambda\mathbf{I}$ je pozitivně definitní (vizte cvičení 4.2). Toto byl původní záměr autorů hřebenové regrese a z tohoto důvodu se také o přičtené matice $\lambda I'$ (nebo obecně λI) hovoří jako o regularizaci. Přímým výpočtem lze odvodit, že

odkaz na cvičení

$$\mathbf{X}_c'^T\mathbf{X}'_c = \begin{pmatrix} N & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{X}_c^T\mathbf{X}_c & \\ 0 & & & \end{pmatrix} \in \mathbb{R}^{p+1,p+1},$$

kde zobrazené nulové prvky matice napravo jsou důsledkem rovnosti $\sum_{j=1}^{N}(x_{j,i} - \overline{x_i}) = 0$ pro všechna $i \in \{1, \ldots, p\}$. Celkem tedy máme, že první složka vektoru $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$ splňuje $(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)})_0 = \frac{\sum_{i=1}^{N} y_i}{N}$ a další složky lze spočíst ze vzorce

$$\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda),+} = ((\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)})_1, \ldots, (\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)})_p)^T = \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_c^T\boldsymbol{Y}.$$

Jelikož díky vystředění příznaků známe hodnotu $(\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)})_0$, nemusíme se již odhadem posunu zabývat.

4.2.1.1 Hřebenová regrese a rozklad na singulární hodnoty

Pro prohloubení pochopení hřebenová regrese využijeme vhodného rozkladu matice $\mathbf{X}_c \in \mathbb{R}^{N,p}$: existují matice $\mathbf{U} \in \mathbb{R}^{N,p}$, $\mathbf{D}, \mathbf{V} \in \mathbb{R}^{p,p}$ takové, že sloupce matice $\mathbf{U}$ tvoří ortonormální soubor, $\mathbf{V}$ je ortogonální, matice $\mathbf{D}$ je diagonální s nezápornými prvky a

$$\mathbf{X}_c = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

Tento rozklad je dán větou A.4 a nazývá se rozkladem na singulární hodnoty (nebo též singulárním rozkladem, *Singular value decomposition (SVD)*).

Transformujme matici $\mathbf{X}_c$ takto

$$\widetilde{\mathbf{X}} = \mathbf{X}_c\mathbf{V} = \mathbf{U}\mathbf{D}.$$

Tuto transformaci lze chápat jako přepis příznaků do báze dané sloupci matice $\mathbf{V}$. Označme dále d_i prvky na diagonále matice $\mathbf{D}$. Těmto prvkům se říká singulární hodnoty. Předpokládejme, že $\lambda > 0$ a označme sloupce matice $\mathbf{U}$ takto: $\mathbf{U} = \begin{pmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_p \end{pmatrix}$. Pro odhad vysvětlované proměnné platí

$$\begin{aligned}
\widehat{\boldsymbol{Y}} = \mathbf{X}_c\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda),+} &= \mathbf{X}_c\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_c^T\boldsymbol{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T\left(\mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T + \lambda\mathbf{I}\right)^{-1}\mathbf{V}\mathbf{D}^T\mathbf{U}^T\boldsymbol{Y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\left(\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D} + \lambda\mathbf{I}\right)^{-1}\mathbf{V}^T\mathbf{V}\mathbf{D}^T\mathbf{U}^T\boldsymbol{Y} \\
&= \mathbf{U}\mathbf{D}\left(\mathbf{D}^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{D}^T\mathbf{U}^T\boldsymbol{Y} \\
&= \mathbf{U}\begin{pmatrix} \frac{d_1^2}{d_1^2+\lambda} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{d_p^2}{d_p^2+\lambda} \end{pmatrix}\mathbf{U}^T\boldsymbol{Y} \\
&= \begin{pmatrix} \frac{d_1^2}{d_1^2+\lambda}\boldsymbol{u}_1 & \cdots & \frac{d_p^2}{d_p^2+\lambda}\boldsymbol{u}_p \end{pmatrix}\begin{pmatrix} \boldsymbol{u}_1^T\boldsymbol{Y} \\ \vdots \\ \boldsymbol{u}_p^T\boldsymbol{Y} \end{pmatrix} \\
&= \sum_{j=1}^{p}\left(\frac{d_j^2}{d_j^2+\lambda}\boldsymbol{u}_j^T\boldsymbol{Y}\right)\boldsymbol{u}_j.
\end{aligned}$$

Tento zápis lze chápat jako vyjádření $\widehat{\boldsymbol{Y}}$ v bázi $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)$. Předpoklad $\lambda > 0$ potřebujeme jen pro případ singulární matice $\mathbf{X}_c^T\mathbf{X}$. Je-li tato matice regulární, pak $\lambda = 0$ stále odpovídá obyčejným nejmenším čtvercům.

Protože $\frac{d_i^2}{d_i^2+\lambda} = 1 - \frac{\lambda}{d_i^2+\lambda}$, tak

$$d_i \geq d_j \quad \Longrightarrow \quad \frac{d_i^2}{d_i^2+\lambda} \geq \frac{d_j^2}{d_j^2+\lambda}.$$

Z toho je vidět, že čím menší je hodnota d_i, tím více je souřadnice u $\boldsymbol{u}_i$ zmenšena[2] (pro nějaké pevné nenulové λ) oproti řešení pomocí obyčejných nejmenších čtverců, tedy pro $\lambda = 0$. Dále je vidět, že čím větší λ, tím větší je zmenšení pro všechny souřadnice.

[2] Neboli „smrsknuta“ - odkud plyne název této třídy metod.

Zbývá podívat se na interpretaci singulárních hodnot d_i. Předpokládejme

$$d_1 \geq d_2 \geq \cdots \geq d_p.$$

Označme sloupce matice $\mathbf{V}$ takto: $\mathbf{V} = \begin{pmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_p \end{pmatrix}$. Označme $\boldsymbol{z}_i = \mathbf{X}_c \boldsymbol{v}_i$. Vektory $\boldsymbol{z}_i$ lze chápat jako nové vektory příznaků vzniklé projekcí na vektory $\boldsymbol{v}_i$ (nejedná se o nic jiného, než o sloupce matice $\widetilde{\mathbf{X}} = \mathbf{X}_c\mathbf{V}$). Uvažujme nyní jiný příznak $\boldsymbol{z}$, který si vyjádříme také v bázi $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$[3]. Budeme chtít porovnat vektor $\boldsymbol{z}$ s vektorem $\boldsymbol{z}_1$. Aby nedocházelo ke škálování, budeme volit $\boldsymbol{z}$ takto

$$\boldsymbol{z} = \mathbf{X}_c \sum_{i=1}^{p} \alpha_i \boldsymbol{v}_i \quad \text{kde } \sum_{i=1}^{p} \alpha_i^2 = 1, \alpha_i > 0.$$

Z rovnosti $\mathbf{X}_c\mathbf{V} = \mathbf{U}\mathbf{D}$ odvodíme $\mathbf{X}_c\boldsymbol{v}_i = d_i\boldsymbol{u}_i$. Z toho plyne $\boldsymbol{z} = \sum_{i=1}^{p} \alpha_i d_i \boldsymbol{u}_i$, a tedy

$$\begin{aligned}
\boldsymbol{z}^T\boldsymbol{z} &= \left(\sum_{i=1}^{p} \alpha_i \boldsymbol{u}_i d_i\right)^T \left(\sum_{j=1}^{p} \alpha_j \boldsymbol{u}_j d_j\right) \\
&= \sum_{i,j} \alpha_i \alpha_j d_i d_j \boldsymbol{u}_i^T \boldsymbol{u}_j \\
&= \sum_{i=1}^{p} \alpha_i^2 d_i^2 \\
&\leq \sum_{i=1}^{p} \alpha_i^2 d_1^2 = d_1^2.
\end{aligned}$$

Z této nerovnosti plyne, že vektor $\boldsymbol{z}_1$ má největší výběrový rozptyl v porovnání s jinými normovanými kombinacemi sloupců matice $\mathbf{X}_c$. Tuto úvahu můžeme opakovat na prostoru bez směru $\boldsymbol{v}_1$ a zjistíme, že analogickou vlastnosti splňuje $\boldsymbol{z}_2$ (budeme-li souřadnice u $\boldsymbol{v}_1$ ignorovat). Vektory $\boldsymbol{v}_i$, které udávají tyto směry nových příznaků $\boldsymbol{z}_i$, se z tohoto důvodu nazývají **hlavními komponentami** (*principal components*). Obrázek 4.2 ilustruje tuto vlastnost hlavních komponent (ve dvou dimenzích).

Tedy celkově lze působení faktoru λ shrnout tak, že čím je větší, tím více je potlačen vliv nevýznamných hlavních komponent (těch náležící malým singulárním hodnotám d_p, d_{p-1}, ...).

Nyní je třeba uvést několik komentářů. Jelikož hlavní komponenty porovnávají výběrový rozptyl transformovaných příznaků, vzniká otázka, zda standardizovat příznaky před použitím této metody. Standardizací se rozumí nastavení výběrového rozptylu všech příznaků na 1. (Spolu s již provedeným vystředěním hovoříme o normalizaci dat.) Pro neznámé příznaky je toto jistě rozumných krokem, ovšem pro příznaky, u kterých víme, že jejich různé rozptyly mají nějaký význam, je standardizace nevhodná, protože o tuto informaci přijdeme.

Druhá poznámka se týká toho, že obecně neexistuje závislost mezi hlavními komponentami a vysvětlovanou proměnnou, jedná se tedy o čistě heuristickou metodu, která je založená na tom, že hlavní komponenty náležící největším singulárním hodnotám (tedy mající největší výběrový rozptyl) obsahují také nejvíce informace o vysvětlované proměnné. Mohli bychom uměle zkonstruovat příklad, kde by vysvětlovaná proměnná závisela až na poslední komponentě (nebylo by to kvůli zašumění dat tak jednoznačné).

[3] Matice $\mathbf{V}$ je ortogonální.

Obrázek 4.2: Ukázka hlavních komponent pro dvoudimenzionální data. Symbolem $\widetilde{\boldsymbol{v_i}}$ značíme vektor $\boldsymbol{v_i}$ o velikosti rovnající se odhadu směrodatné odchylky dat ve směru $\boldsymbol{v_i}$, přesněji $\widetilde{\boldsymbol{v_i}} = \frac{d_i}{\sqrt{N}}\boldsymbol{v_i}$. Bod $\boldsymbol{\mu}$ je průměrem dat. Pro tuto ukázkou nejsou data standardizována, protože ve dvou dimenzích bychom neviděli nic zajímavého (cvičení 4.3).

Použitelnost metody je založená na praktických zkušenostech: hlavní komponenty často obsahují dostatek informace o vysvětlované proměnné. Toto se přímo využívá v metodě regrese pomocí hlavních komponent, kterou uvádíme níže v části 4.2.4.

4.2.1.2 Vychýlení a rozptyl

V následujícím textu si zjednodušíme značení a budeme psát $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} = \widehat{\boldsymbol{w}}_{\text{ridge}(\lambda),+}$. Stále pracujeme s vystředěnými příznaky a předpokládáme, že platí $\boldsymbol{Y} = \mathbf{X}_c\boldsymbol{w}_c + \boldsymbol{\varepsilon}$[4]. Dále předpokládáme $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$.

Spočtěme si nyní střední hodnotu $\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$. Platí

$$\begin{aligned}
\mathbb{E}\,\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} &= \mathbb{E}\left(\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c^T\boldsymbol{Y}\right)\\
&= \mathbb{E}\left(\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c^T\mathbf{X}_c\boldsymbol{w}_c + \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c^T\varepsilon\right)\\
&= \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c^T\mathbf{X}_c\boldsymbol{w}_c\\
&= \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I - \lambda I\right)\boldsymbol{w}_c\\
&= \boldsymbol{w}_c + \lambda\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\boldsymbol{w}_c.
\end{aligned}$$

Pro $\lambda = 0$ tedy opět pozorujeme nestrannost odhadu pomocí nejmenších čtverců. Obecně lze dokázat, že tento výraz se blíží k 0 s rostoucím λ. Tato tendence však nemusí (a často není) monotónní. dukaz? cviceni?

Spočtěme variační matici $\operatorname{var}\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}$:

$$\begin{aligned}
\operatorname{var}\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} &= \operatorname{var}\left(\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c^T\boldsymbol{Y}\right)\\
&= \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c\operatorname{var}\boldsymbol{Y}\left(\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c^T\right)^T\\
&= \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c\sigma^2\mathbf{I}\mathbf{X}_c^T\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-T}\\
&= \sigma^2\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\mathbf{X}_c\mathbf{X}_c^T\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-T}.
\end{aligned}$$

Tento výraz má opět za limitu $\mathbf{0}$ pro λ jdoucí do nekonečna. dukaz? cviceni?

Uvažujme nějaký bod $\boldsymbol{x}_0 \in \mathbb{R}^{p,1}$ a spočtěme vychýlení a roztpyl v tomto bodě. Pro kvadrát vychýlení platí

$$\left(\mathbb{E}\,\widehat{\boldsymbol{Y}_0} - \mathbb{E}\,\boldsymbol{Y}_0\right)^2 = \left(\mathbb{E}\,\boldsymbol{x}_0^T\widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)} - \mathbb{E}\,\boldsymbol{x}_0^T\boldsymbol{w}_c\right)^2 = \left(\boldsymbol{x}_0^T\lambda\left(\mathbf{X}_c^T\mathbf{X}_c + \lambda I\right)^{-1}\boldsymbol{w}_c\right)^2.$$

[4] Značením $\boldsymbol{w}_c$ chceme zdůraznit, že vystředěním příznaků se nám změní i vektor skutečných vah.

Pro rozptyl

$$\operatorname{var} \widehat{vY_0} = \operatorname{var}\left(\boldsymbol{x}_0^T \widehat{\boldsymbol{w}}_{\text{ridge}(\lambda)}\right).$$

Je tedy vidět, že součet kvadrátu vychýlení a rozptylu, který je ve střední chybě modelu (3.7), má pro λ jdoucí do nekončna následující chování: pro $\lambda = 0$ je vychýlení je nulové a rozptyl nenulový a pro $\lambda \to +\infty$ je situace naopak.

doplnit, dodelat, theobal 1970

4.2.2 Laso (*Lasso*)

$\lambda \geq 0$

$$\begin{aligned}\widehat{\boldsymbol{w}}_{\text{lasso}(\lambda)} &= \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \operatorname{RSS}(\boldsymbol{w}) + \lambda \sum_{j=1}^{p} |w_j| \right\} \\ &= \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{p} x_{i,j} w_j\right)^2 + \lambda \sum_{j=1}^{p} |w_j| \right\}\end{aligned} \tag{4.4}$$

toto zmenit jen na poznamku

4.2.3 (*Least angle regression, LAR*)

Least angle regression (LAR)

4.2.4 Regrese pomocí hlavních komponent (*Principal component regression, PCR*)

Na pojem hlavní komponenty jsme již narazili v části 4.2.1.1. Myšlenka regrese pomocí hlavních komponent je založena na použití pouze některých komponent, typicky těch, na kterých mají data největší rozptyl, tedy těch „nejhlavnějších“. Počet použitých komponent je základním paramtrem této úlohy, my jej označíme $M \in N$.

