FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání diplomové práce

Název:	Bayesovské odhadování emise mikroplastů vzdušnou cestou
Student:	Bc. Ondřej Chládek
Vedoucí:	Ing. Ondřej Tichý, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

Šíření mikroplastů v životním prostředí se stává stále více sledovaným jevem, zároveň rostou i možnosti jeho měření a modelování. Většina dosavadních odhadů šíření mikroplastů vzdušnou cestou je založena na tzv. bottom-up přístupu, kde je na základě odhadu emise z mikro-jevu (např. emise automobilu, emise v zemědělství) extrapolován odhad emise pro globální doménu. Cílem práce je využít tzv. top-down přístupu, kde na základě měření (depozice nebo koncentrace) a výpočtu atmosférického modelu můžeme odhadnout časovou a prostorovou distribuci úniku mikroplastů na dané prostorové doméně. Toho lze dosáhnout optimalizací mezi naměřenými hodnotami a mezi numerickými výsledky atmosférického modelu, což lze formulovat jako velmi špatně podmíněnou lineární úlohu. Student se seznámí především s bayesovským přístupem a metodami odhadu parametrů pravděpodobnostních modelů studovaného problému. Odvozený odhadovací algoritmus bude demonstrován na datech z dostupných měření v národních parcích USA.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 25. listopadu 2021 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Diplomová práce

Bayesovské odhadování emise mikroplastů vzdušnou cestou

Bc. Ondřej Chládek

Katedra Aplikované Matematiky
Vedoucí práce: Ing. Ondřej Tichý, Ph.D.

22. června 2022

Poděkování

Chtěl bych poděkovat mému vedoucímu Ing. Ondřeji Tichému Ph.D. za jeho vedení a čas, který věnoval mně při této práci. Bez jeho rad a návrhů bych tuto práci nebyl schopen napsat. Také bych rád poděkoval své přítelkyni za podporu, kterou mi při psaní byla. Ať už jako povzbuzení, nebo i korekturu, pro kterou ve svém harmonogramu našla čas a pomohla mi práci dostat do takového stavu, v jakém je teď. Díky

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principu při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisu, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 citovaného zákona.

V Praze dne 22. června 2022 . . . . . .

Odkaz na tuto práci

Chládek, Ondřej. *Bayesovské odhadování emise mikroplastů vzdušnou cestou.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2022.

Abstrakt

Tato práce je zaměřená na odhad emise mikroplastů po Spojených Státech Amerických. Budeme řešit lineární regresi ve tvaru $y = Mx$, kde za y dosadíme data z měřicích přístrojů umístěných v Národních parcích Spojených Států Amerických. Matice M je tvořena modelem šíření částic v atmosféře. Úkolem je spočítat vektor x, jež reprezentuje odhad emise s použitím apriorního odhadu získaného z literatury. Odhad, který jsem získal výpočtem má 80% korelovanost v Pearsonově korelačním koeficientu a odhaduje 2.20×10^7 částic na metr čtvereční za rok pro data suché depozice.

Klíčová slova mikroplasty, lineární regrese, atmosferický transport částic, variační Bayesův algoritmus

Abstract

This thesis is focused on the estimation of microplastic emission in the United States of America. The linear regression problem $y = Mx$ will be solved, where y is data obtained from measuring stations in national parks in the US. The matrix M is derived from the model of atmospheric particle dispersion. The goal is to obtain vector x, which represents the estimate of microplastic emission, with usage of the estimate from literature as an a priori. The estimate I calculated has an 80 % correlation in the Pearson correlation coefficient and estimates 2.20×10^7 particles per square meter per year for the dry deposition data.

Keywords microplastics, linear regresion, atmospheric particle transport, variational Bayes algorithm

Obsah

Seznam obrázků

Úvod

Plasty jsou dnes jedním z nejčastěji používaných materiálů. Nicméně jejich použití vede i k jejich degradaci na mikroplasty. O mikroplastech slýcháváme poslední dobou stále více. Nevíme jistě, co všechno nám mohou způsobit, ale jejich výskyt se prokazuje na místech jako třeba v placentě [1], nebo i tam, kde se člověk téměř nepohybuje, např. ledovce v Tibetu [2] či na Antarktidě [3]. Na tato místa se tak musí dostávat z jiných míst pomocí vzduchu nebo vody. Nicméně máme zde podobný problém jako s jinými škodlivinami šířící se vzduchem, nejsou vidět. K měření výskytu jsou potřeba přístroje, které nejsou jednoduché na sestavení. Nelze tedy dát do každé ulice jeden a mít měření přímo u zdroje.

Národní parky ve Spojených Státech Amerických jsou pro Američany velkou přírodní památkou. S trochou nadsázky lze říci, že berou vážně každý odpadek vhozený mimo koš a v reakci na studie ohledně výskytu mikroplastů po světě se rozhodli umístit do několika národních parků přístroje na měření výskytu mikroplastů a mikrovláken [4]. Z těchto dat a z veřejně dostupných open-source modelů šíření mikroplastů a mikrovláken vzduchem se budu následně snažit rekonstruovat emisi a její prostorovou distribuci.

V první kapitole si přiblížíme statistickou teorii potřebnou v této práci a ukážeme si jakým způsobem funguje metoda Variačního Bayese. Dozvíme se, jak funguje Bayesovo pravidlo pro aktualizaci našich znalostí, zjistíme, že existují rozdělení, která po aplikaci Bayesova pravidla zůstávají ve stejné formě, jen s jinými tvarovacími parametry. Též si ukážeme jakým způsobem můžeme měřit chybu modelů a jak na základě chyby lze najít lepší model. V neposlední řadě si zadefinujeme algoritmus variačního Bayese (VB) a tento algoritmus si ukážeme na jednoduchém příkladu rozkladu součinu dvou skalárních proměnných.

V druhé kapitole si nejdříve ukážeme problém lineární regrese, který budeme řešit, a jaké jsou limitace dat. Poté si odvodíme nejjednodušší model, který dokáže danou rovnici alespoň na testovacích datech vyřešit. Dále při-

dáme vlastnosti, jako je řídkost a nezápornost, a uvidíme, jakým způsobem to změní náš model.

V třetí kapitole si nejdříve představíme data. Ukážeme si, kde se nachází měřicí přístroje a jak vypadají atmosferická data o přenosu částic. Dále prozkoumáme atlasy odhadů emisí mikroplastů, které jsou k dispozici. Ukážeme si, jakým způsobem se vytvoří apriorní odhad, který využijeme při výpočtu aposteriorních odhadů emisí.

Kapitola 1

Statistická teorie

V dnešním světě potřebujeme často udělat závěr o vybrané skupině jevů. Existuje k tomu nástroj zvaný statistické modelování. Díky této metodě jsme schopni například říci, jestli je lék efektivní, zda nám hospodský naschvál dává pod míru, nebo i zda nám jedna z mnoha dodavatelských firem dlouhodobě nedodává horší součástky, než jejich konkurence. Můžeme tedy řešit jednoduché školské úlohy, jako je-li kostka férová, nebo i náročnější, avšak důležité, jako hledání ztracené vodíkové bomby [5].

1.1 Úvod do statistické teorie

Než se ponoříme do bayesovské statistiky, je potřebné si zadefinovat některé důležité pojmy. V této části budu vycházet především z [6].

Náhodný jev je takový jev z množiny všech možných jevů, kterou typicky značíme Ω, jemuž můžeme přiřadit nějakou pravděpodobnost. Tento jev tedy není deterministický, nebo aspoň nedokážeme zjistit a spočítat všechny parametry, které do tohoto jevu vstupují. Můžeme si dát jako příklad jednoduchý hod kostkou. Pokud bychom se snažili předpovídat, jaké číslo padne na kostce, museli bychom znát směr hodu, rychlost rotace podle všech os, tření, které kosta vytváří a spoustu dalších parametrů. Což je aktuálně velice obtížné.

Pravděpodobnost je zobrazení z prostoru jevů na reálná čísla splňující nezápornost, normalizaci a disjunktní aditivitu. Nezápornost znamená, že žádný jev nemůže mít zápornou pravděpodobnost. Druhé pravidlo normalizace znamená, že pravděpodobnost, že nastane nějaký ze všech možných jevů je 1. Poslední pravidlo mluví o tom, jak pracovat s pravděpodobností, že nastane jeden z vícero jevů. Pokud dané jevy nemají průnik, neboli jsou disjunktní, tak můžeme jejich pravděpodobnosti jednoduše sečíst. Opět příklad na kostce, první pravidlo nám říká, že nemůžeme žádnému číslu přiřadit zápornou pravděpodobnost. Normalizace zaručuje, že pravděpodobnost všech jevů nepřesáhne 1. Disjunktní aditivita zase dává pravidla o tom, jak zjistit pravděpodobnost,

že padne např. sudé číslo. Protože všechny stěny jsou disjunktní, nemůžou padnout zároveň, můžeme jednoduše sečíst pravděpodobnosti pádu stěn se sudými čísly a získáme naši cílenou. Pokud však máme jevy padne sudé číslo nebo padne 2, pak tyto jevy mají průnik a nelze sečíst pravděpodobnost těchto jevů.

Podmíněné pravděpodobnosti budeme využívat v této práci hojně. Máme-li dva jevy s pravděpodobnostmi $P(A)$ a $P(B) > 0$, je pak podmíněná pravděpodobnost $P(A|B)$ definována jako

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Pokud nám po hodu kostkou někdo oznámí, že výsledek je sudé číslo, tak se prostor možných jevů zmenší. To reflektujeme zvýšením pravděpodobnosti pro sudé stěny a nastavením pravděpodobnosti na 0 pro liché stěny.

Náhodná veličina je zobrazení, jež každému jevu přiřadí hodnotu z množiny reálných čísel, toto zobrazení musí však splňovat podmínku měřitelnosti, která vlastně jen říká, že musíme být schopni spočítat $P(X \leq x)$. Pokud však nebudeme používat nějaká podivuhodná zobrazení, nestane se, že bychom tuto podmínku porušili. Náhodnou veličinu si však definujeme právě proto, že chceme umět spočítat $P(X \leq x)$, protože pak můžeme s náhodnými veličinami různě počítat. Dá se třeba zjistit, jaká je očekávaná hodnota, nebo jak moc můžeme čekat jinou hodnotu od té očekávané.

Náhodné veličiny se dělí do dvou kategorií. Pokud je prostor jevů Ω konečný a nebo alespoň spočetně nekonečný, je daný prostor diskrétní. Příkladem budiž naše již oblíbená kostka, má 6 stěn, což je konečný prostor. Příklad spočetně nekonečného prostoru může být třeba hráč šipek v hospodě snažící se trefit terč. U něho nás pak bude zajímat po kolika hodech trefí cíl. Což v závislosti na tom, jak moc se posílil před daným hodem, může být číslo nízké, ale i vysoké. Náš hráč může házet mimo, ale stále se bude pohybovat na množině $\mathbb{N}$, která je spočetná, takže rozdělení je diskrétní. Druhý typ rozdělení je spojité. Sem patří prostor jevů, který není konečný. Příkladem může být samotný hod šipkou na terč, prostor kam všude se může trefit je nekonečný.

Pokud máme velké množství dat, tak nás většinou nezajímají jednotlivé datové body, ale třeba jejich průměr. Podobně bychom si rádi popsali i náhodnou veličinu. Typicky používáme dvě veličiny, *Střední hodnotu* a *Rozptyl*. Střední hodnotu můžeme spočítat následovně:

$$EX = \sum_{x \in \Omega} xp(x) \text{ pro diskrétní náhodné veličiny,}$$
$$EX = \int_{-\infty}^{+\infty} xp(x)dx \text{ pro spojité náhodné veličiny,}$$

kde x je hodnota náhodné veličiny a $p(x)$ pravděpodobnost, že nastane daný jev. Může se stát, že suma nekonverguje nebo integrál není konečný, v tako-

vém případě střední hodnota neexistuje. Střední hodnota nám popisuje průměrnou hodnotu náhodné veličiny. Můžeme to použít např. pro zjištění zda je nějaká pravděpodobnostní hra férová. Např. u rulety, pokud bychom mohli kvůli zjednodušení sázet jen na černou nebo bílou, tak máme 37 pozic, které mohou padnout. 18 z nich nám dá výhru, zbylých 19 je buď jiné barvy, nebo je to nula. Pokud si jevy, jež nám vyplatí odměnu, ohodnotíme 1 a ostatní -1, tak pokud by ruleta byla v náš prospěch, střední hodnota by byla větší než 0. Nicméně tomu tak není, a k naší smůle ruleta není ani férová. Střední hodnota je totiž $\frac{1}{37}$, takže pokud bychom jí hráli dlouhodobě, tak spíš víc ztratíme než vyděláme.

Dalším způsobem, jakým můžeme popsat náhodnou veličinu je rozptyl. Ten nám říká, jak moc můžeme očekávat deviaci od střední hodnoty. Počítá se následovně:

$$varX = E(X - EX)^2 \text{ pro diskrétní náhodné veličiny,}$$
$$varX = \int_{-\infty}^{\infty} (x - EX)^2 dx \text{ pro spojité náhodné veličiny.}$$

Variance existuje za podmínky, že suma v prvním případě konverguje, nebo že integrál je konečný v případě druhém.

V některých případech nám však variance a střední hodnota nebudou stačit. Proto si definujeme *moment*. K-tý moment se dá spočítat jako:

$$EX^k = \sum_{j=1}^{\infty} (x_j)^k p(x_j) \text{ pro diskrétní veličiny,}$$
$$EX^k = \int_{-\infty}^{\infty} f(x) dx \text{ pro spojité veličiny,}$$

kde $f(x)$, resp. $p(x)$ jsou pravděpodobnostní funkce. I pro momenty platí, že suma musí konvergovat, resp. integrál musí být konečný. Můžeme si všimnout podobností se střední hodnotou, střední hodnota je vlastně první moment. Při trošce přemýšlení uvidíme obecný moment i ve varianci, která je vlastně centrovaným druhým momentem. Vyšší momenty v této práci nebudeme potřebovat, nicméně je vhodné znát jak se obecný moment počítá. Pokud budu v práci zmiňovat, že potřebuji vypočítat momenty, tak je tím myšleno střední hodnotu a příp. varianci.

1.2 Bayesovská statistika

Tato sekce bude primárně vycházet z literatury [7]. Pokud máme soubor dat D a chtěli bychom modelovat zdroj těchto dat, rádi bychom věděli, jaké parametry vstupují do procesu generace. Tuto množinu si označíme symbolem θ a při modelování je snaha nalézt takové $\widehat{\theta}$, které jsou co nejvíce podobné skutečnosti. Samozřejmě je velice těžké nalézt všechny parametry, které vstupují

do procesu, proto ten proces nazýváme *náhodný*. On možná náhodný není, ale nemáme dostatek informací nebo výpočetního výkonu, abychom daný proces nazvali deterministický. Tu část, kterou nedokážeme dostatečně dobře modelovat, nazveme šumem. Samotná data D můžou být také zanesená šumem. Ať už to jsou nepřesnosti v návrhu měřicího přístroje, případně při výrobě, nebo třeba kosmické záření, které nám přehodí bit a udělá nám chybné měření [8]. S chybou je tedy vždy potřeba počítat.

Bayesovská statistika se následně snaží odhadnout z dat, nebo spíše z našeho přesvědčení o datech, parametry modelu, který nad daty předpokládáme. Tímto modelem pak modelujeme systém, který generuje datové body D. Budeme se snažit získat $f(\theta|D)$, což je aposteriorní rozdělení, které v sobě váže naše přesvědčení o systému, který generuje D. Nicméně na začátku máme jen apriorní rozdělení $f(\theta)$, ve kterém máme již zmíněné přesvědčení o systému, avšak nijak ovlivněné daty, a model pozorování $f(D|\theta)$. Zde následeně použijeme bayesovu větu, podle které celý tento směr statistiky dostal svůj název, a můžeme tak získat aposteriorní model $f(\theta|D)$, který již obsahuje naše přesvědčení o datech ovlivněné samotnými datovými body. Tato věta zní následovně:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Umožňuje nám tedy získat podmíněnou pravděpodobnost opačnou. Díky této větě jsme schopni aktualizovat naše přesvědčení a adaptovat se na nové jevy nebo datová pozorování. Tato věta samozřejmě funguje jen za nějakých podmínek, ze zlomku plyne, že $P(B) \neq 0$, a také je vhodné aby $P(A)$ i $P(B)$ nebyly rovné 1. Protože pak není moc co zjišťovat, když už jsme si na 100% jistí. Tudíž aby dobře fungovala tato aktualizace, musíme si být alespoň trochu nejistí, což bychom měli být vždy, už jen kvůli šumu.

My si ji však trošku upravíme, nebo spíš použijeme operátor, který nám trošku pomůže zpřehlednit některé vzorce. Je to operátor rovno až na konstantu $\propto$, nebo taky proporciálně rovno. Pokud za A dosadíme θ jako náš model a za B datový bod D a předchozí rovnici tedy přepíšeme jako

$$P(\theta|D) \propto P(D|\theta)P(\theta),$$

čímž dostaneme rovnici, která nám říká, jakým způsobem dostat z apriorního modelu θ na základě dat D aposteriorní model, který závisí na datech. Když nám přijde další datový bod, tak jednoduše prohlásíme předchozí aposteriorní model jako apriorní a provedeme stejný krok. Toto má velkou výhodu pro systémy, kde data přichází postupně, což je nazýváno online aktualizace, nicméně můžeme tento model použít i pro datovou sadu, kde máme všechna data od začátku. V druhém případě model jednoduše dostane všechny datové body jednorázově, proto to nazýváme offline update. V této úpravě nám zmizela část $P(B)$ ve zlomku. Za to může právě operátor $\propto$, protože tato část je kon-

stanta. Zároveň díky tomu, že dokážeme model aktualizovat datovými body, jsme i schopni z výsledného modelu jednoduše dostat stav, který byl $p, p < n$ datových bodů zpátky. To může být vhodné pro případ, že se nám model dostane do nějakého nevhodného stavu, např. sérií špatných datových bodů. Samotný krok aplikace Bayesova pravidla budeme nazývat jako inference.

