FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of master's thesis

Title:	Bayesian model for atmospheric emission estimation
Student:	Bc. Tomáš Kořistka
Supervisor:	Ing. Ondřej Tichý, Ph.D.
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2023/2024

Instructions

The aim of the thesis is to study and improve methods for the estimation of atmospheric emissions from available concentration measurements of a monitored substance, e.g., chemicals, radionuclides, and gaseous substances. The output is an emission profile discretized in time along with the total emissions of the given substance. The emission estimate can be achieved by optimisation between measured values and numerical results of the atmospheric dispersion model. In the thesis, the student will study both, classical optimisation methods and, more importantly, the Bayesian approaches and parameter estimation methods of probability models (Tichý et al., 2016), respectively. The derived algorithms will be tested on synthetic data and subsequently applied to appropriately selected data from a real atmospheric emission, e.g., (Nodop et al., 1998; Masson et al., 2021).

1) Study the construction of the optimisation problem for estimating air emissions.
2) Learn about the Bayesian formulation of the optimisation problem and propose modifications for its application to the selected real dataset.
3) Implement the derived technique and design an experiment for emissions estimation using synthetic and subsequently real data.
4) Discuss the achieved results with respect to previous estimates in the literature and available information about the specific release.

Masson, O., Romanenko, O., Saunier, O., et al. (2021). Europe-wide atmospheric radionuclide dispersion by unprecedented wildfires in the Chernobyl Exclusion Zone,

Electronically approved by Ing. Magda Friedjungová, Ph.D. on 14 January 2023 in Prague.

April 2020. Environmental Science & Technology, 55(20), 13834-13848.

Nodop, K., Connolly, R., & Girardi, F. (1998). The field campaigns of the European Tracer Experiment (ETEX): Overview and results. Atmospheric Environment, 32(24), 4095-4108.

Tichý, O., Šmídl, V., Hofman, R., & Stohl, A. (2016). LS-APC v1.0: a tuning-free method for the linear inverse problem and its application to source-term determination. Geoscientific Model Development, 9(11), 4297-4311.

Master’s thesis

BAYESIAN MODEL FOR ATMOSPHERIC EMISSION ESTIMATION

Bc. Tomáš Kořistka

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: Ing. Ondřej Tichý,Ph.D.
May 3, 2023

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Kořistka Tomáš. *Bayesian model for atmospheric emission estimation.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2023.

Contents

Acknowledgments viii

Declaration ix

Abstract x

Summary xi

Introduction xii

1 Preliminary background 1
- 1.1 Atmospheric dispersion modeling 1
- 1.2 Chernobyl disaster 1
- 1.3 Effect on woods 2
  - 1.3.1 Forest fires 2
  - 1.3.2 Other threats 3
- 1.4 European tracer experiment 3

2 Regression 7
- 2.1 Training data 8
- 2.2 Linear regression 9
  - 2.2.1 Prediction 9
  - 2.2.2 Loss function 10
  - 2.2.3 Partial derivative 10
  - 2.2.4 Hessian matrix 11
  - 2.2.5 Collinearity conundrum 12
  - 2.2.6 Ridge regression 12

3 Bayesian theory 15
- 3.1 Bayes theorem 15
- 3.2 Exponential family of distributions 17
  - 3.2.1 Normal distribution 18
  - 3.2.2 Truncated normal distribution 19
  - 3.2.3 Multivariate normal distribution 20
  - 3.2.4 Gamma distribution 21
- 3.3 Conjugate prior 22
- 3.4 Methods of inference 23
  - 3.4.1 Variational Bayes 24

4 Controlled emission modeling 27
- 4.1 Implementation overview 27
- 4.2 Classic regression 28
  - 4.2.1 Formulation 28
  - 4.2.2 Performance 28

4.3 Bayesian ridge regression . . . 29
4.3.1 Establishing prior distributions . . . 30
4.3.2 Deriving posterior distributions . . . 31
4.3.3 Performance . . . 34
4.4 Bayesian sparse regression . . . 35
4.4.1 Establishing prior distributions . . . 36
4.4.2 Deriving posterior distributions . . . 37
4.4.3 Performance . . . 38
4.5 Bayesian smooth regression . . . 41
4.5.1 Establishing prior distributions . . . 41
4.5.2 Deriving posterior distributions . . . 42
4.5.3 Performance . . . 43
4.6 LDL . . . 45
4.6.1 Establishing prior distributions . . . 46
4.6.2 Deriving posterior distributions . . . 48
4.6.3 Performance . . . 50

5 Uncontrolled emission modeling 55
5.1 Blind run . . . 55
5.2 Classic ridge regression . . . 55
5.2.1 Bayesian ridge regression . . . 56
5.2.2 Bayesian sparse regression . . . 56
5.2.3 Bayesian smooth regression . . . 58
5.2.4 LDL . . . 59

6 Conclusion 71

List of Figures

1.1 Chernobyl exclusion zone. Source: BBC News [11]. 2

1.2 Placement of measuring stations partaking in ETEX. Source: [38] 4

1.3 Molecules of tracers used in the two waves of the European tracer experiment in 1997. Both tracers are non-toxic, non-water-soluble, inert perfluorocarbons. . . . 5

2.1 Annotated data sample of a population. The sample is a small subset of the entire population and is further split into complementary subsets for training, validation and quality assurance of a given model. Source: [46] 8

3.1 An example of the law of total probability. Events $B_1, \dots, B_n$ are mutually exclusive and collectively exhaustive. 16

3.2 Standard normal distribution density function 19

3.3 Truncated normal distribution's PDF with $\mu = -8, \sigma^2 = 2, a = -10, b = 10$ (blue) compared with a standard normal distribution's PDF (red). Source: [75]. 20

3.4 Bivariate normal distribution with marginal distributions. 21

3.5 Gamma function graph [1]. 21

3.6 Graph of Gamma distribution's density function for different combinations of a, b. Adapted from [2] 22

4.1 L-curve of the ETEX dataset for determining λ penalty of ridge regression. . . . 28

4.2 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a ridge regression model (blue). 29

4.3 Comparison of measured data $\boldsymbol{y}$ (y-axis) and linear ridge regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis). 29

4.4 Bayesian ridge regression parameters dependency graph. The graph is directed with multiple cycles. 33

4.5 ω parameter values during 100 iterations using bayesian ridge regression on ETEX data. 34

4.6 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a bayesian ridge regression model (blue) after 100 iterations. 34

4.7 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian ridge regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations. 35

4.8 Sparse bayesian dependency graph 38

4.9 ω parameter values during 100 iterations using bayesian sparse regression on ETEX data. 38

4.10 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of sparse bayesian regression on the ETEX dataset. 39

4.11 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a bayesian sparse model (blue) after 100 iterations. 40

4.12 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian sparse regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations. 40

4.13 ω parameter values during 100 iterations using bayesian smooth regression on ETEX data. 44

4.14 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of smooth bayesian regression on the ETEX dataset. 44
4.15 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a bayesian smooth model (blue) after 100 iterations. 45
4.16 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian smooth regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations. 45
4.17 ω parameter values during 100 iterations using LDL on ETEX data. 51
4.18 Graphs of the change in value of $\boldsymbol{v}$ during iterations of LDL on the ETEX 1.4 dataset. 51
4.19 $\boldsymbol{l}$ parameter values during 100 iterations using LDL on ETEX data. 52
4.20 Change in value of $\boldsymbol{\psi}$ during 100 iterations of LDL on the ETEX dataset. 52
4.21 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by an LDL model (blue) after 100 iterations. 52
4.22 Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL-attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations. 54

5.2 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian ridge regression approximated retrieved $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations. 56
5.4 ω parameter values during 100 iterations using bayesian sparse regression on Chernobyl data. 57
5.5 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of sparse bayesian regression on the Chernobyl fires data. 58
5.6 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of smooth bayesian regression on the Chernobyl fires data. 59
5.8 ω parameter values during 100 iterations using bayesian smooth regression on the Chernobyl fires dataset. 60
5.9 ω parameter values during 100 iterations using LDL on Chernobyl data. 61
5.10 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of LDL on the Chernobyl fires data. 61
5.11 $\boldsymbol{l}$ parameter values during 100 iterations using LDL on Chernobyl data. 62
5.12 Approximate intermediate values of $\boldsymbol{\psi}$ during 100 iterations of LDL on the Chernobyl fires data. 62
5.16 Comparison of $\boldsymbol{y}$ and $\widehat{y}$ after different number of iterations. 65
5.17 Comparison of $\boldsymbol{y}$ and $\widehat{y}$ given different l_0 with 300 iterations and empirically set initial parameters. 66
5.19 Daily emissions estimate of ^{137}Cs between 2.4.2020 and 25.4.2020. 68

List of Tables

3.1 Central and non-central moments of a normal distribution. 19
3.2 Table of selected conjugate priors. Source [81], [82]. 23

4.1 Metrics of different algorithms on the ETEX 1.4 dataset. 53

5.1 Initial values and prior constants set for the first run of LDL on Chernobyl data, set to be non-informative. 60

5.2 Metrics for the LDL algorithm given various number of iterations. 63
5.3 Metrics for the LDL algorithm given various prior constant l_0. 67
5.4 Initial values and prior constants set for the best fit of LDL on Chernobyl data. . 67
5.5 Metrics of different algorithms on the 2020 Chernobyl fires 1.3.1 dataset. 69

I would like to express my sincere gratitude to my supervisor, Ing. Ondřej Tichý, PhD., for providing me with guidance, support, and encouragement throughout my research. His insights and feedback have been invaluable in shaping this thesis.

I would also like to thank the staff and faculty members of Czech Technical University in Prague's Faculty of Information Technology for their assistance and cooperation during the course of my research. Their expertise and resources were critical to the success of this project.

In addition, I would like to thank my family and friends for their unwavering support, shelter and encouragement, both during my studies and throughout my life. Their love and encouragement have been the foundation of my success.

Thank you, from the bottom of my heart for your contributions and undying support.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on May 3, 2023 ....

Abstract

Atmospheric pollution is a major environmental issue that has significant impacts on public health, ecosystems, and climate change. Accurate estimation of atmospheric emissions from various sources is crucial for effective environmental management and policy-making. The goal of this thesis is to establish an emission profile (a time series) of a substance and approximate the total emission of the given substance. Since traditional approaches to parameter estimation, such as linear regression's ordinary least squares method fall short due to the number of parameters, their approximations and uncertanties introduced by assumptions, approximations and limitations of both models and data, causing an instability of the model. These shortcoming are addresssed through a Bayesian approach, by modeling parameters through probabilistic distributions and capturing these uncertanties and approximations in this fashion. Namely, models from the Variational Bayes family are developed and examined on both controlled (ETEX) and uncontrolled emissions (Chernobyl fires), producing approximations of emission profiles discretized in time and total emissions of the substances. The best performing model, LDL with finely tuned hyperparameters, yields results comparable with existing described models.

Keywords machine learning, Bayesian modeling, variational Bayes, atmospheric dispersion modeling, emission estimation

Abstrakt

Znečištění atmosféry je významným ekologickým problémem, který má významný vliv na veřejné zdraví, ekosystémy a změnu klimatu. Přesné odhadování emisí do ovzduší z různých zdrojů je klíčové pro efektivní environmentální řízení a tvorbu politiky. Cílem této práce je odhadnout emisní profil (časovou řadu) uniklé látky a přibližně odhadnout celkovou emisi této látky. Tradiční přístupy k odhadování parametrů, jako například metoda obyčejných nejmenších čtverců v modelu lineární regrese, selhávají kvůli počtu parametrů, jejich aproximacím a nejistotě způsobené předpoklady, aproximacemi a omezeními jak modelů, tak dat, což způsobuje nestabilitu modelu. Tyto nedostatky jsou řešeny bayesovským přístupem, při kterém jsou parametry modelovány prostřednictvím pravděpodobnostních distribucí zachycujících neurčitosti. Konkrétně jsou vytvořeny modely z rodiny modelů založených na variační Bayesově metodě a jsou následně vyhodnoceny na datech řízené (ETEX) a neřízené (požáry v Černobylu) emise, kde výstupem jsou emisní profily diskretizované v čase a celková emise. Nejvýkonnější model poskytuje výsledky srovnatelné s existujícími popsanými modely.

Klíčová slova strojové učení, Bayesovské modelování, variační Bayesova metoda, modelování atmosférické disperze, odhad emise

Summary

Motivation

Atmospheric pollution is a major environmental issue that has significant impacts on public health, ecosystems, and climate change. Accurate estimation of atmospheric emissions from various sources is crucial for effective environmental management and policy-making. One of the many issues raised in this domain is the estimation of a time profile of a (usually hazardous) substance's emission (a time series) based on an atmospheric model and a set of measurements. Acquiring such knowledge could help in risk assessment and further modeling of the substance's spread in the atmosphere.

Goals

The goal of the thesis is to research existing Bayesian models, to develop and evaluate them firstly on a well-documented and measured experimental data with ground truth values, to test these developed models on a real recorded dispersion event and to compare them with existing solutions.

Method

Both traditional methods (linear regression and ridge regression) and methods based on the Bayes theorem were applied to this issue to provide a comparable baseline, with a major focus on the latter. Namely, Variational Bayes models were used, from a simple bayesian ridge regression to a complex LDL model. To further boost their performance, several heuristic steps were undertaken, such as positivity enforcement of models' normally distributed variables via truncation and softening. To compare models, simple and widely used metrics were used - mean square error, mean absolute error and root mean square error, to estimate the model's quality. Models were evaluated using a grid search, with a further fine-tuning of the best performing model, whose results was in the latter part of the thesis (Chernobyl data) used to provide result for comparison with other existing models.

Results

Out of the many models that were examined, implemented and their results compared and interpreted, the best performing one's provides results comparable with previously described and published models, despite the limited scope of the data utilised.

Structure

This thesis consists of 5 major parts. The first chapter establishes a baseline knowledge of terms and knowledge of game theory and machine learning. The second delves into generative adversarial networks and their many offshoots, which are the bulk of this thesis, followed by a dive into their foothold in the domain of medical imaging in the third chapter. Following is the description of implemented models and data operations in chapter 4. The entire thesis is then concisely summed in last, fifth chapter.