Připomeňme si použití SVD rozkladu (věta A.4): $\mathbf{X}_c = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Příznaky ve směrech hlavních komponent opět označme $\boldsymbol{z}_i = \mathbf{X}_c \boldsymbol{v}_i$. Předpokládejme stále, že singulární hodnoty jsou seřazené, tedy $d_1 \geq d_2 \geq d_3 \ldots$. PCR volí prvních M nových příznaků $\boldsymbol{z}_i$. Označme tedy

$$\mathbf{V}_M = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_M \end{pmatrix} \in \mathbb{R}^{p,M}.$$

Náš model se tedy změní následujícím způsobem

$$\boldsymbol{Y} = \mathbf{X}_c \boldsymbol{w} + \boldsymbol{\varepsilon} = \mathbf{X}_c \mathbf{V}_M \mathbf{V}_M^T \boldsymbol{w} + \boldsymbol{\varepsilon} = \mathbf{Z}_M \boldsymbol{w}' + \boldsymbol{\varepsilon},$$

kde $\mathbf{Z}_M = \mathbf{X}_c \mathbf{V}_M$ a $\mathbf{V}_M^T \boldsymbol{w} = \boldsymbol{w}'$. Řešme tuto úlohu metodou nejmenší čtverců:

$$\widehat{\boldsymbol{w}'}_{\text{OLS}} = \left(\mathbf{Z}_M^T \mathbf{Z}_M\right)^{-1} \mathbf{Z}_M^T Y. \tag{4.5}$$

Z rovnosti $\boldsymbol{z}_i = \mathbf{X}_c \boldsymbol{v}_i = d_i \boldsymbol{u}_i$ plyne

$$\boldsymbol{z}_i^T \boldsymbol{z}_j = \begin{cases} d_i^2 & \text{pokud } i = j, \\ 0 & \text{jinak.} \end{cases}$$

Obrázek 4.3: Hlavní komponenta nemusí být tou „nejlepší“.

Tedy matice $\mathbf{Z}_M^T\mathbf{Z}_M$ je diagonální a my vlastně řešíme M nezávislých problémů jednorozměrné lineární regrese. Z rovnosti (4.5) plyne

$$\left(\widehat{\boldsymbol{w}'}_{\mathrm{OLS}}\right)_i = \frac{\boldsymbol{z}_i^T\boldsymbol{Y}}{\|\boldsymbol{z}_i\|^2} = \frac{\boldsymbol{z}_i^T\boldsymbol{Y}}{d_i^2}.$$

Zbýva použít rovnost $\mathbf{V}_M^T\boldsymbol{w} = \boldsymbol{w}'$, abychom dostali hledaný odhad vah

$$\widehat{\boldsymbol{w}}_{\mathrm{PCR}(M)} = \mathbf{V}_M\widehat{\boldsymbol{w}'}_{\mathrm{OLS}} = \sum_{i=1}^{M}\frac{\boldsymbol{z}_i^T\boldsymbol{Y}}{d_i^2}\boldsymbol{v}_i.$$

Komponenty nesouvisejí s vysvětlovanými proměnnými.

Souvislost s vaz. extrem, vypocet rozptylu, NIPALS algoritmus?

obrázek z winequality-red.csv

4.3 Cvičení

Cvičení 4.1: Odvoďte rovnosti (4.3).

Cvičení 4.2: Dokažte, že matice $\mathbf{X}_c'^T\mathbf{X}_c' + \lambda\mathbf{I}'$ je pozitivně definitní.

Cvičení 4.3: Mějme dvoudimenzionální data, tedy dva příznaky. Jak dopadne PCA, pokud taková data normalizujeme?

Kapitola 5

Jádrové metody

Doposud jsme uvažovali pouze jednoduchý lineární model vysvětlované proměnné Y zahrnující lineární kombinace p vstupních proměnných $X_1, \ldots, X_p$ ve tvaru

$$Y = f(\boldsymbol{X}) + \varepsilon,$$

kde

$$f(\boldsymbol{X}) = w_0 + w_1 X_1 + \ldots + w_p X_p = \boldsymbol{X}^T \boldsymbol{w}.$$

Principiálně jsme tak schopni modelovat pouze lineární funkci ve vstupních proměnných. V této kapitole si ukážeme metody, jak rozšířit naše možnosti za obzor linearity. Základní úvaha spočívá v nahrazení složek vstupních proměnných novými proměnnými, které vzniknou transformacemi vstupních proměnných. Jako rozumný model potom budeme opět uvažovat lineární model ovšem nyní již v těchto nových proměnných.

Pro $M \in \mathbb{N}$ vezměme M funkcí $\varphi_1, \ldots, \varphi_M$ z $\mathbb{R}^p$ do $\mathbb{R}$ reprezentujících transformace X a nazvěme je *bázové funkce*. Z těchto funkcí můžeme vytvořit jednu vícekomponentní funkci $\boldsymbol{\varphi} : \mathbb{R}^p \to \mathbb{R}^M$ vztahem $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_M)^T$. Jako základní model vztahu Y a $\boldsymbol{X}$ budeme uvažovat model

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{5.1}$$

kde neznámou funkci f hledáme ve tvaru

$$f(\boldsymbol{X}) = \sum_{j=1}^{M} w_j \varphi_j(\boldsymbol{X}) = \boldsymbol{\varphi}(\boldsymbol{X})^T \boldsymbol{w}. \tag{5.2}$$

Výhodou tohoto přístupu je, že jakmile jsou jednou bázové funkce zvoleny, je model lineární v těchto nových proměnných a můžeme použít všechny klasické v předchozích kapitolách analyzované metody odhadu neznámého vektoru parametrů $\boldsymbol{w}$. Mějme nyní náhodný výběr z výše uvedeného náhodného modelu určený N páry typu $(Y_i, \boldsymbol{x}_i)$, kde Y_i je náhodná veličina ukazující výsledek vysvětlované proměnné v i-tém bodě $\boldsymbol{x}_i = (x_{i;1}, \ldots, x_{i;p})^T$. Hodnoty bázových funkcí v bodech trénovacích dat si poskládáme do matice $\boldsymbol{\Phi} \in \mathbb{R}^{N,M}$. To znamená, že i-tý řádek tvoří složky vektoru $\boldsymbol{\varphi}(\boldsymbol{x}_i)$, $\boldsymbol{\Phi}_{ij} = \varphi_j(\boldsymbol{x}_i)$.

Při metodě nejmenších čtverců tak budeme minimalizovat

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} \left(Y_i - \boldsymbol{\varphi}(\boldsymbol{x}_i)^T \boldsymbol{w}\right)^2 = \|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 .$$

V případě, že má matice $\mathbf{\Phi}$ plnou hodnost, je analogicky k (3.5) řešením

$$\hat{\boldsymbol{w}} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\boldsymbol{Y}. \tag{5.3}$$

Predikce hodnoty Y_0 v bodě $\boldsymbol{x}_0$ je potom určena vztahem

$$\hat{Y}_0 = \boldsymbol{\varphi}(\boldsymbol{x}_0)^T\hat{\boldsymbol{w}}.$$

Pro matici $\mathbf{\Phi}$, která nemá plnou hodnost, můžeme použít hřebenovou regresi a minimalizovat

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = \sum_{i=1}^{N}\left(Y_i - \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{w}\right)^2 + \lambda\sum_{j=1}^{M} w_j^2 = \|\boldsymbol{Y} - \mathbf{\Phi}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2.$$

Řešení pro $\lambda > 0$ vždy existuje a je

$$\hat{\boldsymbol{w}}_\lambda = \left(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I}_M\right)^{-1}\mathbf{\Phi}^T\boldsymbol{Y},$$

kde $\mathbf{I}_M$ je jednotková matice $M \times M$.

Mezi obvyklé volby bázových funkcí patří:

- $\varphi_i(\boldsymbol{x}) = x_i$ – bázové funkce tvořené jednotlivými příznaky,
- $\varphi_i(\boldsymbol{x}) = x_i^2$, $\varphi_j(\boldsymbol{x}) = x_k x_\ell$ – bázové funkce z mocnin příznaků a jejich různé součiny, odpovídá polynomiální regresi,
- $\varphi_i(\boldsymbol{x}) = \log(x_i), \sqrt{x_i}, \sin(x_i)$ atd. – bázové funkce vzniklé nelineárními transformacemi jednotlivých příznaků,
- $\varphi_i(\boldsymbol{x}) = \mathbb{1}_{(a,b)}(x_i)$, kde $\mathbb{1}_A(x) = 1$ pokud $x \in A$ a $\mathbb{1}_A(x) = 0$ pokud $x \notin A$ – bázové funkce z indikátorů množin, umožňuje rozdělení prostoru příznaků na kousky a fitování v každém kousku zvlášť,
- $\varphi_i(\boldsymbol{x}) = h(\|\boldsymbol{x} - \boldsymbol{x}_i\|)$, kde $\boldsymbol{x}_i$ je i-tý trénovací bod a h je nějaká funkce – *radiální bázové funkce* centrované v bodech trénovací množiny.

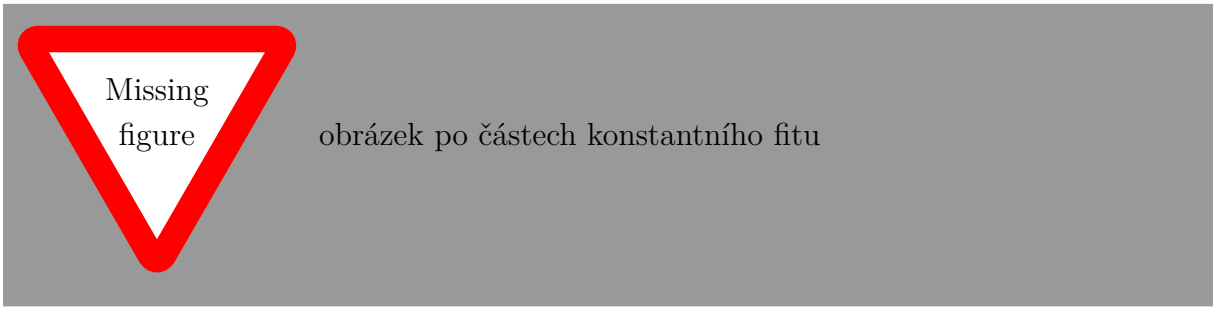
Pokud nemáme žádné speciální znalosti o systému, který modelujeme, typicky na počátku volíme velké množství bázových funkcí a hřebenová regrese (nebo jiná metoda z části 4) je velmi rozumná.

5.1 Spline křivky

Nyní se zaměříme na situaci, kdy máme pouze jeden příznak, tj. $p = 1$, ale očekáváme, že neznámá funkce f v modelu (5.1) je značně nelineární. Jako bázové funkce budeme nyní volit různé mocniny vstupní proměnné x v kombinaci s charakteristickými funkcemi intervalů, které tvoří rozklad $\mathbb{R}$. Výsledkem bude po částech polynomiální model, který potom rozšíříme o požadavky spojitosti v uzlových bodech dělících intervalů a získáme tak model spline křivek.

Mějme K dělících bodů $\xi_1 < \xi_2 < \ldots < \xi_K$ a uvažujme $K+1$ intervalů $I_1 = (-\infty, \xi_1]$, $I_i = (\xi_{i-1}, \xi_i]$ pro $i = 2, \ldots, K$, a $I_{K+1} = (\xi_K, +\infty)$. Zjevně platí

$$\mathbb{R} = \bigcup_{i=1}^{K+1} I_i.$$



Obrázek 5.1: Po částech konstantní fit.

5.1.1 Po částech polynomiání model

Uvažujme nejprve bázové funkce $\varphi_j(x) = \mathbb{1}_{I_j}(x)$ pro každé $j = 1, \ldots, K+1$. Podívejme se na odhad $\hat{\boldsymbol{w}}$ metodou nejmenších čtverců v modelu (5.1) pro

$$f(X) = \sum_{j=1}^{K+1} w_i \varphi_i(X).$$

Pro každé $i = 1, \ldots, N$ a $j = 1, \ldots, K+1$ platí $\varphi_j(x_i) = 1$ pokud $x_i \in I_j$ a $\varphi_j(x_i) = 0$ pokud $x_i \notin I_j$. To znamená, že vektor $\boldsymbol{\varphi}(x_i)$ má pouze jednu složku odpovídající intervalu do kterého x_i patří rovnou 1 a ostatní složky rovné 0. Matice $\boldsymbol{\Phi}$ má tedy v každém řádku pouze jednu jedničku a jinak nuly. Platí tedy

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \begin{pmatrix} \#I_1 & 0 & \cdots & 0 \\ 0 & \#I_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \#I_{K+1} \end{pmatrix},$$

kde $\#I_j$ značí počet trénovacích bodů, které se nacházejí v intervalu I_j. Pro inverzní matici tedy platí

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \begin{pmatrix} \frac{1}{\#I_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\#I_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\#I_{K+1}} \end{pmatrix}$$

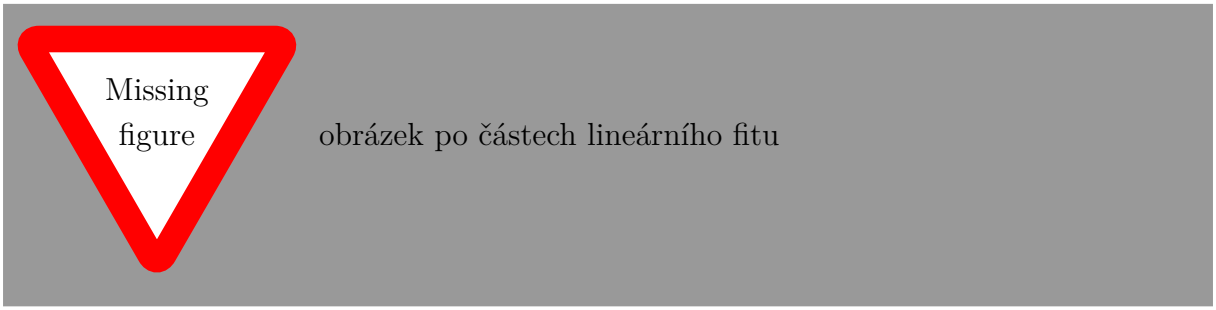
a pro odhad $\hat{\boldsymbol{w}}$ daný vztahem (5.3) získáváme

$$\hat{w}_j = \frac{1}{\#I_j} \sum_{i=1}^{N} \mathbb{1}_{I_j}(x_i) Y_j = \bar{Y}_j,$$

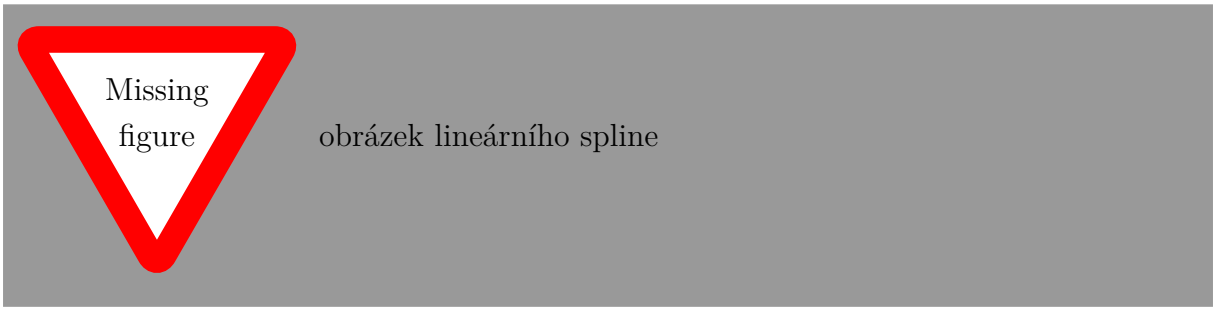
kde $\bar{Y}_j$ značí výběrový průměr $\boldsymbol{Y}$ pouze přes body obsažené v intervalu I_j. Příklad po částech konstantního proložení je na obrázku 5.1.