1.3 Konjugovaná rozdělení

Vlastnost rozdělení, kterou má smysl rozebrat před další částí, je konjugovanost. Rozdělení je určováno nějakými tvarovacími parametry s. Mějme aposteriorní rozdělení f. Toto rozdělení vzniklo inferencí z dat a tvarováním apriorního modelu, je tedy závislé nejen na D, ale i na startovních tvarovacích parametrech apriorního modelu $f_0(\theta|s_0)$. Z výpočetních i programátorských důvodů se hodí, když tvar aposteriorního rozdělení je stejný, jako tvar apriorního rozdělení f_0. K tomu si potřebujeme zadefinovat suficientní statistiku s jako vektor dat a počáteční hodnoty s_0

$$s = s(D, s_0), s \in R^q, q < \infty.$$

Speciálním případem je $s(, s_0) \equiv s_0$. Suficientní statistika s určuje pořadí, jakým model projde inferencí. Pokud nám po aplikaci bayesova pravidla vznikne stejný typ rozdělení, tak lze aposteriorní rozdělení zapsat ve tvaru apriorního rozdělení

$$f_0(\theta|s) = f(D|\theta)f(\theta|s_0).$$

Takové rozdělení jsou nazývány v angličtině jako self-replicating [9] neboli sebe-replikující, v češtině bych ale volil alternativní pojmenování – konjugovaná rozdělení. Této vlastnosti je vhodné využít, snižuje nám totiž složitost aposteriorního modelu. Převážně proto, že budeme pracovat stále se stejným typem rozdělení a zjednodušíme si výpočet momentů. Příkladem budiž *normální* a *Gamma* rozdělení.

1.4 Proměnné parametry

Při použití Bayesova vzorce pro aktualizaci modelu můžeme narazit na problémy, že nevíme, na jaké hodnoty nastavit některé neměnné parametry modelu. Modely, které popisují jednoduché případy, špatné vstupní parametry mohou ustát, ale pokud bychom modelovali něco složitějšího, např. šíření mikroplastů v ovzduší, tak chyba v nastavení variancí může způsobit, že model nezkonverguje, případně jeho konvergence povede k nastavenému apriornu. Obojí je pro nás špatné, kvůli první chybě nám vyjdou hodnoty, které jsou často vysoké a nedávají smysl, v druhém případě zase se může stát, že data nám naše apriorní přesvědčení nijak nebo téměř nijak neovlivní. Parametry pak budou velmi blízko apriorních hodnot, což úplně není chtěná záležitost.

Zde by se hodilo mít model tvořený více parametry, které jsme schopni v průběhu měnit, ať už v závislosti na čase a nebo v závislosti na datech. Z určitých důvodů by se nám hodilo mít parametry nezávislé mezi sebou. Hlavní důvod pro tuto vlastnost je, že zvládneme parametry od sebe jednoduše rozdělit. Protože jak víme ze Sekce 1.1, jsou-li náhodné veličiny nezávislé, tak jejich součin je pravděpodobnost jevu sdruženého. Tím pádem pokud bychom chtěli získat ze sdruženého jevu nějakou část, říkejme tomu marginála, jde to jednoduchým dělením. Na náš problém se to převede následovně:

$$f(\theta|D) \approx \widehat{f(\theta_1|D)}\widehat{f(\theta_2|D)}\dots\widehat{f(\theta_n|D)},$$

s n parametry θ. Počet těchto parametrů je závislý od výběru modelu a volbě modeláře. Všechny tyto marginály však závisí tvarovacími parametry na momentech ostatních parametrů. Tedy ne nutně všech, ale každý moment je použit alespoň při výpočtu jedné z marginál. Proto je potřeba k úspěšné optimalizaci parametrů mít možnost spočítat všechny potřebné momenty [7].

Pro samotný výpočet momentů parametrů budeme potřebovat získat podmíněné marginální rozdělení $f(\theta_p|D)$. To získáme vyintegrováním všech ostatních parametrů z modelu θ. Pro jednoduchý případ dvou parametrů θ_1, θ_2, kde nás zajímá např. marginála $f(\theta_1|D)$ to znamená vyintegrovat podle θ_2

$$f(\theta_1|D) = \int_{\Theta_2} f(\theta_1, \theta_2|D)d\theta_2.$$

Po marginalizaci nám již zbyde jen spočítat momenty rozdělení $g(\theta)$. Moment se počítá jako střední hodnota funkce:

$$E_{f(\theta|D)}g(\theta) = \int_{\Theta^*} g(\theta)f(\theta|D)d\theta.$$

Nicméně v praxi se volí taková rozdělení, které jsou konjugovaná a zároveň mají známé momenty. Díky tomu nemusíme v každém kroku počítat integrál, ale jednoduše použijeme již známé vzorce na výpočet momentů. To znatelně zrychluje a zjednodušuje výpočet.

V literatuře [7, 10] se obvykle jako aposteriorní bodový odhad $g(\theta)$ používá značení $\widehat{g(\theta)}$, které převezmu. Použijeme to převážně pro zjednodušení zápisu předchozí levé strany rovnice

$$\widehat{g(\theta)} = E_{f(\theta|D)}g(\theta).$$

1.5 Chyba modelu

Víme už jak vyjádřit jednotlivé marginály daného modelu, jak však nalézt tu nejlepší? Hodilo by se mít možnost měřit chybu odhadovaného modelu

$\breve{f}(\theta|D)$ oproti opravdovému modelu dat $f(\theta|D)$. Dokázali bychom pak vybírat ty nejlepší modely $\tilde{f}$ ze všech možných. A přesně k tomu nám poslouží Kullback-Leiblerova divergence [11]. Ta je definována následovně:

$$KLD(f(\theta|D)||\breve{f}(\theta|D)) = \int_{\Theta^*} f(\theta|D) ln \frac{f(\theta|D)}{\breve{f}(\theta|D)} d\theta.$$

Této metrice se občas také říká cross-entropie. Má několik důležitých vlastností:

1. $KLD\Big(f(\theta|D)||\breve{f}(\theta|D)\Big) >= 0$
2. $KLD\Big(f(\theta|D)||\breve{f}(\theta|D)\Big) = 0$ pokud $f(\theta|D) = \breve{f}(\theta|D)$ skoro všude
3. $KLD\Big(f(\theta|D)||\breve{f}(\theta|D)\Big) = \infty$ pokud pro všechny prvky $f(\theta|D) > 0$ a zároveň pro všechny prvky platí $\breve{f}(\theta|D) = 0$
4. Obecně $KLD\Big(f(\theta|D)||\breve{f}(\theta|D)\Big) \neq KLD\Big(\breve{f}(\theta|D)||f(\theta|D)\Big)$
5. Obecně také KLD nesplňuje trojúhelníkovou nerovnost

Kullback-Leiblerova divergence nám tedy dává možnost srovnání modelů. Nejlepší model $\tilde{f}$ ze všech modelů $\dot{f}$ je pak takový, který má nejmenší hodnotu KLD. Neboli

$$\tilde{f}(\theta|D) = arg_{\Theta^*} min KLD\Big(\dot{f}(\theta|D)||f(\theta|D)\Big).$$

1.6 Optimalizace parametrů

Víme už, jak vypadá nejlepší model. Jak takový model nalézt? Možností je nespočet, můžeme zkusit použít gradientní sestup, evoluční mechanismy, různé heuristiky a další. Někdy dokonce existuje analytický postup jak nalézt nejlepší řešení, to se využívá např. v modelování pomocí lineární regrese. Problém který však budeme řešit, má nepříjemné vlastnosti, popsané v Sekci 2.1, kvůli kterým nebudeme mít možnost využít exaktní minimalizaci. V této Sekci bych rád stručně popsal metodu gradientního sestupu, protože se docela často používá. Následně si rozebereme do detailu optimalizaci, která se používá v metodě VB. Obě metody mají podobnosti, gradientní sestup je však jednodušší na popsání a pochopení.

1.6.1 Metoda Gradientního sestupu

Pokud budeme chtít najít minimum pomocí metody gradientního sestupu, budeme v každém kroku chtít jít po směru největšího spádu. Můžeme si to

představit, jako když jsme zrovna vylezli na horu a chceme sestoupit co nejrychleji. Ale máme superschopnosti a můžeme skákat i z vysokých výšek bez obav. Tohle tedy většinou neplatí, ale pokud by to zrovna platilo, tak si vybereme směr, kde ztratíme co nejvíc výškových metrů. V každém kroku sestupu si zvolíme směr a vzdálenost, kterou tímto směrem půjdeme. Jak zjistit směr? Derivací:

$$x_{n+1} = x_n + \nu \nabla f(x_n),$$

kde $f(x)$ vyjadřuje funkční prostor, na kterém se pohybujeme a ν uraženou vzdálenost v kroku. Pokud budeme vzdálenost v každém kroku nastavovat *rozumně*, což může být dost náročné, protože můžeme skončit výše, než jsme začali, tak skončíme v nějakém ďolíku, matematicky minimu. Zde je však velký problém, pokud funkce není *hezká*, těchto minim existuje mnoho a můžeme skončit v nějakém, které není to nejlepší. Všechny další variace gradientního sestupu se pak snaží *rozumně* nastavovat vzdálenost ν, ať už přidáním hybnosti, kdy si pamatujeme jak moc se v každém kroku pohneme a ovlivníme tím, jak moc se pohneme příště, postupným zmenšováním a tak podobně. Možností je spousta, fantazii se meze nekladou. V této práci použijeme optimalizaci pomocí iterační Bayesovy metody, neboli IVB.

1.6.2 Iterační optimalizace v metodě Variačního Bayese

Věta 1.1 *Pokud je $f(\theta|D)$ aposteriorní rozdělení dat D definované vektorem parametrů θ, které můžeme rozdělit do p sub-vektorů:*

$$\theta^T = [\theta_1^T, \theta_2^T, \ldots, \theta_p^T]^T$$

Buď $\breve{f}(\theta|D)$ odhad rozdělení omezený podmíněně nezávislými proměnnými $\theta_1, \theta_2, \ldots, \theta_p$ takový, že platí:

$$\breve{f}(\theta|D) = \breve{f}(\theta|D) = \Pi_{i=1}^{p} \breve{f}(\theta_i|D)$$

Pak minimum KLD:

$$\breve{f}(\theta|D) = arg_{\Theta^*} min KLD\Big(\breve{f}(\theta|D)||f(\theta|D)\Big)$$

se nalezne jako:

$$\tilde{f}(\theta_i|D) \propto exp\Big(E_{\tilde{f}(\theta_{/i}|D)}(lnf(\theta, D))]\Big)$$

kde $\theta_{/i}$ značí komplement θ_i v θ a $\tilde{f}(\theta_{/i}|D) = \Pi_{j=1, j\neq i}^{q} \tilde{f}(\theta_j|D)$.

Důkaz konvergence je možné nalézt [7]. Můžeme si všimnout, že celý proces hledání nejlepšího rozdělení $\tilde{f}(\theta_i|D)$ je plně deterministický. Odhadnuté rozdělení $\tilde{f}(\theta_i|D)$ je pak marginála a $\tilde{f}(\theta|D)$ je pak výsledný model, který nám tato optimalizace vrátí.

Samotný výpočet ve VB pak funguje iterativně, kdy se střídají marginály, které se zrovna počítají. Uvedeme si jednoduchý příklad pro $p = 2$. V n-tém kroku, kde $n >= 1$, prvně spočítáme marginálu θ_1 následovně:

$$\tilde{f}^{(n)}(\theta_1|D) = \int_{\Theta_2^*} \tilde{f}^{(n-1)}(\theta_2|D) ln f(\theta_1, \theta_2, D) d\theta_2$$

Tento výsledek použijeme ve výpočtu pro θ_2 při stále stejném n:

$$\tilde{f}^{(n)}(\theta_2|D) = \int_{\Theta_1^*} \tilde{f}^{n}(\theta_1|D) ln f(\theta_1, \theta_2, D) d\theta_1$$

Ještě si musíme nějak zadefinovat $f(0)$. Zde je volba na nás. A protože se první počítá parametr θ_1, stačí zvolit pouze parametr $\theta_2^{(0)}$. Tento cyklus v každém kroku snižuje KLD, v každém kroku se tak nejlepšímu modelu blížíme. Volba pořadí výpočtu parametrů záleží na modelářovi, můžeme dokonce i nějaký parametr v n-té iteraci přepočítat vícekrát, nicméně typicky se každá marginála v jednom kroku počítá právě jednou. Složitost tohoto algoritmu je tedy pn, kde n je námi zvolený počet iterací a p je počet parametrů modelu. Tomuto výpočtu říkáme Iterační Variační Bayes.

Při vyjadřování marginál však využijeme pár věcí definovaných dříve. Víme, že θ se skládá z p nezávislých parametrů, což předpokládáme, že se dá zapsat jako součin pravděpodobností:

$$f(\theta|D) = f(\theta_1, \theta_2, \ldots, \theta_p) = \Pi_{i=1}^{p} f(\theta_i|D)$$

Také víme, z podmínky normality pravděpodobnosti, že součet všech pravděpodobností je 1. Tedy u spojité pravděpodobnosti je to integrál přes definiční obor. To zní povědomě, pojďme si rozepsat jeden z předchozích kroků:

$$\begin{aligned}
\tilde{f}^{(n)}(\theta_2|D) &= \int_{\Theta_1^*} \tilde{f}^{(n)}(\theta_1|D) ln f(\theta_1, \theta_2, D) d\theta_1 \\
&= \int_{\Theta_1^*} \tilde{f}^{(n)}(\theta_1|D) ln\Big(\Pi_{i=1}^{2} f(\theta_i, D)\Big) d\theta_1 \\
&= \int_{\Theta_1^*} \tilde{f}^{(n)}(\theta_1|D) ln\Big(f(\theta_1, D)\Big) d\theta_1 + \int_{\Theta_1^*} \tilde{f}^{(n)}(\theta_1|D) ln\Big(f(\theta_2, D)\Big) d\theta_1 \\
&= \int_{\Theta_1^*} \tilde{f}^{(n)}(\theta_1|D) ln\Big(f(\theta_1, D)\Big) d\theta_1 + ln\Big(f(\theta_2, D)\Big) \int_{\Theta_1^*} \tilde{f}^{(n)}(\widehat{\theta_1}|D) d\theta_1 \\
&= \int_{\Theta_1^*} \tilde{f}^{(n)}(\theta_1|D) ln\Big(f(\theta_1, D)\Big) d\theta_1 + ln\Big(f(\theta_2, D)\Big) \\
&\propto ln\Big(f(\theta_2, D)\Big)
\end{aligned}$$

První člen nám kompletně zmizel. To je kvůli operátoru $\propto$, nikde se v něm totiž nevyskytuje parametr θ_2 a vůči funkci vlevo je to konstanta. A jak víme,

tento operátor zahazuje konstanty, takže jednoduše zmizí. U druhého členu si můžeme všimnout, že nám vymizel integrál závislý na θ_1. Integrace přes všechny možné hodnoty v Θ_1^* funkce $f(\theta_1|D)$ je rovná 1. Pro vyjádření θ_2 potřebujeme členy rovnic, které jsou závislé jen na samotném členu θ_2. Všechny ostatní členy, které nezávisí na θ_2, při tvorbě rovnic jednoduše zahodíme.

1.7 Algoritmus Variačního Bayese

Po přípravě se dostáváme k samotnému algoritmu VB. Algoritmus se dá rozdělit do 7 kroků [7].

Krok 1: Formuluj bayesovský model. V tomto kroku zvolíme rozdělení $f(\theta, D)$. Zde volíme model dat $f(D|\theta)$ a apriorní rozdělení $f(\theta)$

Krok 2: Rozděl parametry θ do p částí. Ověř, zda po rozdělení jsou parametry nezávislé, neboli že platí v jednoduchém případě $p = 2$

$$lnf(\theta_1, \theta_2, D) = g(\theta_1, D)^T h(\theta_2, D).$$

Kde $g(\theta_1, D)$ a $h(\theta_2, D)$ jsou vektory konečných kompatibilních rozměrů. Toto pouze značí, že model lze separovat v logaritmu. Pokud tento krok platí, parametry jsou separovatelné a můžeme pokračovat dalším krokem. Pokud ne, musíme daný model přeformulovat.

Krok 3: Vyjádři marginály. V tomto kroku si vyjádříme pro každý parametr jak vypadá jeho marginála. Pro náš $p = 2$ příklad to bude vypadat následovně

$$\begin{aligned}
\tilde{f}(\theta_1|D) &\propto exp\Big(E_{\tilde{f}(\theta_2|D)} lnf(\theta_1, \theta_2, D)\Big) \\
&\propto exp\Big(lnf(\theta_1, D) + E_{\tilde{f}(\theta_2|D)} lnf(\theta_2, D)\Big) \\
&\propto exp\Big(g(\theta_1, D)\widehat{h(\theta_2, D)}\Big) \\
\tilde{f}(\theta_2|D) &\propto exp\Big(E_{\tilde{f}(\theta_1|D)} lnf(\theta_1, \theta_2, D)\Big) \\
&\propto exp\Big(E_{\tilde{f}(\theta_1|D)} lnf(\theta_1, D) + lnf(\theta_2, D)\Big) \\
&\propto exp\Big(\widehat{g(\theta_1, D)} h(\theta_2, D)\Big)
\end{aligned}$$

Krok 4: Nalezni v marginálách podobnost k nějakému typickému rozdělení. Zde využijeme vlastnosti popsané v Sekci 1.3. Každý parametr $\theta_i, i < p$ vyjádříme jako jeho apriorní rozdělení g pomocí tvarovacích parametrů. Pokud toto není možné, můžeme stále použít symbolickou nebo numerickou integraci, nicméně složitost výpočtu je mnohem náročnější, než když se nám podaří nalézt apriorní rozdělení. Tento krok je často nejtěžší část celého procesu, protože vyžaduje znalosti toho, jak vypadají typická rozdělení. Nejčastěji se setkáme s *Normálním* rozdělením a rozdělením *Gamma*. Připodobnění k Normálnímu rozdělení budeme aplikovat, pokud uvidíme 2 členy, lineární a kvadratický. Gamma je též typická 2 členy, ale tentokrát to je lineární a logaritmický.