Introduction

Atmospheric pollution is a major environmental issue that has significant impacts on public health, ecosystems, and climate change. Accurate estimation of atmospheric emissions from various sources is crucial for effective environmental management and policy-making. One of the many issues raised in this domain is the estimation of a time profile of a (usually hazardous) substance's emission (a time series) based on an atmospheric model and a set of measurements. Acquiring such knowledge could help in risk assessment and further modeling of the substance's spread in the atmosphere.

There are many approaches to the construction of models in this domain, such as the common ordinary least squares or its modifications, but these solutions are hardly ever tractable or stable due to the vast number of influences, their approximations and uncertainties present in both atmospheric models and the measured data. In general, traditional methods of emission estimation based on engineering calculations or direct measurements are often expensive, time-consuming, limited in their spatial and temporal coverage, and unable to model the aforementioned uncertanties.

Bayesian modeling offers a promising approach for atmospheric emission estimation that can overcome some of these limitations. Bayesian models use statistical inference to combine prior knowledge with observed data, providing a systematic framework for uncertainty quantification and decision-making.

The goal of this thesis is to examine the issue posed in modeling of atmospheric emission, to examine the capabilities of models from the Variational Bayes family, develop and compare select models and measure tangible results on both controlled and uncontrolled emissions. Results are two-fold: an approximation of an emission profile discretised in time, and of total emission of a given substance to determine the model's practical application to model emissions, such as hazardous material dispersion.

Chapter 1

Preliminary background

1.1 Atmospheric dispersion modeling

Source-receptor relationship refers to the relationship between the emission source of a pollutant and the location where the pollutant is detected, also known as the receptor location. This relationship is important for understanding the transport and fate of pollutants in the environment, as well as for identifying the sources of pollution and developing effective control strategies [3].

The source-receptor relationship can be affected by a variety of factors, including the emission rate and location of the pollutant source, the atmospheric conditions, and the distance and direction between the source and receptor locations. Pollutants can be transported long distances from their source, and may undergo chemical and physical transformations during transport that can affect their impact on human health and the environment.

In this text, two datasets will be used: ETEX [4] and Chernobyl 2020 fires dataset, where the former is, thanks to its nature of being a controlled experiment, a reliable dataset for development and model fine-tuning, whereas the latter is used to validate the quality of developed models. Both models consists of actual spatiotemporal measurements at various times in various locations (measuring stations), and an atmospheric model which encompasses meteorological conditions, source and receptor characteristics, such as the pollutant concentration, temperature; emission parameters such as height and location of the source, exic velocity, . . . , physical properties of the pollutant, terrain properties, etc. [5].

1.2 Chernobyl disaster

The Chernobyl accident[6] in April 1986 was a catastrophe on a hitherto unimaginable level, resulting in drastic and permanent consequences both locally and globally. The accident was caused by a combination of incompetent, poorly trained staff and a flawed reactor design[1] [8] [9].

More than 9.6 tonnes of radioative material leaked from the power plant as a result. Most of the material was deposited in close proximity to the site as dust and debris, but some lighter material was carried by wind over Eurasia, mainly Ukraine, Russia and Europe proper. This event is considered to be the worst nuclear power plant accident in history, with the only contender for the spotlight being the 2011 Fukushima meltdown[10].

[1] Which was luckily unique and as such other nuclear reactors do not have the design flaw in question[7].

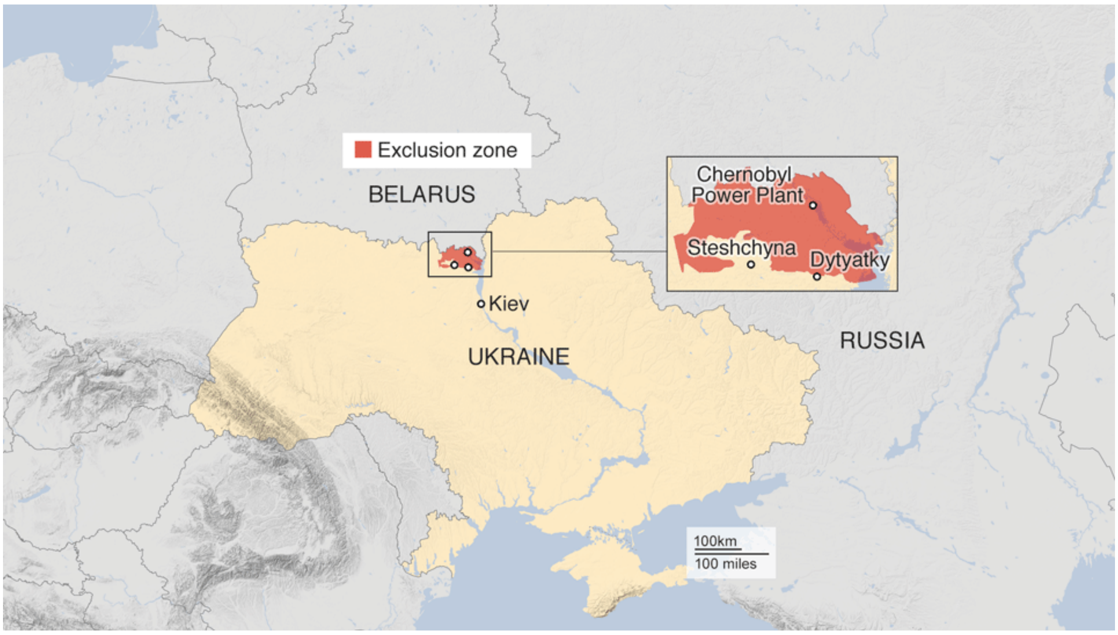


Figure 1.1 Chernobyl exclusion zone. Source: BBC News [11].

1.3 Effect on woods

One of the consequences of the Chernobyl incident was the die-off[2] of the flora and fauna, leaving a large amount of flammable material in the form of tree husks and litter (Layers of leaves and other material[13]) in the area [14]. This is caused by a combination of absence of forestry in the restricted areas and difficult decomposition in the altered conditions in the contaminated area. Further, most of the radioactive material has been deposited into the soil, or has migrated into soil since the accident, where most of it remains. In the absence of trees, the contaminants would most likely disperse as dust or dissolve in water.

However, water-soluble salts of cesium and strontium, are taken up by plants' root systems instead of potassium and calcium salts, founding the aforementioned litter on the top layer of the soil [15].

1.3.1 Forest fires

Forest fires can kindle both naturally and through a manmade error or intention[3]. What causes them to start naturally, and also spread in either case, are dry areas, a consequence of climate change. Whether occuring naturally or inadvertenly caused by humans, the number of wildfires has been increasing with blazing speed in recent history [17], with [18].

According to [19], based on the original report [20] released in 2022, the number of extreme fires will see an increase of up to 14 % by 2030. Further, not only the number of wildfires is assumed to increase, but severity of the fires is on the rise as well [18].

Smoke is the ultimate actor in the problem. As fires rage across the contaminated forests, smoke carries and reintroduces radioactive material originating in the power plant meltdown from the burning area bound and contained in soil, trees, etc. up into the atmosphere, where they're no longer bound and can spread over distances both short and long.

[2] "A sudden sharp drop in the numbers of plants or animals in a group" per Merriam-Webster dictionary [12].

[3] Although according to Center for Climate and Energy Solutions [16], 80 % of fires in USA are caused by people.

Multiple wildfires have swept across the contaminated regions of both Ukraine and Belarus, in the years 2003[21], 2008 [22], 2010 [23], [24] and 2020, to name a few [15]; and are expected to occur with larger impact in the future, as the surface temperature increased in the area, fueling wildfire in the Chernobyl exclusion zone [15].

The 2020 fires caused a release of about 1 billion (10^9) lower than the original caused by the reactor meltdown [25]. Fortunately, due to the large size of the burning particles, long range transport was minimal, which was confirmed by measuring stations scattered across Europe [25]. Assessments on particular radionuclides were made, such as [26], [13].

Characteristics of the fires, including daily emissions of ^{137}Cs (see [27] for details), ^{90}Sr (see [28]) and other radioactive material estimated and subsequent source-term scenario is established and reported by [29].

1.3.2 Other threats

Other concerns include mechanical disturbance of the soil by passing persons and vehicles, as was the case during the Russian troops movement at the brink of the invasion of Ukraine in late February 2022[14] through the Red Forest, whose impact on the levels of radiation is disputed by some [30].

Another threat is the presence of roughly 200 tonnes of inaccessible unburnt fuel at the bottom of the fourth reactor [31], which poses a potential risk of leaking if unattended or unprotected from the elements.

1.4 European tracer experiment

If there was another lesson to be learnt and wisdom to be gained from the nuclear tragedy besides the need for proper design, quality control, construction, safety measures and qualified personnel[4], it is a toolset to deal with the consequences, in the form of failsaves, response teams and impact assessment and prevention, the hindmost which was the spark to accelerate the field of atmospheric dispersion modeling. It wasn't just the threat of (another) breakdown of a nuclear power plant that motivates the study of this field. [32] provides a detailed breakdown and timeline of air pollution and show the (figuratively, hyperbolically, not mathematically) exponential growth of this issue throughout mankind's history.

Going back, air pollution has been recognised as a threat since at least 400 BC [32], at which point in time even the most profound philosophers of the time had a very loose grasp on the nature of matter and the rules and laws that govern them to recognize and address the threats there imposed. With the industrial revolution on the brink of the 19th century came air pollutants en masse, further increasing with the growth of the industries[32]. In 1967, the US Clean Air Act was enacted[33] and air pollution was recognised as an international problem[5].

The main obstacle for atmospheric dispersion models was both data and computational power. Whilst the latter is diminished naturally with progress of time, as technology improves[6], data collection and engineering is domain driven. Data most commonly used was data resulting from the Chernobyl accident, which lacked standardisation, was not open [36] and lacked quality control[7] to be reliably used to create models, but mainly to compare different models created by different institutions. To that end, the European Tracer Experiment (ETEX) [4] has been proposed, undertaken and evaluated.

The main goal of the ETEX was to study the transport and dispersion of pollutants in the atmosphere, particularly in the case of a large-scale accidental release of hazardous materials,

[4] Which is quite an extensive list.

[5] Credit where credit is due, multiple countries have enacted similar bills prior to the US - e.g. the UK Clean Air Act in 1956 [34], as well as many other acts and bills prior in countries all over the world.

[6] Even if in such a ”narrow” scope observed as the Moore's law [35].

[7] Something software engineering inherited.

such as radioactive materials or industrial chemicals or caused by resuspension [37] of deposited hazardous material.

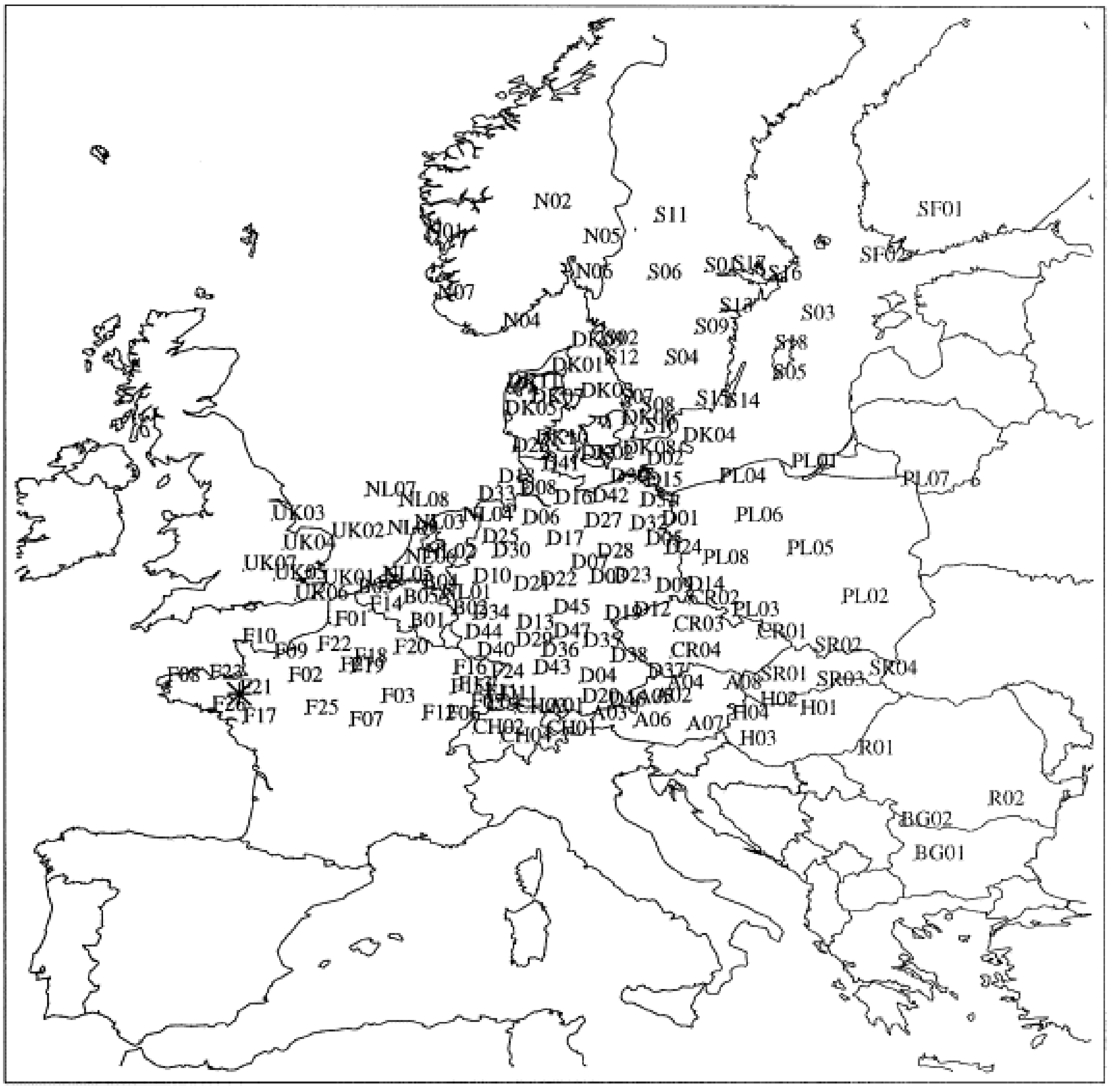


Figure 1.2 Placement of measuring stations partaking in ETEX. Source: [38]

The experiment consisted of emitting a trackable tracer and measuring its concentration at measuring stations scattered across Europe, plus three aircrafts. The tracer released had to meet a list of criteria in order to fulfill its purpose:

- non-toxicity
- non-depositing
- non-water-soluble
- inert[8]
- environmentally safe

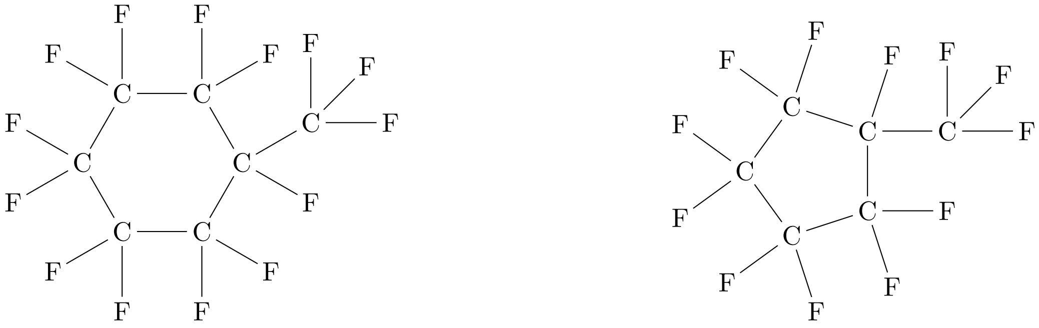
[8] Chemically inactive.