Další možností je použít po částech lineární model. K výše zavedeným bázovým funkcím $\varphi_1, \ldots, \varphi_{K+1}$ přidejme v takovém případě přidáme ještě dalších $K+1$ bázových funkcí definovaných pro každé $j = 1, \ldots, K+1$ vztahem

$$\varphi_{K+1+j}(x) = x\varphi_j(x) = x\, \mathbb{1}_{I_j}(x).$$



Obrázek 5.2: Po částech lineární fit.



Obrázek 5.3: Po částech lineární fit.

Příklad po částech lineárního proložení je na obrázku 5.2.

V mnoha případech je rozumné předpokládat, že neznámá funkce f je spojitá. V takovém případě je po částech lineární model nedostatečný, protože je v okolí dělících bodů ξ_i nespojitý. V takovém případě je vhodné omezit volnost modelu podmínkami spojitosti v dělících bodech. Tyto podmínky můžeme zapsat jako

$$f(\xi_{j-}) = \lim_{x \to \xi_{j-}} f(x) = \lim_{x \to \xi_{j+}} f(x) = f(\xi_{j+}),$$

pro každé $j = 1, \ldots, K$. Po dosazení tvaru (5.2) a konkrétním vyjádření bázových funkcí dostaneme K rovnic:

$$w_j + \xi_j w_{K+1+j} = w_{j+1} + \xi_{j+1} w_{K+2+j}$$

pro každé $j = 1, \ldots, K$. Úlohu minimalizace $\mathrm{RSS}(\boldsymbol{w})$ spolu s těmito podmínkami je možno řešit jako klasickou úlohu vázaných extrémů vedoucí na Lagrangeovu funkci.

Druhou možností je zaintrodukovat tyto podmínky do volby báze tak, aby je všechny bázové funkce splňovali a zároveň zůstala zachována plná variabilita modelu. Lze ukázat (viz cvičení 5.1), že takovéto bázové funkce jsou

$$\varphi_1(x) = 1, \quad \varphi_2(x) = x, \quad \varphi_{j+1}(x) = (x - \xi_j)_+ \text{ pro } j = 1, \ldots, K,$$

kde $(x)_+ = x$ pokud $x > 0$ a $(x)_+ = 0$ pokud $x \leq 0$. Ukázka spojitého po částech lineárního proložení je na obrázku 5.3.

Obvykle je vhodné předpokládat, že neznámá funkce není jenom spojitá, ale že je ještě mnohem hladší. Například, že je diferencovatelná v každém bodě a že její derivace je spojitá, a podobně i pro derivace vyšších řádů. Jednoduchý model, který respektuje

tuto hladkost, lze vytvořit analogickým způsobem jako spojitý po částech lineární model akorát s použitím vyšších mocnin.

Označme L nejvyšší mocninu polynomů, kterými chceme lokálně prokládat. Jako požadavek dostatečné hladkosti zvolme spojitost všech derivací (i nulté - což je spojitost funkce) až do řádu $L-1$. Lze ukázat, že odpovídající množina bázových funkcí splňující tyto podmínky má velikost $L+K$ a je tvořena funkcemi

$$\varphi_j(x) = x^{j-1},\ j = 1, \ldots, L \quad \text{a} \quad \varphi_{j+L}(x) = (x-\xi_j)_+^L \text{ pro } j = 1, \ldots, K.$$

Funkce f ve tvaru (5.2) s těmito bázovými funkcemi se obecně nazývá *spline křivka* nebo jenom *spline*. Pro nejčastější volbu $L = 3$ se používá název *kubický spline*. Kubický spline je tedy po částech kubický polynom, který je spojitý a má spojitou první i druhou derivaci. Říká se, že kubický spline je spline nejnižšího stupně, pro který nejsou jednotlivé uzlové body $\xi_1, \ldots, \xi_K$ pro lidské oko rozpoznatelné.

5.1.2 Normální spline

Jedním z největších problémů polynomiálních modelů je extrapolace. Mimo rozsah trénovacích bodů převáží nejvyšší mocnina proloženého polynomu a regresní křivka tak s velkou pravděpodobností přestane odpovídat skutečné funkci f. Tento problém se obvykle řeší přidáním požadavku linearity regresní křivky v krajních intervalech $I_1 = (-\infty, \xi_1]$ a $I_{K+1} = (\xi_K, +\infty)$. Tento požadavek lze pro spline stupně L zapsat pomocí rovnic požadující nulovost všech derivací od druhé výše v bodě ξ_1 zleva a v bodě ξ_K zprava:

$$f^{(2)}(\xi_{1-}) = f^{(3)}(\xi_{1-}) = \ldots = f^{(L)}(\xi_{1-}) = 0,$$
$$f^{(2)}(\xi_{K+}) = f^{(3)}(\xi_{K+}) = \ldots = f^{(L)}(\xi_{K+}) = 0.$$

Pro kubický spline lze snadno ukázat, viz cvičení 5.2, že plná množina bázových funkcí, které jsou již s těmito podmínkami kompatibilní, je určena K funkcemi

$$\varphi_1(x) = 1, \quad \varphi_2(x) = x, \quad \varphi_{j+2}(x) = \frac{(x-\xi_j)_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_j} - \frac{(x-\xi_j)_+^3 - (x-\xi_{K-1})_+^3}{\xi_K - \xi_{K-1}} \tag{5.4}$$

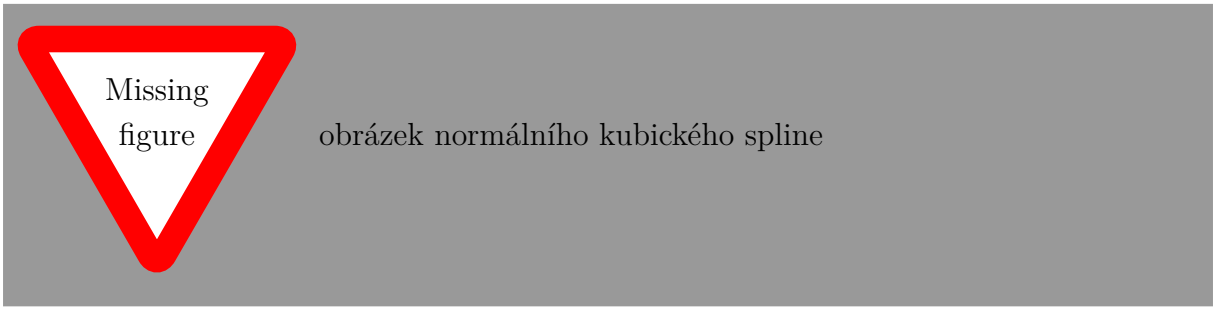
pro každé $j = 1, \ldots, K-2$. Funkce f ve tvaru (5.2) s těmito bázovými funkcemi se obecně nazývá *normální kubický spline*. Ukázka proložení normálním kubickým splinem je na obrázku 5.4.

5.1.3 Smoothing spline

Doposud jsme se nezabývali problémem selekce uzlových bodů $\xi_1, \ldots, \xi_K$. Tento problém je ale podstatný neboť vnáší velkou volnost a tedy nejistotu do procesu volby modelu. Zkusme se nyní zabývat metodou, která se tomuto problému elegantně vyhne.

Začněme s na první pohled obecnějším problémem. Označme jako $C^2(\mathbb{R})$ množinu všech spojitých funkcí na $\mathbb{R}$ se spojitou první a druhou derivací. Zkusme se zabývat problémem nalezení $f \in C^2(\mathbb{R})$, která minimalizuje

$$\mathrm{RSS}(f, \lambda) = \sum_{i=1}^{N} \big(Y_i - f(x_i)\big)^2 + \lambda \int_{-\infty}^{+\infty} |f''(t)|^2 \, \mathrm{d}t. \tag{5.5}$$



Obrázek 5.4: Normální kubický spline.

První část $\mathrm{RSS}(f, \lambda)$ je běžná penalizace z metody nejmenších čtverců a druhá část je penalizace na globální míru velikosti druhé derivace. Jelikož druhou derivaci můžeme interpretovat jako míru „zatáčení“ funkce – odpovídá druhá část penalizaci za globální „zatáčení“ funkce f. Čím více a na delším úseku je funkce f „zvlněná“, tím větší je hodnota integrálu. Naopak, čím více se f blíží lineární funkci, tím je hodnota integrálu menší.

Lze ukázat (viz např. [**?**]), že funkce, která minimalizuje (5.5) je normální kubický spline s uzlovými body ve všech bodech trénovacích množiny, tj. $K = N$ a $\xi_i = x_i$ pro každé $i = 1, \ldots, N$. Takovouto funkci nazýváme *smoothing spline*.

Opět tedy uvažujeme model (5.1) pro

$$f(x) = \sum_{j=1}^{N} w_j \varphi_j(X)$$

s bázovými funkcemi $\varphi_1, \ldots \varphi_N$ definovanými vztahy (5.4). Odhady složek vektoru $\boldsymbol{w}$ najdeme minimalizací $\mathrm{RSS}(f, \lambda)$, který si nyní opět upravíme do maticového tvaru. Protože derivace je lineární, platí

$$f''(x) = \sum_{j=1}^{N} w_j \varphi_j''(X).$$

Po dosazení do (5.5) tak můžeme psát

$$\mathrm{RSS}(f, \lambda) = \|\boldsymbol{Y} - \boldsymbol{\Phi w}\|^2 + \lambda \boldsymbol{w}^T \boldsymbol{\Omega w},$$

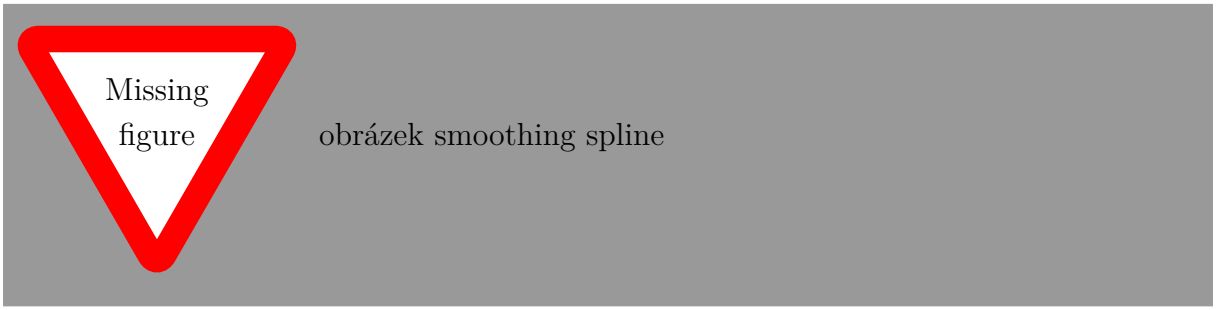
kde $\boldsymbol{\Omega}$ je matice $N \times N$ se složkami

$$\Omega_{ij} = \int_{-\infty}^{+\infty} \varphi_i''(t) \varphi_j''(t)\, \mathrm{d}t.$$

Řešení je možné napsat explicitně jako

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{\Omega}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{Y}.$$

Ukázka proložení smoothing spline křivkou je na obrázku 5.5.



Obrázek 5.5: Smoothing spline

5.2 Duální reprezentace pomocí jádrové funkce

V této části se pokusíme zobecnit vyjádření (5.2) funkce f pomocí bázových funkcí v modelu (5.1) do vyjádření, které bude určeno pouze jedinou tzv. jádrovou funkcí. Opět uvažujme obecný počet příznaků p a obecný počet bázových funkcí M.

Uvažujme problém hřebenové regrese, kdy chceme minimalizovat výraz

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = (\boldsymbol{Y} - \boldsymbol{\Phi w})^T(\boldsymbol{Y} - \boldsymbol{\Phi w}) + \lambda \boldsymbol{w}^T\boldsymbol{w}.$$

Jak víme z části 4.2.1, při hledání řešení této úlohy se klade gradient $\mathrm{RSS}_\lambda(\boldsymbol{w})$ roven $\mathbf{0}$, což vede na rovnici

$$-2\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi w}) + 2\lambda \boldsymbol{w} = 0$$

neboli

$$\boldsymbol{\Phi}^T\boldsymbol{Y} = \boldsymbol{\Phi w} + \lambda \boldsymbol{w},$$

z čehož plyne, že řešení pro $\lambda > 0$ vždy existuje a lze ho vyjádřit jako

$$\hat{\boldsymbol{w}}_\lambda = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \mathbf{I}_M\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{Y}, \tag{5.6}$$

kde $\mathbf{I}_M$ je jednotková matice $M \times M$.

Alternativně můžeme pro $\lambda > 0$ předchozí rovnici přepsat jako

$$\boldsymbol{w} = \frac{1}{\lambda}\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi w}) = \boldsymbol{\Phi}^T\boldsymbol{\alpha}, \tag{5.7}$$

kde jako vektor $\boldsymbol{\alpha} \in \mathbb{R}^N$ jsme označili

$$\boldsymbol{\alpha} = \frac{1}{\lambda}(\boldsymbol{Y} - \boldsymbol{\Phi w}).$$

Všimněte si, že předchozí dva vztahy určují jednoznačný vztah mezi $\boldsymbol{\alpha}$ a $\boldsymbol{w}$.