Krok 5: Vyjádři momenty marginál. Pokud se nám v předchozím kroku podařilo identifikovat nějaké typické rozdělení, tento krok bude jednoduchý. Stačí nám se akorát podívat do tabulek, jak se potřebné momenty počítají.

Krok 6: Pusť Iterační VB na rovnicích z předchozích kroků. Zde známe momenty, známe tvarovací parametry funkcí, můžeme tedy použít IVB na výpočet odhadu. Zvolíme si vhodný počet iterací, případně nastavíme zastavovací kritérium. Poté pustíme iteraci popsanou v Sekci 1.6.2. Pokud máme p různých sub-parametrů ve vektoru θ, tak obecně potřebujeme zvolit $p-1$ různých počátečních parametrů. Ten poslední se přepočítá v první iteraci IVB, tudíž není třeba ho volit.

Krok 7: Vrať marginály VB. Poté co doběhne předchozí krok, vrátíme žádané marginály zjistěné v běhu. Výhodou tohoto procesu je, že tyto aproximované marginální rozdělení jsou již ve tvaru rozdělení, které hledáme. Často nás bude na výsledku zajímat pouze jedna nebo subset parametrů θ, případně momenty rozdělení, které jsou tvarovány těmito parametry.

1.8 Příklad použití VB

Pojďme si ukázat, jak by se dal Variační Bayes použít pro co nejvíce jednoduchý příklad. Jako pěkný příklad jsem vybral problém skalární dekompozice součinu:

$$d = ax + \epsilon$$

Kde a a x jsou neznámé proměnné, d jsou naše data a ϵ je šum se střední hodnotou 0. Model dat se dá tedy zapsat jako

$$f(d|a,x) = N(ax, r_\epsilon)$$

Jako rozdělení modelující naše data tedy zvolíme rozdělení *Normální* se střední hodnotou ax a rozptylem r_ϵ. Apriorní parametry můžeme také zvolit jako *Normální* rozdělení se střední hodnotou 0. Tedy:

$$f(a) = N(0, r_a) \propto exp\Big(-\frac{1}{2}\frac{a^2}{r_a}\Big)$$
$$f(x) = N(0, r_x) \propto exp\Big(-\frac{1}{2}\frac{x^2}{r_x}\Big)$$

Všimněme si, že zde nastavujeme střední hodnotu na 0. Tím se snažíme vyjádřit, že apriorno je neinformativní. Apriorno má též varianci r_a, resp. r_x, ty nastavíme na nějaké hodnoty blízké šumu r_e. Jsme v ukázkovém testovacím příkladě, takže si můžeme dovolit minimálně pro ukázku mít informace, které v reálném výpočtu mít nebudeme. Pro demonstrativní příklady však ukážu i část, kdy nastavíme parametry „ošklivě“, abychom věděli, co se může stát.

1.8.1 Odvození rovnic

Model sdružené věrohodnosti vypadá následovně:

$$f(\theta, D) = f(a, x, d) \propto exp\Big(-\frac{1}{2}\big(\frac{(ax-d)^2}{r_e} + \frac{a^2}{r_a} + \frac{x^2}{r_x}\big)\Big)$$

Dále si máme napsat aposteriorní rozdělení obou marginál:

$$\begin{aligned}
\tilde{f}(a|D) &\propto exp\Big(E_{\tilde{f}(x|d)}\big(lnf(a,x,d)\big)\Big) \\
&\propto exp\Big(\big(-\frac{1}{2}a^2(\widehat{x^2}r_x^{-1} + r_a^{-1}) - a(d\widehat{x}r_e^{-1})\Big)\Big) \\
\tilde{f}(x|D) &\propto exp\Big(E_{\tilde{f}(a|d)}\big(lnf(a,x,d)\big)\Big) \\
&\propto exp\Big(\big(-\frac{1}{2}x^2(\widehat{a^2}r_a^{-1} + r_x^{-1}) - x(d\widehat{a}r_e^{-1})\Big)\Big)
\end{aligned}$$

V dalším kroku máme najít podobnost k nějakému typickému rozdělení. Z předchozích rovnic vidíme, že jsou si podobné s rozdělením definovaném v apriornu. Obě marginály totiž mají 2 členy, kvadratický a lineární, u čehož jsme si již řekli, že to je typické *Normální* rozdělení.

$$\begin{aligned}
f(a|d) = N(\mu_a, \sigma_a) &\propto exp\Big(-\frac{1}{2}\frac{(a-\mu_a)^2}{\sigma_a}\Big) \\
f(x|d) = N(\mu_x, \sigma_x) &\propto exp\Big(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x}\Big)
\end{aligned}$$

Tvarovací parametry jsou tedy $\mu_a, \sigma_a, \mu_x, \sigma_x$. Ty si můžeme z marginál vyjádřit následovně:

$$exp\Big(-\frac{1}{2}\frac{a^2 - 2\mu_a a + \mu_a^2}{\sigma_a}\Big) \propto exp\Big(\big(-\frac{1}{2}a^2(\widehat{x^2}r_x^{-1} + r_a^{-1}) - a(d\widehat{x}r_e^{-1})\Big)\Big)$$

Poté porovnáme kvadratické členy:

$$\begin{aligned}
-\frac{1}{2}a^2\sigma_a^{-1} &= -\frac{1}{2}a^2(\widehat{x^2}r_x^{-1} + r_a^{-1}) \\
\sigma_a^{-1} &= (\widehat{x^2}r_x^{-1} + r_a^{-1}) \\
\sigma_a &= (\widehat{x^2}r_x^{-1} + r_a^{-1})^{-1}
\end{aligned}$$

A lineární členy:

$$\begin{aligned}
-\frac{1}{2}2\mu_a a\sigma_a^{-1} &= ad\widehat{x}r_e^{-1} \\
\mu_a &= d\widehat{x}r_e^{-1}\sigma_a
\end{aligned}$$

Pro tvarovací parametry x bude postup podobný, uvedu jen výsledky:

$$\sigma_x = \left(\widehat{a^2} r_a^{-1} + r_x^{-1}\right)^{-1}$$
$$\mu_x = d\widehat{a} r_e^{-1} \sigma_x$$

Máme tedy rovnice pro tvarovací parametry $\mu_a, \sigma_a, \mu_x, \sigma_x$. Pro výpočet však potřebujeme $\widehat{a^2}, \widehat{a}, \widehat{x^2}, \widehat{x}$. Což jsou momenty normálního rozdělení.

$$\widehat{a} = \mu_a$$
$$\widehat{a^2} = \mu_a^2 + \sigma_a$$
$$\widehat{x} = \mu_x$$
$$\widehat{x^2} = \mu_x^2 + \sigma_x$$

Můžeme pokračovat na krok 6, kde pustíme iteraci VB. Počítání začneme počítáním parametru např. x. Musíme tedy zvolit startovací σ_a^0, μ_a^0. Poté střídavým výpočtem aktualizujeme parametry. Graficky by se to dalo vyjádřit následovně:

$$\widehat{x} \to \widehat{a} \to \widehat{x} \to \dots$$

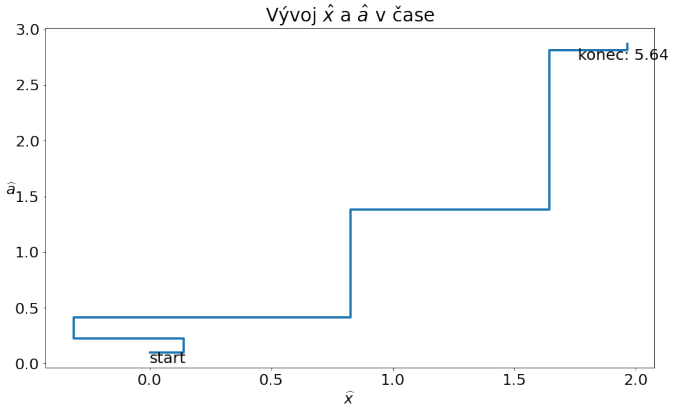
Toto opakujeme do konvergence, případně to můžeme zastavit po určitém počtu kroků.

Krok 7 nám už jen říká, že máme vrátit momenty marginál. To jsou již naše hledaná $\widehat{x}$ a $\widehat{a}$.

Odvodili jsme si celý algoritmus pro skalární dekompozici. Tento problém není nějak náročný, hlavně pokud se můžeme pohybovat po $\mathbb{R}$. Je však dostatečně názorný pro prezentaci této metody. Zde iterace vede k nějakému lokálnímu minimu. Těch je u tohoto problému nekonečně mnoho, tudíž bude záležet na tom, odkud odstartujeme. Pseudokód pak vypadá následovně:

Algoritmus 1 Variační bayess pro dekompozici součinu

inicializuj d, r_e, r_a, r_x
nastav startovací μ_a, σ_a
dokud není splněna podmínka konvergence
  aktualizuj $\widehat{x}, \widehat{x^2}$
  aktualizuj $\widehat{a}, \widehat{a^2}$
konec
vrať $\widehat{x}, \widehat{a}$



Obrázek 1.1: VB pro *rozumně* nastavené začáteční hodnoty

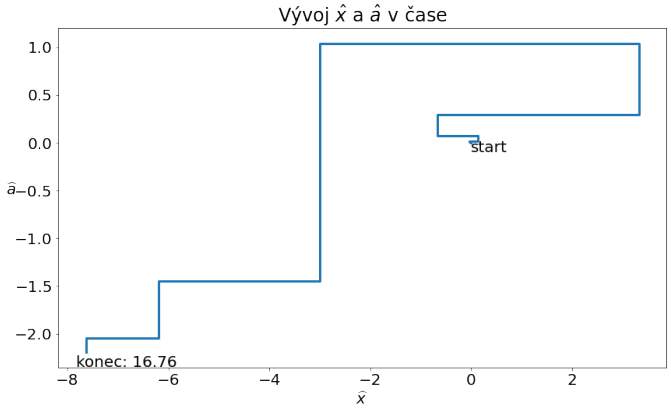
1.8.2 Chování Variačního Bayese pro problém skalární dekompozice

V této části práce si vyzkoušíme předchozí algoritmus v praxi. Uvidíme, jak reaguje na relativně *rozumné* startovací parametry a také zkusíme algoritmu předhodit nějaké parametry *špatné*.

Pojďme tedy nejdříve zkusit nastavit naše parametry nějak *hezky*. Zvolme si $d = 6, r_e = 1, r_a = 0.3, r_x = 0.3$. Dále podle pseudokódu si máme zvolit vhodné startovací $\mu_a^{(0)}, \sigma_a^{(0)}$. Pojďme tedy jako $\mu_a^{(0)}$ zvolit třeba 0.1 a $\sigma_a^{(0)}$ jako 1. Zastavovací kritérium zvolíme jako vzdálenost $\widehat{a} * \widehat{x} - d$, tato vzdálenost musí být menší než $sqrt(r_e)$. Pustíme VB a v každém kroku si budeme pamatovat hodnoty $\widehat{a}, \widehat{x}$. Sledujeme vývoj parametrů $\widehat{a}, \widehat{x}$ a dobu konvergence.

Obrázek 1.1 ukazuje vývoj hodnot $\widehat{a}, \widehat{x}$ v čase. Můžeme si všimnout, že v začátku algoritmus zdánlivě „tápe“, ale v jistém momentě začne správně minimalizovat hodnoty, až se dostane za zastavovací kritérium. Výsledek $\widehat{x}\widehat{a}$ je pak zanesen do grafu, můžeme si všimnout, že odhad skončil u čísla 5.64. Což je velice dobrý odhad, když počítáme s tím, že máme $r_e = 1$. Též jsme se k němu dostali relativně rychle, po 5 iteracích IVB. Jen poznámka, datových bodů v grafu je dvakrát počet iterací + 1 za počáteční hodnoty, protože si ukládám datové body po každém pod-výpočtu. Proto je počet iterací IVB 5.

V dalším kroku zkusíme hodnoty nastavit *špatně*. Počáteční μ_a^0 nastavíme na číslo blízké nule. Nemůžeme ho však nastavit na 0, protože ve výpočtu μ_x se používá při násobení a algoritmus by skončil v lokálním extrému $(0, 0)$. Zvolil jsem tedy hodnotu 10^{-4}. Počáteční σ_a^0 jsem také dost utáhl, na hodnotu 0.1, tím algoritmus trošku rozhodíme, protože říkáme, že jsme si jistí naším apriornem. Odhadované d dejme výše, např. 20. A protože algoritmu nechceme pomoci, nastavíme r_e na hodnotu vysokou, např. 25. Pokud chceme aby algoritmus zkonvergoval, ale zároveň mu to chvilku trvalo, musíme r_x, r_a



Obrázek 1.2: VB pro *špatně* nastavené počáteční parametry a data

nastavit na nízké hodnoty, minimum potřebné ke konvergenci se ukázalo být kolem 2.8. Pokud bychom nastavili nižší hodnoty, dostáváme se do problému popsanému již dříve, algoritmus nám zkonverguje k apriornu, které jsme však záměrně nastavili špatně. Zde se ale snažím demonstrovat, že i se špatnými daty a ne úplně optimálním startem, jsme se schopni dostat k rozumnému výsledku. Počet iterací byl relativně malý, algoritmus se pod podmínku zastavení dostal po 10 krocích. Obrázek 1.2 ukazuje průběh výpočtu, můžeme si všimnout, že chvilku trvalo, než se algoritmus dostal z apriorna, jakmile se však dostal do podobného řádu jako d, zkonvergoval velice rychle.

Ukázali jsme si, že Variační Bayes funguje dobře i ve chvílích, kdy mu předhodíme relativně špatné apriorní $\theta^{(0)}$ a špatná data. Tento model však není tolik robustní, počítáme zde např. s pevným šumem u odhadované proměnné, což může způsobit nekonvergenci v extrémních případech. Nicméně pro uvedení do problematiky je toto docela pěkný příklad, jak tato metoda funguje.

Kapitola 2

Bayesovský model lineární regrese

2.1 Motivace a konstrukce modelu

Mikroplasty nejsou lokální problém, byly nalezeny v místech, kde se člověk téměř nepohybuje. Mohli bychom tedy ze série měření nějak zjistit, jaká byla emise mikroplastů ve větší vzdálenosti? Představte si, že máte vektor y složený z p různých měření na senzorech:

$$y^T = [y_1, y_2, \ldots, y_p]^T$$

Dále také máme spočítané citlivosti měření na aktivitu zdroje v daném čase, ze kterého můžeme sestrojit model šíření částic M. Tento model je matice o rozměru $p \times n$, kde p je rozměr y a n je rozměr času. V této matici pro náš problém se často vyskytují nulové řádky a sloupce. Též je velmi častý výskyt čísel, která jsou velice malá. V závislosti na počasí může vítr foukat např. jen jedním směrem a pak je jasné, že některá místa budou mít napočítané citlivosti nulové. Případně nějaké místo může být vzdálené vůči místu emise, tím pádem budou napočítané citlivosti velice nízké.

Když máme tedy model teoretického šíření částic, můžeme zkusit najít vektor x o rozměrech n takový, že splňuje rovnici

$$y = Mx$$

Tento problém je docela známý, říká se mu lineární regrese. A řešení je teoreticky jednoduché, stačí znát nebo zjistit M^{-1}. Akorát že to je mnohem složitější, jak si ukážeme níže.

2.1.1 Motivace pro VB

Při hledání řešení pro nejmenší čtverce bychom řešili *normální* rovnici, která by pro náš případ vypadala následovně:

$$x = (M^T M)^{-1} M^T y$$

Tato rovnice má jednu důležitou podmínku, kvůli které se obyčejná lineární regrese nedá aplikovat na všechny problémy tohoto typu. Matice $M^T M$ musí být regulární. Regulárnost je podmínka k existenci inverze k maticím, definovat si ji můžeme třeba hodností matice. Pokud má čtvercová matice $M^T M$ rozměrů n hodnost $h(M^T M) = n$, pak je regulární.

Naše matice $M^T M$ regulární nebude. Je to z jednoduchého důvodu. V matici M se vyskytují nulové řádky i sloupce. Pojďme si ukázat, jakým způsobem nám to ovlivňuje matici $M^T M$. Mějme tedy matici M s alespoň jedním nulovým sloupkem.

$$M = \begin{pmatrix} m_{11} & 0 & \dots & m_{1p} \\ \vdots & 0 & \ddots & \vdots \\ m_{n1} & 0 & \dots & m_{np} \end{pmatrix}$$

Potom matice $M^T M$ obsahuje nulový řádek:

$$M^T M = \begin{pmatrix} m_{11} & \dots & m_{n1} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ m_{1p} & \dots & m_{np} \end{pmatrix} \begin{pmatrix} m_{11} & 0 & \dots & m_{1p} \\ \vdots & 0 & \ddots & \vdots \\ m_{n1} & 0 & \dots & m_{np} \end{pmatrix} = \begin{pmatrix} m_{11}^2 & \dots & m_{n1} m_{1p} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ m_{1p} m_{n1} & \dots & m_{np}^2 \end{pmatrix}$$

Pokud má matice jeden nulový řádek, pak její hodnost zcela jistě není rovna n. Tím pádem nemůže být regulární a nemůže ani existovat její inverze.