- easily detectable

A suitable family of compounds was found for long distance transportation - perfluorocarbons[4] (PFCs), meeting all the aforementioned constraints. 168 stations partook in the experiment across 18 countries (figure 1.2).

There were two runs of the experiment altogether, first in 1997 and second in late 1997. To meet the last constraint, ease of detection, the two runs could not risk any sort of interference. Using the same tracer in the two consecutive emissions could provide false measurements - during the second emission, particles released in the first could still be present in the atmosphere and as such give wrong measurements. To that end, a different tracer was used in the second emission. During both experiments, the emission lasted 12 hours, releasing 340 and 490 kg of tracer particels, respectively.

During the first run, perfluoromethylcyclohexane (PMCH) (see figure 1.3a) was released, whilst for the second run worked with perfluoromethylcyclopentane (see figure 1.3b). The sampling stations scattered across Europe worked with a sampling time of 3 hours over 72 hours, resulting in 23 samples per station.



(a) Perfluoromethylcyclohexane molecule (C_7F_{14})

(b) Perfluoromethylcyclopentane molecule (C_6F_{12})

Figure 1.3 Molecules of tracers used in the two waves of the European tracer experiment in 1997. Both tracers are non-toxic, non-water-soluble, inert perfluorocarbons.

Chapter 2

Regression

Regression analysis is a statistical technique for investigating and modeling the relationship between variables[39].

The goal is to find a functional relationship that can predict the value of a dependent variable based on the values of a set of independent variables. This relationship is inferred from a limited subset of data.

The term ”dependent variable” is primarily used in statistics, and is interchangeable with terms such as response, explained variable, signal, amongst many others. Its counterpart(s), the dependent variable, is referred to as feature, input variable, regressor, explanatory variable, estimator, etc. [40].

Let $\mathcal{X}$ be the feature space, also called input space [41] or instance space [42], of p dimensions, $\mathcal{Y}$ the label space [41][42]. Points in the feature space are called feature vectors (tuples) of p dimensions:

$$\boldsymbol{x} = (x_1, x_2, \dots, x_p)^\intercal \in \mathbb{R}^p. \tag{2.1}$$

The feature space contains all combinations of the p features, but not all values or combinations are possible - for instance, if one dimension measures a person’s height, its value cannot be negative[1].

The label space $\mathcal{Y}$ is the set of possible values of the output variable Y_i. This output domain determines the possible responses of the model. In the case of a classification task, the label space $\mathcal{Y}$ is discrete, whilst in the case of regression, determined by annotated samples

$$\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n). \tag{2.2}$$

In the language of machine learning, regression belongs to the family of supervised tasks[43] - the model’s parameters are set based on a set of pairs $(Y_i, \boldsymbol{x_i})$. For instance, if the task is to model (and subsequently, predict) fuel consumption (e.g. liters per 100 kilometers) of vehicles, the feature vector might consist of parameters such as weight, engine type, engine power, volume, emission data, type of tyres, car brand, age of car, mileage, average speed, ..., and the label (output) would be the average fuel consumption.

Finally, let Ω be a set of observed instances, i.e.

$$\Omega \in \mathcal{X} \times \mathcal{X} \tag{2.3}$$

. The goal of regression is to find a suitable function $f : \mathcal{X} \to \mathcal{X}$ which predicts correct labels for priorly unknown data [44]. Importantly, the function is not guaranteed to be injective 2.1, i.e. for two different input feature vectors, the output can be identical.

[1] Further, a minimal possible height of a person could be measured and determined.

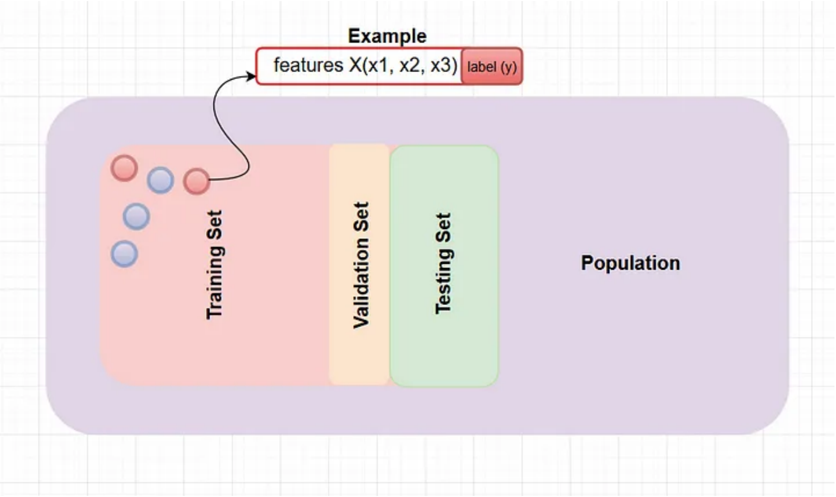


Figure 2.1 Annotated data sample of a population. The sample is a small subset of the entire population and is further split into complementary subsets for training, validation and quality assurance of a given model. Source: [46]

▶ **Definition 2.1** (Injective function [45]). *A function $f : A \to B$ is called injective (one-to-one), if $\forall x, y \in D_f, f(a) = f(b) \implies a = b$.*

2.1 Training data

The goal of any model is to capture the nature of a population as closely as possible given a sample of said population.If the entire population was available as data, then the model could be created from that and it would be absolutely accurate. However, that is almost never possible and so the models have to be made with a limited, but hopefully representative sample of the population. This representative sample is referred to as the training set (or training data). In the case of linear regression, which belongs to the supervised learning family of algorithms in machine learning, that training set consists of data points $\boldsymbol{x_i}$ annotated with labels Y_i. Usually, for the sake of evaluation of models, their accuracy, and general quality assurance, this representative set is split into smaller subsets (as depicted in figure 2.1):

1. training set, upon which the model's parameters are approximated for closest fit of predictions $\widehat{Y}_i$ and labels Y_i
2. optional validation set
3. testing set, which is used to acquire an unbiased evaluation of the model

The question that naturally arises is in what ratio to split the data into the individual subsets. There are two competing concerns, both concerned with the variance of either (train and test set) variance. A trade-off has to be found between the high variance caused by a small size of the subset [47]. A common rule-of-thumb is to split the data in an 80:10:10 fashion for training set, validation set and test set. If validation does not take place, the split is in an 80:20 ratio. In the case of a limited sample size, cross validation is often employed to stabilise the training and evaluation process [48].

2.2 Linear regression

Linear regression, much like the parent term regression, bleeds from statistics to machine learning and the world of artifical intelligence directly.

Linear regression assumes a linear relationship between the dependent variable and the set of independent variables, i.e. the dependent variable can be modeled by a linear combination of the independent variables[39].

The value of the dependent variable Y at $\boldsymbol{x} = (x_1, x_2, \dots, x_p)^\intercal$ is

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_p x_p + \epsilon = w_0 + \sum_{i=1}^{p} w_i x_i + \epsilon, \tag{2.4}$$

where ϵ is a random variable which encompasses the information about Y not explainable by the p features. As a result, Y is also a random variable. It is assumed that the noise ϵ has an expected value of 0, $\mathbb{E}[\epsilon] = 0$ [49]. This assumption is often much stronger in that the noise ϵ_i is assumed to a normal distribution with mean $\mu = 0$ and variance σ^2, and is indepedent and identically distributed, i.e.

$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{2.5}$$

In the terminology of statistics, linear regression models the conditional distribution of response y_i by observed predictor values $X_i = \boldsymbol{x_i}$ [50]. Since ϵ_i is a random variable, and values $(x_1, x_2, \dots, x_p)^\intercal$ and $(w_0, w_1, \dots, w_p)^\intercal$ are deterministic, Y_i is simply a linear transformation of the random variable ϵ_i

$$Y_i \sim \mathcal{N}\left(w_0 + \sum_{i=1}^{p} w_i x_i, \sigma^2\right), \tag{2.6}$$

and the predicted value $\widehat{Y_i}$ is the expected value of the distribution, $\widehat{Y_i} = \mathbb{E}[Y_i] = w_0 + \sum_{i=1}^{p} w_i x_i$. Design matrix $\boldsymbol{X}$ is gained by placing all entries $\boldsymbol{x_i}$ in a matrix in row-wise fashion:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}. \tag{2.7}$$

w_0 is called an intercept (or a constant), and represents the mean value of the dependent variable Y when all the independent variables are 0 - the baseline for the model. The vector $\boldsymbol{w} = (w_1, w_2, \dots, w_p)^\intercal$ is referred to as the weight vector in machine learning technology, or vector of regressor coefficients in statistics[2].

$$\boldsymbol{x} = (x_0, x_1, \dots, x_p)^\intercal, \boldsymbol{w} = (w_0, w_1, \dots, w_p)^\intercal \tag{2.8}$$

With that, the equation can be written in vector notation:

$$Y = \mathbf{w}^\intercal \boldsymbol{x} + \epsilon. \tag{2.9}$$

2.2.1 Prediction

The goal of a linear regression task is to predict the value Y as precisely as possible.

Given an approximation $\hat{\mathbf{w}}$ of the vector $\mathbf{w}$, the predicted value for point $\boldsymbol{x}$ is

$$\widehat{Y} = \hat{\mathbf{w}}^\intercal \boldsymbol{x}. \tag{2.10}$$

[2] Where commonly it is denoted with β rather than $\mathbf{w}$ [39].

Thanks to the assumption of $\mathbb{E}[\mathbf{w}] = 0$ the expected value of Y is

$$\mathbb{E}[Y] = \mathbf{w}^\intercal \boldsymbol{x}, \tag{2.11}$$

and therefore the aforementioned prediction $\widehat{Y}$ is a point estimate of $\mathbb{E}[Y]$ at $\boldsymbol{x}$.

2.2.2 Loss function

Given a prediction $\widehat{Y}_i$ for the point $\boldsymbol{x}_i$ and the label Y_i, the error of the prediction can be calculated using an approriate metric function - the choice of a metric differs case from case, and is dependent on both the task and the data.

▶ **Definition 2.2** (Metric [51]). *A metric is any non-negative function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the following conditions for $\forall x, y, z \in \mathcal{X}$:*

- $f(x, y) = f(y, x)$ *(symmetry)*
- $f(x, y) = 0 \leftrightarrow x = y$ *(identity)*
- $f(x, z) + f(z, y) \geq f(x, z)$ *(triangle inequality)*

2.2.2.1 Common loss functions

- Mean bias error (MBE): $L(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)$
- Mean absolute error (MAE): $L(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$
- Mean squared error (MSE): $L(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
- Root mean squared error (RMSE): $L(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$
- Root mean squared logarithmic error (RMSLE): $L(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(Y_i + 1) - \log(\widehat{Y}_i + 1))^2}$

Loss functions can be used as performance metrics, alongside information criteria and other estimators, such as precision, recall or F1-score in classification tasks.

Usually, for the sake of fitting a model, the mean squared error loss function is used. The optimisation problem thereby proposed is the minimilisation of the residual sum of squares (RSS) 2.12:

$$RSS(w) = \sum_{i=1}^{N} L_2(Y_i, \boldsymbol{w}^\intercal \boldsymbol{x_i}) = \sum_{i=1}^{N}(Y_i - \boldsymbol{w}^\intercal \boldsymbol{x_i})^2 = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}||^2. \tag{2.12}$$

2.2.3 Partial derivative

▶ **Definition 2.3** (Norm [52],[53]). *Given a vector space $\boldsymbol{V}$, norm is a function $\|\cdot\| : \boldsymbol{V} \to \mathbb{R}_0$ such that $\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{V}$ and*

1. $\|\boldsymbol{x}\| = 0 \implies \boldsymbol{x} = 0$ *(positive definiteness)*
2. $\|\alpha \boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|$ *(homogenity)*
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ *(triangle inequality)*

The norm of a mathematical object is a quantity that in some (possibly abstract) sense describes the length, size, or extent of the object[54]. Commonly used norms are, for $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{C}^n$,

- $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$ (absolute value norm)
- $\|\boldsymbol{x}\|_2 = \sqrt[2]{\sum_{i=1}^{n} x_i^2}$ (euclidean norm)
- $\|\boldsymbol{x}\|_\infty = \max |x_i|\, i \in 1, \ldots, n$ (maximum norm)

▶ **Definition 2.4** (Point neighbourhood [52])**.** *Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, let $\mathbf{x} \in \mathbb{R}^n, \delta \in \mathbb{R}^+$, a δ-environment of point $\boldsymbol{x}$ is a set*

$$H_\delta(\boldsymbol{x} = \{\boldsymbol{b} \in \mathbb{R}^n | \, \|\boldsymbol{x} - \mathbf{b}\| < \delta\}. \tag{2.13}$$

▶ **Definition 2.5** (Partial derivative[52])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n variables.*

A partial derivative of function f in the direction x_i at point $\boldsymbol{b} = (b_1, b_2, \ldots, b_n) \in D_f$ such that $\exists H(\boldsymbol{b})$ is

$$\frac{\partial f}{\partial x_i}(\boldsymbol{b}) = \lim_{h \to 0} \frac{f(b_1, b_2, \ldots, b_i + h, \ldots, b_n) - f(b_1, b_2, \ldots, b_i, \ldots, b_n)}{h} \tag{2.14}$$

if such a limit exists.