Dosaďme (5.7) do vztahu pro $\mathrm{RSS}_\lambda(\boldsymbol{w})$:

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = \left\|\boldsymbol{Y} - \boldsymbol{\Phi\Phi}^T\boldsymbol{\alpha}\right\|^2 + \lambda\boldsymbol{\alpha}^T\boldsymbol{\Phi\Phi}^T\boldsymbol{\alpha} =: \mathrm{RSS}_\lambda(\boldsymbol{\alpha}).$$

Z tohoto vyjádření a jednoznačnosti mezi $\boldsymbol{\alpha}$ a $\boldsymbol{w}$ plyne, že minimalizace $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ vůči $\boldsymbol{\alpha}$ je ekvivalentní s minimalizací $\mathrm{RSS}_\lambda(\boldsymbol{w})$ vůči $\boldsymbol{w}$. Matici $\boldsymbol{\Phi\Phi}^T$ nazýváme *Gramova matice* a značíme $\mathbf{G} = \boldsymbol{\Phi\Phi}^T$. Můžeme tedy psát

$$\mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \left\|\boldsymbol{Y} - \mathbf{G}^T\boldsymbol{\alpha}\right\|^2 + \lambda\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha}.$$

Řešení minimalizace $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ vůči $\boldsymbol{\alpha}$ získáme položením gradientu rovno nule nebo přímo z rovnice (5.7) a vztahu mezi $\boldsymbol{\alpha}$ a $\boldsymbol{w}$ jako

$$\lambda\boldsymbol{\alpha} = \boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}.$$

Tedy pro každé $\lambda > 0$ platí

$$\hat{\boldsymbol{\alpha}}_\lambda = (\mathbf{G} + \lambda\mathbf{I}_N)^{-1}\boldsymbol{Y}. \tag{5.8}$$

Definujme nyní *jádrovou funkci* $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ vztahem

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{M} \varphi_j(\boldsymbol{x})\varphi_j(\boldsymbol{y}). \tag{5.9}$$

Pro složky Gramovy matice $\mathbf{G}$ platí

$$G_{ij} = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)_{ij} = \sum_{k=1}^{M} \Phi_{ik}\Phi_{kj}^T = \sum_{k=1}^{M} \Phi_{ik}\Phi_{jk} = \sum_{k=1}^{M} \varphi_k(\boldsymbol{x}_i)\varphi_k(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Je tedy možné je všechny vyjádřit pomocí hodnot jádrové funkce v bodech trénovací množiny.

Podívejme se nyní na predikci Y_0 v bodě $\boldsymbol{x}_0$. Při určeném odhadu $\hat{\boldsymbol{w}}$ je predikce určena jako

$$\hat{Y}_0 = \boldsymbol{\varphi}(\boldsymbol{x}_0)^T\hat{\boldsymbol{w}}.$$

Při známém odhadu $\hat{\boldsymbol{\alpha}}$ dosadíme do předchozího vztahu $\hat{\boldsymbol{w}} = \boldsymbol{\Phi}^T\hat{\boldsymbol{\alpha}}$ a dostaneme

$$\hat{Y}_0 = \boldsymbol{\varphi}(\boldsymbol{x}_0)^T\boldsymbol{\Phi}^T\hat{\boldsymbol{\alpha}} = \sum_{i=1}^{M}\sum_{j=1}^{N} \varphi_i(\boldsymbol{x}_0)\Phi_{ij}^T\hat{\alpha}_j = \sum_{i=1}^{M}\sum_{j=1}^{N} \varphi_i(\boldsymbol{x}_0)\Phi_{ji}\hat{\alpha}_j = \sum_{i=1}^{M}\sum_{j=1}^{N} \varphi_i(\boldsymbol{x}_0)\varphi_i(\boldsymbol{x}_j)\hat{\alpha}_j.$$

Použitím definice jádrové funkce K konečně získáme

$$\hat{Y}_0 = \sum_{j=1}^{N} \hat{\alpha}_j K(\boldsymbol{x}_0, \boldsymbol{x}_j). \tag{5.10}$$

Obecně budeme model (5.1), kde neznámou funkci f hledáme ve tvaru

$$f(\boldsymbol{X}) = \sum_{j=1}^{N} \alpha_j K(\boldsymbol{X}, \boldsymbol{x}_j), \tag{5.11}$$

nazývat *jádrovým modelem* (*vector machine* a někdy také *kernel machine*).

Celou úlohu tedy můžeme reformulovat pouze pomocí jádrové funkce K, která obsahuje bázové funkce implicitně. Na definiční vztah (5.9) je možné pohlížet jako na skalární součin vektorů $\boldsymbol{\varphi}(\boldsymbol{x})$ a $\boldsymbol{\varphi}(\boldsymbol{y})$. Jádrová funkce K je tak zvaně *pozitivně semidefinitní*, což znamená, že pro každé N a pro libovolné body $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^p$ je její Gramova matice $\mathbf{G}$ se složkami $G_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ positivně semidefinitní matice. Toto tvrzení je zřejmé z vyjádření $\mathbf{G} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$. Lze ukázat, že v jistém smyslu platí i opak. To znamená, že ke každé positivně semidefinitní funkci $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ existuje nějaký Hilbertův[1] prostor funkcí H a zobrazení $\boldsymbol{\varphi} : \mathbb{R}^p \to H$ tak, že $K(\boldsymbol{x}, \boldsymbol{y}) = \langle\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{y})\rangle$, kde $\langle\cdot,\cdot\rangle$ značí skalární součin v H, detaily např. v [**?**, kapitola 3].

[1] Lineární vektorový prostor se skalárním součinem, který je navíc úplný.

Příklad 5.1: Pro $p = 2$ uvažujme jádrovou funkci $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + 1)^2$ a pokusme se ji vyjádřit ve tvaru (5.9) pro nějaké bázové funkce. Snadno upravíme:

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + 1)^2 = (x_1y_1 + x_2y_2 + 1)^2 = x_1^2y_1^2 + x_2^2y_2^2 + 1 + 2x_1x_2y_1y_2 + 2x_1y_1 + 2x_2y_2.$$

Zjevně tedy $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\varphi}^T(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{y})$ pro

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \Big(x_1^2, x_2^2, 1, \sqrt{2}x_1y_1, \sqrt{2}x_1, \sqrt{2}x_2\Big).$$

△

Jak bylo zmíněno, validní jádrová funkce odpovídající nějaké množině bázových funkcí musí být pozitivně semidefinitní. Otázka je, jak můžeme získat validní jádrovou funkci v případě, že se rozhodneme s jádrovou funkcí začít. Mezi obvyklé volby jádrové funkce patří tzv. *Gaussovské jádro* definované vztahem

$$K_\gamma(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2},$$

kde parametr $\gamma > 0$ určuje „šířku“ jádrové funkce. Poznamenejme, že takováto volba odpovídá modelu radiálních bázových funkcí (viz výše) s volbou $h(x) = \mathrm{e}^{-\gamma x^2}$ a proto bývá Gaussovské jádro někdy nazýváno *RBF jádro*. Další možnou volbou jádra je polynomiální jádro ve tvaru $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + 1)^d$ pro $d > 0$.

Obecně je možné získávat validní jádrové funkce z jiných pomocí různých skládání. Uveďme nyní tvrzení, které popisuje jakým způsobem je možné jádrové funkce skládat a tvořit, a pro další detaily a důkaz odkazujeme čtenáře na knihu [**?**].

Tvrzení 5.2: Nechť K_1 a K_2 jsou dvě validní jádrové funkce na $\mathbb{R}^p \times \mathbb{R}^p$, $a \in \mathbb{R}^+$, f libovolná reálná funkce na $\mathbb{R}^p$, ϕ zobrazení z $\mathbb{R}^p$ do $\mathbb{R}^s$, K_3 validní jádrová funkce na $\mathbb{R}^s \times \mathbb{R}^s$ a $\mathbf{B}$ symetrická reálná pozitivně semidefinitní matice $p \times p$. Potom jsou všechny následující funkce validní jádrové funkce:

a) $K(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y}) + K_2(\boldsymbol{x}, \boldsymbol{y})$,

b) $K(\boldsymbol{x}, \boldsymbol{y}) = aK_1(\boldsymbol{x}, \boldsymbol{y})$,

c) $K(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y}) \cdot K_2(\boldsymbol{x}, \boldsymbol{y})$,

d) $K(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x})f(\boldsymbol{y})$,

e) $K(\boldsymbol{x}, \boldsymbol{y}) = K_3\Big(\phi(\boldsymbol{x}), \phi(\boldsymbol{y})\Big)$,

f) $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\mathbf{B}\boldsymbol{y}$,

g) $K(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{K_1(\boldsymbol{x},\boldsymbol{y})}$.

Na závěr uveďme ještě jednu možnou výhodu jádrového modelu. Tou je fakt, že v odhadu (5.8) je třeba invertovat matici $N \times N$ zatímco v odhadu (5.6) je třeba invertovat matici $M \times M$. Je známo, že k inverzi matice $N \times N$ je třeba $\mathcal{O}(N^3)$ kroků a k inverzi matice $M \times M$ analogicky $\mathcal{O}(M^3)$ kroků. Pokud tedy máme oproti velikosti trénovací množiny velké množství bázových funkcí, je jádrový přístup výhodný i z důvodu nižších výpočetních nároků. Jak uvidíme dále, je navíc možné modifikovat jádrový model tak, aby bylo pouze několik složek $\hat{\boldsymbol{\alpha}}$ nenulových a tedy aby se výrazně urychlila i tzv. vybavovací fáze, při které provádíme predikci v nových bodech.

5.3 Support vector machines v lineární regresi

Jak jsme viděli v předchozí části, výhodou jádrového modelu (5.11) je možnost vyhnout se specifikaci množiny bázových funkcí respektive možnost využití velmi rozsáhlé množiny bázových funkcí.

V případě, že máme velkou množinu trénovacích dat, může být problémem nejen tzv. trénovací část, při které počítáme odhad $\hat{\boldsymbol{\alpha}}$ inverzí matice podle vztahu (5.8), ale také samotná predikce $\hat{Y}_0$ hodnoty Y_0 v bodě $\boldsymbol{x}_0$ podle vztahu (5.10). Je proto zajímavé zkusit se zabývat možností získání odhadu $\hat{\alpha}$ takového, který má pouze několik složek nenulových. Při predikci s takovýmto odhadem se potom použijí pouze trénovací body odpovídající nenulovým složkám $\hat{\alpha}$ - tzv. *support vectors*. Existuje několik metod jak takovéto *řídké* modely získat. Ve této části si předvedeme jeden z nejčastějších přístupů nazývaný *support vector machines* neboli zkráceně *SVM*.

Základem SVM je lehce modifikovaný jádrový model (5.11). Pro odvozování se ovšem odrazíme od modelu bázových funkcí (5.2) do kterého explicitně přidáme intercept. Budeme tedy uvažovat

$$f(\boldsymbol{X}) = w_0 + \sum_{j=1}^{M} w_j \varphi_j(\boldsymbol{X}) = w_0 + \boldsymbol{\varphi}(\boldsymbol{X})^T \boldsymbol{w}, \tag{5.12}$$

kde $\boldsymbol{w} = (w_1, \ldots, w_p)$. Ve všech předchozích případech jsme pro získání odhadu $\hat{\boldsymbol{w}}$ a $\hat{w}_0$ minimalizovali residuální součet čtverců, který odpovídá kvadratické ztrátové funkci $L(Y, f(\boldsymbol{X})) = \Big(Y - f(\boldsymbol{X})\Big)^2$. Při hřebenové regresi to znamená minimalizovat výraz

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = \sum_{i=1}^{N} \Big(Y_i - \boldsymbol{\varphi}(\boldsymbol{x}_i)^T \boldsymbol{w} - w_0\Big)^2 + \lambda \left\|\boldsymbol{w}\right\|^2 = \sum_{i=1}^{N} L\Big(Y_i, f(\boldsymbol{x}_i)\Big) + \lambda \left\|\boldsymbol{w}\right\|^2.$$

Minimalizace tohoto výrazu je ekvivalentní minimalizaci obecného výrazu

$$J = C \sum_{i=1}^{N} L\Big(Y_i, f(\boldsymbol{x}_i)\Big) + \frac{1}{2} \left\|\boldsymbol{w}\right\|^2,$$

kde L je kvadratická ztrátová funkce a $C = 1/(2\lambda)$.

Při SVM přístupu zvolíme rozdílnou ztrátovou funkci L_ε nazývanou *epsilon necitlivá ztrátová funkce* a definovanou pro libovolné $\varepsilon > 0$ vztahem

$$L_\varepsilon\Big(Y, f(\boldsymbol{X})\Big) = \max\{|Y - f(\boldsymbol{X})| - \varepsilon, 0\}. \tag{5.13}$$

Výraz, který budeme minimalizovat, je potom

$$\begin{aligned} J_\varepsilon(w_0, \boldsymbol{w}) &= C \sum_{i=1}^{N} L_\varepsilon\Big(Y_i, f(\boldsymbol{x}_i)\Big) + \frac{1}{2} \left\|\boldsymbol{w}\right\|^2 \\ &= C \sum_{i=1}^{N} \max\{|Y_i - f(\boldsymbol{x}_i)| - \varepsilon, 0\} + \frac{1}{2} \left\|\boldsymbol{w}\right\|^2. \end{aligned} \tag{5.14}$$

Na rozdíl od situace s kvadratickou ztrátovou funkcí je nyní minimalizace mnohem obtížnější, neboť J_ε není diferencovatelná podle složek $\boldsymbol{w}$. Řešení tedy nemůžeme získat

položením gradientu rovno nule. Existuje několik způsobů, jak se s tímto problémem vypořádat. Jedním z nich, který zde provedeme, je reformulace pomocí vázaného extrému.