Matice M však nemusí obsahovat jen nulové sloupky, ale může obsahovat i čísla blízká nule. Inverze pak sice existuje, ale je velmi nestabilní. To se může při odhadování znásobit a způsobit velké chyby. Pro tyto případy existuje mnoho modifikací pro lineární regresi, jedním z nich je právě metoda VB.

2.2 Testovací data

Testovací data, která jsem zvolil pro účely zkoumání kvalit navržených algoritmů, reflektují vlastnosti dat reálných. Matice M bude obsahovat náhodná čísla s jedním nulovým sloupcem druhým v pořadí. V reálných datech je jich sice více, nicméně pro testovací účely to bude stačit. Rozměry jsem volil malé, počet měření $p = 6$, velikost vektoru x zase 11. Vektor x, což je vektor emisí, který se snažíme odhadovat, obsahuje hodnoty 0 a 1. Tyto hodnoty symbolizují intenzitu emise daný den. Vektor y následně zjistíme vynásobením Mx a přidáním normálního šumu s rozptylem 0.3. Na těchto datech budu prezentovat, jak se navržené modely chovají.

2.3 Základní model Bayesovské lineární regrese

Pro rekapitulaci, řešíme problém lineární regrese:

$$y = Mx$$

Vektor y tedy dostaneme součinem. Využijeme podobný model jako v sekci 1.8, což je *Normální* rozdělení s $\mu = Mx$. Varianci v neznáme, proto je vhodné a výhodné ji odhadovat. Předpokládejme, že jednotlivé prvky ve vektoru y jsou nekorelované. Kovariační matici pak zvolíme jako $\omega^{-1}I_p$, kde I_p je jednotková matice velikosti p. To jednoduše znamená, že předpokládáme pro všechny prvky stejnou varianci šumu. Model dat tedy bude vypadat následovně:

$$\begin{aligned} f(y|x,\omega) &= N(Mx, \omega^{-1}I_p) \\ &\propto |\omega^{-1}I_p|^{-\frac{1}{2}} exp\Big(-\frac{1}{2}(y-Mx)^T\omega(y-Mx)\Big) \end{aligned}$$

Dále musíme zvolit rozdělení modelující naše parametry x a ω. Vektor x budeme modelovat pomocí *Normálního* rozdělení:

$$f(x) = N(0_n, I_n) \propto exp\Big(-\frac{1}{2}(x^T I_n x)\Big)$$

Pro začátek tedy použijeme nulový vektor jako parametr μ a jednotkovou matici jako korelační matici.

2.3.1 Model šumu

Jakým způsobem modelovat ω? *Normální* rozdělení doteď fungovalo, pojďme ho zkusit taky. Předpis pro skalár ω tedy bude vypadat následovně:

$$f(\omega) = N(0, I) \propto exp\Big(-\frac{1}{2}(\omega^T I\omega)\Big)$$

Sdružená pravděpodobnost $f(y, x, \omega)$:

$$f(y,x,\omega) \propto |\omega^{-1}I_n|^{-\frac{1}{2}} exp\Big(-\frac{1}{2}(y-Mx)^T\omega(y-Mx) - \frac{1}{2}(x^T Ix) - \frac{1}{2}(\omega^T I\omega)\Big)$$

A marginály:

$$\tilde{f}(x|y) \propto exp\left(-\frac{1}{2}(y-Mx)^T\widehat{\omega}(y-Mx) - \frac{1}{2}(x^T I_p x)\right)$$

$$\begin{aligned} \tilde{f}(\omega|y) &\propto |\omega^{-1}I_n|^{-\frac{1}{2}} exp\left(-\frac{1}{2}\omega(y^Ty - 2yM\widehat{x} + \widehat{x^Tx}) - \frac{1}{2}(\omega^T I\omega)\right) \\ &\propto exp\left(\frac{n}{2}ln(\omega) - \frac{1}{2}\omega(y^Ty - 2yM\widehat{x} + \widehat{x^Tx}) - \frac{1}{2}(\omega^T I\omega)\right) \end{aligned}$$

Parametr x má lineární a kvadratický člen, tudíž podle 1.7 kroku 4 víme, že to můžeme připodobnit k *normálnímu* rozdělení. Nicméně u parametru ω se objevil i člen logaritmický. To k normálnímu rozdělení nelze připodobnit. Musíme tedy přeformulovat funkci modelující ω. Všimněme si, že v marginále logaritmický člen je původem z modelu dat. To platí i o členu lineárním. Kvadratický člen je zase tvořen normálním rozdělením, které jsme vybrali jako model ω. S prvními dvěma členy nemůžeme nic udělat, tedy pokud bychom nechtěli dělat zásadní změny v modelu dat. Což by znamenalo pravděpodobně i zvolit jiný model pro parametr x. Avšak z 1.7 víme, že logaritmický a lineární člen má rozdělení *Gamma*. Pojďme tedy ω napsat jako *Gamma* rozdělení:

$$f(\omega) = G(a_0, b_0) \propto \frac{b_0^{a_0}}{\Gamma(a_0)} \omega^{a_0 - 1} exp(-b_0\omega)$$

Kde Γ je *Gamma* funkce v bodě a. Změní se nám sdružené rozdělení $f(y, x, \omega)$:

$$f(y, x, \omega) \propto |\omega^{-1} I_n|^{-\frac{1}{2}} exp\Big(-\frac{1}{2}(y - Mx)^T \omega (y - Mx) - \frac{1}{2}(x^T I_p x) \\ -b_0\omega + (a_0 - 1) ln(\omega)\Big)$$

Marginála pro x zůstane stejná, x není obsaženo v rozdělení modelující ω. Marginála pro ω pak vypadá následovně:

$$\tilde{f}(\omega|y) \propto exp\Bigg(-b_0\omega + (a_0 - 1) ln(\omega) + \frac{n}{2} ln(\omega) \\ -\frac{1}{2}\omega(y^T y - 2\widehat{x^T} My + x^T \widehat{M^T M} x)\Bigg)$$

Zde vidíme již jen lineární a logaritmické členy. Tudíž můžeme jednoduše připodobnit zpátky k rozdělení *Gamma*. Nevhodná volba rozdělení je tedy opravena a můžeme pokračovat na další kroky.

2.3.2 Tvarovací parametry

Máme tedy už vhodně vybraná rozdělení modelující x a ω. Teď je na čase zjistit, jak se budou aktualizovat tvarovací parametry μ_x, σ_x a a, b. Začněme parametry μ_x a σ_x

$$\sigma_x : -\frac{1}{2}x^T I x - \frac{1}{2}\widehat{\omega} x^T M^T M x = -\frac{1}{2} x^T \sigma_x^{-1} x \\ \sigma_x = \Big(I + \widehat{\omega} M^T M\Big)^{-1}$$

$$\mu_x : -\frac{1}{2}\widehat{\omega}\Big(2x^T M^T y\Big) = -\frac{1}{2}\Big(-2x^T\sigma^{-1}\mu_x\Big)$$
$$\mu_x = \sigma_x M^T y\widehat{\omega}$$

Dále a, b:

$$b : b_0\omega = b\omega - \frac{1}{2}\omega\Big(y^T y - 2\widehat{x}^T My + x^T\widehat{M^T M}x\Big)$$
$$b_0 = b - \frac{1}{2}\Big(y^T y - 2\widehat{x}^T My + x^T\widehat{M^T M}x\Big)$$
$$b = b_0 + \frac{1}{2}Tr\Big(y^T y - 2\widehat{x}^T My + \widehat{xx^T}M^T M\Big)$$

$$a : (a-1)ln(\omega) = (a_0-1)ln(\omega) + \frac{n}{2}ln(\omega)$$
$$a = a_0 + \frac{n}{2}$$

kde $Tr(A)$ značí stopu matice A.

Aktualizaci tvarovacích parametrů máme odvozenou, pro výpočty však potřebujeme zjistit momenty. Z rovnic vidíme, že potřebujeme první a druhý moment pro x a první moment ω. Díky tomu, že jsme identifikovali typická rozdělení v obou parametrech, stačí se nám podívat do literatury, např. [7]. Momenty x jsou:

$$\widehat{x} = \mu_x$$
$$\widehat{xx^T} = \sigma_x + \mu_x\mu_x^T.$$

První moment ω je pak:

$$\widehat{\omega} = \frac{a}{b}.$$

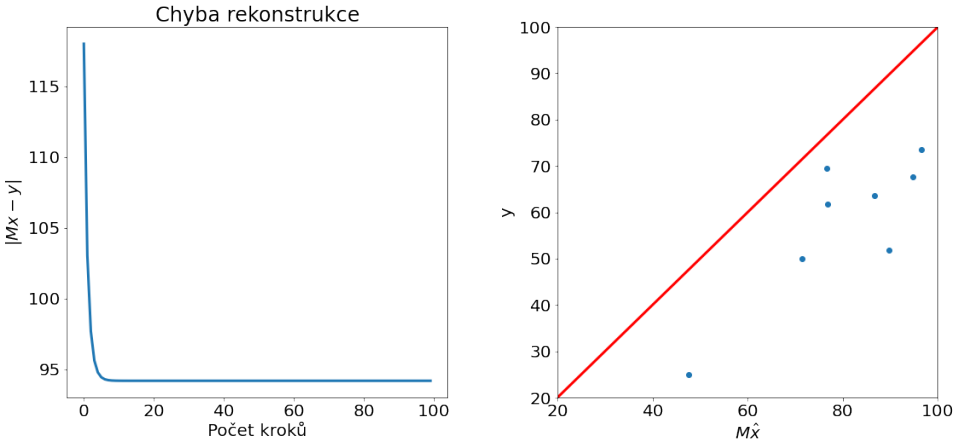
Dále nás čeká optimalizace iteračním algoritmem. Iteraci začneme výpočtem x, je tedy potřeba zvolit počáteční a_0, b_0. Parametry máme dva, budeme je tedy v každém kroku střídat. Na konci nás zajímá především první moment x, který potřebujeme k výpočtu rekonstrukce $M\widehat{x}$ pro měření výkonnosti modelu a zároveň je to výsledek, který hledáme.

2.3.3 Výsledky

Pojďme prozkoumat, jak se tento nejjednodušší model chová. K testování použiji hodnoty popsané v 2.2. Vektor x_{true}, jež budeme odhadovat bude:

$$x_{true} = \begin{pmatrix}0 & 1 & 1 & 0 & 1 & 1\end{pmatrix}^T$$

Matice M je náhodně vygenerovaná matice obsahující čísla od 0 do 40, s nulovým druhým sloupcem. Výpočet začneme výpočtem x, potřebujeme tedy zvolit



(a) Vývoj chyby rekonstrukce

(b) Bodový graf y vs $M\widehat{x}$

Obrázek 2.1: Metriky prvního modelu pro bayesovskou regresi. Obrázek a ukazuje jakým způsobem klesá chyba rekonstrukce v čase. Na obrázku b zase můžeme vidět jak vypadají rekonstruovaná data $M\widehat{x}$ oproti datům y

počáteční tvarovací parametry pro ω a samotné $\omega^{(0)}$. Oba apriorní tvarovací parametry nastavíme na 10^{-10}. Omega pak zvolíme následující metodou:

$$\omega = max(M^T M).$$

Je to z jednoduchého důvodu. Úplně první výpočet počítá tvarovací parametr σ_x. Ten se pro připomenutí počítá následovně:

$$\sigma_x = \left(I + \widehat{\omega} M^T M\right)^{-1}.$$

Volbou špatného ω bychom mohli algoritmus vychýlit, proto zvolíme takové ω, jež v první rovnici znormuje do velikosti matice I.

Z Obrázku 2.1 si můžeme všimnout, že v čase klesá chyba rekonstrukce, po nějaké době se však ustálí. Algoritmus totiž nepracuje se zastavovacím kritériem, ale s pevně daným počtem kroků, kterých je 100. Tento přístup byl zvolen, aby i ve stavu, který špatně konverguje (což se na reálných datech může stát), se iterace zastavila. Z bodového grafu si zase můžeme všimnout, že model našel takové $\widehat{x}$, které docela koreluje s původním x_{true}. Nicméně to vypadá, že odhad virtuální emise je větší, než je skutečnost. Toto opravíme v sekci 2.4. Nalezené $\widehat{x}$ vypadá následovně:

$$\widehat{x} = \begin{pmatrix} 0.55 & 0.8 & 0.98 & 0.49 & 0.87 & 0.85 \end{pmatrix}^T$$

což potvrzuje teorii o tom, že odhad je vyšší, než by měl.

2.4 Řídkost

Do modelu přidáme řídkost. Motivace za tím je taková, že pokud model nemá dostatek dat, tak by měl spíš táhnout k apriornu, které jsme mu nastavili. Od této úpravy si slibujeme stabilnější výsledky a rychlejší pokles chyby. Jak toho docílíme? Upravíme si apriorní model modelující x:

$$f(x) = N(0, V^{-1}) \propto exp\Big(-\frac{1}{2} x^T V x \Big)$$

Přibyla nám tedy matice V. Tu můžeme nastavit staticky, ale to bychom si nepomohli. Neplatilo by totiž, že pro prvky, u kterých si model není jistý, jako výsledek použijeme spíše apriorno. Přibyl nám tedy další parametr V. Jak tato matice vypadá? Je to kovariační matice pro x, u této proměnné jsme předpokládali nekorelovanost mezi sebou, proto byla původně kovariační matice zvolena jako I. To se nezmění, prvky budou též jen na diagonále. A protože hledané $\widehat{x}$ má velikost n, budeme hledat n parametrů v tvořící V:

$$V = \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & v_n \end{pmatrix}$$

Nepřibyl nám tedy jen jeden parametr, ale n. Všechny parametry budou mít stejný předpis. Ten si zapíšeme pro obecné v_j. Poučeni z předchozího kroku nebudeme zkoušet *normální* rozdělení ale zkusíme rovnou rozdělení *Gamma*. Shrňme si tedy modely parametrů:

$$\begin{aligned} f(x) &= N(0, V^{-1}) \propto exp\Big(-\frac{1}{2} x^T V x \Big) \\ f(\omega) &= G(a_0, b_0) \propto \frac{b_0^{a_0}}{\Gamma(a_0)} \omega^{a_0 - 1} exp(-b_0 \omega) \\ f(v_j) &= G(\alpha_0, \beta_0) \propto \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \omega^{\alpha_0 - 1} exp(-\beta_0 \omega), \text{ kde } 1 \leq j \leq n \end{aligned}$$

V marginálách parametrů x, ω se změní pouze x:

$$\tilde{f}(x|y, V) \propto exp\left(-\frac{1}{2} (y - Mx)^T \widehat{\omega} (y - Mx) - \frac{1}{2} (x^T V x) \right)$$

$$\tilde{f}(\omega|y) \propto exp\left(-b_0 \omega + (a_0 - 1 + \frac{n}{2}) ln(\omega) - \frac{1}{2} \omega Tr(y^T y - 2\widehat{x^T} M y + \widehat{x x^T} M^T M) \right)$$

Marginálu pro v_j si trošku rozebereme:

$$\tilde{f}(v_j|y) \propto exp\Big((\alpha-1)ln(v_j) + \frac{1}{2}ln(v_j) - \beta v_j - \frac{1}{2}\widehat{x_j^2}v_j\Big)$$

Člen pocházející z věrohonosti od x je jen j-tý prvek z vektoru x. Proč tomu tak je, ukáže následující příklad pro $n = 3$:

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} v_1 & 0 & 0 \\ 0 & v_2 & 0 \\ 0 & 0 & v_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 v_1 x_1 \\ x_2 v_2 x_2 \\ x_3 v_3 x_3 \end{pmatrix} = \begin{pmatrix} x_1^2 v_1 \\ x_2^2 v_2 \\ x_3^2 v_3 \end{pmatrix}$$

Tím, že matice V má prvky pouze na diagonále, se nám zjednoduší vyjádření pro x_j. To je pro nás velmi výhodné, snižuje nám to výpočetní složitost. V marginále vidíme logaritmický a lineární člen, naše volba apriorna byla rozumná a můžeme to zpátky připodobnit ke *Gammě*.

Vzhledem k tomu, že se neměnila marginála pro parametr ω, zůstanou stejné i rovnice pro tvarovací parametry. U tvarovacího parametru σ_x dojde k malé změně:

$$\sigma_x : -\frac{1}{2}x^T I x - \frac{1}{2}\widehat{\omega}x^T M^T M x = -\frac{1}{2}x^T \sigma_x^{-1} x$$
$$\sigma_x = \left(\widehat{V} + \widehat{\omega}M^T M\right)^{-1}$$

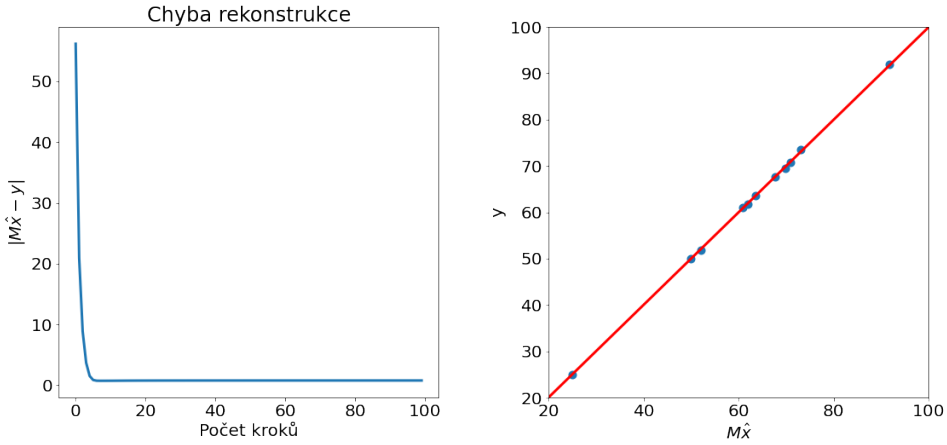
Jediné, co se změnilo, bylo nahrazení I za matici $\widehat{V}$. Parametr μ_x zůstane zatím stejný, na $\widehat{V}$ nezávisí. Parametry α_j a β_j zjistíme připodobněním následovně:

$$\alpha_{v_j} : \alpha_{v_j} ln(v_j) = \alpha_{0,v_j} ln(v_j) + \frac{1}{2}ln(v_j)$$
$$\alpha_{v_j} = \alpha_{0,v_j} + \frac{1}{2}$$

$$\beta_{v_j} : \beta_{v_j} v_j = \beta_{0,v_j} v_j + \frac{1}{2}\widehat{x_j^2}v_j$$
$$\beta_{v_j} = \beta_{0,v_j} + \frac{1}{2}\widehat{x_j^2}$$

Máme tedy odvozenou rovnici pro obecné $\widehat{v_j}$. Pokud budeme počítat všechny $\widehat{v_j}$ po sobě, tak nezáleží na pořadí. K výpočtu pak potřebujeme první moment $\widehat{v_j}$. Je to *Gamma* rozdělení, tudíž první moment zjistíme následovně:

$$\widehat{v_j} = \frac{\alpha_j}{\beta_j}.$$



(a) Vývoj rekonstrukční chyby

(b) Bodový graf y vs $M\widehat{x}$

Obrázek 2.2: Metriky modelu s řídkostí pro bayesovskou regresi

Pro získání matice $\widehat{V}$ nám již jen stačí jednotlivé $\widehat{v_j}$ naskládat na diagonálu.