The partial derivative describes the rate of change in the respective direction, considering all other variables (directions) constant. This definition describes only partial derivative in the direction of one of the variables, but there exists a generalised definition for partial derivative in an arbitrary direction [52]

▶ **Definition 2.6** (Gradient of a function [52])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n variables, such that all partial derivatives of f are finite in the point $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$. Gradient of function f at $\boldsymbol{b}$ is defined as a vector of partial derivative in each of the n directions:*

$$\nabla f(\boldsymbol{b}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{b}), \frac{\partial f}{\partial x_2}(\boldsymbol{b}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{b}) \right). \tag{2.15}$$

Akin to the univariate's first derivative, which describes whether a function is, at a given point, increasing or decreasing and the rate of thischange, the gradient represents the direction of the fastest rate of increase of the function f at a given point $\boldsymbol{b}$. Any point $\boldsymbol{b}$, for which $\nabla f(\boldsymbol{b}) = \boldsymbol{b}$ or for which the gradient $\nabla f(\boldsymbol{b})$ does not exist, is considered a critical point and can hold a local extremum, as the function is neither stricly increasing or decreasing on its neighbourhood.

2.2.4 Hessian matrix

Let $f : D_f \to \mathbb{R}$, $D_f \subset \mathbb{R}^n$ and assuming that all second partial derivations of f in $\boldsymbol{b}$ exist, then the Hessian matrix of f in $\boldsymbol{b}$ is equal to

$$\nabla^2 f(\boldsymbol{b}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\boldsymbol{b}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{b}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{b}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{b}) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(\boldsymbol{b}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{b}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{b}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{b}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\boldsymbol{b}) \end{pmatrix}. \tag{2.16}$$

The Hessian matrix is the multidimensional counterpart of the one dimensional second derivative, which describes the concavity/convexity of a given function at a given point.

▶ **Theorem 2.7** (Sufficient condition for local minimum)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n variables and $\boldsymbol{x}^* \in \mathbb{R}^n$ such that $\nabla f(\boldsymbol{x}^* = 0$, and f has all second derivations continuous in some neighbourhood of $\boldsymbol{x}^*$.*

If $\forall \boldsymbol{s} \in \mathbb{R}^n, \boldsymbol{s} \neq \mathbf{0}$

$$\boldsymbol{s} \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{s} > 0, \tag{2.17}$$

then the function f attains a strict local minima in point $\boldsymbol{x}^*$.

The property described in theorem 2.7 is called the positive semidefinitness of the matrix $\nabla^2 f$ in x^*. Minimising the residual sum of squares 2.12 analytically gives critical points described by its gradient

$$\nabla RSS = -2\sum_{i=1}^{N}(\boldsymbol{Y}_i - \boldsymbol{w}^\intercal x_i)x_i = -2\boldsymbol{X}^\intercal(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}), \tag{2.18}$$

by setting it to 0 in the form of normal equations

$$\boldsymbol{X}^\intercal\boldsymbol{Y} = \boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{w}. \tag{2.19}$$

The Hessian matrix $\nabla^2_{RSS}(\boldsymbol{w}) = 2\boldsymbol{X}^\intercal\boldsymbol{X}$ is positively semidefinite for any given $\boldsymbol{w}$, and as such, under theorem 2.7, it holds that any point $\boldsymbol{w}$ solving normal equations 2.19 is a local minimum [52]. If $\boldsymbol{X}^\intercal\boldsymbol{X}$ is regular, there is a unique solution[3]

$$\widehat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}. \tag{2.20}$$

2.2.5 Collinearity conundrum

If the matrix $\boldsymbol{X}^\intercal\boldsymbol{X}$ is not regular, there is a risk of numeric instability and poor approximation of the inversion. In such a case, different methods are used for minimisation, such as an iterative gradient descent [55] [56] (not considered in this text) or regularisation via a penalty by ridge regression 2.2.6.

If the features are independent, $\boldsymbol{X}^\intercal\boldsymbol{X}$ is regular and there exists exactly one solution $(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}$, otherwise there are infinitely many $\boldsymbol{w}$ and $\boldsymbol{w}'$ such that $\boldsymbol{w} \neq \boldsymbol{w}', \boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}') = 0$. The issue is also caused when they're "nearly" linearly dependent - multicollinearity. This issue can be tackled by introducing a regularising component, which defines the type of regularisation, with the most widely used being [57]:

- Lasso (L_1)
- Tichonov (L_2)
- Elastic-net - a combination of Lasso and Tichonov [58]

Tichonov's regularisation will be discussed moving forward.

2.2.6 Ridge regression

Ridge regression (also called L_2 or Tichonov regularisation[4]) resolves the issue of matrix singularity by introducing a penalty to the weight vector $\mathbf{w}$ in the form of a quadratic norm of the vector [60]:

$$RSS_\lambda(\mathbf{w}) = ||\mathbf{Y} - \mathbf{X}\mathbf{w}||^2 + \lambda\sum_{i=1}^{p} w_i^2, \tag{2.21}$$

where the parameter $\lambda \in \mathbb{R}, \lambda \geq 0$ is the regularisation parameter, determining how much is the regularisation enforced. With $\lambda = 0$, the model simplifies to plain linear regression 2.2. As the value of λ increases, the pressure on $\mathbf{w}$ to minimise increases. It should also be noted that the intercept w_0 is exempt from penalisation. A ridge regression model is typically fit using an optimization algorithm such as gradient descent or closed-form solutions such as the normal equation.

[3] $(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$ exists thanks to regularity

[4] Named after the Soviet mathematician and geophysicist Andrey Nikolayevich Tikhonov[59].

One of the advantages of ridge regression is that it can handle multicollinearity, a situation where two or more predictor variables are highly correlated with each other. In this case, the ordinary least squares (OLS) method can produce unstable and unreliable estimates of the model coefficients. Ridge regression can overcome this problem by shrinking the coefficients towards zero and reducing their variance.

Chapter 3

Bayesian theory

3.1 Bayes theorem

Bayes theorem is nowadays an implicit part of any course relating to probability and statistics, pitting its approach in juxtaposition with the classic, frequentist approach. Where the frequentist approach ties the probability of an event happening given by its relative frequency, bayesian approach bases the probability as a degree of belief, which is updated by gaining more data [61]. It stems from conditional probability, whereby for two events (in the discrete case, works analogously for the continuous case described later) A, B the condition of event A occuring given that event B occurred is defined as

▶ **Definition 3.1** (Conditional probability)**.**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0. \tag{3.1}$$

▶ **Definition 3.2** (Partition of a sample space Ω [62])**.** *A set of mutually disjoint ($\forall i \neq j \implies B_j \cap B_j = \emptyset$) events $B_1, \dots, B_n$ is called a partition of the set Ω if*

$$\Omega = \bigcup_{i=1}^{n} B_i. \tag{3.2}$$

▶ **Theorem 3.3** (Law of total probability [62])**.** *Let $B_1, B_2, \dots, B_n$ be a partition of Ω such that $\forall i : P(B_i) > 0$. Then for each event A it holds that*

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \times P(B_i). \tag{3.3}$$

$B_1, B_2, \dots, B_n$ are collectively exhaustive, covering the entire sample space Ω. This is graphically deducible for A and $B_1, B_2, \dots, B_n$ in figure 3.1.

▶ **Theorem 3.4** (Bayes' theorem for events [62])**.** *Let $B_1, B_2, \dots, B_n$ be a partition of Ω such that $\forall i : P(B_i) > 0$. let A be an event with $P(A) > 0$. Then it holds that*

$$P(B_j|A) = \frac{P(A|B_j) \times P(Bj)}{\sum_{i=1}^{n} P(A|B_i) \times P(B_i)}. \tag{3.4}$$

Bayes' theorem formulates the probability that a hypothesis B_j is true, given new evidence (observation) A. In its fundamental and simple formulation, it gives an incredibly strong apparatus, which, amongst other capabilities, is easily lenient to online computations.

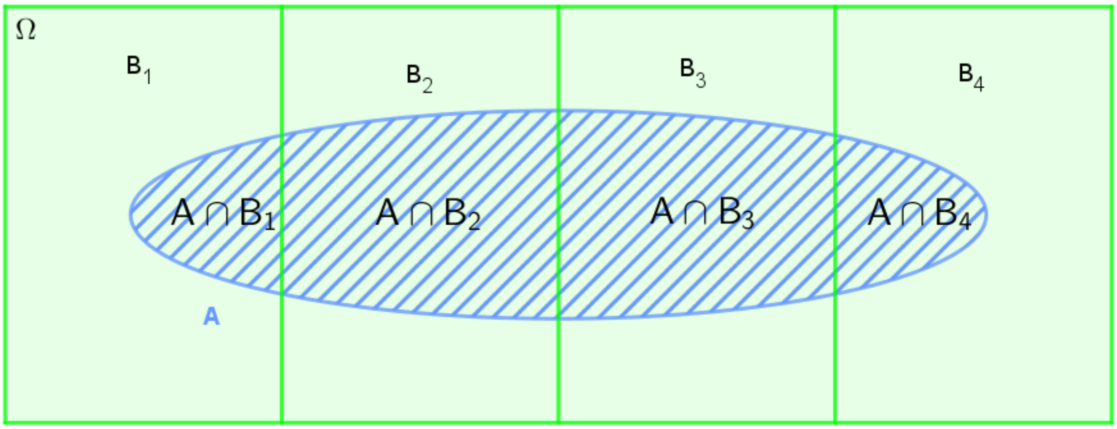


Figure 3.1 An example of the law of total probability. Events $B_1, \dots, B_n$ are mutually exclusive and collectively exhaustive.

In the Bayes' theorem 3.4, $P(B_j)$ is called the prior probability, $P(B_j|A)$ the posterior probability and $P(A|B_j)$ is the probability of observing new evidence A given the hypothesis B_j is true - a likelihood.

The aforementioned formulation of Bayes' theorem 3.4 and its associated definitions and theorems (conditional probability 3.1, partition of a sample space 3.3, ...) have their counterparts for the continuous case and work in the same fashion, with the former (for events) defined and explained as a series of stepping stones to the continuous case used henceforth.

▶ **Definition 3.5** (Conditional probability density function [63])**.** *Suppose that X, Y have a joint continuous distribution and let $Y \in \mathbb{R}$ such that $f_Y(y) > 0$. The conditional density of X given $Y = y$ is defined for*

$$\forall x \in \mathbb{R} : f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \tag{3.5}$$

where

- $f_{X,Y}(x,y)$ *is the joint probability density*
- $f_Y(y)$ *is the marginal distribution of* Y*.*

▶ **Theorem 3.6** (Bayes theorem for continuous variables [64])**.** *Let x, θ be random variables with probability density functions $f(x|\theta), \pi(\theta)$, respectively. Then it holds that*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}, f(x) > 0, \tag{3.6}$$

where

- $\pi(\theta|x)$ *is the posterior conditional density function of* $\boldsymbol{X}$
- $\pi(\theta)$ *is the prior density function*
- $f(x|\theta)$ *is the model (or data likelihood)*
- $f(x)$ *is the marginal density of* $\boldsymbol{X}$ *(also called evidence).*

The denominator $f(x) = \int f(x,\theta)\mathrm{d}y = \int f(x|\theta)\pi(\theta)\mathrm{d}\theta$ is independent of the estimated parameter θ and fills the role of a normalisation factor the posterior density - it can be simplified using proportionality - neglecting of this normalisation factor

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta). \tag{3.7}$$

Proportionality $\propto$, first mentioned in 3.7 and used heavily in the rest of the text, is a binary operator, indicating that the left value is proportional to the right value (differing only in a constant), i.e. $x \propto y$ expresses that $\exists c, x = cy$. Proportionality is often used to ignore cumbersome normalisation constants independend of a modeled variable, but retaining the same behaviour as to the change in value of the modeled variable - both values, x, y change in a consistent and predictable fashion in relation to each other.

The normalisation factor $f(x)$ ensures that the posterior distribution sums to 1 over all possible values[1]. Ignoring this factor will make modeling more tractable and computable, as the computability of integrals is dubious and difficult in the cases where the integral can be calculated.

Circling back to the likelihood mentioned in 3.4 and 3.6 was not described. A likelihood function is a function that measures the goodness of fit between a statistical model and observed data. It is the probability of observing the data, given the model and its parameters.

Here, the main focus will be on the parameters, as the model itself (distributions of variables) will not be subject to much change, but it will be the parameters that will be estimated for a good fit. The likelihood function is obtained by treating the observed data as fixed and varying the parameters of the model. The parameter values that maximize the likelihood function are considered to be the most likely values of the parameters given the observed data [65] [66]. Likelihood is not a density function, and as such, is not normalised.

3.2 Exponential family of distributions

The exponential family is a widely used family of probability distributions in statistics and machine learning. It is a special class of probability distributions that has a particular mathematical form that allows for efficient computation and modeling. In this section, we will discuss the properties and applications of the exponential family.

The exponential family is characterized by a probability density function that has the form:

▶ **Definition 3.7** (Exponential family probability density function)**.** *Let y be a random variable conditioned by a random variable x and a parameter θ. Exponential family of distributions contains distributions with probability density function with the following form*

$$f(y|x,\theta) = h(y,x)g(\theta)\exp\left(\eta(\theta)^\intercal T(y,x)\right), \tag{3.8}$$

where

- *$\eta(\theta)$ is a natural parameter*
- *$T(y,x)$ is a sufficient statistics*
- *$h(y,x)$ is a known function*
- *$g(\theta)$ is a normalisation function.*

The form is chosen for a convenient use and easy algebraic operations, as will become apparent in further sections.