Pro každý bod $\boldsymbol{x}_i$ trénovací množiny zavedeme dvojici nových tzv. *uvolněných* (slack) proměnných ξ_i^+ a ξ_i^- spolu s následujícími podmínkami:

$$\xi_i^+ \geq 0, \tag{5.15}$$
$$\xi_i^- \geq 0, \tag{5.16}$$
$$\xi_i^+ \geq Y_i - f(\boldsymbol{x}_i) - \varepsilon, \tag{5.17}$$
$$\xi_i^- \geq f(\boldsymbol{x}_i) - \varepsilon - Y_i. \tag{5.18}$$

Proměnné ξ_i^+ a ξ_i^- jsou tedy podle (5.15) a (5.16) vždy nezáporné. Z (5.17) dále plyne, že $\xi_i^+ > 0$ pokud je Y_i v bodě $\boldsymbol{x}_i$ nad ε pásem $\Big(f(\boldsymbol{x}) - \varepsilon, f(\boldsymbol{x}) + \varepsilon\Big)$, tj. $Y_i > f(\boldsymbol{x}_i) + \varepsilon$, a pokud $\xi_i^+ = 0$, tak je Y_i v bodě $\boldsymbol{x}_i$ na, uvnitř nebo pod ε pásem, tj. $Y_i \leq f(\boldsymbol{x}_i) + \varepsilon$. Podle podmínky (5.18) pak analogicky $\xi_i^- > 0$ pokud je Y_i v bodě $\boldsymbol{x}_i$ pod ε pásem, tj. $Y_i < f(\boldsymbol{x}_i) - \varepsilon$, a pokud $\xi_i^- = 0$, tak je Y_i v bodě $\boldsymbol{x}_i$ na, uvnitř nebo nad ε pásem, tj. $Y_i \geq f(\boldsymbol{x}_i) - \varepsilon$. Z výše uvedených podmínek navíc ihned plyne následující omezení pro ztrátovou funkci (5.13) v argumentu Y_i a $\boldsymbol{x}_i$:

$$L_\varepsilon\Big(y_i, f(\boldsymbol{x}_i)\Big) \leq \xi_i^+ + \xi_i^-. \tag{5.19}$$

Označme $\boldsymbol{\xi}^+ = (\xi_1^+, \ldots, \xi_N^+)$ a $\boldsymbol{\xi}^- = (\xi_1^-, \ldots, \xi_N^-)$. Minimalizace J_ε je ekvivalentní vázané minimalizaci

$$\tilde{J}_\varepsilon(w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-) = C \sum_{i=1}^{N} (\xi_i^+ + \xi_i^-) + \frac{1}{2} \|\boldsymbol{w}\|^2 \tag{5.20}$$

s vazbami (5.15) – (5.18). Tato ekvivalence plyne z toho, že (5.19) implikuje $J_\varepsilon \leq \tilde{J}_\varepsilon$ a pro $\xi_i^+ = \max\{Y_i - f(\boldsymbol{x}_i) - \varepsilon, 0\}$ a $\xi_i^- = \max\{f(\boldsymbol{x}_i) - \varepsilon - Y_i, 0\}$ platí $J_\varepsilon = \tilde{J}_\varepsilon$.

5.3.1 Duální formulace

Minimalizace tohoto vázaného extrému se nejčastěji řeší duální reformulací, viz teorie v části C.2. K tomu nejprve sestrojíme odpovídající Lagrangeovu funkci $\mathscr{L}$ závislou na proměnných a také na Lagrangeových multiplikátorech. Označme a_i^+ resp. a_i^- multiplikátory příslušející vazbám (5.17) resp. (5.18) a μ_i^+ resp. μ_i^- multiplikátory příslušející vazbám (5.15) resp. (5.16). Označme dále $\boldsymbol{a}^\pm = (a_1^\pm, \ldots, a_N^\pm)$ a $\boldsymbol{\mu}^- = (\mu_1^\pm, \ldots, \mu_N^\pm)$. Lagrangeova funkce je potom dána vztahem

$$\begin{aligned}\mathscr{L}(w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \boldsymbol{a}^+, \boldsymbol{a}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) &= C \sum_{i=1}^{N} (\xi_i^+ + \xi_i^-) + \frac{1}{2} \|\boldsymbol{w}\|^2 \\ &- \sum_{i=1}^{N} a_i^+ \Big(\xi_i^+ - Y_i + f(\boldsymbol{x}_i) + \varepsilon\Big) - \sum_{i=1}^{N} a_i^- \Big(\xi_i^- + Y_i - f(\boldsymbol{x}_i) + \varepsilon\Big) - \sum_{i=1}^{N} (\mu_i^+ \xi_i^+ + \mu_i^- \xi_i^-).\end{aligned}$$

Pro bod $(\hat{w}_0, \hat{\boldsymbol{w}}, \hat{\boldsymbol{\xi}}^+, \hat{\boldsymbol{\xi}}^-, \hat{\boldsymbol{a}}^+, \hat{\boldsymbol{a}}^-, \hat{\boldsymbol{\mu}}^+, \hat{\boldsymbol{\mu}}^-)$, ve kterém se nachází minimum $\mathscr{L}$ platí

$$\nabla_{\hat{w}_0, \hat{\boldsymbol{w}}, \hat{\boldsymbol{\xi}}^+, \hat{\boldsymbol{\xi}}^-} \mathscr{L} = \boldsymbol{0}$$

a Karush-Kuhn-Tucker (KKT) podmínky

$$\xi_i^+ \geq 0, \qquad \xi_i^- \geq 0, \tag{5.21}$$
$$\xi_i^+ - Y_i + f(\boldsymbol{x}_i) + \varepsilon \geq 0, \qquad \xi_i^- + Y_i - f(\boldsymbol{x}_i) + \varepsilon \geq 0, \tag{5.22}$$
$$\mu_i^+ \geq 0, \qquad \mu_i^- \geq 0, \tag{5.23}$$
$$\mu_i^+ \xi_i^+ = 0, \qquad \mu_i^- \xi_i^- = 0, \tag{5.24}$$
$$a_i^+ \geq 0, \qquad a_i^- \geq 0, \tag{5.25}$$
$$a_i^+(\xi_i^+ - Y_i + f(\boldsymbol{x}_i) + \varepsilon) = 0, \qquad a_i^-(\xi_i^- + Y_i - f(\boldsymbol{x}_i) + \varepsilon) = 0. \tag{5.26}$$

Lze snadno ukázat, že v našem případě jsou splněny Slaterovy podmínky a platí tedy silná dualita. To znamená, že

$$\begin{aligned}\min \tilde{J}_\varepsilon(w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-) &= \min \mathscr{L}(w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \boldsymbol{a}^+, \boldsymbol{a}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) \\ &= \max \tilde{\mathscr{L}}(\boldsymbol{a}^+, \boldsymbol{a}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-)\end{aligned}$$

při příslušných omezujících podmínkách na proměnné v argumentech, kde $\tilde{\mathscr{L}}$ je duální Lagrangeova funkce definovaná vztahem

$$\tilde{\mathscr{L}}(\boldsymbol{a}^+, \boldsymbol{a}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) = \inf_{w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-} \mathscr{L}(w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \boldsymbol{a}^+, \boldsymbol{a}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-).$$

K nalezení $\tilde{\mathscr{L}}$ postupně položíme parciální derivace $\mathscr{L}$ podle proměnných $w_0, \boldsymbol{w}, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-$ rovny nule.

$$\frac{\partial \mathscr{L}}{\partial w_j} = w_j - \sum_{i=1}^N a_i^+ \varphi_j(\boldsymbol{x}_i) + \sum_{i=1}^N a_i^- \varphi_j(\boldsymbol{x}_i) = 0,$$

což implikuje

$$w_j = \sum_{i=1}^N (a_i^+ - a_i^-) \varphi_j(\boldsymbol{x}_i). \tag{5.27}$$

Dále

$$\frac{\partial \mathscr{L}}{\partial w_0} = -\sum_{i=1}^N a_i^+ + \sum_{i=1}^N a_i^- = 0,$$

neboli

$$\sum_{i=1}^N (a_i^+ - a_i^-) = 0. \tag{5.28}$$

Derivace podle ξ_i^+ dává

$$\frac{\partial \mathscr{L}}{\partial \xi_i^+} = C - a_i^+ - \mu_i^+ = 0,$$

což implikuje

$$\mu_i^+ = C - a_i^+. \tag{5.29}$$

Nakonec, derivace podle ξ_i^- dává

$$\frac{\partial \mathscr{L}}{\partial \xi_i^-} = C - a_i^- - \mu_i^- = 0$$

a tedy

$$\mu_i^- = C - a_i^-. \tag{5.30}$$

Hessova matice druhých parciálních derivací bude mít diagonální elementy příslušejících druhým derivacím podle w_i rovny 1 a všechny ostatní elementy nulové. Je tedy positivně semidefinitní v každém bodě a proto jsou nalezené kritické body splňující výše odvozené rovnice body minima.

Dosaďme nyní výše odvozené vztahy do Lagrangeovy funkce $\tilde{\mathscr{L}}$. Po krátkém počítání dostaneme

$$\tilde{\mathscr{L}}(\boldsymbol{a}^+, \boldsymbol{a}^-, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{N}(a_i^+ - a_i^-)\varphi_j(\boldsymbol{x}_i)\varphi_j(\boldsymbol{x}_k)(a_k^+ - a_k^-) \\ - \varepsilon\sum_{i=1}^{N}(a_i^+ + a_i^-) + \sum_{i=1}^{N}(a_i^+ - a_i^-)Y_i.$$

Všimněme si, že duální Lagrangeova funkce $\tilde{\mathscr{L}}$ na proměnných $\boldsymbol{\mu}^+, \boldsymbol{\mu}^-$ explicitně vůbec nezávisí. Zavedeme-li opět jádrovou funkci $K : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ vztahem (5.9), můžeme psát

$$\tilde{\mathscr{L}}(\boldsymbol{a}^+, \boldsymbol{a}^-) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}(a_i^+ - a_i^-)K(\boldsymbol{x}_i, \boldsymbol{x}_k)(a_k^+ - a_k^-) - \varepsilon\sum_{i=1}^{N}(a_i^+ + a_i^-) + \sum_{i=1}^{N}(a_i^+ - a_i^-)Y_i. \tag{5.31}$$

Tuto duální Lagrangeovu funkci je nyní třeba maximalizovat vzhledem k podmínkám (5.23) a (5.25), tj. vzhledem k

$$a_i^+ \geq 0, a_i^- \geq 0, \mu_i^+ \geq 0, \mu_i^- \geq 0.$$

Z výše odvozených vztahů (5.28), (5.29) a (5.30) mezi $\mu_i^\pm$ a $a_i^\pm$ plyne, že duální úloha znamená maximalizovat $\tilde{\mathscr{L}}(\boldsymbol{a}^+, \boldsymbol{a}^-)$ určenou vztahem (5.31) za podmínek

$$0 \leq a_i^+ \leq C, \quad 0 \leq a_i^- \leq C \quad \text{a} \quad \sum_{i=1}^{N}(a_i^+ - a_i^-) = 0. \tag{5.32}$$

Řešení této úlohy není možné nalézt explicitně. Jedná se ale o standardní úlohu kvadratického programování, která je numericky dobře řešitelná. Označme řešení této úlohy jako $\hat{\boldsymbol{a}}^+$ a $\hat{\boldsymbol{a}}^-$. Platí tedy

$$\hat{\boldsymbol{a}}^+, \hat{\boldsymbol{a}}^- = \operatorname*{argmax}_{0 \leq a_i^+, a_i^- \leq C, \sum_{i=1}^{N}(a_i^+ - a_i^-) = 0} \tilde{\mathscr{L}}(\boldsymbol{a}^+, \boldsymbol{a}^-).$$

Máme-li $\hat{\boldsymbol{a}}^+$ a $\hat{\boldsymbol{a}}^-$ můžeme vyjádřit $\hat{\boldsymbol{w}}$ pomocí vztahu (5.27) jako

$$\hat{w}_j = \sum_{i=1}^{N}(\hat{a}_i^+ - \hat{a}_i^-)\varphi_j(\boldsymbol{x}_i).$$

Pro predikci $\hat{Y}_0$ hodnoty Y_0 v bodě $\boldsymbol{x}_0$ tak podle (5.12) platí

$$\hat{Y}_0 = \hat{w}_0 + \sum_{j=1}^{M}\hat{w}_j\varphi_j(\boldsymbol{x}_0) = \hat{w}_0 + \sum_{j=1}^{M}\sum_{i=1}^{N}(\hat{a}_i^+ - \hat{a}_i^-)\varphi_j(\boldsymbol{x}_i)\varphi_j(\boldsymbol{x}_0) \tag{5.33}$$

Spolu s definicí (5.9) jádrové funkce tak dostáváme finální vyjádření v termínech $\hat{\boldsymbol{a}}^+$, $\hat{\boldsymbol{a}}^-$ a $\hat{w}_0$:

$$\hat{Y}_0 = \hat{w}_0 + \sum_{i=1}^{N}(\hat{a}_i^+ - \hat{a}_i^-)K(\boldsymbol{x}_0, \boldsymbol{x}_i), \tag{5.34}$$

což tvoří analogii k jádrovému modelu (5.11), kde $\alpha_i = (a_i^+ + a_i^-)$.

Před tím, než se budeme zabývat vyjádřením $\hat{w}_0$ pomocí $\hat{\boldsymbol{a}}^+$ a $\hat{\boldsymbol{a}}^-$, pojďme se nejprve podívat na možné hodnoty $\hat{\boldsymbol{a}}^+$ a $\hat{\boldsymbol{a}}^-$. Pro bod extrému $(\hat{w}_0, \hat{\boldsymbol{w}}, \hat{\boldsymbol{\xi}}^+, \hat{\boldsymbol{\xi}}^-, \hat{\boldsymbol{a}}^+, \hat{\boldsymbol{a}}^-, \hat{\boldsymbol{\mu}}^+, \hat{\boldsymbol{\mu}}^-)$ platí kromě (5.27) – (5.30) také KKT podmínky (5.21) – (5.26). Z první části (5.26) plyne, že

$$\hat{a}_i^+ = 0 \quad \text{nebo} \quad \hat{\xi}_i^+ - Y_i + \hat{f}(\boldsymbol{x}_i) + \varepsilon = 0.$$

V prvním případě $\hat{a}_i^+ = 0$ je z (5.24)

$$\hat{\mu}_i^+\hat{\xi}_i^+ = (C - \hat{a}_i^+)\hat{\xi}_i^+ = C\hat{\xi}_i^+ = 0$$

a tedy $\hat{\xi}_i^+ = 0$. To z (5.22) znamená $\hat{f}(\boldsymbol{x}_i) + \varepsilon \geq Y_i$ neboli, že se bod Y_i nachází na nebo pod horním okrajem ε pásu. Obdobně $\hat{a}_i^- = 0$ znamená $Y_i \geq \hat{f}(\boldsymbol{x}_i) - \varepsilon$ a tedy, že se bod Y_i nachází na nebo nad spodním okrajem ε pásu.