Pro urychlení výpočtu však uděláme malou úpravu. Vzhledem k tomu, že v praktické části budu na výpočty používat knihovnu Numpy v jazyce Python, je vhodné většinu výpočtů vektorizovat tak, aby byly rychleji dokončené. Proto než rovnice prohlásíme za dostatečně upravené, celý výpočet si zvektorizujeme následovně:

$$\alpha = \alpha_0 \mathbf{1}_n^T + \frac{1}{2}\mathbf{1}_n^T$$
$$\beta = \beta_0 \mathbf{1}_n^T + \frac{1}{2} diag(\widehat{xx^T})$$

Kde $\mathbf{1}_n$ je vektor jedniček rozměru n a $diag()$ je operátor, který z matice vybere pouze diagonálu. Výsledné $\widehat{V}$ pak zjistíme jako $V = diag(\alpha\beta^{-1})$.

2.4.1 Výsledky

Algoritmu potřebujeme nově přidat α_0 a β_0 jako počáteční parametry apriorna. Taktéž potřebujeme pro naše výpočty $V^{(0)}$. Tvarovací parametry zvolíme jako malá čísla, zhruba 10^{-10}. Startovací $V^{(0)}$ zvolíme jako I, protože při volbě $\omega(0)$ se snažíme matice $M^T M$ a V znormovat na podobnou velikost. Jiná volba, především v jiných řádech by mohla algoritmus dostat do nějakého lokálního minima a tomu bychom se rádi vyvarovali.

Z Obrázků 2.2 si můžeme všimnout, že došlo k velkému zlepšení. Samotná chyba $|M\widehat{x} - y|$ v posledním kroku je 3. Což je dobré, vzhledem k šumu, který

ovlivňuje y. V bodovém grafu vidíme krásnou přímku s minimem deviací, což je přesně to, co bychom chtěli vidět. Znamená to, že naše rekonstrukce je velice dobrá. Nalezené $\widehat{x}$ vypadá následovně:

$$\widehat{x} = \begin{pmatrix} -0.01 & 0 & 1 & -0.01 & 1.01 & 1.01 \end{pmatrix}^T$$

Můžeme si všimnout, že v místech, kde v x_{true} byly nuly, model s řídkostí odhadnul nuly. Jediné, co model nezvládl, je druhý prvek. Ale to je očekáváné, matice M má totiž druhý sloupec nulový a v této sekci jsme model v případě nedostatku dat utahovali k apriornu. Pokusíme se to však opravit v Sekci 2.6. Nicméně objevila se nám tu malinká chybička. Dostali jsme se s čísly do záporu. To nám zde nevadí, ale na reálných datech to moc smysl nedává.

2.5 Nezápornost

Při modelování pracujeme při hledání $\widehat{x}$ s *Normálním* rozdělením. To však může nabývat i hodnot, které pro tuto doménu nedávají smysl. Když odhadujeme emisi, nemůže být záporná. To z logiky problému nedává smysl. Pokud si model myslí, že by měla být emise záporná, tak skoro jistě záporná není, ale též nebude vysoká. Máme několik způsobů, jak toto opravit. Prvním je zvolit takové rozdělení, jež nabývá hodnot od 0 do nekonečna. Druhá možnost je omezit jednotlivé prvky x_i zdola 0. Což je sice jednoduší na implementaci, nicméně je to heuristika. Budeme tedy muset změnit rozdělení pro x.

Naštěstí toho nebudeme muset měnit moc. Existuje totiž *Omezené Normální rozdělení*. To je rozdělení, které má tvar *Normálního*, ale existuje jen na námi zvoleném definičním oboru (support). *Omezené Normální* rozdělení s tvarovacími parametry μ, σ na omezeném definičním oboru $a < x \leq b$ vypadá následovně [7]:

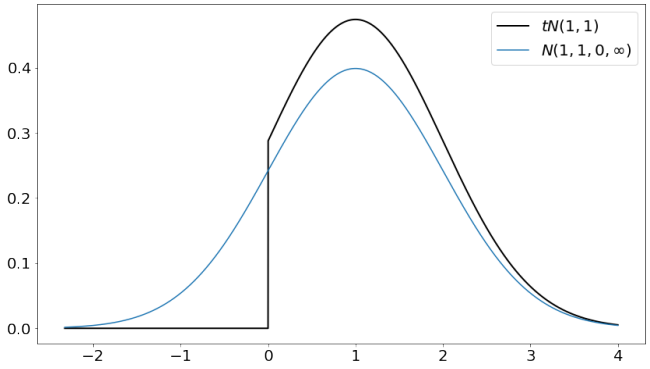
$$f(x|\mu,\sigma,a,b) = \frac{\sqrt{2}exp\left(-\frac{1}{2}\frac{\left(x-\mu\right)^2}{\sigma}\right)}{\sqrt{\pi\sigma}\Big(erf(\alpha) - erf(\beta)\Big)}\chi_{(a,b]}(x)$$

kde α a β je definováno následovně:

$$\alpha = \frac{a-\mu}{\sqrt{2\sigma}}$$
$$\beta = \frac{b-\mu}{\sqrt{2\sigma}}$$

Momenty se pak počítají jako:

$$\widehat{x} = \mu - \sqrt{\sigma}\varphi(\mu,\sigma)$$
$$\widehat{x^2} = \sigma + \mu\widehat{x} - \sqrt{\sigma}\kappa(\mu,\sigma)$$



Obrázek 2.3: Srovnání pravděpodobnostní funkce pro $N(1,1)$ a $tN(1,1,0,\infty)$

S funkcemi φ a κ:

$$\varphi(\mu,\sigma) = \frac{\sqrt{2}\Big(exp(-\beta^2) - exp(-\alpha^2)\Big)}{\sqrt{\pi}\Big(erf(\beta) - erf(\alpha)\Big)}$$

$$\kappa(\mu,\sigma) = \frac{\sqrt{2}\Big(b\ exp(-\beta^2) - a\ exp(-\alpha^2)\Big)}{\sqrt{\pi}\Big(erf(\beta) - erf(\alpha)\Big)}$$

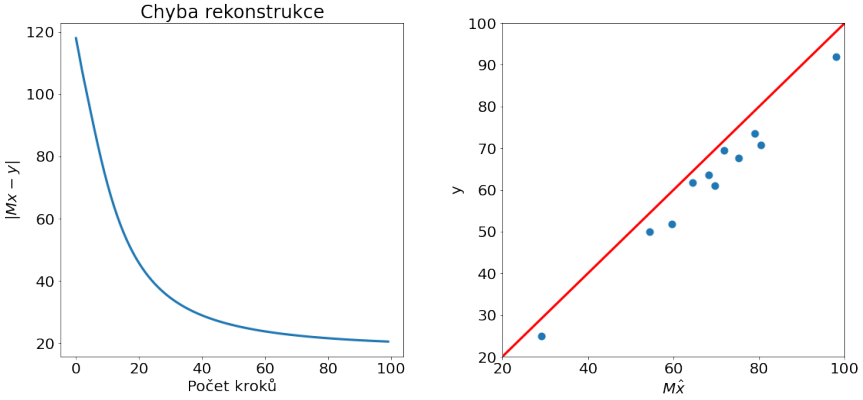
kde erf je error funkce, definovaná následovně:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Víme tedy, jak matematicky popsat toto rozdělení. Tvar rozdělení můžeme vidět v Obrázku 2.3. Je zde vidět rozdíl mezi obyčejným Normálním rozdělením a Omezeným Normálním rozdělením. Normální rozdělení je nám již známé, pro lepší ilustraci má střední hodnotu 1. Omezené rozdělení má stejné parametry, akorát je na definičním oboru $0 < x \leq \infty$. Můžeme si všimnout vyššího kopečku typického pro normální rozdělení, tím se kompenzuje to, že není definované pro hodnoty menší než 0.

Pro náš výpočet budeme uvažovat $a = 0$, $b = \infty$. Emise totiž nemůžou být záporné, vrchní limit však není. Ve výpočtu modelu se změní akorát způsob, jakým počítáme momenty $\widehat{x}, \widehat{x^2}$. Když se totiž podíváme na rovnici pro toto rozdělení, můžeme si všimnout, že v základu je stejná jako pro Normální rozdělení.

V Obrázku 2.4 můžeme vidět, jak se výpočet chová s tímto rozdělením. Je vidět pomalejší pokles rekonstrukční chyby, to je způsobeno omezením volnosti modelu. Ale my budeme raději, když dostaneme horší výsledky, které se



(a) Vývoj rekonstrukční chyby (b) Bodový graf y vs $M\widehat{x}$

Obrázek 2.4: Metriky modelu s řídkostí a Omezeným normálním rozdělením pro bayesovskou regresi

dají vysvětlit, než kdybychom měli prezentovat výsledky co v realitě nedávají smysl. Odhadnuté $\widehat{x}$ pak vypadá následovně:

$$\widehat{x} = \begin{pmatrix} 0.03 & 0.8 & 1 & 0.02 & 0.98 & 1 \end{pmatrix}^T$$

Dokonce se nám zlepšil i odhad pro druhý prvek. V bodovém grafu jsme se trošku zhoršili, ale na reálných datech by takový výsledek byl stále dobrý.

2.6 Nenulové apriorno pro x

Máme model, který obstojně dokáže vyřešit problém lineární regrese a nevadí mu tolik typické problémy pro lineární regresi. V atmosferickém modelování však máme často velmi špatně podmíněné problémy, které mají spoustu lokálních extrémů, a zároveň máme k dispozici apriorní odhad emise. Aktuálně jsme naší apriorní znalost o x nijak nepoužili, řekli jsme si, že nevíme nic, a použili nulové apriorno. To ale může našemu modelu i škodit, hlavně v případech reálných dat, kde model v nejednoznačných případech bude tíhnout spíše k nulovým emisím. Kdybychom byli schopni nastavit apriorní emisi na nějaký odhad získaný z literatury, mohli bychom si přilepšit. Ale pozor, toto může být dvousečné. Pokud bychom nastavili apriorní střední hodnotu x nějak extrémně špatně, může se stát, že nulové apriorno, které jsme používali doteď, bude lepší odhad.

Protože nás čeká znova odvození skoro všech parametrů, je vhodné si připomenout, jaké parametry máme a jakými apriorními rozděleními je modelu-

jeme:

$$f(y) = N(Mx, \omega^{-1}I) \propto exp\Big(-\frac{1}{2}\omega\Big(y - Mx\Big)^T\Big(y - Mx\Big)\Big)$$

$$f(x) = tN(x_0, V^{-1}, 0, \infty) \propto exp\Big(-\frac{1}{2}\Big(x - x_0\Big)^T V\Big(x - x_0\Big)\Big)$$

$$f(\omega) = G(a_0, b_0) \propto exp\Big(-b_0\omega + \Big(a_0 - 1\Big)ln\Big(\omega\Big)\Big)$$

$$f(v_j) = G(\alpha_0, \beta_0) \propto exp\Big(-\beta_0 v_j + \Big(\alpha_0 - 1\Big)ln\Big(v_j\Big)\Big)$$

Marginála pro $\widehat{\omega}$ vypadá:

$$\tilde{f}(\omega|y) \propto exp\Big(-b_0\omega + \Big(a_0 - 1\Big)ln\Big(\omega\Big) - \frac{1}{2}\omega Tr\Big(y^T y - 2M\widehat{x}y + \widehat{x^T x}M^T M\Big)\Big)$$

Tato marginála zůstala stejná. Pojďme se dále věnovat parametru x:

$$\tilde{f}(x|y, V^{-1}) \propto exp\Big(-\frac{1}{2}\Big(x - x_0\Big)^T\widehat{V}\Big(x - x_0\Big) - \frac{1}{2}\widehat{\omega}\Big(y - Mx\Big)^T\Big(y - Mx\Big)\Big)$$

$$\propto exp\Big(-\frac{1}{2}\Big(x^T\widehat{V}x\Big) - x^T\widehat{V}x_0 - \frac{1}{2}\widehat{\omega}\Big(-2x^T M^T y - x^T M^T Mx\Big)\Big)$$

Marginála obsahuje stále jen členy lineární a kvadratické. Můžeme tedy připodobnit k *Normálnímu* rozdělení:

$$\sigma_x : -\frac{1}{2}x^T Vx - \frac{1}{2}\widehat{\omega}x^T M^T Mx = -\frac{1}{2}x^T\sigma_x x$$

$$\Big(\widehat{V} + \widehat{\omega}M^T M\Big)^{-1} = \sigma_x$$

$$\mu_x : x^T\widehat{V}x_0 + \widehat{\omega}x^T M^y = x^T\sigma^{-1}\mu_x$$

$$\sigma_x\Big(\widehat{\omega}M^T y + \widehat{V}x_0\Big) = \mu_x$$

Jediná změna je tedy v parametru μ_x. Výpočet momentů nalezneme v Sekci 2.5.

Dále nás čeká předělat přepočet parametrů v_j. Marginála vypadá následovně:

$$\tilde{f}(v_j|\widehat{x}) \propto exp\Big(-\beta_0 v_j + (\alpha - 1)ln(v_j) + \frac{1}{2}ln(v_j) - \frac{1}{2}v_j(\widehat{x_j x_j} - 2\widehat{x_j}x_{0,j} + x_{0,j}^2)\Big)$$

Tvarovací parametry Gamma rozdělení, kterému je marginála podobná:

$$\alpha_{v_j} : \alpha_{v_j} ln(v_j) = \alpha_{0,v_j} ln(v_j) + \frac{1}{2} ln(v_j)$$
$$\alpha_{v_j} = \alpha_{0,v_j} + \frac{1}{2}$$

$$\beta_{v_j} : \beta_{v_j} v_j = \beta_{0,v_j} v_j + \frac{1}{2}(\widehat{x_j x_j} - 2\widehat{x_j} x_{0,j} + x_{0,j}^2) v_j$$
$$\beta_{v_j} = \beta_{0,v_j} + \frac{1}{2}(\widehat{x_j x_j} - 2\widehat{x_j} x_{0,j} + x_{0,j}^2)$$

Parametr α_{v_j} se nám nezměnil. Zato parametr β_{v_j} se o dost zkomplikoval. Zbývá nám již jen převést rovnici pro obecné β_{v_j} na operace s vektory, kvůli již dříve zmiňované rychlosti v knihovně Numpy. To bude vypadat následovně:

$$\beta = \beta_0 \mathbf{1} + \frac{1}{2}\Big(diag(\widehat{xx^T}) - 2\widehat{x} \odot x_0 + x_0 \odot x_0\Big)$$

Kde $\odot$ je násobení po prvcích. Výsledné $\widehat{V}$ pak zjistíme jako $\widehat{V} = diag(\alpha \cdot \beta^{-1})$.

Pojďme si vyzkoušet, jestli naše snaha nebyla zbytečná. x_0 získáme náhodným zašuměním x_{true}, musíme si totiž nějak nasimulovat, že naše x_0 nemusí být perfektní. Vyšel nám následující vektor:

$$x_0 = \Big(0.08 \quad 0.87 \quad 1.02 \quad 0.22 \quad 0.95 \quad 1.08\Big)^T$$

Není úplně perfektní, ale to jsme ani nechtěli. V našem problému totiž apriorní odhady nebudou úplně perfektní.