Non-central 3.9 and central 3.10 moments of a function describe characteristics of a distribution [67] [63].

[1] As a random variable distribution should!

▶ **Definition 3.8** (Nth non-central and central moments [62]).

$$\mu_n = \mathbb{E}[X^n] \tag{3.9}$$

$$\sigma_n = \mathbb{E}[(X - \mathbb{E}[X])^n]. \tag{3.10}$$

In general, moments define parameters of a distribution, such as location, shape and scale [68]. It is important to keep in mind that moments of a given random variable X do not always exist - such is the case if the corresponding sum or integral do not converge. Further, the first non-central moment of random variable X corresponds to expected value $\mathbb{E}[X]$, whilst the second non-central moment corresponds to its variance var .

3.2.1 Normal distribution

Normal (univariate) distribution, also called Gaussian distribution is considered to be the most important and widely utilised probability distribution [69]. It is a continuous distribution for real-valued values, defined as follows:

▶ **Definition 3.9** (Normal distribution [62]). *A random variable X follows a normal distribution with parameters $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}$, denoted $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ if its probability density function is*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu - x)^2}{2\sigma^2}\right) \textit{ for } x \in (-\infty, \infty). \tag{3.11}$$

The probability density function is symmetric about its mean μ, around which values are spread according to its variance σ^2.

▶ **Theorem 3.10** (Standardisation of a normal distribution [62]). *Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then*

$$Z = \frac{X - \mu}{\sigma} \tag{3.12}$$

follows a standard normal distribution, i.e. $Z \sim \mathcal{N}(0, 1)$.

Its widespread use is attributed to the Central limit theorem (CLT), which is build on top of the laws of large numbers described the limit theorems. In essence, per the laws of large numbers, with increasing sample size from a population, the sample mean approaches the true population mean. The coined term ”Law of large numbers” can be broken down into a weak and strong case, which differ only in the strength of their convergence.

▶ **Theorem 3.11** (Central limit theorem [62]). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite expected values $\mathbb{E}[X_i] = \mu, |\mu| < \infty$ and finite variances $varX_i = \sigma^2, 0 < \sigma^2 < \infty$. Then*

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \tag{3.13}$$

Normal distribution tends to be the go-to distribution, as natural phenomena are observed to follow it[2], as well as per the central limit theorem, the mean of of a large number of independent and identically distributed random variables approaches a normal distribution. Further, computational benefits of a normal distribution are the ease of tractability, simple characterisation through only two parameters[3] and the ability to infer parameters of a population based on sample data[70].

In the case of a normal distribution, all its moments are finite - proof via recursive moment calculation in [71]. Further, the mode and median values are equal to mean value μ.

The first non-central moment is therefore equal to μ, the second central moment is equal to σ^2. As such, their computation is trivial, which will prove useful in posterior parameter shape inference (see 4).

[2] Within certain boundaries, which is addressed in 3.2.2

[3] Which in it of themselves have crystal clear interpretation

Order	Central moment	Non-central moment
1	0	μ
2	σ^2	$\mu^2 + \sigma^2$
3	0	$\mu^3 + 3\mu\sigma^2$
4	$3\sigma^4$	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$

Table 3.1 Central and non-central moments of a normal distribution.

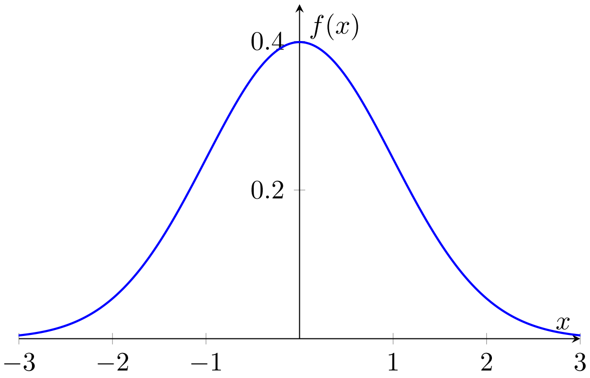


Figure 3.2 Standard normal distribution density function

3.2.2 Truncated normal distribution

Where normal distribution falls short of reality is, well, most places. A normal distribution, whilst being concentrated around its mean value μ in degree based on its variance σ^2, is unbound, i.e. defined for all real numbers. Theoretically, sampling from a normal distribution can result in any number, but such a sample is extremely improbable as the value deviates further from μ. In reality, most variables that are modeled with normal distributions cannot physically attain such values - even though the height of a population can be described by a normal distribution with given μ, σ^2, there is a minimal value (in the very worst case 0), and a maximal value, both governed by the laws of physics and biology[4].

Such behaviour might be satisfactory, following the British statistician George Box's aphorism "All models are wrong, but some are useful"[74], describing the limitations of models to capture complexity. That might not always be a detriment, as many realities can be modeled within a reason with fairly straightforward, simple, practical models, as compared to models which would reflect the nature of a given reality perfectly, but would render any use impossible.

With normal distributions, an alternative to accepting the potential shortcomings of attaining non-sensical values (negative height, impossibly high weight, ...), is to set bounds for the distribution. This alters the distribution and shifts probabilities, as will become clear shortly. A truncated normal distribution 3.12 is derived from a normal distribution by setting (either or both) lower and upper bounds.

▶ **Definition 3.12** (Probability density function of truncated normal distribution [76]). *Probability density function of a truncated normal distribution for a scalar variable x on interval $[a, b]$ is defined*

$$f(x, \mu, \sigma, a, b) = \begin{cases} 0 & \textit{if } x \leq a; \\ \frac{\sqrt{2}\exp\left(-\frac{1}{2\sigma}(x-\mu)^2\right)}{\sqrt{\pi\sigma}(\operatorname{erf}\left(\frac{b-\mu}{\sqrt{2\sigma}}\right)-\operatorname{erf}\left(\frac{a-\mu}{\sqrt{2\sigma}}\right))} & \textit{if } a < x < b; \\ 0 & \textit{if } b \leq x, \end{cases} \tag{3.14}$$

[4] Such as the square-cube law[72] [73]

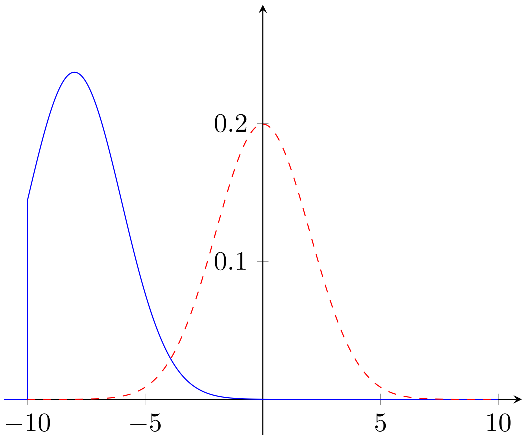


Figure 3.3 Truncated normal distribution's PDF with $\mu = -8, \sigma^2 = 2, a = -10, b = 10$ (blue) compared with a standard normal distribution's PDF (red). Source: [75].

where

$$\begin{aligned} \operatorname{erf}(t) &= \frac{2}{\sqrt{\pi}} \int_0^t \exp(-u^2) \mathrm{d}u, \\ \alpha &= \frac{a-\mu}{\sqrt{2\sigma}}, \beta = \frac{b-\mu}{\sqrt{2\sigma}}. \end{aligned} \tag{3.15}$$

A visualisation of the truncated normal distribution's probability density function in tandem with a standardised normal distribution's density is plotted in figure 3.3. Due to the truncation, the density shifts around and results in an increase concentration around the mean value, as probability of a distribution has to add to 1 (in this case, integrate to 1). Multivariate case of a truncated normal distribution is usually numerically approximated [77].

Fortunately, in later calculations using conjugate priors (see 3.3), a normal distribution can be directly substituted by a truncated normal distribution, with a slightly more complex calculation of moments, of which the first two 3.16, 3.17 are needed later on. Multivariate case is approximated by [78].

$$\widehat{x} = \mu - \sigma \frac{\sqrt{2}(\exp(-\beta^2) - \exp(-\alpha^2))}{\pi(\operatorname{erf}\beta - \operatorname{erf}\alpha)} \tag{3.16}$$

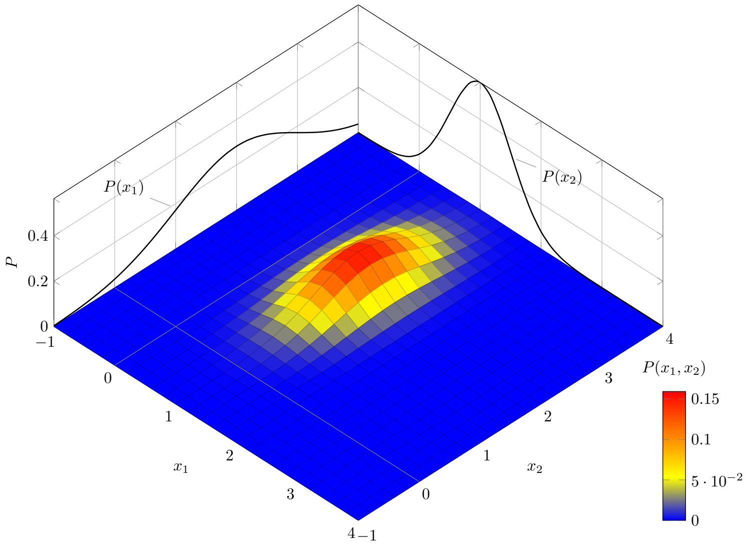
$$\widehat{x}^2 = \sigma + \mu\widehat{x} - \sigma \frac{\sqrt{2}(b\exp(-\beta^2) - a\exp(-\alpha^2))}{\pi(\operatorname{erf}(\beta) - \operatorname{erf}(\alpha))}. \tag{3.17}$$

3.2.3 Multivariate normal distribution

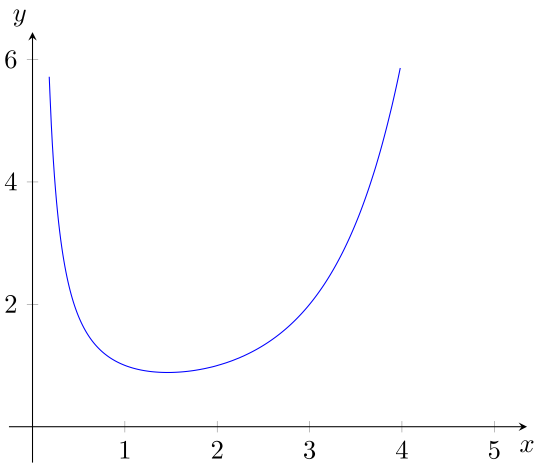
Normal multivariate distribution is a generalization of the normal univariate distribution (see 3.2.1). The distribution is fully characterized by its mean vector and covariance matrix. The mean vector specifies the average value of each component of the random vector, and the covariance matrix specifies the degree of linear association between the components. The probability density function (PDF) of a multivariate normal distribution is given by 3.13.

▶ **Definition 3.13** (Multivariate normal distribution)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n,n}$ be a positively semidefinite matrix. A random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ follows an n-dimensional normal distribution with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ ($\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), if for each $\boldsymbol{c} \in \mathbb{R}^n$ it holds that*

$$\boldsymbol{c}^\intercal \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{c}^\intercal \boldsymbol{\mu}, \boldsymbol{c}^\intercal \Sigma \boldsymbol{c}). \tag{3.18}$$



■ **Figure 3.4** Bivariate normal distribution with marginal distributions.



■ **Figure 3.5** Gamma function graph [1].

The probability density function of a multivariate normal distribution is given by 3.19

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp{(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}, \tag{3.19}$$

and for the bivariate ($n = 2$) case, the density function is plotted in figure 3.4, alongside both marginal distributions.

3.2.4 Gamma distribution

▶ **Definition 3.14** (Gamma function [63]). *Gamma function Γ is defined for any $p > 0$ by the relation*

$$\Gamma(p) = \int_0^\infty x^{p-1} \exp{(-x)} \mathrm{d}x. \tag{3.20}$$

Basic properties of the gamma function are, for $p > 0$ and $n \in \mathbb{N}$

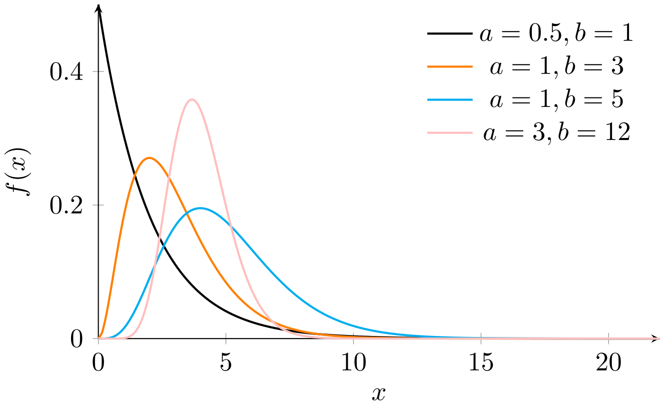


Figure 3.6 Graph of Gamma distribution's density function for different combinations of a, b. Adapted from [2]

- $\Gamma(p+1) = p\Gamma(p)$
- $\Gamma(1) = 1$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- $\Gamma(n) = (n-1)!$
- $\Gamma(p) = \Gamma(p-1)\Gamma(p-1)$.

As such, the gamma function can be perceived as an interpolation of the factorial function for non-natural[5] parameters [79][63]. To compute, numeric algorithms have to be used in most cases [79].