V případě $\hat{a}_i^+ > 0$ a tedy $\hat{\xi}_i^+ - Y_i + \hat{f}(\boldsymbol{x}_i) + \varepsilon = 0$ z podmínky $\hat{\xi}_i^+ \geq 0$ dostáváme $\hat{f}(\boldsymbol{x}_i) + \varepsilon \leq Y_i$, tj. že se bod Y_i nachází na nebo nad horním okrajem ε pásu. Obdobně, pokud $\hat{a}_i^- > 0$ a tedy $\hat{\xi}_i^- + Y_i - \hat{f}(\boldsymbol{x}_i) + \varepsilon = 0$, pak se bod Y_i nachází na nebo pod spodním okrajem ε pásu.

Dále platí, že není možné aby $\hat{a}_i^+ > 0$ a současně $\hat{a}_i^- > 0$. V takovém případě by totiž podle předchozích úvah muselo současně platit

$$\hat{\xi}_i^+ - Y_i + \hat{f}(\boldsymbol{x}_i) + \varepsilon = 0 \quad \text{a} \quad \hat{\xi}_i^- + Y_i - \hat{f}(\boldsymbol{x}_i) + \varepsilon = 0.$$

Sečteme-li tyto dvě rovnice dostaneme $\hat{\xi}_i^+ + \hat{\xi}_i^- + 2\varepsilon = 0$, což vzhledem k nezápornosti $\hat{\xi}_i^+, \hat{\xi}_i^-$ a pozitivitě ε nemůže nastat.

Jak jsme tedy zjistili, buď jsou oba koeficienty $\hat{a}_i^+$ i $\hat{a}_i^-$ nulové, nebo je pouze jeden z nich nenulový. Body $\boldsymbol{x}_i$ odpovídající této druhé situaci se nazývají *nosné body* případně *support vectors*. Při predikci podle vztahu (5.34) tedy sčítáme pouze přes nosné body.

Závěrem se zabývejme určením $\hat{w}_0$. Nejprve uvažujme situaci, kdy existuje $i = 1, \ldots, N$ takové, že $0 < \hat{a}_i^+ < C$ nebo $0 < \hat{a}_i^- < C$. Zabývejme se nyní prvním z těchto dvou případů. Z (5.24) opět dostaneme $\hat{\mu}_i^+\hat{\xi}_i^+ = (C - \hat{a}_i^+)\hat{\xi}_i^+ = 0$ a tedy $\hat{\xi}_i^+ = 0$. Z (5.26) pak dostáváme

$$Y_i = \hat{f}(\boldsymbol{x}_i) + \varepsilon = \hat{w}_0 + \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \varepsilon.$$

To znamená, že

$$\hat{w}_0 = Y_i - \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \varepsilon.$$

Analogicky, pokud $0 < a_i^- < C$, dostaneme

$$\hat{w}_0 = Y_i - \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \varepsilon.$$

Pokud výše uvedené i neexistuje, není $\hat{w}_0$ určeno obecně jednoznačně a můžeme ho zvolit libovolně v rámci omezení (5.21)–(5.26), což pro $\hat{a}_i^+ = 0$ znamená

$$\hat{w}_0 \geq Y_i - \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \varepsilon,$$

pro $\hat{a}_i^- = 0$

$$\hat{w}_0 \leq Y_i - \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \varepsilon,$$

pro $\hat{a}_i^+ = C$

$$\hat{w}_0 \leq Y_i - \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \varepsilon$$

a konečně pro $\hat{a}_i^- = C$

$$\hat{w}_0 \geq Y_i - \sum_{j=1}^{N}(\hat{a}_j^+ - \hat{a}_j^-)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \varepsilon.$$

5.4 Cvičení

Cvičení 5.1: Dokažte, že bázové funkce ... odpovídají spojitému lineárnímu modelu - tj. lineární spline křivce.

Cvičení 5.2: Dokažte, že bázové funkce ... odpovídají modelu normálního spline.

Příloha A

Lineární algebra

A.1 Hodnost matice

Hodnost matice $\mathbf{A} \in \mathbb{R}^{n,m}$ definujeme jako počet lineárně nezávislých řádků (nebo ekvivalentně sloupců) a značíme $h(\mathbf{A})$. Platí-li, $h(\mathbf{A}) = \min\{m, n\}$, říkáme, že $\mathbf{A}$ má plnou hodnost.

Věta A.1: Pro libovolnou matici $\mathbf{A}$ platí $h(\mathbf{A}) = h(\mathbf{A}^T\mathbf{A})$.

Důkaz. Nechť vektor $\boldsymbol{x}$ je kolmý na všechny řádky matice $\mathbf{A}$. To znamená, že $\mathbf{A}\boldsymbol{x} = \mathbf{0}$, neboť vektor $\mathbf{A}\boldsymbol{x}$ je lineární kombinací řádků matice $\mathbf{A}$ s koeficienty určenými složkami vektoru $\boldsymbol{x}$. Vynásobením maticí $\mathbf{A}^T$ zleva dostaneme $\mathbf{A}^T\mathbf{A}\boldsymbol{x} = \mathbf{A}^T\mathbf{0} = \mathbf{0}$, což znamená, že $\boldsymbol{x}$ je rovněž kolmý na všechny řádky matice $\mathbf{A}^T\mathbf{A}$. Na druhou stranu, platí-li $\mathbf{A}^T\mathbf{A}\boldsymbol{x} = \mathbf{0}$, tj. vektor $\boldsymbol{x}$ je kolmý na všechny řádky matice $\mathbf{A}^T\mathbf{A}$, pak $\boldsymbol{x}^T\mathbf{A}^T\mathbf{A}\boldsymbol{x} = \mathbf{0}$. To ale znamená $(\mathbf{A}\boldsymbol{x})^T(\mathbf{A}\boldsymbol{x}) = \|\mathbf{A}\boldsymbol{x}\|^2 = 0$ a tudíž $\mathbf{A}\boldsymbol{x} = \mathbf{0}$, čili $\boldsymbol{x}$ je kolmý i na všechny řádky matice $\mathbf{A}$. Zjistili jsme tedy, že každý vektor, který je kolmý na všechny řádky matice $\mathbf{A}$, je zároveň kolmý na všechny řádky matice $\mathbf{A}^T\mathbf{A}$ a naopak. To znamená, že lineární obaly řádků matic $\mathbf{A}$ a $\mathbf{A}^T\mathbf{A}$ musí být stejné, protože jinak by existoval vektor z jednoho z těchto obalů kolmý na všechny vektory z druhého obalu. Tudíž matice $\mathbf{A}$ a $\mathbf{A}^T\mathbf{A}$ mají stejné počty lineárně nezávislých řádků. □

A.2 Rozklady matic

A.2.1 QR rozklad

Věta A.2: Nechť $\mathbf{A} \in \mathbb{R}^{n,m}$ je matice s lineárně nezávislými sloupci a $n \geq m$. Potom existují matice $\mathbf{Q} \in \mathbb{R}^{n,m}$ s navzájem kolmými sloupci a horní trojúhelníková matice $\mathbf{R} \in \mathbb{R}^{m,m}$ s kladnými diagonálními prvky tak, že

$$A = QR.$$

Následující sekce je jen zkopírovaná z mojí DP

A.2.2 Givensova rotace

Givensova rotace je základní ortogonální transformace, která může sloužit k vynulování daného prvku v matici. Givensova rotace v rovině i-té a j-té složky (bez újmy na obecnosti volíme $i < j$) o úhel θ je definována maticí

$$\mathbf{G}_{i,j,\theta} \stackrel{def}{=} \begin{pmatrix} I_{i-1} & 0 & 0 & 0 & 0 \\ 0 & \cos\theta & 0 & \sin\theta & 0 \\ 0 & 0 & I_{j-i-1} & 0 & 0 \\ 0 & -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 0 & I_{m-j} \end{pmatrix} \in \mathbb{R}^{m,m},$$

kde $I_n \in \mathbb{R}^{n,n}$ je identita.

Vezměme $\mathbf{x} \in \mathbb{R}^m$ a proveďme Givensovu rotaci toho vektoru, tedy $\tilde{\mathbf{x}} = G_{i,j,\theta}\mathbf{x}$. Transformace je dána vztahy

$$\begin{aligned} \tilde{x}_i &= x_i \cos\theta + x_j \sin\theta \\ \tilde{x}_j &= -x_i \sin\theta + x_j \cos\theta \\ \tilde{x}_l &= x_l, \text{ pro } l \neq i, j \end{aligned} \quad .$$

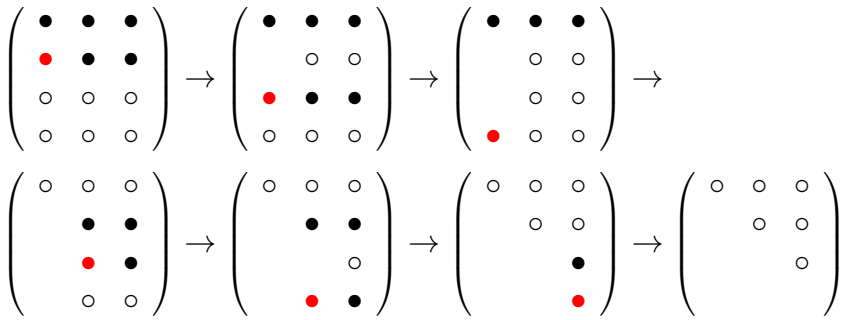
Chceme-li tedy vynulovat $\tilde{x}_j$, je nutno úhel θ splňovat

$$\tan\theta = \frac{x_j}{x_i}.$$

Ortogonální matice Q z QR rozkladu matice A je tedy získána postupným složením matic Givensových transformací:

$$\begin{aligned} A &= Q \cdot \begin{pmatrix} R \\ 0 \end{pmatrix} \\ \begin{pmatrix} R \\ 0 \end{pmatrix} &= \underbrace{G_{i_k,j_k,\theta_k} \dots G_{i_2,j_2,\theta_2} G_{i_1,j_1,\theta_1}}_{Q^T} A. \end{aligned}$$

Schematické znázornění postupného nulování prvků při QR rozkladu je znázorněno na následujícím příkladu.



Při výpočtu hodnot $\cos\theta$ je nutno provést odmocninu, což je numericky nežádoucí proces. V [**?**] lze najít algoritmus, který nevyžaduje výpočet odmocniny u Givensovy rotace.

A.2.3 Singulární rozklad matice *Singular value decomposition*

Věta A.3: Symetrická matice $\mathbf{B} \in \mathbb{R}^{n,n}$ je diagonalizovatelná, tedy existuje matice $\mathbf{V} \in \mathbb{R}^{n,n}$ taková, že

$$\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1},$$

kde $\mathbf{D} \in \mathbb{R}^{n,n}$ je diagonální a na její diagonále jsou vlastní čísla matice $\mathbf{B}$ (včetně těch nulových). Matici V lze navíc volit ortogonální, tedy celkem

$$\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^{T}.$$

Věta A.4: Mějme $\mathbf{A} \in \mathbb{R}^{n,m}$. Existují matice $\mathbf{U} \in \mathbb{R}^{n,n}$, $\mathbf{D} \in \mathbb{R}^{n,m}$ a $\mathbf{V} \in \mathbb{R}^{m,m}$ takové, že

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{T} \tag{A.1}$$

a $\mathbf{U}$ a $\mathbf{V}$ jsou ortogonální a matice $\mathbf{D}$ je diagonální s nezápornými prvky.

Nečtvercovou diagonální maticí $\mathbf{D}$ myslíme takovou matici, pro níž platí $i \neq j \Rightarrow (\mathbf{D})_{i,j} = 0$.

Důkaz. Matice $\mathbf{A}^T\mathbf{A} \in \mathbb{R}^{m,m}$ je symetrická. Nechť r značí její hodnost (platí $r \leq m$ i $r \leq n$). Dle věty A.3 existuje r jejích nenulových vlastních čísel $\lambda_1, \lambda_2, \ldots, \lambda_r$ a příslušné (normované) vlastní vektory $v_1, v_2, \ldots, v_r$ tvořící ortonormální soubor (příslušné sloupce matice $\mathbf{V}$ z věty A.3). Máme

$$\|\mathbf{A}v_i\|^2 = v_i^T\mathbf{A}^T\mathbf{A}v_i = v_i^T\lambda_i v_i = \lambda_i\|v_i\|^2 = \lambda_i.$$

Jelikož je matice $\mathbf{A}^T\mathbf{A}$ pozitivně semidefinitní, její vlastní čísla jsou nezáporná, a můžeme tedy definovat $\sigma_i = \|\mathbf{A}v_i\| = \sqrt{\lambda_i}$. Definujme dále $u_i = \frac{\mathbf{A}v_i}{\sigma_i}$. Platí

$$u_i^T u_j = \frac{v_i^T\mathbf{A}^T\mathbf{A}v_j}{\sigma_i\sigma_j} = \frac{\lambda_j v_i^T v_j}{\sigma_i\sigma_j}$$

a tedy soubor $u_1, u_2, \ldots, u_r$ je ortonormální. Přepišme rovnosti $u_i^T A v_j = v_i^T v_j \frac{\lambda_j}{\sigma_i}$ do maticového tvaru:

$$\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_r^T \end{pmatrix} \mathbf{A} \begin{pmatrix} v_1 v_2 \cdots v_r \end{pmatrix} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r). \tag{A.2}$$

Vektory u_i jsou dimenze n, můžeme tedy najít vektory $u_{r+1}, \ldots, u_n$ tak, aby soubor $u_1, \ldots, u_n$ byl ortonormální. Podobně doplníme i soubor vektorů v_i na ortonormální soubor $v_1, \ldots, v_m$. Maticově dostaneme

$$\underbrace{\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{pmatrix}}_{\in\mathbb{R}^{n,n}} \mathbf{A} \underbrace{\begin{pmatrix} v_1 v_2 \cdots v_m \end{pmatrix}}_{\in\mathbb{R}^{m,m}} = \mathbf{D},$$

kde matice $\mathbf{D}$ je dimenze $n \times m$ mající v levém horním rohu matici $\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ a jinde nuly. Výsledný rozklad získáme položením

$$U = \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{pmatrix} \quad \text{a} \quad V = \begin{pmatrix} v_1 v_2 \cdots v_m \end{pmatrix}. \qquad \square$$

Všimněme si, že ve výsledné maticové rovnosti (A.1) je mnoho nul. Rovnice (A.2) v důkazu je někdy ta, která se hodí více a často se na ni odkazuje pojmem „redukované SVD“.

Nenulovým prvkům matice $\mathbf{D}$, označeným v důkazu σ_i, se říká *singulární hodnoty*. Velmi často se předpokládá, že jsou již seřazeny: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \ldots$.