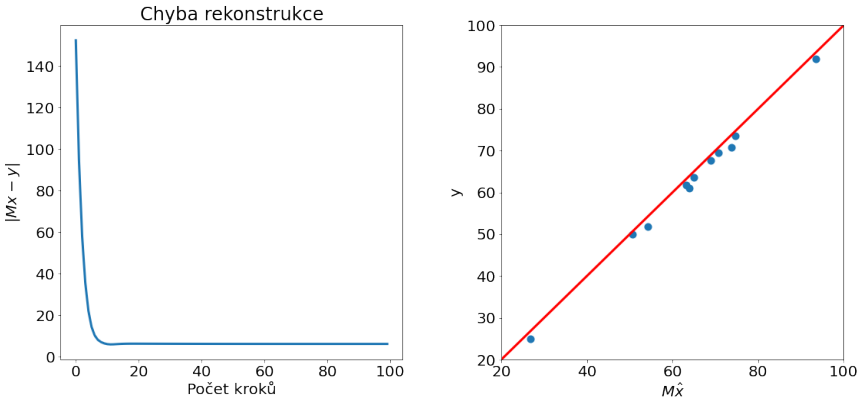
Obrázek 2.5 ukazuje výkonnost tohoto modelu. Můžeme si všimnout, že konvergence je zase rychlejší. Máme však i vlastnost nezápornosti a nenalezneme tedy nereálné hodnoty. Nalezené $\widehat{x}$ vypadá takto:

$$\widehat{x} = \Big(0.03 \quad 1.13 \quad 1.02 \quad 0.05 \quad 0.94 \quad 1\Big)^T$$

Je vidět, že $\widehat{x}$ je ovlivněné naším dodaným x_0. Ještě si však ukážeme, že je možné výpočet špatným x_0 zkazit. Použitím více zašuměného x_0:

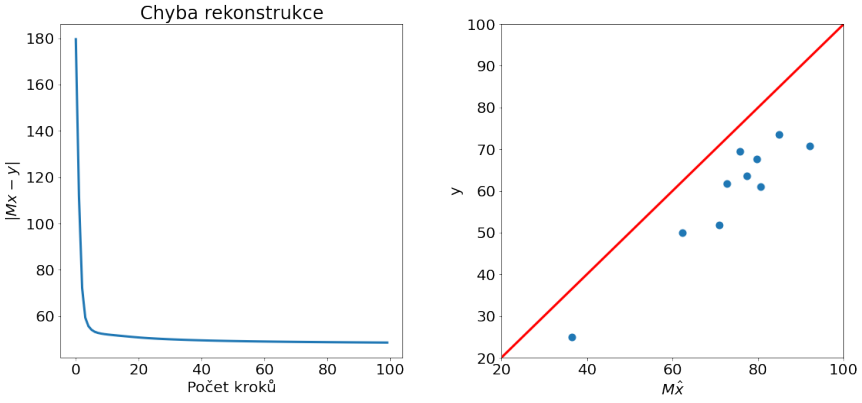
$$x_0 = \Big(0.41 \quad 2.38 \quad 1 \quad 1.9 \quad 1.54 \quad 1.08\Big)^T$$

které vůbec nepopisuje původní vektor x_{true} dostaneme výsledky mnohem horší. To můžeme vidět na Obrázku 2.6. Ačkoliv podle bodového grafu to zas tak špatně nevypadá, protože stále vidíme náznak přímky a na reálných datech bude bodový graf vypadat mnohem hůře, graf chyby rekonstrukce ukazuje, že chyba, kterou jsme udělali, není úplně malá. Je tedy potřeba k x_0 přistupovat velmi opatrně, protože je možné, že se dostaneme k výsledku, který kvalitou odrazí kvalitu našeho x_0.



(a) Vývoj rekonstrukční chyby (b) Bodový graf y vs $M\widehat{x}$

Obrázek 2.5: Metriky finálního modelu pro bayesovskou regresi



(a) Vývoj rekonstrukční chyby (b) Bodový graf y vs $M\widehat{x}$

Obrázek 2.6: Metriky finálního modelu pro bayesovskou regresi se špatným x_0 jako vstup

Kapitola 3

Výsledky

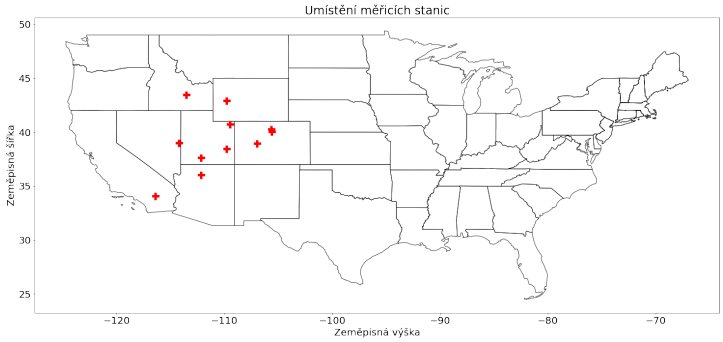
V předchozí kapitole jsme si odvodili model, který v této kapitole použijeme. Nejdříve však pár slov o datech, které budou reprezentovat vektor y a atmosférickém modelu šíření částic, což je naše matice M.

3.1 Data

Ve Spojených Státech Amerických jsou národní parky pro Američany součást jejich domovské hrdosti. Proto když se začalo mluvit o tom, že mikroplasty se objevují téměř všude po světě, tak zpozornili a instalovali na 11 míst přístroje měřicí depozici mikroplastů a mikrovláken [4]. Tyto přístroje umí měřit depozici za sucha, kde interval měření je měsíční, tak i za deště, kdy byl interval měření týden. Ze všech vzorků 98% obsahovalo nějaké množství mikroplastů nebo mikrovláken. U mikrovláken se dá kvůli velikostem, které byly nalezeny, říci, že pochází převážně z regionální emise (do 1000 km), protože jsou relativně velké, od 20 µm do zhruba 3 mm. Mikroplasty však jejich velikost od 4 µm do 25 µm řadí mezi částice, které se můžou šířit na velké vzdálenosti. Díky tomuto předpokladu můžeme modelovat emise na velké vzdálenosti.

Pro suchou depozici je k dispozici 103 měření z časového úseku 2018-04-01 až 2020-02-12 ve formátu YYYY-MM-DD. Pro mokrou depozici máme měření více, 203, z období 2017-08-13 až 2019-02-05. V [4] tvrdí, že 98% dat obsahovalo nějaké množství mikroplastů, ale data obsahovala zhruba 15% nulových datových bodů pro suchou depozici a 50% pro mokrou. Tyto chyby jsou často způsobené různými metodikami měření, např. některé stanice absenci měření značí nulou a jiné zase takovýto záznam do dat neuvedou vůbec. Proto jsem se rozhodl tyto nulové záznamy neuvažovat.

Též chci modelovat roční emisi. Proto jsou datové body omezené na jeden rok. Uvažuji data od 2018-10-28 do 2019-09-27 pro data suché depozice a 2018-02-01 až 2019-01-31 pro depozici mokrou. Chtěl jsem nejdříve modelovat celý rok 2019 na datech suché depozice a rok 2018 na datech mokré, ale mezi



Obrázek 3.1: Umístění měřicích stanic

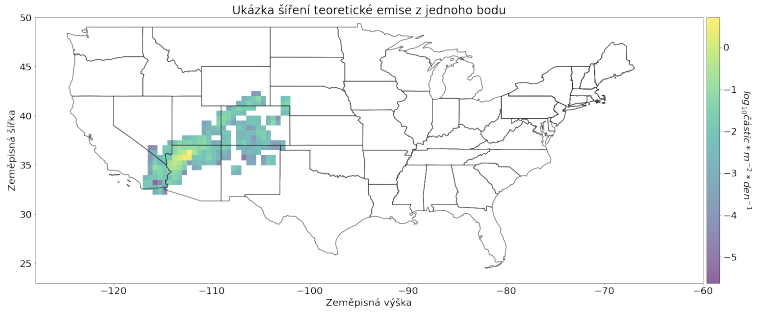
červencem a prosincem 2019 je velmi málo datových bodů v datasetu suchých měření. To může být chyba v datech, případně mohlo zrovna tento rok hodně pršet a nemohly se tedy dělat suchá měření. U mokrých dat jsem taktéž vybral období, které bylo nejvíce obsáhlé.

V Obrázku 3.1 si můžeme všimnout, kde jsou jednotlivé měřicí stanice rozmístěné. Stanice nejsou rozmístěné rovnoměrně. Také nám chybí pokrytí východní části Spojených Států Amerických. Proto nebudu počítat emise pro východní část USA. Nemá smysl počítat emisi z druhé strany kontinentu, pokud tam nemáme měřicí přístroje. Totiž citlivosti, které jsou v modelu šíření částic, jsou velice nízké, výpočet bývá nestabilní a výsledkům nemůžeme tolik věřit. Měřicí stanice budu nadále zakreslovat do mapy odhadnuté emise. Pokud ve výsledcích bude někde vysoká emise, ale bude v bodě vzdáleném od měřicích stanic, je to pravděpodobně způsobeno nestabilním výpočtem.

Každý záznam má uvedeno počet částic na metr čtvereční na den. Atmosferický model však pracuje s různými velikostmi částic, protože menší částice jsou schopné dostat se dále než ty větší. Proto použijeme rozdělení z literatury [12], ze kterého zjistíme, kolik částic v jaké kategorii je. Kategorie se dělí na částice <10 µm, 10-25 µm, 25-50 µm, 50-100 µm a 100-250 µm. Rozdělení je pak různé pro suchou i mokrou depozici. Kategorie jsou číslované od 1 do 5 v pořadí od nejmenšího.

3.2 Model šíření částic

Jako citlivostní model byl použit částicový disperzní model FLEXPART 10.4 [13]. Vstupem jsou mu meteorologická data s časovým intervalem 1h. Výstu-



Obrázek 3.2: Model šíření jednotkové částice z oblasti 0.5°×0.5°

pem je pak atmosferický citlivostní model šíření částic. Model FLEXPART je open-source projekt vyvinutý v roce 1995. Validita modelu byla poprvé ověřována ve studii v roce 1998 [14], kdy zkoušeli kontrolovaně vypouštět částice a zjistili jak moc se liší skutečnost od modelu. Aktuální verze 10.4 dokáže pracovat se třemi rozlišeními 1°×1°,0.5°×0.5° a 0.1°×0.1°, podle toho, jak častá data máme na vstupu. Tato verze také umí pracovat s turbulencemi, prouděním a nově i právě s měřením depozice. Rozlišení, které pro výpočet používáme, je 0.5°×0.5°. Pro každou kategorii uvedenou v 3.1 je pak potřeba napočítat samostatný model šíření. Též je potřeba ještě rozdělit suché a mokré šíření. Samotné nastavení parametrů modelu je uvedeno v [12].

Z atmosférického modelu šíření můžeme dostat 2 pohledy. První je teoretické šíření jednotkové částice z jednoho bodu v prostoru. Toto se pěkně vizualizuje a ukazuje nám to, kam všude by se částice dostaly, pokud by byly vypuštěny z daného bodu. Toto je vhodné pro vizualizaci, protože rozměry jsou zeměpisná šířka a výška, můžeme to vidět v Obrázku 3.2. Druhý pohled je ten, který budeme používat k výpočtu. Zde jsou uvedené teoretické citlivosti pro každou měřicí stanici v závislosti na emisi v dané oblasti. To nemůžeme hezky vizualizovat, protože na jedné ose je citlivost na specifické stanici a na druhé je čas.

3.3 Apriorní emise

V našem modelu potřebujeme jako jeden ze vstupů apriorní odhad emise x_0. U něho nepředpokládáme přesná čísla, protože je neznáme, a také je zpřesnění tohoto odhadu naším cílem. Zdrojem dat je [15, 16, 17]. Dle literatury jsou

největší zdroje mikroplastů a mikrovláken v atmosféře částice vzniklé transportací, zemědělstvím, částice z rozkladu plastů v oceánu a prachové částice spojené s lidskou populací.

Data k zemědělské emisi pochází ze studie [16]. V této studii agregovali výsledky více studií zaměřených na zemědělství. U zemědělství nejsou nutně všechny částice vznikem od zemědělské aktivity. Částice se do půdy můžou dostat například pomocí odpadních vod, odhazováním odpadků, či povodněmi, které s sebou vezmou plasty. Původ pak není čistě jen ze zemědělské činnosti. Nás však zajímá, ze kterých míst se částice dostanou do atmosféry, a když máme na mysli původ zemědělství, tak tím myslíme zemědělskou půdu. Jinak samozřejmě zemědělská aktivita jako taková také generuje odpad, největšími zdroji je odírání částic z pneumatik a plastový odpad obsažený v kompostu.

Zdroj z dopravy je způsoben především odíráním pneumatik při brždění a jízdě motorových vozidel. Ve Spojených Státech je auto typický způsob dopravy [18], proto je očekávané, že tento sektor produkuje velké množství mikroplastů. Opět i zde nás však zajímá jen emise ze země do atmosféry, takže všechny částice nutně nemusí pocházet z výše zmíněného způsobu.

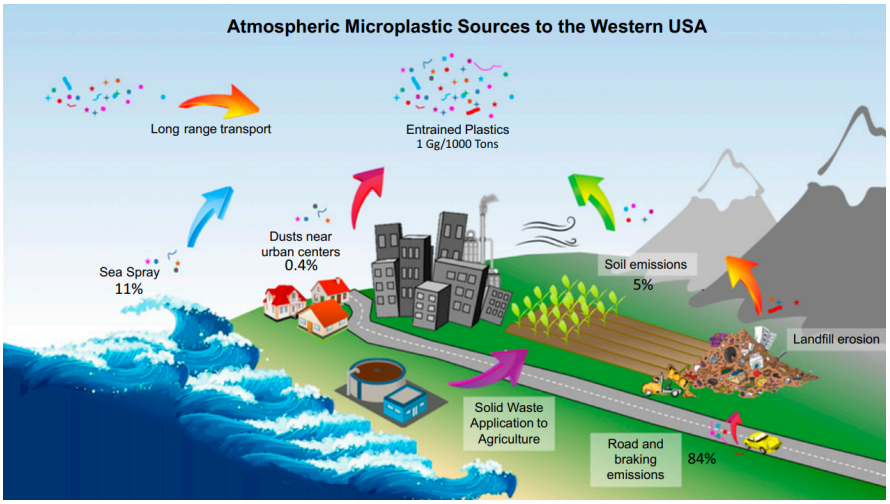
Dalším zdrojem emise, který budeme uvažovat, je oceán. Oceán samotný tedy žádné plastové částice neprodukuje, emise z něj je způsobena odpady, které v něm člověk nechal. Ty se pak vlivem externích jevů rozkládají na menší částice a pokud jsou dostatečně malé, můžou být větrem erodovány do atmosféry.

Poslední kategorií, z pohledu hrubých odhadů emisí však minimální, je prach ze zastavěných oblastí. Tím se myslí veškeré částice emitované z měst a vesnic.

3.3.1 Tvorba apriorního odhadu

Představili jsme si hlavní kategorie, které budou tvořit náš odhad. Data k těmto odhadům mají různá rozlišení, my je potřebujeme všechny sjednotit tak, abychom je mohli použít pro náš model. Dalším problémem jsou různé jednotky, ve kterých jsou samotná data uvedená. Některé studie uvádějí odhady v částicích na plochu, některé v hmostnosti na plochu, kde plocha i hmotnost se napříč datasety různí. Abych zamezil chybám v převodu jednotek, rozhodl jsem se všechny hodnoty v každém datasetu znormalizovat na hodnoty mezi 0 a 1. Tímto dostaneme místo jednotek, které použil autor, jednotky bezrozměrné. Dateset tedy vydělím maximem z datasetu. Hodnoty tedy nebudou prezentovat reálná čísla, ale bezrozměrná čísla reprezentující poměr emise k maximu z domény.

Dále potřebujeme získat váhy, které nám řeknou, jak jsou jednotlivé kategorie zastoupené. Použijeme váhy, které jsou z [15]. Ty můžeme vidět v Obrázku 3.3, který je ze stejné literatury. Můžeme si všimnout, že ~84% všech emisí tvoří doprava, ~11% se do atmosféry dostává z oceánu a jen ~0.4% je

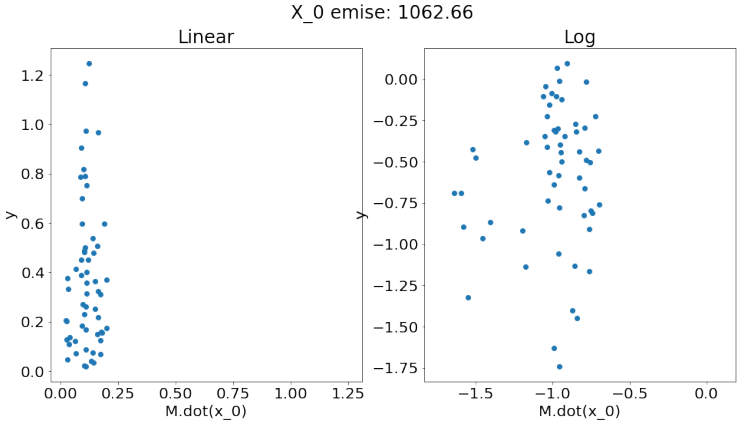


Obrázek 3.3: Procentuální zastoupení mikroplastů dle typu emise. Zdroj obrázku [15]

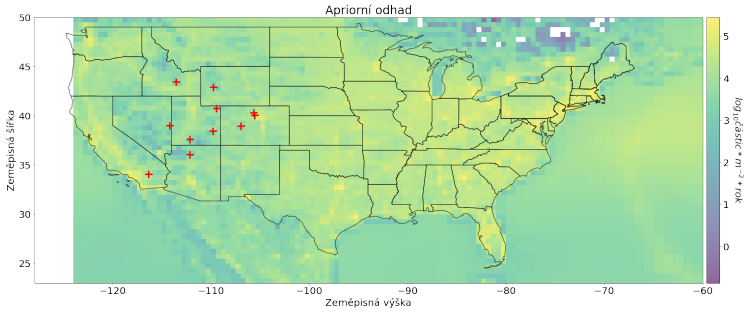
původem z měst. Ze zemědělské půdy se pak dostává do atmosféry zhruba ~5% částic.

Protože však máme bezrozměrná čísla, měli bychom je pro algoritmus zpátky převést na hodnoty, které mají podobný rozměr, jako naše data. Můžeme opět zvolit cestu přepočtu z literatury, ale já se rozhodl zkusit trošku jiný přístup. Pro získání koeficientu násobení jsem si zvolil výpočet na průměrné doméně s nulovým x_0. To znamená, že si spočítáme průměrné M z domény, kterou chceme počítat. Poté použijeme toto M pro úvodní výpočet celkové průměrné emise na naší zvolené doméně. Následně tímto průměrem vynásobíme naše nalezené x_0 a měli bychom se dostat na hrubý odhad, který je v jednotkách, jež potřebujeme.