Definition 3.15. *Non-negative random variable X has a gamma distribution with parametrs $a > 0, b > 0$, denoted $X \sim \mathcal{G}(a, b)$, if it has a continuous distribution with probability density*

$$f_X(X) = \frac{a^b}{\Gamma(b)} x^{a-1} \exp(-bx). \tag{3.21}$$

The two parameters a, b define the scale and shape, respectively, as can be also visible in figure 3.6. Gamma distribution's mean value and variance are very simple:

$$\mathbb{E}[X] = \frac{b}{a}, \text{var}X = \frac{b}{a^2}. \tag{3.22}$$

3.3 Conjugate prior

"Although Bayes' theorem, the cornerstone of Bayesian Statistics, is mathematically simple, its implementation can prove troublesome [80]". This issue has already been mentioned - it is caused by the normalising denominator, integrating over estimated parameter's domain, which might result in the product of a prior distribution and a likelihood function not being integrable. Two approaches are described by [80], one focused on deriving pairs of likelihood functions and prior distributions that provide tractable solutions, whilst the other approaches the issue numerically. The former gives shape to families of the so-called conjugate priors. Both approaches have their disadvantages, with the former narrowing down possible distributions to use for modeling to retain the desired properties, whilst the later can be computionally expensive, especially as the size of the problem grows.

[5] Natural numbers are positive integers.

▶ **Definition 3.16** (Conjugate prior distribution [81]). *Let $y|x,\theta$ follow a distribution from the exponential family of distributions 3.2. Prior distribution θ with hyperparameter Ξ, and the distribution ν is conjugated to if, if its probability density function has the form*

$$\pi(\theta) = q(\Xi,\nu)g(\theta)^{\nu}\exp\left(\eta(\theta)^{\intercal}\Xi\right), \tag{3.23}$$

where

- $\nu \in \mathbb{R}^{+}$
- $q(\Xi,\nu)$ *is a known function*
- $g(\theta)$ *is a normalisation function.*

Conjugate priors are prior probability distributions that belong to the same family as the posterior probability distribution after observing data. When the prior distribution and the posterior distribution belong to the same family, it makes the mathematical computations for Bayesian inference more straightforward. In other words, if the prior and posterior distributions are conjugate, then the updated posterior distribution can be calculated using a closed-form expression, rather than numerical integration. An extensive list of conjugate priors is derived and described by [80]. The table 3.2 records a handful of distributions and their respective conjugate prior distributions [81], with an extensive list provided by [82].

Data model (likelihood)	Conjugate prior distribution
Normal with known variance	Normal
Normal with unknown variance	Normal inverse-gamma
Normal with known mean μ	Gamma
Bernoulli	Beta
Poisson	Gamma
Multinomial	Dirichlet

Table 3.2 Table of selected conjugate priors. Source [81], [82].

The Bayes theorem is commonly interpreted as an inference of unknown parameters based on prior knowledge and (newly) collected data. To combine the expertise of prior knowledge and information gathered from observations, a likelihood function is used.

The use of conjugate priors simplifies Bayesian inference because the posterior distribution can be obtained analytically [83]. Additionally, the resulting posterior distribution has the same functional form as the prior distribution, which makes it easier to interpret the results. However, it is important to note that using conjugate priors may not always be the most appropriate choice, and in some cases, non-conjugate priors may be necessary to model the data accurately.

The exponential family of distributions described in 3.2 contains a large scope of conjugate priors, of which many are commonly used distributions.

3.4 Methods of inference

There are many different approaches to the problem at hand under the umbrella of Bayesian modeling. These include [84], but are not limited to:

1. Bayesian Model Averaging [85]
2. Importance sampling
3. Laplace approximation

4. Markov Chain Monte Carlo (MCMC)

5. Variational Bayes,

and their many variations. Of the extensive list, the focus will be on Variational Bayes, given its tractability and ease of computability.

3.4.1 Variational Bayes

The goal is to compute the posterior distribution over a set of unknown parameters given some observed data. However, in many cases, the posterior distribution cannot be computed analytically, and so some form of approximation is necessary. Variational Bayes is one such approximation method that seeks to find a tractable approximation to the true posterior distribution by optimizing a simpler distribution, known as the variational distribution. The idea is to choose a family of distributions that is simple enough to be tractable but flexible enough to approximate the true posterior distribution.

The optimisation problem is formulated as the minimisation of the Kullback-Leibler divergence 3.17 between the true posterior distribution and the variational distribution, which is the expected amount of information lost by substituting the true posterior distribution with its variational approximation:

▶ **Definition 3.17** (Kullback-Leibler divergence for continuous variables[86])**.** *Let $p(x)$ and $q(x)$ be two probability distributions of a continuous random variable x.*

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x_i) \ln \frac{p(x_i)}{q(x_i)} \mathrm{d}x. \tag{3.24}$$

The Kullback-Leibler divergence is not a metric, as it violates the symmetry prerequisite, i.e. $D_{KL}(p||q) \neq D_{KL}(q||p)$ [87].

▶ **Theorem 3.18** (Variational Bayes [84])**.** *Let $f(\theta|D)$ be the posterior probability density function of multivariate parameter θ. The parameter θ is partitioned into $\theta = (\theta'_1, \theta'_2, \ldots, \theta'_q)$. Let weird $\breve{f}(\theta|D)$ be an approximate probability density function restricted to the set of conditionally independent distributions on $\theta_1, \theta_2, \ldots, \theta_q$:*

$$\breve{f}(\theta|D) = \breve{f}(\theta_1, \theta_2, \ldots, \theta_q) = \prod_{i=1}^{q} \breve{f}(\theta_i|D). \tag{3.25}$$

Then the minimum of the Kullback-Leibler distance,

$$\tilde{f}(\theta|D) = \arg\min_{\breve{f}(.)} D_{KL}(\breve{f}(\theta|D)||f(\theta|D)), \tag{3.26}$$

is reached for

$$\tilde{f}_i(\theta|D) \propto \exp\left(\mathbb{E}_{\tilde{f}(\theta_{\setminus i}|y)}[\ln f(\theta_1, \ldots, \theta_q, y)]\right), i = 1, \ldots, q, \tag{3.27}$$

where $\theta_{\setminus i}$ stands for all variables from a given set (θ), excluding the ith variable (θ_i), i.e. $\theta_{\setminus i} := \theta \setminus \{\theta_i\}$.

The approximating distribution is usually chosen from a family of distributions that is tractable, such as a Gaussian distribution. The Variational Bayes algorithm starts by specifying a prior distribution and a likelihood function, and then derives an expression for the posterior distribution. The posterior distribution is then approximated by an approximating distribution, which is usually a Gaussian distribution. The parameters of the approximating distribution are estimated under the hood by minimizing the Kullback-Leibler divergence with the true posterior distribution. This is done using an iterative optimization algorithm.

Further, per [84], the conditionally independent elements of 3.27 are (VB)-marginals. Variational Bayes attempts to break through with a practical algortihm compared to the application of sampling techniques and approximation methods in mixture models [88].

Variational Bayes has several advantages over other Bayesian inference methods. It is computationally efficient and can be used to estimate the posterior distribution of large datasets. It also provides a way to estimate the posterior distribution in cases where the likelihood function is intractable or difficult to compute [89].

As will become apparent when describing individual models later on, the approximation boils down to the following approach [90]:

1. Select appropriate models for approximated variables.
2. For each parameter derive a posterior distribution based on the joint probability distribution.
3. Assign non-informative prior values to the variables.
4. Iteratively update approximations within a fixed number of iterations (or, alternatively, specify a terminating condition - ideally convergence).

Chapter 4

Controlled emission modeling

4.1 Implementation overview

Models were implemented in Python 3.11 [91], which was chosen due to the author's familiarity with the language, as well as the wide range of libraries and packages for computation, machine learning and modeling, of which the following used are the most noteworthy[1]:

1. Jupyter notebook [92]
2. SciPy [93], as well as scikit [94]
3. NumPy [95]
4. SymPy [96]
5. Pandas [97]
6. Matplotlib [98]

The project, which is in part described in this thesis, including data manipulation, preprocessing, tests, interactive notebooks for presentation, etc. is publically available from the faculty's oficial gitlab at the following link: koristo1/dpr. The repository also includes the ETEX dataset, but, due to its sheer size, not the data of Chernobyl measurements.

Following sections describe a single model at a time, and can be divided into two groups:

1. frequentist models - linear and ridge regression
2. VB models - bayesian ridge, sparse, smooth regression and LDL,

with the latter following a similar structure:

1. Define prior distributions
2. Infer likelihood function
3. Infer posterior distributions and their shaping parameters
4. Describe shaping parameter functional dependencies
5. Describe calculation as pseudocode
6. Evaluate models on ETEX data.

Finally, all models are compared in table 4.1 on the ETEX dataset.

[1] in alphabetical order.

4.2 Classic regression

4.2.1 Formulation

To provide a contrast to the various bayesian regression methods shown later, classic linear (and ridge) regression were also evaluated as representatives of the classic, frequentist vanilla methods.

In the case of ETEX data, the matrix $\boldsymbol{X}^\intercal\boldsymbol{X}$ is singular, and as such does not have an inversion. This poses an impenetrable barrier for an analytical solution via ordinary least squares. Not all is lost, introduction of regularisation to the model solves this issue.

Ridge regression introduces a hyperparameter, hereforth referred to as λ, which gauges the pressure on the size of coefficients $\boldsymbol{\beta}$. Given its hyperparametric nature, it is model and data specific, and requires tuning. One such method to find a well-behaved[2] λ is the L-curve method (figure 4.1, comparing the norm of a regularised solution against the norm of the corresponding residual norms. The actual value of λ resides in the ”elbow” of the graphs. That is, in layman’s terms, the point where the curve changes from the vertical segment to the horizontal segment (roughly $(50, 1.55)$ in the graph 4.1). Through this analysis, the penalty term λ is set to 10^{-4}.

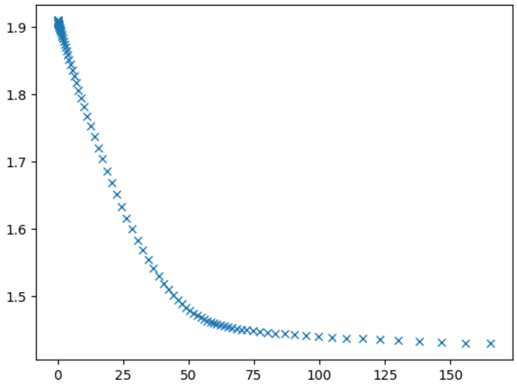


Figure 4.1 L-curve of the ETEX dataset for determining λ penalty of ridge regression.

4.2.2 Performance

Despite its simplicity and non-iterative nature, ridge regression has substantial results. The activity of the approximated $\widehat{\boldsymbol{\beta}}$ lies in the correct interval $[60, 70]$ and the value spike to more than half of the real $\boldsymbol{\beta}$ values, as can be plainly observed in figure 4.2. However, there is also significant activity outside of the window, and dips below 0 once or twice.

The reconstruction given by ridge regression with $\lambda = 10^{-4}$ is shown in figure 4.3. There is a general understatement of the values, visible from the upward (in the sense of increasing value) tendency of the data, as well as two clear outliers. There are also a few points below the red dotted regression line, which correspond to $\widehat{\boldsymbol{\beta}}$ being non-zero outside the correct activity window, as described prior.

The metrics[3] measured for this models are

- $MSE = 7.1020 \times 10^{-4}$

[2]

[3] A summary and comparison of all models can be found in table 4.1.

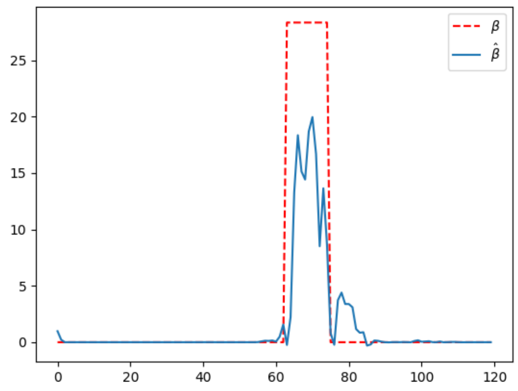


Figure 4.2 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a ridge regression model (blue).

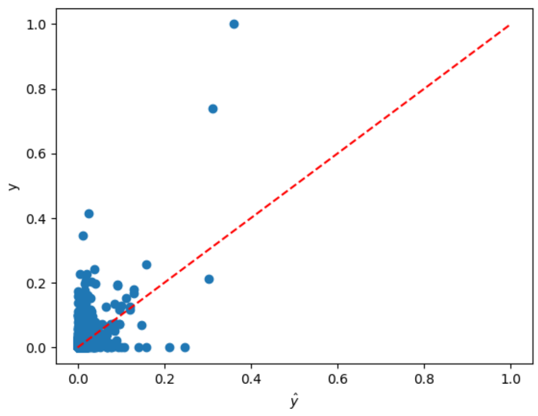


Figure 4.3 Comparison of measured data $\boldsymbol{y}$ (y-axis) and linear ridge regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis).

- $MAE = 65.2597 \times 10^{-4}$
- $RMSE = 266.4950 \times 10^{-4}$

Performance could be further improved by data preprocessing, standardisation to support uniform regularisation of each individual feature. A plethora of other approaches related to linear regression could be applied to the data, such as the method of gradient descent, different forms of regularisations, random forest regression, etc. However, that is not the goal of the focus of this thesis.

4.3 Bayesian ridge regression

This model is the entry point to Bayesian regression models, others following expand on it by adding more variables and parameters. It is the Bayesian counterpart of ”classic” ridge regression, which models the data followingly

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon. \tag{4.1}$$

The random vector of noise ε is assumed to have a mean value of 0[4] and to be normally distributed with variance ω^{-1} [5]

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n), \boldsymbol{\varepsilon_i} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \omega^{-1}\right) \tag{4.2}$$

Consequently, as ε_i is a random variable, so is $\boldsymbol{y}$, as it is a linear transformation (addition of $\boldsymbol{X}\boldsymbol{\beta}$).