Příloha B

Pravděpodobnost a matematická statistika

V této kapitole připomeneme a rozšíříme některé známé pojmy teorie pravděpodobnosti a matematické statistiky. Zájemce o pokročilejší analýzu dané problematiky odkazujeme především na knihy [**?**] a [**?**], ze kterých je většina částí tohoto dodatku volně převzata. Jako referenční zdroje k pojmům základní teorie pravděpodobnosti potom doporučujeme knihy [**?**, **?**].

B.1 Základní pojmy

Uvažujme nějaký experiment (pokus, děj), jehož výsledky nejsme schopni predikovat na základě informací, které máme k dispozici. V takovém případě říkáme, že jsou výsledky náhodné. Při matematickém popisu takového experimentu rozlišujeme možné různé výsledky a nazýváme je *elementárními jevy*. Množinu všech elementárních jevů značíme Ω a nazýváme ji *prostor elementárních jevů* nebo také *výběrový prostor*. Při matematické formulaci tvrzení o výsledcích experimentu se obvykle zajímáme o různé podmnožiny prostoru elementárních jevů. Ukazuje se, že je velmi vhodné požadovat, aby tyto „zajímavé“ množiny tvořili tak zvanou σ-algebru. Mějme tedy nějakou sigma algebru $\mathcal{F}$ podmnožin Ω. Její prvky nazýváme *náhodné jevy*. Náhodným jevům potom přiřazujeme pravděpodobnost pomocí nějaké pravděpodobnostní míry $\mathbb{P}$. Celkově se trojice $(\Omega, \mathcal{F}, \mathbb{P})$ nazývá *pravděpodobnostní prostor*.

Náhodná veličina X je zobrazení z Ω do $\mathbb{R}$, které splňuje podmínku měřitelnosti, tj. pro každé $x \in \mathbb{R}$ platí $\{X \leq x\} = X^{-1}((-\infty, x]) \in \mathcal{F}$. Rozdělení náhodné veličiny je potom určeno distribuční funkcí $F_X : \mathbb{R} \to \mathbb{R}$ definovanou pro každé $x \in \mathbb{R}$ vztahem

$$F_X(x) = \mathbb{P}(X \leq x).$$

Tato funkce je neklesající zprava spojitá s limitami 0 v $-\infty$ a 1 v $+\infty$. Podle dodatečných vlastností distribuční funkce F_X rozlišujeme diskrétní a spojité náhodné veličiny.

- *Diskrétní náhodná veličina* X má distribuční funkci F_X, která je po částech konstantní. F_X se tedy skládá z nejvýše spočetně mnoha skoků v bodech $x_1, x_2, \ldots$. Přitom platí, že velikost skoku F_X v bodě x_i určuje pravděpodobnost $\mathbb{P}(X = x_i)$,

tj.

$$\mathbb{P}(X = x_i) = F_X(x_i) - \lim_{x \to x_{i-}} F_X(x)$$

pro každé přípustné i.

- *Spojitá náhodná veličina* X má distribuční funkci F_X, kterou lze vyjádřit integrálem

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

pro všechna $x \in \mathbb{R}$, kde f_X je nezáporná funkce nazývaná *hustota pravděpodobnosti* náhodné veličiny X.

Je-li $B \subset \mathbb{R}$ Borelovská množina[1], tak pravděpodobnost, že hodnota veličiny X bude v množině B, můžeme v případě diskrétní veličiny resp. spojité veličiny s hustotou f_X spočíst jako

$$\mathbb{P}(X \in B) = \sum_i \mathbb{P}(X = x_i) \quad \text{resp.} \quad \mathbb{P}(X \in B) = \int_B f_X(x)\,\mathrm{d}x.$$

Střední hodnotu náhodné veličiny X, která je diskrétní resp. spojitá s hustotou f_X, definujeme vztahy

$$\mathbb{E}\,X = \sum_i x_i\,\mathbb{P}(X = x_i) \quad \text{resp.} \quad \mathbb{E}\,X = \int_{\mathbb{R}} x f_X\,\mathrm{d}x,$$

pokud daná suma resp. integrál existují. Je-li $g : \mathbb{R} \to \mathbb{R}$ měřitelná funkce, pak $g(X)$ je nová náhodná veličina. Její střední hodnotu můžeme spočítat přímo ze znalosti rozdělení náhodné veličiny X. Čili pro diskrétní resp. spojitou veličinu dostáváme

$$\mathbb{E}\,g(X) = \sum_i g(x_i)\,\mathbb{P}(X = x_i) \quad \text{resp.} \quad \mathbb{E}\,g(X) = \int_{\mathbb{R}} g(x) f_X\,\mathrm{d}x,$$

pokud daná suma resp. integrál existují. Uvažujme $k \in \mathbb{N}$. Hodnotu $\mathbb{E}\,X^k$ nazýváme *k-tý obecný moment* náhodné veličiny X a hodnotu $\mathbb{E}(X - \mathbb{E}\,X)^k$ potom *k-tý centrální moment* X. *Rozptyl* náhodné veličiny X je definován jako $\operatorname{var} X = \mathbb{E}(X - \mathbb{E}\,X)^2$, tj. jedná se o druhý centrální moment X.

B.2 Příklady jednorozměrných spojitých rozdělení

V následujících příkladech se vyskytuje tzv. *gamma funkce* Γ, která je pro každé $a > 0$ definována vztahem

$$\Gamma(a) = \int_0^{+\infty} x^{a-1}\mathrm{e}^{-x}\,\mathrm{d}x.$$

Mezi její základní vlastnosti patří

$$\Gamma(a+1) = a\Gamma(a), \quad \Gamma(1/2) = \sqrt{\pi} \quad \text{a} \quad \Gamma(n) = (n-1)!$$

pro každé $a > 0$ a $n \in \mathbb{N}$.

[1] Borelovské množiny v $\mathbb{R}$ tvoří nejmenší σ-algebru, která obsahuje všechny otevřené intervaly v $\mathbb{R}$. Detaily viz např. [**?**, Section 1.1]. Poznamenejme, že Borelovské množiny jsou v podstatě všechny „představitelné“ množiny v $\mathbb{R}$ a tedy z praktického pohledu nepředstavuje toto omezení žádný problém.

B.2.1 Normální rozdělení

Náhodná veličina X má *normální rozdělení* s parametry $\mu \in \mathbb{R}$ a $\sigma^2 > 0$, značíme $X \sim \mathrm{N}(\mu, \sigma^2)$, jestliže má spojité rozdělení s hustotu pravděpodobnosti

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

pro každé $x \in \mathbb{R}$. Parametry μ a σ^2 udávají přímo střední hodnotu a rozptyl X, tj.

$$\mathbb{E}\, X = \mu \quad \text{a} \quad \operatorname{var} X = \sigma^2.$$

Normální rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$, tj. $\mathrm{N}(0, 1)$, nazýváme *standardní normální rozdělení*. Pro úplnost definujme také normální rozdělení $N(\mu, 0)$ s parametry μ a $\sigma^2 = 0$ jako diskrétní rozdělení při kterém náhodná veličina $X \sim \mathrm{N}(\mu, 0)$ nabývá s pravděpodobností 1 hodnoty μ. Opět zjevně platí $\mathbb{E}\, X = \mu$ a $\operatorname{var} X = 0$.

Následující věta ukazuje, že normální rozdělení je invariantní vůči lineárním transformacím.

Věta B.1: Nechť $a, b \in \mathbb{R}$ a $X \sim \mathrm{N}(\mu, \sigma^2)$. Potom $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

Důkaz. Označme $Y = a + bX$. Pro $b = 0$ dostaneme $Y = a$ a tedy $Y \sim N(a, 0)$. Předpokládejme tedy $b > 0$ a určeme distribuční funkci Y:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(a + bX \le y) = \mathbb{P}\left(X \le \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

pro každé $y \in \mathbb{R}$, kde F_X je distribuční funkce náhodné veličiny X. Jelikož hustota pravděpodobnosti je derivací distribuční funkce, pomocí derivace složené funkce dostáváme

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} = \frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}\left(\frac{y-a}{b}\right)} \frac{\mathrm{d}\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \\ &= \frac{1}{b} f_X\left(\frac{y-a}{b}\right) = \frac{1}{\sqrt{2\pi}b\sigma} \mathrm{e}^{-\frac{1}{2\sigma^2}\left(\frac{y-a}{b}-\mu\right)^2} \\ &= \frac{1}{\sqrt{2\pi}(b\sigma)} \mathrm{e}^{-\frac{1}{2(b\sigma)^2}(y-(a+b\mu))^2}. \end{aligned}$$

To ovšem znamená $Y \sim \mathrm{N}\big(a + b\mu, (b\sigma)^2\big)$. Případ $b < 0$ se dokáže zcela analogicky. □

B.2.2 Rozdělení χ^2

Nezáporná náhodná veličina X má rozdělení χ^2 („chí kvadrát“) s $n \in \mathbb{N}$ stupni volnosti, píšeme $X \sim \chi^2_n$, jestliže má spojité rozdělení s hustotou

$$f_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} \mathrm{e}^{-\frac{x}{2}}$$

pro $x > 0$ a $f_X(x) = 0$ pro $x \le 0$. Lze ukázat, že

$$\mathbb{E}\, X = n \quad \text{a} \quad \operatorname{var} X = 2n.$$

χ^2 rozdělení se nejčastěji objevuje jako součet kvadrátů standardních normálních rozdělení ve smyslu následující věty, jejíž důkaz je uveden např. v [**?**].

Věta B.2 ([?, Věta 4.13])**:** Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny se standardním normálním rozdělením $\mathrm{N}(0,1)$. Pak náhodná veličina $Y = X_1^2 + \ldots + X_n^2$ má rozdělení χ^2_n.

B.2.3 Studentovo rozdělení

Buď $n \in \mathbb{N}$. Náhodná veličina X má *Studentovo t rozdělení* o n stupních volnosti, píšeme $x \sim \mathrm{t}_n$, jestliže má spojité rozdělení s hustotou

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

pro každé $x \in \mathbb{R}$. Střední hodnota je definována pro $n > 1$ a platí $\mathbb{E}\, X = 0$. Rozptyl je definován a konečný pro $n > 2$, přičemž $\operatorname{var} X = n/(n-2)$.

Studentovo rozdělení se často objevuje jako podíl standardního normálního rozdělení a χ^2 rozdělení.

Věta B.3 ([?, Věta 4.22])**:** Nechť $X \sim \mathrm{N}(0,1)$ a $Z \sim \chi^2_n$ jsou nezávislé náhodné veličiny. Pak náhodná veličina

$$T = \frac{X}{\sqrt{Z/n}}$$

má Studentovo t rozdělení s n stupni volnosti, tj. $T \sim \mathrm{t}_n$.

B.2.4 Rozdělení F

Buďte $m, n \in \mathbb{N}$. Náhodná veličina X má F *rozdělení* (Fisherovo-Snedecorovo) s m a n stupni volnosti, píšeme $X \sim \mathrm{F}_{m,n}$, jestliže má spojité rozdělení s hustotou

$$f_X(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{m/2} x^{\frac{m}{2}-1}\left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}$$

pro $x > 0$ a $f_X(x) = 0$ pro $x \leq 0$. Je-li $n > 2$, existuje střední hodnota $\mathbb{E}\, X = n/(n-2)$ a je-li $n > 4$, existuje konečný rozptyl

$$\operatorname{var} X = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}.$$

F rozdělení se často objevuje jako podíl dvou χ^2 rozdělení.

Věta B.4 ([?, Věta 4.28])**:** Nechť $X \sim \chi^2_m$ a $Y \sim \chi^2_n$ jsou nezávislé náhodné veličiny. Pak náhodná veličina

$$Z = \frac{X/m}{Y/n}$$

má F rozdělení s m a n stupni volnosti, tj. $Z \sim \mathrm{F}_{m,n}$.

B.3 Náhodné vektory a jejich charakteristiky

Nejprve se budeme zabývat rozšířením pojmů střední hodnoty a variance na vícerozměrné situace. Pro $n \in \mathbb{N}$ uvažujme náhodné veličiny $X_1, \dots, X_n$ na stejném pravděpodobnostním prostoru[2]. Vektor $\boldsymbol{X} = (X_1, \dots, X_n)^T$ potom nazýváme *náhodným vektorem*. Obdobně, jsou-li $m, n \in \mathbb{N}$ a $Z_{i,j}$ pro $i = 1, \dots, m; j = 1, \dots, n$ náhodné veličiny na stejném pravděpodobnostním prostoru, pak matici $\mathbf{Z} = (Z_{i,j})$ nazýváme *náhodnou maticí*. Náhodný vektor je tak vlastně náhodnou maticí o rozměrech $n \times 1$.

Pravděpodobnostní rozdělení vektoru $\boldsymbol{X}$ popisujeme pomocí sdružené distribuční funkce $F_{\boldsymbol{X}} : \mathbb{R}^n \to \mathbb{R}$ definované vztahem

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n)$$

pro každé $\boldsymbol{x} \in \mathbb{R}^n$, kde $\boldsymbol{x} = (x_1, \dots, x_n)^T$. Říkáme, že náhodný vektor $\boldsymbol{X}$ má sdružené spojité rozdělení s hustotou pravděpodobnosti $f_{\boldsymbol{X}}$, jestliže platí

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\boldsymbol{X}}(\boldsymbol{u}) \, \mathrm{d}u_1 \dots \mathrm{d}u_n$$

pro každé $\boldsymbol{x} \in \mathbb{R}^n$. Analogicky můžeme definovat předchozí pojmy i pro náhodné matice.

Existují-li střední hodnoty veličin $X_1, \dots, X_n$, pak vektor $\mathbb{E}\,\boldsymbol{X} = (\mathbb{E}\,X_1, \dots, \mathbb{E}\,X_n)^T \in \mathbb{R}^n$ nazýváme *střední hodnotou vektoru* $\boldsymbol{X}$. Existují-li střední hodnoty veličin $Z_{i,j}$ pro všechny $i = 1, \dots, m$ a $j = 1, \dots, n$, pak matici $\mathbb{E}\,\mathbf{Z} = (\mathbb{E}\,Z_{i,j}) \in \mathbb{R}^{m,n}$ nazýváme *střední hodnotou matice* $\mathbf{Z}$. Střední hodnotu tedy definujeme po složkách.