Tím že máme již nějaký odhad, můžeme zkusit zjistit, jak dobrý je tento odhad. Pokud totiž vynásobíme tento odhad maticí M, dostaneme rekonstruované y. To si můžeme dát do bodového grafu tak, jak jsme to viděli u testovacích dat. Výsledky můžeme vidět v Obrázku 3.4. Můžeme si všimnout, že data moc nekorelují, odhadujeme nižší hodnoty oproti realitě. Ale tohle je zatím jen náš apriorní odhad, který nijak nesouvisí s daty, jež máme. Obrázek 3.5 ukazuje samotné x_0 na mapě USA. Tento apriorní odhad jsem konstruoval pro celou doménu, pro kterou jsem měl k dispozici model šíření částic M. Mapy odhadů pak již obsahují doménu zmenšenou, ale přišlo mi zajímavé vykreslit apriorní odhad celý.



Obrázek 3.4: Metriky hrubého apriorního odhadu. Obrázky sledují chybu rekonstrukce $M\hat{x}$ oproti původnímu y v lineární resp. logaritmické škále



Obrázek 3.5: Mapa apriorního odhadu

3.4 Statistické porovnání výsledků

Než budeme mít nějaké výsledky, bylo by dobré mít možnost výsledky statisticky ohodnotit. Sice v obrázku můžeme vidět korelaci, to však nelze statisticky porovnávat, což bychom chtěli. Umožnilo by nám to porovnat výsledky mezi sebou, nebo výsledky od apriorna. Použijeme pro to 3 metriky z [19], první je Pearsonův korelační koeficient. Ten je definován následovně:

$$R(a,b) = \frac{n\sum_{i=1}^{n} a_i b_i - \sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i}{\sqrt{n\sum_{i=1}^{n} a_i^2 - (\sum_{i=1}^{n} a_i)^2}\sqrt{n\sum_{i=1}^{n} B_i^2 - (\sum_{i=1}^{n} b_i^2)}}$$

kde n je velikost vektorů a či b. Tato metrika zjednodušeně říká, jak moc se bodový graf y a Mx podobá přímce se směrnicí 1. Nabývá hodnot od -1 do 1, hodnotu, kterou bychom rádi viděli, je právě 1.

Další metrikou bude rMSE, česky normalizovaná odmocnina ze střední kvadratické chyby. Předpis pro ni je:

$$nRMSE(a,b) = \frac{\sqrt{\sum_{i=1}^{n} \frac{1}{n}(b_i - a_i)^2}}{max(a) - min(a)}$$

kde jako vektor a musíme dosadit vektor dat y. U této metriky je pro ideální stav nula, proto budeme chtít vidět co nejnižší čísla.

Poslední metrikou kterou budeme uvažovat je MFB, v angličtině jako mean fractional bias. Tato metrika je vhodná, protože dává stejné váhy přehnaným i nedostatečným odhadům. Její předpis je:

$$MFB(a,b) = \frac{1}{n}\frac{\sum_{i=1}^{n}(b_i - a_i)}{\sum_{i=1}^{n}\frac{a_i + b_i}{2}}$$

Rozsah nabývá od -200% do 200%, ideální stav je při hodnotě 0.

Když máme tyto metriky definované, můžeme se podívat, jak vypadá náš apriorní odhad x_0. Pro každý bod z domény spočítáme rekonstrukci y, tyto vektory sečteme a vydělíme velikostí domény. Tento vektor následně použijeme v metrikách jako vektor b. Výsledky jsou následující:

$$\begin{aligned} R &= -0.4 \\ nRMSE &= 25.3 \\ MFB &= 0.03 \end{aligned}$$

Můžeme si všimnout, že metriky tohoto odhadu nejsou ideální, krom metriky MFB, která vypadá dobře.

3.5 Suchá depozice

Dostáváme se tedy k výsledkům odhadů emisí. Jako první si rozebereme suchou depozici, též ji budu značit jako dataset DRY. Z Kapitoly 2 víme, že jako vstup potřebujeme několik parametrů. x_0 jsme již získali. Dále potřebujeme úvodní tvarovací parametry pro ω a všechna v_j. Všechny nastavíme na hodnotu 10^{-10}. Také kvůli pořadí výpočtu potřebujeme zvolit úvodní $V(0)$ a $\omega(0)$. To zvolíme stejně, jako jsme to volili v Kapitole 2, pro připomenutí, $V(0) = I$ a $\omega(0) = 1/max(M^T M)$.

Též si vstupní M a y znormalizujeme pomocí maxima z M. Toto je ekvivalentní úprava, výsledek zůstane stále v částicích na metr čtvereční za den. Slibuji si od toho stabilnější výpočet v začátku, kde volba špatných parametrů může model poslat do lokálního minima. Jak již bylo zmíněno, matici M omezíme na zmíněné období od 2018-9-28 do 2019-9-27. Pro urychlení výpočtu můžeme z vektoru y odstranit měření, která neobsahují tato data. Doménu pro výpočet jsem zvolil na rozsah -124 až -98 v ose zeměpisné šířky a 28 až 50 v ose zeměpisné výšky. Nemá smysl počítat dál, citlivosti v matici M jsou velice nízké, protože nám chybí měřicí stanice ve východní části USA.

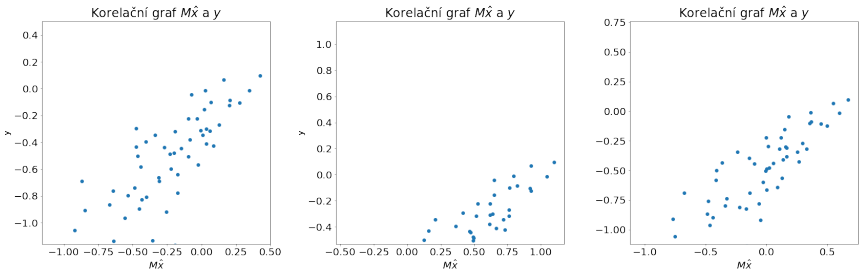
Pro každou dlaždici v doméně spočítáme odhad emise $\widehat{x}$. Protože jsme však v každém kroku výpočtu uvažovali, že všechna depozice je z aktuálně počítané dlaždice, tak je potřeba ještě výsledek vydělit velikostí domény. Výsledek můžeme vynásobit maticí M a podívat se, jak moc rekonstruované $M\widehat{x}$ koreluje s y z dat. To je vidět na Obrázku 3.6. Můžeme si všimnout, že pro velikosti částic 1, 3 a 4 je rekonstrukce slušná, vidíme podobnost s přímkou se směrnicí 1. Odhad mikroplastů velikosti 2 je trošku přehnaný a odhad pro velikost 5 moc nekoreluje. Je možné, že jsou částice již tak veliké, že zvolená doména byla moc velká a citlivosti v matici M jsou pro vzdálené hodnoty malé.

V Obrázku 3.7 zase můžeme vidět součet všech odhadnutých emisí. V Příloze B Obrázcích B.1, B.2, B.3, B.4, B.5 zase můžeme vidět jak se celková emise rozpadne na jednotlivé kategorie. V Obrázku B.6 zase můžeme vidět graf denních emisí v logaritmickém měřítku. Data se zdají být realistická, na denních grafech si můžeme všimnout, že pro některé dny chybí měření, model pak jako odhad použije dodané x_0. Pro tyto případy je vhodné, že aspoň nějaké x_0 máme, protože jinak bychom dostali jako odhad 0.

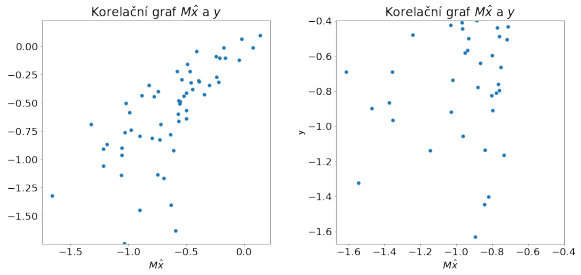
Můžeme se ještě podívat, jak model zlepšil úvodní x_0. Vybereme si např. dataset DRY-1, metriky vypadají následovně:

$$R = 0.811$$
$$nRMSE = 0.399$$
$$MFB = 0.011$$

Můžeme si všimnout, že metriky jsou docela pěkné, pearsonův korelační koeficient naznačuje korelaci, což značí rozumný odhad. Znormovaná střední kvadratická chyba je též velice nízká a u MFB se toho moc od x_0 nezměnilo. To naznačuje, že odhad získaný výpočtem je lepší, než apriorní odhad.

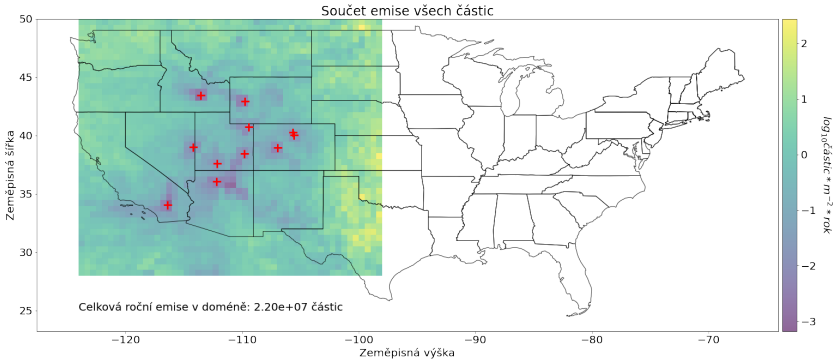


(a) DRY 1 (b) DRY 2 (c) DRY 3

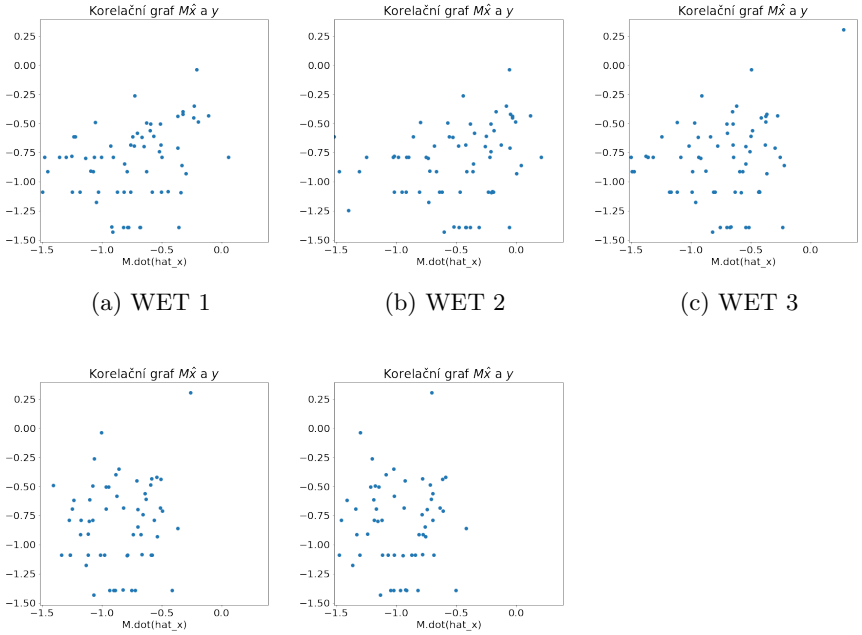


(d) DRY 4 (e) DRY 5

Obrázek 3.6: Bodový graf log chyb pro suchou depozici. Čísla značí skupiny velikosti částic definované v 3.1



Obrázek 3.7: Odhadnutá emise DRY



(a) WET 1 (b) WET 2 (c) WET 3

(d) WET 4 (e) WET 5

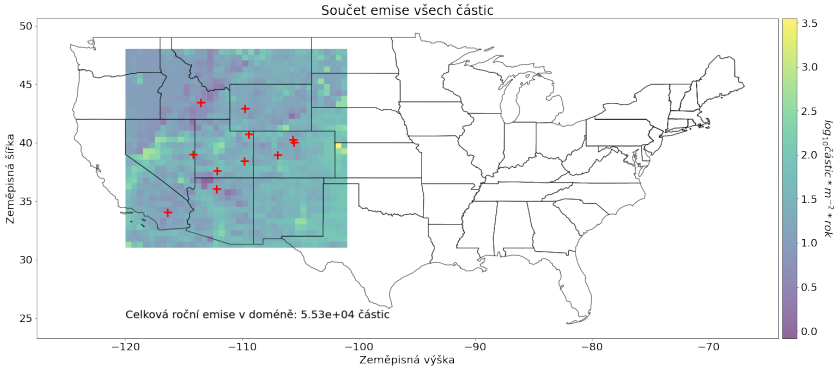
Obrázek 3.8: Bodový graf log chyb mokré depozice.

3.6 Mokrá depozice

Pro mokrou depozici použijeme stejné parametry jako pro suchou depozici. Tato data budu též nazývat jako data WET. Data jsou z jiného časového období, proto použijeme takové, které má nejvíce měření a začíná na začátku měsíce. Proto jsem se rozhodl použít období 01-02-2018 až 31-01-2019. Velikost domény však musí být trošku menší, částice se totiž pravděpodobně nešíří stejně dobře jako za sucha, a výpočet v okrajových bodech domény je nestabilní. Omezíme ho tedy na $-120°$ až $-101°$ zeměpisné výšky a $31°$ až $48°$ zeměpisné šířky. Výpočet bude probíhat stejným stylem, proto je potřeba udělat všechny kroky stejně, jako v Sekci 3.5.

Bodové grafy v Obrázku 3.8 naznačují, že výsledky nebudou úplně nejlepší. Obrázek 3.9 vypadá docela rozumně, u dat mokré depozice po prozkoumání Obrázků B.7 až B.11 to vypadá, že do součtu přispívají v podobných řádech. Můžeme si tedy udělat lepší obrázek o celkové emisi.

Po spočítání námi sledovaných metrik u datasetu WET-1 se dostaneme



Obrázek 3.9: Odhadnutá emise pro mokrou depozici

k následujícím výsledkům:

$$R = 0.75$$
$$nRMSE = 0.12$$
$$MFB = -0.01$$

Korelace vyšla vyšší, než jsem očekával, a ostatní metriky vykazují také dobrá čísla, proto odhad může odpovídat skutečnosti. Nicméně tato velikost částic vyšla nejlépe, pokud se podíváme např. na WET-4:

$$R = 0.40$$
$$nRMSE = 0.11$$
$$MFB = -0.02$$

vidíme, že korelace poklesla, což značí špatnou dekompozici. Odhad této velikosti mikroplastů je však velice obtížný, kvůli nízkým hodnotám citlivosti v matici M.

3.7 Shrnutí výsledků

Máme tedy odhady emisí pro suchou (DRY) i mokrou (WET) depozici a také pro jednotlivé velikosti částic. Tato čísla můžeme sečíst a získat tím odhad

roční emise na metr čtvereční pro naši doménu. Dostaneme se k odhadům

$$2.20 * 10^7 \text{ částic} \times m^{-2} \times rok \text{ pro suchou depozici}$$
$$5.53 * 10^4 \text{ částic} \times m^{-2} \times rok \text{ pro mokrou depozici.}$$

Tato čísla by šla dále přepočítat na tuny za rok pro účel validace výsledků a vyvozování závěrů o stavu přírody. To už je však mimo rozsah této práce.

Závěr

V této práci byl odvozen vzorec pro řešení lineární regrese pomocí metody VB. Mezi jeho hlavní vlastnosti patří řídkost, nezápornost a nenulové apriorno. Řídkost model ovlivňuje ve chvílích, kdy nemáme dostatek dat, tím, že volí hodnoty apriorního odhadu. Nezápornost potřebujeme kvůli výsledkům, emise nemůže být záporná. V neposlední řadě model využívá nenulové apriorno kvůli stabilizaci výpočtu v místech, kde není dostatek dat.

Tento model byl následně aplikován na data z depozice mikroplastů. Data jsou dosti řídká, interval měření pro suchou depozici je měsíc, pro mokrou je to týden. Měřicí stanice také pokrývají plochu Spojených Států Amerických nerovnoměrně, proto je obtížné odhadovat emisi z východu USA. I přes to se mi podařilo zlepšit apriorní odhad, který byl doposud znám, podle mnou zvolených metrik.

Na tuto práci by šlo navázat více způsoby. Prvním je složitější model, modelu bychom mohli přidat např. hladkost. Emise totiž obvykle v čase nejdříve postupně narůstá a pak také postupně klesá. Aktuální model s tímto nepočítá, denní grafy jsou tak plné vrcholků následovaných prudkým poklesem emise. Druhý způsob vylepšení odhadu jsou lepší data. Pokud bychom měli přesnější data, případně lépe rozmístěné měřicí stanice po doméně, mohli bychom dojít k lepšímu výsledku. Taktéž by šlo validovat získané odhady s jinými pracemi a odhad extrapolovat na větší doménu.