4.3.1 Establishing prior distributions

The target variable $\boldsymbol{y}$ follows a normal distribution conditioned on two parameters, $\boldsymbol{\beta}$ and ω, resulting in the model 4.3:

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \omega) = \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \omega^{-1}\boldsymbol{I}_p\right) = (2\pi)^{-\frac{p}{2}}|\omega^{-1}\boldsymbol{I}_p|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal \omega(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right). \tag{4.3}$$

Instead of defining $\boldsymbol{\beta}, \omega$, and their intermediate values and approximations as scalar values, as is the case in linear regression, they are represented by their respective distributions. Prior distributions of $\boldsymbol{\beta}$ and ω are selected from the exponential family, specifically to be conjugate with the model $f(\boldsymbol{y}|\boldsymbol{\beta}, \omega)$. Distribution of $\boldsymbol{\beta}$ is chosen as normal distribution, with mean value $\mu_{\boldsymbol{\beta}} = 0$ and variance $\sigma^2_{\boldsymbol{\beta}} = \boldsymbol{I}_n$

$$f(\boldsymbol{\beta}) = \mathcal{N}\left(0, \boldsymbol{I}_n\right) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{I}_n|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\intercal \boldsymbol{I}_n \boldsymbol{\beta}\right). \tag{4.4}$$

For ω, the Gamma distribution is selected as the most suitable, with prior shaping parameters c_0, d_0 with values set non-informatively to $c_0 = d_0 = 10^{-10}$

$$f(\omega) = \mathcal{G}(c_0, d_0) = \frac{c_0^{d_0}}{\Gamma(d_0)}\omega^{c_0 - 1} \exp\left(-d_0\omega\right) \tag{4.5}$$

The joint likelihood (under the previously established assumption of conditional independence) is a product of the individual prior probabilities and the model of $\boldsymbol{y}$

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\beta}, \omega) = f(\boldsymbol{y}|\boldsymbol{\beta}, \omega) f(\boldsymbol{\beta}) f(\omega) = (2\pi)^{-\frac{p}{2}}|\omega^{-1}\boldsymbol{I}_p|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal \omega(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \times \\ (2\pi)^{-n\frac{1}{2}}|\boldsymbol{I}_n|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\intercal \boldsymbol{I}_n \boldsymbol{\beta}\right) \times \frac{c_0^{d_0}}{\Gamma(d_0)}\omega^{c_0 - 1} \exp\left(-d_0\omega\right). \end{aligned} \tag{4.6}$$

This model can be simplified by applying a logarithmic function to it. This step will be later undone by applying its inverse, the exponential function, resulting in an identity. This allows to utilise the following selected properties of logarithm

- product property: $\ln(xy) = \ln(x) + \ln(y)$
- power property: $\ln(x^a) = a\ln(x)$,

[4] A very common and weak assumption.

[5] The variance of ε_i is denoted ω^{-1} for the sake of simplification in later equations, and as such is just a convenient notation.

$$\ln f(\boldsymbol{y}, \boldsymbol{\beta}, \omega) = \ln f(\boldsymbol{y}|\boldsymbol{\beta}, \omega) + \ln f(\boldsymbol{\beta}) + \ln f(\omega). \tag{4.7}$$

To finish the logarithmic form of the joint probability 4.6, logarithms of individual functions are as follows, utilising basic (and linear) algebra

$$\ln f(\boldsymbol{y}) = -\frac{p}{2}\ln(2\pi) + \frac{p}{2}\ln(\omega) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal \omega(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{4.8}$$

where the second addendum of 4.8 $\ln(|\omega^{-1}\boldsymbol{I}_p|^{-\frac{1}{2}})$ is explained in 4.9, exploiting the determinant of a diagonal matrix and the aforementioned logarithmic properties.

$$\ln(|\omega^{-1}\boldsymbol{I}_p|^{-\frac{1}{2}}) = \frac{1}{2}\ln(|\omega\boldsymbol{I}_p|) = \frac{1}{2}\ln(\prod_{i=1}^{p}\omega) = \frac{1}{2}\sum_{i=1}^{p}\ln(\omega) = \frac{p}{2}\ln(\omega) \tag{4.9}$$

$$\ln f(\boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{I}_n|) - \frac{1}{2}(\boldsymbol{\beta}^\intercal\boldsymbol{I}_n\boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}(\boldsymbol{\beta}^\intercal\boldsymbol{I}_n\boldsymbol{\beta}) \tag{4.10}$$

$$\ln f(\omega) = \ln\left(\frac{c_0^{d_0}}{\Gamma(d_0)}\omega^{c_0-1}\exp(-d_0\omega)\right) = \ln\left(\frac{c_0^{d_0}}{\Gamma(d_0)}\right) + (c_0 - 1)\ln\omega - d_0\omega \tag{4.11}$$

Altogether, the function 4.7 expanded is as follows:

$$\begin{aligned}\ln f(\boldsymbol{y}, \boldsymbol{\beta}, \omega) = -\frac{p}{2}\ln(2\pi) + \frac{p}{2}\ln(\omega) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal\omega(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\\ -\frac{n}{2}\ln(2\pi) - \frac{1}{2}(\boldsymbol{\beta}^\intercal\boldsymbol{I}_n\boldsymbol{\beta})\\ +\ln\left(\frac{c_0^{d_0}}{\Gamma(d_0)}\right) + (c_0 - 1)\ln\omega - d_0\omega.\end{aligned} \tag{4.12}$$

4.3.2 Deriving posterior distributions

Having the (complex) model of joint likelihood of prior distributions $\boldsymbol{\beta}, \omega$ and model $\boldsymbol{y}$, the posterior distributions are derived using the Variational Bayes theorem for each parameter θ_i: $\boldsymbol{\beta}, \omega$.

Using proportionality ($\propto$), it is possible to perceive any term that is independent of parameter θ_i in the expression of $\tilde{f}(\theta_i|\boldsymbol{y})$ as a constant, and ignore any such term in the marginal distribution. The exponential function is applied to the θ_i-based proportion of the logarithmic joint probability, effectively cancelling out the logarithm along the way

$$\begin{aligned}\tilde{f}(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp(-\frac{1}{2}(\widehat{\omega}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})) - \frac{1}{2}(\boldsymbol{\beta}^\intercal\boldsymbol{I}_n\boldsymbol{\beta})) =\\ \exp(\widehat{\omega}\boldsymbol{y}^\intercal\boldsymbol{X}\boldsymbol{\beta} - \frac{1}{2}\widehat{\omega}\boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\beta} - \frac{1}{2}(\boldsymbol{\beta}^\intercal\boldsymbol{I}_n\boldsymbol{\beta})) =\\ \exp(-\frac{1}{2}(\boldsymbol{\beta}^\intercal(\widehat{\omega}\boldsymbol{X}^\intercal\boldsymbol{X} + \boldsymbol{I}_n)\boldsymbol{\beta} - 2\widehat{\omega}\boldsymbol{y}^\intercal\boldsymbol{X}\boldsymbol{\beta})).\end{aligned} \tag{4.13}$$

As it was established that $\boldsymbol{\beta}$ follows a multivariate normal distribution, i.e. $\boldsymbol{\beta} \sim \mathcal{N}\left(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right)$, and so its probability density function can be written as

$$f(\boldsymbol{\beta}) = (2\pi)^{-\frac{n}{2}} |\Sigma_{\boldsymbol{\beta}}^{-1}| \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \mu_{\boldsymbol{\beta}})^{\intercal} \Sigma_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \mu_{\boldsymbol{\beta}})\right). \tag{4.14}$$

The argument of the exponential function can be expanded to clearly separate a quadratic, a linear and a constant term:

$$f(\boldsymbol{\beta}) = (2\pi)^{-\frac{n}{2}} |\Sigma_{\boldsymbol{\beta}}^{-1}| \exp\left(-\frac{1}{2}(\boldsymbol{\beta}^{\intercal} \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} - 2\mu_{\boldsymbol{\beta}}^{\intercal} \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} + \mu_{\boldsymbol{\beta}}^{\intercal} \Sigma_{\boldsymbol{\beta}}^{-1} \mu_{\boldsymbol{\beta}})\right), \tag{4.15}$$

from which the shaping parameters $\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}$ can be inferred. Under proportionality, any term that is not in relation with $\boldsymbol{\beta}$ is considered constant, therefore the equality of the two equations under proportionality is based on the equality of its quadratic (4.16) and linear (4.17 terms, respectively

$$\Sigma_{\boldsymbol{\beta}}^{-1} = \widehat{\omega} \boldsymbol{X}^{\intercal} \boldsymbol{X} + \boldsymbol{I}_n \tag{4.16}$$

$$\mu_{\boldsymbol{\beta}}^{\intercal} \Sigma_{\boldsymbol{\beta}}^{-1} = \widehat{\omega} \boldsymbol{y} \boldsymbol{X}. \tag{4.17}$$

Resulting shaping parameters $\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are just a few adjustments away

$$\Sigma_{\boldsymbol{\beta}} = (\widehat{\omega} \boldsymbol{X}^{\intercal} \boldsymbol{X} + \boldsymbol{I}_n)^{-1} \tag{4.18}$$

$$\mu_{\boldsymbol{\beta}} = (\widehat{\omega} \boldsymbol{X}^{\intercal} \boldsymbol{X} + \boldsymbol{I}_n)^{-1} \widehat{\omega} \boldsymbol{X}^{\intercal} \boldsymbol{y}. \tag{4.19}$$

As for ω, it follows a Gamma distribution, i.e. $f(\omega) = \frac{c^d}{\Gamma(d)} \omega^{c-1} \exp\left(-d\omega\right)$, and can be rewritten in the exponential family's parametrisation:

$$f(\omega) = \exp\left((c-1)\ln(\omega) - d\omega + \ln(\Gamma(c)) + c\ln(d)\right., \tag{4.20}$$

and its marginal distribution has the following shape

$$\begin{aligned} \tilde{f}(\omega|\boldsymbol{y}) \propto \exp\left(\ln\left(|\omega^{-1}\boldsymbol{I}_p|^{-\frac{1}{2}}\right) - \frac{1}{2}\omega(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\intercal}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (c_0 - 1)\ln(\omega) - d_0\omega\right) = \\ \exp\left(\left(\frac{p}{2} + c_0 - 1\right)\ln(\omega) - \frac{1}{2}(\boldsymbol{y}\boldsymbol{y}^{\intercal} - 2y^{\intercal}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta}^{\intercal}\widehat{\boldsymbol{X}^{\intercal}\boldsymbol{X}}\boldsymbol{\beta} + 2d_0)\omega\right). \end{aligned} \tag{4.21}$$

The form 4.21 shows the linear and logarithmic terms, as in the case of $\boldsymbol{\beta}$, can be set equal to the linear and logarithmic terms of 4.20

$$c - 1 = \frac{p}{2} + c_0 - 1 \tag{4.22}$$

$$-d = -d_0 - \frac{1}{2}(\boldsymbol{y}\boldsymbol{y}^{\intercal} - 2\boldsymbol{y}^{\intercal}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta}^{\intercal}\widehat{\boldsymbol{X}^{\intercal}\boldsymbol{X}}\boldsymbol{\beta}). \tag{4.23}$$

In conclusion, the parameters c, d are updated in the following fashion

$$c = c_0 + \frac{p}{2} \tag{4.24}$$

$$d = d_0 + \frac{1}{2}(\boldsymbol{y}\boldsymbol{y}^{\intercal} - 2\boldsymbol{y}^{\intercal}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta}^{\intercal}\widehat{\boldsymbol{X}^{\intercal}\boldsymbol{X}}\boldsymbol{\beta}). \tag{4.25}$$

Both posterior distributions have their shaping parameters explicitly stated, and thanks to the conjugate prior nature of them, the type of prior and posterior distributions does not change. The shaping parameters define the prior distributions for a subsequent calculation, paving way to an iterative algorithm. The figure 4.4 shows the dependencies between individual parameters (and data $\boldsymbol{y}, \boldsymbol{X}$). It is an oriented graph, where the edge between nodes $a \to b$ represents dependency of a on b, i.e. the value of b is a part of calculation of a. The graph cannot be topologically sorted

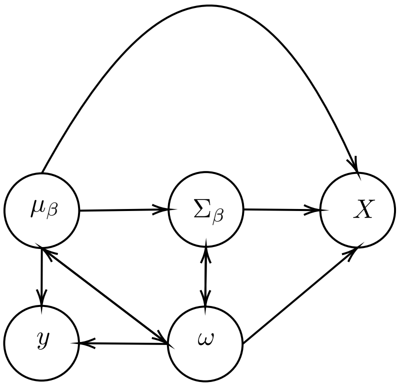


Figure 4.4 Bayesian ridge regression parameters dependency graph. The graph is directed with multiple cycles.

Algorithm 1: Bayesian ridge regression algorithm

```
input  : $\boldsymbol{y} \in \mathbb{R}^{n,1}$, $\boldsymbol{X} \in \mathbb{R}^{n,p}$
output: $\mu_{\boldsymbol{\beta}} \in \mathbb{R}^{p,1}$
init
    $\omega \leftarrow \max(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$
for $i \leftarrow 1$ to 100 do
    $\Sigma \leftarrow (\omega \boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{I}_n)^{-1}$
    $\mu \leftarrow \Sigma(\omega \boldsymbol{X}^\intercal \boldsymbol{y})$
    $\boldsymbol{\beta}, \boldsymbol{\beta\beta} \leftarrow$ Bottom truncated normal distribution's first two moments$(\mu, \mathrm{diag}(\Sigma)^{\frac{1}{2}}))$
    $\mathrm{var}_{\boldsymbol{\beta\beta}} \leftarrow \boldsymbol{\beta\beta} - \boldsymbol{\beta}^2$
    $\boldsymbol{\beta\beta}^\intercal \leftarrow \boldsymbol{\beta\beta}^\intercal + \mathrm{var}_{\boldsymbol{\beta\beta}}$
    $c \leftarrow c_0 + \frac{1}{2}n$
    $d \leftarrow d_0 + \frac{1}{2}((\boldsymbol{y}^\intercal \boldsymbol{y}) - 2\boldsymbol{y}^\intercal X \mu_{\boldsymbol{\beta}} + \mathrm{Tr}(\boldsymbol{\beta\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}))$
    $\omega \leftarrow \frac{c}{d}$
end
```

as it contains directed cycles, and as such, there is not one way to order the calculations in a linear fashion. Therefore, initial prior values will have to be set for some, if not all parameters.