Věta B.5: Nechť $\mathbf{A} \in \mathbb{R}^{p,q}$, $\mathbf{B} \in \mathbb{R}^{p,m}$ a $\mathbf{C} \in \mathbb{R}^{n,q}$ jsou matice. Potom

$$\mathbb{E}(\mathbf{A} + \mathbf{BZC}) = \mathbf{A} + \mathbf{B}(\mathbb{E}\,\mathbf{Z})\mathbf{C}.$$

Důkaz. Označme $\mathbf{W} = \mathbf{A} + \mathbf{BZC}$. Potom platí $W_{i,j} = A_{i,j} + \sum_{k=1}^{m}\sum_{\ell=1}^{n} B_{i,k}Z_{k,\ell}C_{\ell,j}$ pro každé $i = 1, \dots, p$ a $j = 1, \dots, q$. Z linearity střední hodnoty plyne

$$\mathbb{E}\,W_{i,j} = \mathbb{E}\,A_{i,j} + \mathbb{E}\sum_{r=1}^{m}\sum_{s=1}^{\ell} B_{i,r}Z_{r,s}C_{s,j} = A_{i,j} + \sum_{k=1}^{m}\sum_{\ell=1}^{n} B_{i,k}(\mathbb{E}\,Z_{k,\ell})C_{\ell,j}.$$

Celkem tedy $\mathbb{E}\,\mathbf{W} = (\mathbb{E}\,W_{i,j}) = (A_{i,j}) + (\sum_{k=1}^{m}\sum_{\ell=1}^{n} B_{i,k}(\mathbb{E}\,Z_{k,\ell})C_{\ell,j}) = \mathbf{A} + \mathbf{B}(\mathbb{E}\,\mathbf{Z})\mathbf{C}$. □

Důsledek B.6: Nechť $\boldsymbol{a} \in \mathbb{R}^p$ je vektor a $\mathbf{B} \in \mathbb{R}^{p,n}$ je matice. Potom

$$\mathbb{E}(\boldsymbol{a} + \mathbf{B}\boldsymbol{X}) = \boldsymbol{a} + \mathbf{B}\,\mathbb{E}\,\boldsymbol{X}.$$

Nyní se zabývejme charakteristikami druhého řádu. Je-li $\mathbb{E}\,X_i^2 < \infty$ pro každé $i = 1, \dots, n$, říkáme, že $\boldsymbol{X}$ má konečné druhé momenty a definujeme *varianční matici* vektoru $\boldsymbol{X}$ po složkách jako

$$(\operatorname{var}\boldsymbol{X})_{i,j} = \operatorname{cov}(X_i, X_j)$$

pro každé $i, j = 1, \dots, n$. Připomeňme, že $\operatorname{cov}(X_i, X_j)$ značí kovarianci veličin X_i a X_j, která je definována vztahem

$$\operatorname{cov}(X_i, X_j) = \mathbb{E}(X_i - \mathbb{E}\,X_i)(X_j - \mathbb{E}\,X_j) = \mathbb{E}\,X_iX_j - \mathbb{E}\,X_i\,\mathbb{E}\,X_j.$$

[2]

Pro $i = j$ dostáváme $\operatorname{cov}(X_i, X_i) = \operatorname{var} X_i$ a tudíž se na diagonále kovarianční matice $\operatorname{var} \boldsymbol{X}$ nacházejí rozptyly veličin X_i.

Je zřejmé, že varianční matici můžeme vektorově zapsat jako

$$\operatorname{var} \boldsymbol{X} = \mathbb{E}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})^T.$$

Použijeme-li nyní vlastnosti střední hodnoty z věty B.5 dostaneme

$$\begin{aligned}\mathbb{E}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})^T &= \mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}^T - (\mathbb{E}\,\boldsymbol{X})\boldsymbol{X}^T - \boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T + \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T\right)\\ &= \mathbb{E}\,\boldsymbol{X}\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T + \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T\\ &= \mathbb{E}\,\boldsymbol{X}\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T.\end{aligned}$$

Celkově jsme tedy dokázali následující tvrzení.

Věta B.7: Varianční matice splňuje

$$\operatorname{var} \boldsymbol{X} = \mathbb{E}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})^T = \mathbb{E}\,\boldsymbol{X}\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T.$$

Další důležitou vlastností varianční matice je vztah k lineárním transformacím. Výsledek je hodně podobný situaci pro rozptyl lineární kombinace jednorozměrné náhodné veličiny X, kde platí $\operatorname{var}(a + bX) = b^2 \operatorname{var} X$.

Věta B.8: Buď $\boldsymbol{a} \in \mathbb{R}^p$ vektor a $\mathbf{B} \in \mathbb{R}^{p,n}$ matice. Potom

$$\operatorname{var}(\boldsymbol{a} + \mathbf{B}\boldsymbol{X}) = \mathbf{B}(\operatorname{var} \boldsymbol{X})\mathbf{B}^T.$$

Důkaz. Postupným užitím předchozích vět dostaneme:

$$\begin{aligned}\operatorname{var}(\boldsymbol{a} + \mathbf{B}\boldsymbol{X}) &= \mathbb{E}\left(\boldsymbol{a} + \mathbf{B}\boldsymbol{X} - \mathbb{E}(\boldsymbol{a} + \mathbf{B}\boldsymbol{X})\right)\left(\boldsymbol{a} + \mathbf{B}\boldsymbol{X} - \mathbb{E}(\boldsymbol{a} + \mathbf{B}\boldsymbol{X})\right)^T\\ &= \mathbb{E}\left(\boldsymbol{a} + \mathbf{B}\boldsymbol{X} - \boldsymbol{a} - \mathbb{E}\,\mathbf{B}\boldsymbol{X}\right)\left(\boldsymbol{a} + \mathbf{B}\boldsymbol{X} - \boldsymbol{a} - \mathbb{E}\,\mathbf{B}\boldsymbol{X}\right)^T\\ &= \mathbb{E}(\mathbf{B} - \mathbb{E}\,\mathbf{B}\boldsymbol{X})(\mathbf{B} - \mathbb{E}\,\mathbf{B}\boldsymbol{X})^T = \operatorname{var}(\mathbf{B}\boldsymbol{X})\\ &= \mathbb{E}\,\mathbf{B}\boldsymbol{X}(\mathbf{B}\boldsymbol{X})^T - \mathbb{E}(\mathbf{B}\boldsymbol{X})\,\mathbb{E}(\mathbf{B}\boldsymbol{X})^T\\ &= \mathbb{E}\,\mathbf{B}\boldsymbol{X}\boldsymbol{X}^T\mathbf{B}^T - \mathbf{B}\,\mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T\mathbf{B}^T = \mathbb{E}\,\mathbf{B}(\boldsymbol{X}\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T)\mathbf{B}^T\\ &= \mathbf{B}(\mathbb{E}\,\boldsymbol{X}\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{X}^T)\mathbf{B}^T = \mathbf{B}(\operatorname{var}\boldsymbol{X})\mathbf{B}^T.\end{aligned}$$

□

Velmi důležitou vlastností varianční matice je její pozitivní semidefinita (PSD). Pod tímto pojmem nyní budeme chápat implikaci, že pro každé $\boldsymbol{c} \in \mathbb{R}^n$ platí $\boldsymbol{c}^T \operatorname{var}(\boldsymbol{X})\boldsymbol{c} \geq 0$. Tj. platí, že jakmile je varianční matice pozitivně definitní, je také pozitivně semidefinitní.

Věta B.9: Varianční matice $\operatorname{var} \boldsymbol{X}$ náhodného vektoru $\boldsymbol{X}$ je symetrická a pozitivně semidefinitní. Pokud navíc žádná složka $\boldsymbol{X}$ není afinní kombinací ostatních složek (tj. neexistuje $\boldsymbol{a} \in \mathbb{R}^n$, tak že $\boldsymbol{a}^T\boldsymbol{X} = c$ pro nějaké $c \in \mathbb{R}$ s pravděpodobností 1), tak je $\operatorname{var} \boldsymbol{X}$ pozitivně definitní.

Důkaz. Symetrie je zjevná z definice: $(\operatorname{var} \boldsymbol{X})_{i,j} = \operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i) = (\operatorname{var} \boldsymbol{X})_{j,i}$. Buď $\boldsymbol{a} \in \mathbb{R}^n$. Potom:

$$\boldsymbol{a}^T \operatorname{var} \boldsymbol{X}\boldsymbol{a} = \boldsymbol{a}^T\,\mathbb{E}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})^T\boldsymbol{a} = \mathbb{E}\left(\boldsymbol{a}^T(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})^T\boldsymbol{a}\right).$$

Označíme-li $Y = \boldsymbol{a}^T(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})$ jakožto novou náhodnou veličinu vzniklou lineární kombinací složek $\boldsymbol{X}$, dostaneme

$$\boldsymbol{a}^T \operatorname{var} \boldsymbol{X}\boldsymbol{a} = \mathbb{E}\, YY^T = \mathbb{E}\, Y^2 \geq 0.$$

Navíc je zřejmé, že nulovost tohoto výrazu je ekvivalentní nulovosti veličiny Y s pravděpodobností 1, tj. $\mathbb{E}\, Y^2 = 0$ je ekvivalentní $\mathbb{P}(Y = 0) = 1$. To je ale možné pouze tehdy, pokud $\mathbb{P}(\boldsymbol{a}^T(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) = 0) = 1$, což znamená $\boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{a}^T\,\mathbb{E}\,\boldsymbol{X} = c$ s pravděpodobností 1. □

B.4 Vícerozměrné normální rozdělení

Nyní si zaveďme rozšíření normálního rozdělení na náhodné vektory.

Definice B.10: Buď $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ náhodný vektor, $\boldsymbol{\mu} \in \mathbb{R}^n$ vektor a $\mathbf{V} \in \mathbb{R}^{n,n}$ symetrická PSD matice. Říkáme, že $\boldsymbol{X}$ má n-rozměrné normální rozdělení s parametry $\boldsymbol{\mu}$ a $\mathbf{V}$, píšeme $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, právě když pro každé $\boldsymbol{c} \in \mathbb{R}^n$ platí $\boldsymbol{c}^T\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{c}^T\boldsymbol{\mu}, \boldsymbol{c}^T\mathbf{V}\boldsymbol{c})$.

Věta B.11 ([**?**, Věta 4.10])**:** Buď $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ náhodný vektor, $\boldsymbol{\mu} \in \mathbb{R}^n$ vektor a $\mathbf{V} \in \mathbb{R}^{n,n}$ symetrická PD matice. Potom $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$ právě tehdy, když $\boldsymbol{X}$ má spojité sdružené rozdělení s hustotou

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\,|\mathbf{V}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Věta B.12: Buď $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$. Potom $\mathbb{E}\,\boldsymbol{X} = \boldsymbol{\mu}$ a $\operatorname{var}\boldsymbol{X} = \mathbf{V}$.

Důkaz. TBA □

Důsledek B.13: Buď $\boldsymbol{X} = (X_1, \ldots, X_n)^T \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$. $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny právě tehdy, když je matice $\mathbf{V}$ diagonální.

Důkaz. TBA □

Důsledek B.14: Buď $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2), \ldots, X_n \sim \mathrm{N}(\mu_n, \sigma_n^2)$ nezávislé náhodné veličiny. Potom $X_1 + \ldots + X_n \sim \mathrm{N}(\mu_1 + \ldots + \mu_n, \sigma_1^2 + \ldots + \sigma_n^2)$.

Důkaz. TBA □

Věta B.15: Buď $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$ a $\boldsymbol{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{m,n}$. Potom $\boldsymbol{Y} = \boldsymbol{a} + \mathbf{B}\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}^T)$.

Důkaz. TBA □

Z vícerozměrného normálního rozdělení vznikají různými transformacemi rozdělení z části B.2.

Věta B.16: Buď $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$ a nechť $\mathbf{A} \in \mathbb{R}^{n,n}$ je symetrická PSD. Je-li $\mathbf{A}\mathbf{V} \neq \mathbf{0}$ a idempotentní, pak $\boldsymbol{X}^T\mathbf{A}\boldsymbol{X}$ má rozdělení χ^2_k, kde $k = \operatorname{Tr}\mathbf{A}\mathbf{V}$.

Důkaz. TBA □

Příloha C

Vázané extrémy

C.1 Nevázané extrémy

Nejprve si připopmeňme situaci, kdy hledáme extrém funkce $f : \mathbb{R}^n \to \mathbb{R}$.

C.2 Nerovnostní vazby

Uvažujme situaci, kdy chceme

C.2.1 Primární přístup

C.2.2 Duální přístup

Řešení vybraných cvičení

Řešení cvičení 2.1: DFT ;-) TBA

Řešení cvičení 4.2: Nechť $\boldsymbol{x} \in \mathbb{R}^{p+1,1}$ je nenulový. Spočtěme

$$\begin{aligned}\boldsymbol{x}^T \left(\mathbf{X}_c'^T \mathbf{X}_c' + \lambda \mathbf{I}'\right) \boldsymbol{x} &= \boldsymbol{x}^T \mathbf{X}_c'^T \mathbf{X}_c' \boldsymbol{x} + \lambda \boldsymbol{x}^T I' \boldsymbol{x} \\ &= \|\mathbf{X}_c' \boldsymbol{x}\| + \lambda \sum_{i=1}^{p} (\boldsymbol{x})_i^2 .\end{aligned}$$

Tedy pokud existuje $i > 0$ takové, že $(\boldsymbol{x})_i \neq 0$, máme $\boldsymbol{x}^T \left(\mathbf{X}_c'^T \mathbf{X}_c' + \lambda \mathbf{I}'\right) \boldsymbol{x} > 0$. Pokud tento případ nenastane, pak nutně $(\boldsymbol{x})_0 \neq 0$ a z rovnosti

$$\mathbf{X}_c'^T \mathbf{X}_c' = \begin{pmatrix} N & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{X}_c^T \mathbf{X}_c & \\ 0 & & & \end{pmatrix}$$

dostaneme $\boldsymbol{x}^T \left(\mathbf{X}_c'^T \mathbf{X}_c' + \lambda \mathbf{I}'\right) \boldsymbol{x} \geq N (\boldsymbol{x})_0^2 > 0$.

Řešení cvičení 5.1: TBA

Řešení cvičení 5.2: TBA

Seznam použitých zkratek

RSS Residual sum of squares

RFE Recursive feature elimination

LAR Least angle regression

PCR Principal component regression

OLS Ordinary least squares

SVD Singular value decomposition

projít položky a sjednotit co je první a co je druhé

Rejstřík českých termínů

Rejstřík anglických termínů

Literatura

[1] J. Anděl, *Základy matematické statistiky*, matfyzpress, Praha, 2007.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] G. M. Furnival and R. W. Wilson, Jr., *Regressions by leaps and bounds*, Technometrics, 16 (1974), pp. 499–511.

[4] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Oxford University Press, 3rd ed., 2001.

[5] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.

[6] A. Klenke, *Probability Theory*, Springer, 2nd ed., 2014.

[7] X. S. Ni and X. Huo, *Enhanced leaps-and-bounds methods in subset selections with additional optimality tests*. https://www.informs.org/content/download/55245/522655/file/enhanced%20leaps%20and%20bounds.pdf, 2005.

[8] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, John Wiley & Sons, Inc., 2 ed., 2003.

[9] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.