Bibliografie

1. RAGUSA, Antonio et al. Plasticenta: First evidence of microplastics in human placenta. *Environment International*. 2021, roč. 146, s. 106274.
2. ZHANG, Yulan; GAO, Tanguang; KANG, Shichang; ALLEN, Steve; LUO, Xi; ALLEN, Deonie. Microplastics in glaciers of the Tibetan Plateau: Evidence for the long-range transport of microplastics. *Science of The Total Environment*. 2021, roč. 758, s. 143634.
3. GONZÁLEZ-PLEITER, Miguel; EDO, Carlos; VELÁZQUEZ, David; CASERO-CHAMORRO, María Cristina; LEGANÉS, Francisco; QUESADA, Antonio; FERNÁNDEZ-PIÑAS, Francisca; ROSAL, Roberto. First detection of microplastics in the freshwater of an Antarctic Specially Protected Area. *Marine Pollution Bulletin*. 2020, roč. 161, s. 111811.
4. BRAHNEY, Janice; HALLERUD, Margaret; HEIM, Eric; HAHNENBERGER, Maura; SUKUMARAN, Suja. Plastic rain in protected areas of the United States. *Science*. 2020, roč. 368, č. 6496, s. 1257–1260.
5. MOODY, D. H. 40th Aniversary of Palomares. *Faceplate*. 2006, roč. 10, č. 2, s. 15–19.
6. GRINSTEAD, Charles Miller; SNELL, James Laurie. *Introduction to probability*. American Mathematical Soc., 1997.
7. ŠMÍDL, Václav; QUINN, Anthony. *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.
8. KOLASINSKI, WA; BLAKE, JB; ANTHONY, JK; PRICE, WE; SMITH, EC. Simulation of cosmic-ray induced soft errors and latchup in integrated-circuit computer memories. *IEEE Transactions on Nuclear Science*. 1979, roč. 26, č. 6, s. 5087–5091.
9. PETERKA, Václav. Bayesian approach to system identification. In: *Trends and Progress in System identification*. Elsevier, 1981, s. 239–304.

10. TICHÝ, Ondřej; ŠMÍDL, Václav; HOFMAN, Radek; STOHL, Andreas. LS-APC v1. 0: a tuning-free method for the linear inverse problem and its application to source-term determination. *Geoscientific Model Development*. 2016, roč. 9, č. 11, s. 4297–4311.

11. KULLBACK, Solomon; LEIBLER, Richard A. On information and sufficiency. *The annals of mathematical statistics*. 1951, roč. 22, č. 1, s. 79–86.

12. EVANGELIOU, Nikolaos; TICHÝ, Ondřej; ECKHARDT, Sabine; ZWAAFTINK, Christine Groot; BRAHNEY, Janice. Sources and fate of atmospheric microplastics revealed from inverse and dispersion modelling: From global emissions to deposition. *Journal of Hazardous Materials*. 2022, roč. 432, s. 128585.

13. PISSO, Ignacio et al. The Lagrangian particle dispersion model FLEXPART version 10.4. *Geoscientific Model Development*. 2019, roč. 12, č. 12, s. 4955–4997.

14. STOHL, Andreas; HITTENBERGER, Markus; WOTAWA, Gerhard. Validation of the Lagrangian particle dispersion model FLEXPART against large-scale tracer experiment data. *Atmospheric Environment*. 1998, roč. 32, č. 24, s. 4245–4264.

15. BRAHNEY, Janice; MAHOWALD, Natalie; PRANK, Marje; CORNWELL, Gavin; KLIMONT, Zbigniew; MATSUI, Hitoshi; PRATHER, Kimberly Ann. Constraining the atmospheric limb of the plastic cycle. *Proceedings of the National Academy of Sciences*. 2021, roč. 118, č. 16, s. e2020719118.

16. BÜKS, Frederick; KAUPENJOHANN, Martin. Global concentrations of microplastics in soils–a review. *Soil*. 2020, roč. 6, č. 2, s. 649–662.

17. EVANGELIOU, Nikolaos; GRYTHE, Henrik; KLIMONT, Zbigniew; HEYES, Chris; ECKHARDT, Sabine; LOPEZ-APARICIO, Susana; STOHL, Andreas. Atmospheric transport is a major pathway of microplastics to remote regions. *Nature communications*. 2020, roč. 11, č. 1, s. 1–11.

18. HU, Patricia; SCHMITT, Rolf R; SCHWARZER, Julianne; MOORE, William H et al. Transportation Statistics Annual Report 2021. 2021.

19. EVANGELIOU, Nikolaos et al. Changes in black carbon emissions over Europe due to COVID-19 lockdowns. *Atmospheric Chemistry and Physics*. 2021, roč. 21, č. 4, s. 2675–2692.

Příloha A

Seznam použitých zkratek

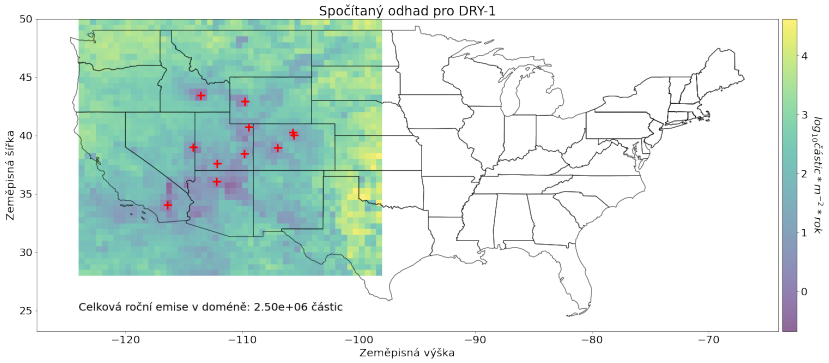
IVB iterační variační Bayes

VB variační Bayes

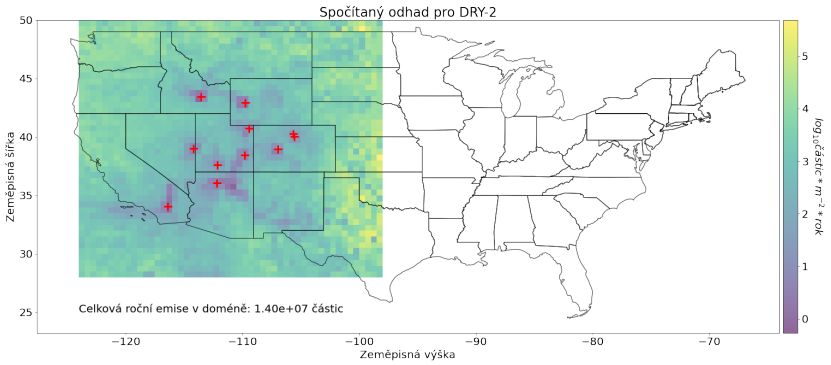
USA Spojené Státy Americké

Příloha B

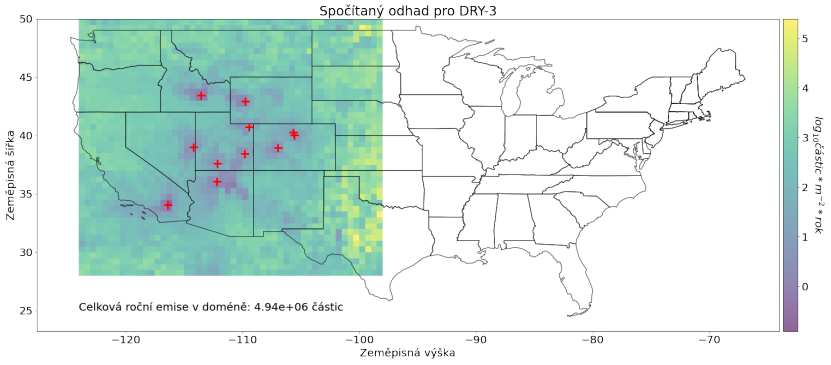
Přiložené obrázky



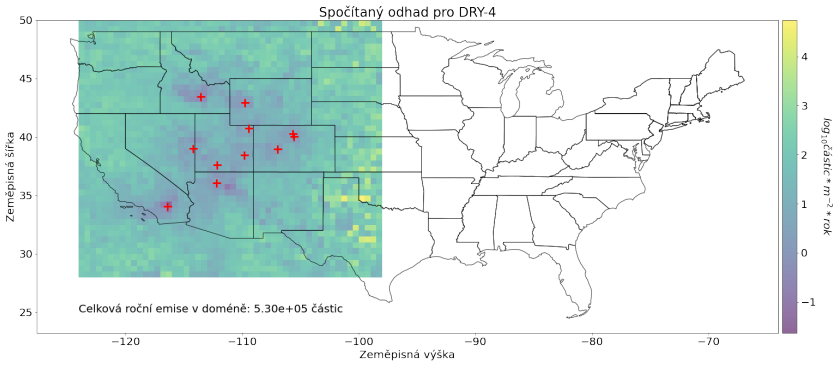
Obrázek B.1: DRY 1



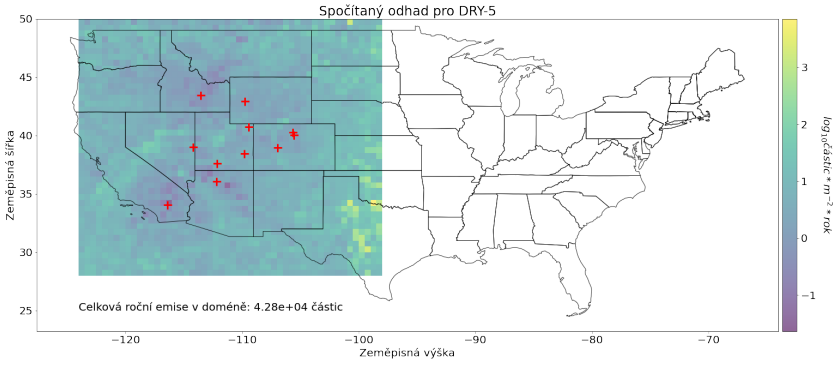
Obrázek B.2: DRY 2



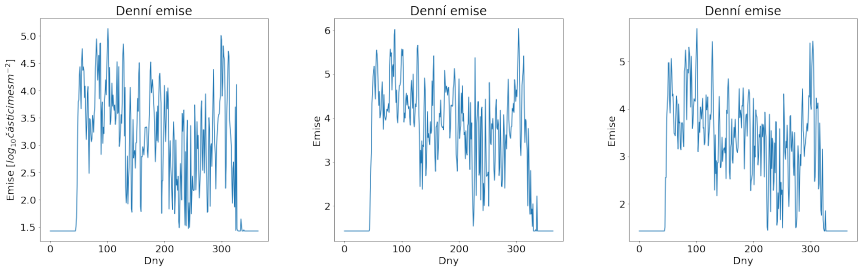
Obrázek B.3: DRY 3



Obrázek B.4: DRY 4



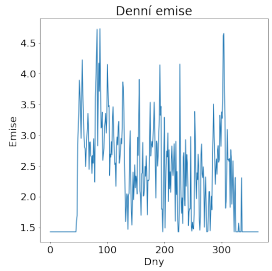
Obrázek B.5: DRY 5



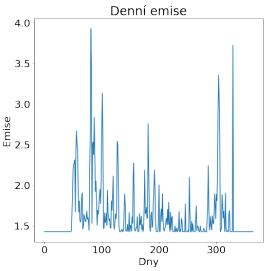
(a) DRY 1

(b) DRY 2

(c) DRY 3

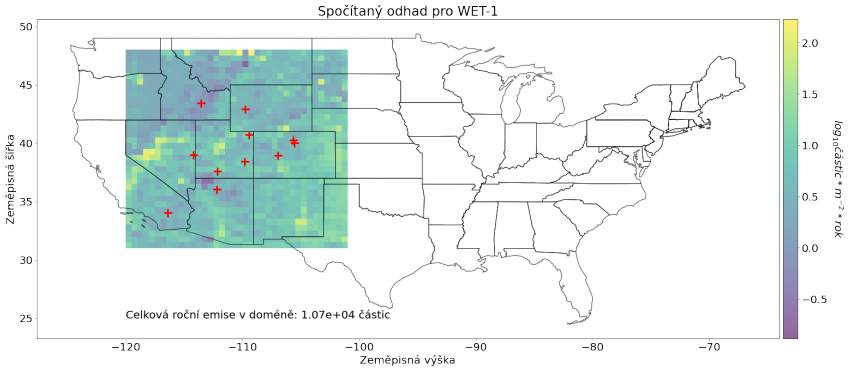


(d) DRY 4

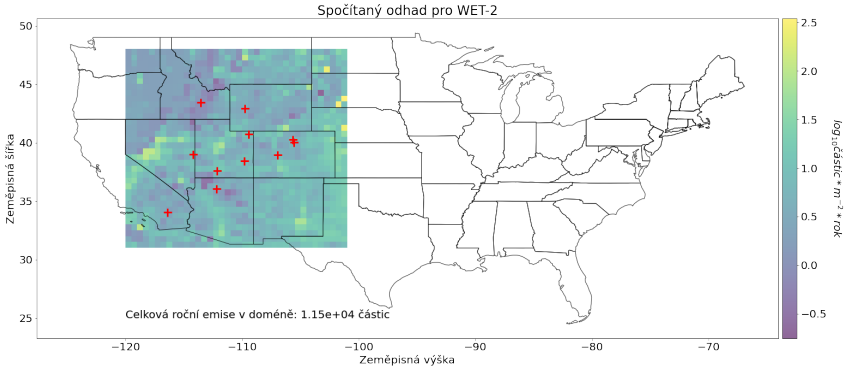


(e) DRY 5

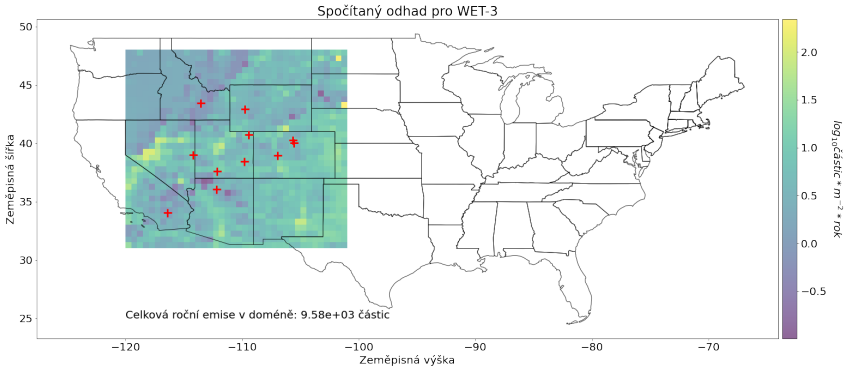
Obrázek B.6: Denní emise pro DRY v logaritmickém měřítku



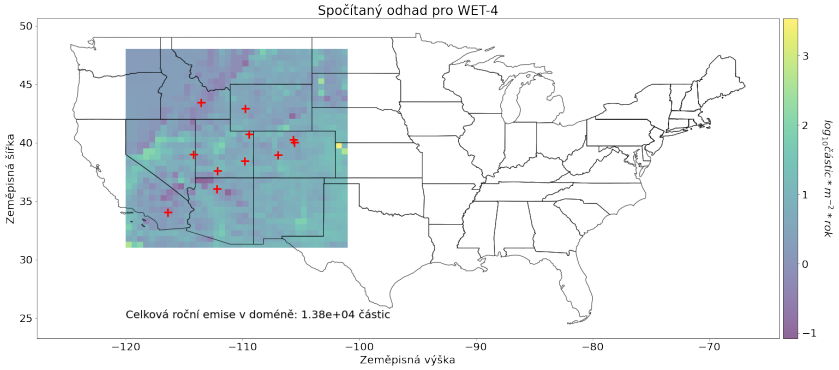
Obrázek B.7: WET 1



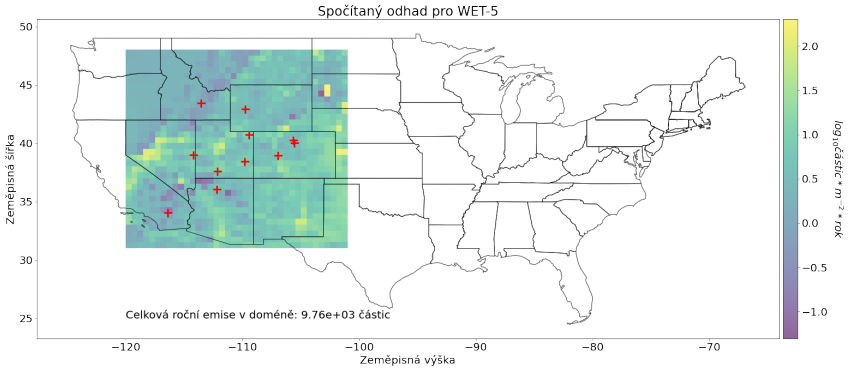
Obrázek B.8: WET 2



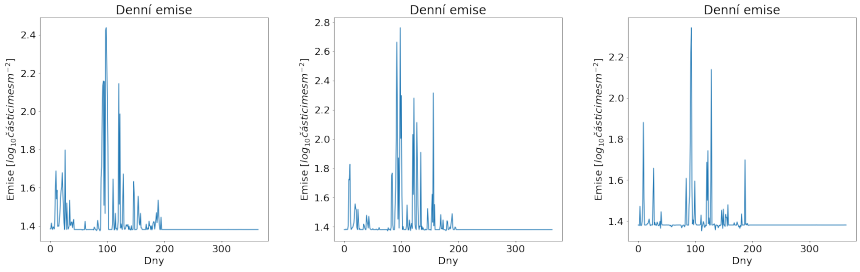
Obrázek B.9: WET 3



Obrázek B.10: WET 4



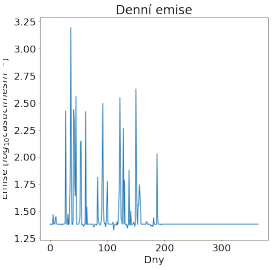
Obrázek B.11: WET 5



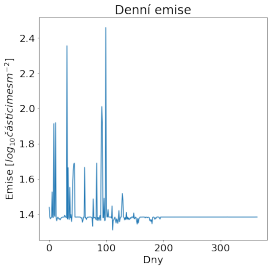
(a) WET 1

(b) WET 2

(c) WET 3



(d) WET 4



(e) WET 5

Obrázek B.12: Denní emise pro WET v logaritmickém měřítku

Příloha C

Obsah přiloženého CD

```
src ..... zdrojové kódy implementace
text ..... text práce
  thesis.pdf ..... text práce ve formátu PDF
```