With the posterior shaping parameters expressed in relation to the prior distributions; data $\boldsymbol{y}, \boldsymbol{X}$ and initial prior values, the shaping parameters are updated iteratively according to the algorithm expressed by pseudocode 1. The order of the parameter calculation approximation is theoretically arbitrary, as the only change introduced by a different order would be the need for different initial values. However, as $\boldsymbol{\beta}$ is the modeled variable[6], it is more feasible to follow the "importance hierachy" of the interdependent parameters.

Previously stated, in the expression of a marginal distribution, the marginal distribution works with the expected values of all other parameters other than the parameter whose marginal distribution is being expressed,$\mathbb{E}[\boldsymbol{\beta}] = \mu_{\boldsymbol{\beta}}, \mathbb{E}[\omega] = \frac{c}{d}$.

To further boost the model's quality, two approaches are incorporated. Firstly, normal distributions are replaced with truncated normal distributions, with a lower bound of 0 and no upper bound. This change is to ensure non-negativity, to better correspond with the theoretical model under the hood[7]. This small change brings a cascade of changes, which are luckily isolated into

[6] technically, $\widehat{\boldsymbol{y}} = \boldsymbol{X\beta}$ is the target variable, however, it is entirely derived from $\boldsymbol{\beta}$.

[7] A similar approach should be employed in the case of any non-negative random variable, such as length or

the calculation of the relevant mean value and variance as described in 3.2.2. The second change is the introduction of a slight softening of $\boldsymbol{\beta}$.

4.3.3 Performance

The algorithms were implemented, tested, debugged and fine-tuned on data from the European tracer experiment 1.4, where both the ground truth of the time vector $\boldsymbol{\beta}$ and the target variable $\boldsymbol{y}$ are known, and as such it provides a potent framework for development. There were 100 iterations altogether; this number was determined empirically from previous experiments in the field. The initial value of ω was set to $(\max(\boldsymbol{X}^\intercal\boldsymbol{X}))^{-1}$, $\boldsymbol{\beta}$ did not need any initial value as it is the first to be computed in the algorithm 1, and any initial value would be discarded rightaway.

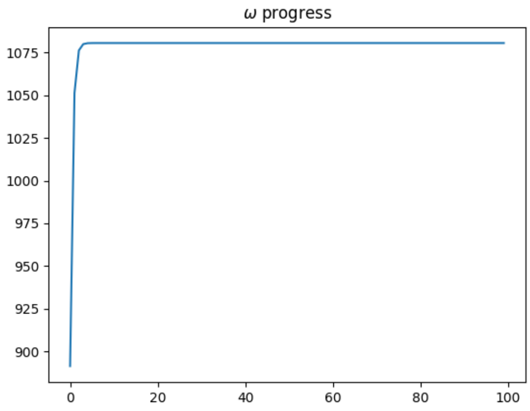


Figure 4.5 ω parameter values during 100 iterations using bayesian ridge regression on ETEX data.

As can be seen in figure 4.5, the parameter ω, which determines the variance matrix of $\boldsymbol{y}$, quickly converges from the initial value to ≈ 1100, within the first 5 iterations of the algorithm.

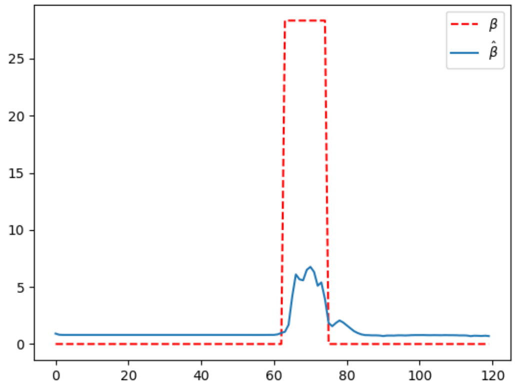


Figure 4.6 Comparison of real β from ETEX (red) with $\widehat{\beta}$ as approximated by a bayesian ridge regression model (blue) after 100 iterations.

weight.

Figure 4.7 compares the value of the reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on the x-axis against the real $\boldsymbol{y}$ n on the y-axis. The red dashed line models the perfect regression line, upon which points $(\hat{y}_i, y_i)$ would lie if the reconstruction fit the data perfectly. Clearly, that is not the case, as most reconstructed values are much lower than the real ones, as can also be seen in figure 4.6, where the main activity between 60 and 80 on the x-axis is a far cry from the reconstructed values. This simple algorithm managed to grasp the nature of the process, showing a degree of sensitivity in correct interval, but the values are far from ideal. It is also worth to notice that outside the activity in the $[60, 70]$ interval, where $\boldsymbol{\beta} = 0, \widehat{\boldsymbol{\beta}}$ is slightly[8] larger than 0, corresponding to some deviations in the direction of x-axis in the comparing figure 4.7.

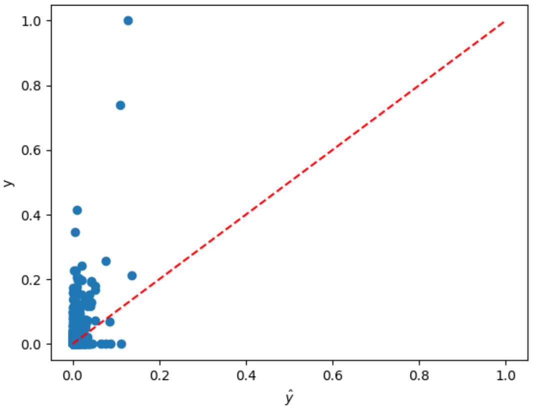


Figure 4.7 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian ridge regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

The model's performance is assessed using mean square error, root mean square error and mean absolute error between the reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\beta}$ and ground truth $\boldsymbol{y}$:

- $MSE = 9.2399 \times 10^{-4}$
- $MAE = 71.2809 \times 10^{-4}$
- $RMSE = 303.9722 \times 10^{-4}$

and will be useful for comparison with other models.

4.4 Bayesian sparse regression

Sparse regression allows the model to change the variance of the parameter β by replacing the static covariance matrix $\boldsymbol{I}_n$ with a diagonal matrix with $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ on its diagonal, where $v_i = \mathcal{G}(a_i, b_i)$. Certain parts of the model stay the same, whilst others change due to the propagation of this newly added set of parameters. Parameters $v_1, \ldots, v_n$ are assumed to be also conditionally independent of all other parameters, and so the joint probability is in fashion similar to the previous model.

[8] A very vague description

4.4.1 Establishing prior distributions

The prior distribution of $\boldsymbol{\beta}$ is set as follows

$$f(\boldsymbol{\beta}|\boldsymbol{v}) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & v_n \end{pmatrix}\right), \tag{4.26}$$

the distribution is still zero centered, but now the covariance matrix is parametric and can change, amplifying or dampening $\boldsymbol{\beta}$ as set during the latter iterative algorithm. v_i is iid having a prior Gamma distribution with parameters a_0, b_0:

$$f(v_i) = \frac{a_0^{b_0}}{\Gamma(b_0)} v_i^{a_0-1} \exp\left(-b_0 v_i\right), \tag{4.27}$$

the covariance matrix $\boldsymbol{V}$ is regular if $\forall i : v_i \neq 0$, and therefore under that condition has an inverse $\boldsymbol{V}^{-1}$. The prior distribution of ω is left unchanged, as is the model $\boldsymbol{y}$. The two prior parameters a_0, b_0 are set to non-informative 10^{-10}.

The joint probability function is, under the assumption of independence,

$$f(\boldsymbol{y}, \boldsymbol{\beta}, \omega, \boldsymbol{v}) = f(\boldsymbol{y}|\boldsymbol{\beta}, \omega) f(\boldsymbol{\beta}|\boldsymbol{v}) f(\omega) \prod_{i=1}^{n} f(v_i) = (2\pi)^{-\frac{p}{2}} |\omega^{-1} \boldsymbol{I}_p|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal \omega (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \times$$
$$(2\pi)^{-n\frac{1}{2}} |\boldsymbol{V}^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\intercal \boldsymbol{V} \boldsymbol{\beta}\right) \times \frac{c_0^{d_0}}{\Gamma(d_0)} \omega^{c_0-1} \exp\left(-d_0 \omega\right) \times \prod_{i=1}^{n} \left(\frac{a_0^{b_0}}{\Gamma(b_0)} v_i^{a_0-1} \exp\left(-b_0 v_i\right)\right). \tag{4.28}$$

The newly added term $\ln f(v_i)$ can be rewritten in the following fashion

$$\ln f(v_i) = \ln\left(\frac{a_0^{b_0}}{\Gamma(b_0)} v_i^{a_0-1} \exp\left(-b_0 v_i\right)\right) = \ln\left(\frac{a_0^{b_0}}{\Gamma(b_0)}\right) + (a_0 - 1)\ln v_i - b_0 v_i, \tag{4.29}$$

the probability of $\boldsymbol{v}$ is then just a simple product (again, thanks to the conditional independence assumption)

$$\ln\left(\prod_{i=1}^{n} f(v_i)\right) = \sum_{i=1}^{n} \left(\ln\left(f(v_i)\right)\right) = \sum_{i=1}^{n} \left(\ln\left(\frac{a_0^{b_0}}{\Gamma(b_0)}\right) + (a_0 - 1)\ln v_i - b_0 v_i\right). \tag{4.30}$$

The $\boldsymbol{\beta}$ prior model has gotten slightly more complicated along the way, its probability density function $f(\boldsymbol{\beta})$ changes to

$$f(\boldsymbol{\beta}|\boldsymbol{v}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}^\intercal \boldsymbol{V} \boldsymbol{\beta})\right), \tag{4.31}$$

and its logarithmic form is as follows

$$\ln f(\boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{\beta}^\intercal \boldsymbol{V} \boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{n}\ln(v_i) - \frac{1}{2}(\boldsymbol{\beta}^\intercal \boldsymbol{V} \boldsymbol{\beta}). \tag{4.32}$$

It is of note that the covariance matrix $\boldsymbol{V}$ is diagonal, and as such its determinant is equal to the product of the diagonal elements and therefore the following holds

$$\ln|\boldsymbol{V}| = \ln\prod_{i=1}^{n} v_i = \sum_{i=1}^{n}\ln(v_i). \tag{4.33}$$

$f(\omega)$ and $f(\boldsymbol{y})$ remain unchanged, as do the prior constants c_0, d_0. The expansion of the logarithmic form of $f(\boldsymbol{y}, \boldsymbol{\beta}, \omega, \boldsymbol{v})$ is, put together, described by the following

$$\begin{aligned} \ln f(\boldsymbol{y}, \boldsymbol{\beta}, \omega, \boldsymbol{v}) = -\frac{p}{2} \ln|\omega| - \frac{1}{2}\omega \boldsymbol{y}\boldsymbol{y}^\intercal + \omega \boldsymbol{y}^\intercal \boldsymbol{X}\boldsymbol{\beta} + \omega(\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\beta}) - \frac{n}{2}\ln(2\pi) \\ +\frac{1}{2}\sum_{i=1}^{n} \ln(v_i) - \frac{1}{2}(\boldsymbol{\beta}^\intercal \boldsymbol{V}\boldsymbol{\beta}) + \ln\left(\frac{c_0^{d_0}}{\Gamma(d_0)}\right) + (c_0 - 1)\ln\omega - d_0\omega \\ +\sum_{i=1}^{n}\left(\ln\left(\frac{a_0^{b_0}}{\Gamma(b_0)}\right) + (a_0 - 1)\ln v_i - b_0 v_i\right). \end{aligned} \tag{4.34}$$

4.4.2 Deriving posterior distributions

The marginal distributions of $\boldsymbol{\beta}$ is near identical to the one in 4.3, with the only difference being the substitution $\boldsymbol{I}_n \to \boldsymbol{V}^{-1}$. Making use of proportionality, the marginal distribution is then

$$\begin{aligned} \tilde{f}(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp(-\frac{1}{2}\widehat{\omega}(\boldsymbol{y} - \boldsymbol{X}\beta)(\boldsymbol{y} - \boldsymbol{X}\beta)^\intercal - \frac{1}{2}(\boldsymbol{\beta}^\intercal \widehat{\boldsymbol{V}}\beta)) = \\ \exp(\widehat{\omega}\boldsymbol{y}^\intercal \boldsymbol{X}\beta - \frac{1}{2}(\widehat{\omega}\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\beta) - \frac{1}{2}(\boldsymbol{\beta}^\intercal \widehat{\boldsymbol{V}}\beta)) = \\ \exp(-\frac{1}{2}(\boldsymbol{\beta}^\intercal(\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \widehat{\boldsymbol{V}})\beta - 2(\widehat{\omega}\boldsymbol{y}^\intercal \boldsymbol{X})\beta))). \end{aligned} \tag{4.35}$$

$\boldsymbol{\beta}$ still follows a normal distribution, $\boldsymbol{\beta} \sim \mathcal{N}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$[9], using the form that separates the linear and quadratic terms of the exponential function exp, and setting the linear and quadratic terms equal to the linear and quadratic terms of the marginal distribution, the following pair of equations gives way to the shaping parameters

$$\Sigma_{\boldsymbol{\beta}}^{-1} = \widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{V} \tag{4.36}$$

$$\mu_{\boldsymbol{\beta}}^\intercal \Sigma_{\boldsymbol{\beta}}^{-1} = \widehat{\omega}\boldsymbol{y}\boldsymbol{X}. \tag{4.37}$$

Expressing the two shaping parameters of $\boldsymbol{\beta}$ yields the following results

$$\Sigma_{\boldsymbol{\beta}} = (\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{V})^{-1} \tag{4.38}$$

$$\mu_{\boldsymbol{\beta}} = \widehat{\omega}\boldsymbol{y}^\intercal \boldsymbol{X}\Sigma_{\boldsymbol{\beta}} = (\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{V})^{-1}\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{y}. \tag{4.39}$$

Therefore, in each iteration the covariance is added (in place of iN). As for the marginal distribution of v_i, that is where things get complicated. Given there is no information of whether $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ has a joint probability distribution[10], they're modeled individually $\forall i$. The marginal for a given v_i is

$$\tilde{f}(v_i|\boldsymbol{y}) \propto \exp((a_0 - 1 + \frac{1}{2})\ln(v_i) - b_0 + \frac{1}{2}(\widehat{\boldsymbol{\beta}\boldsymbol{\beta}^\intercal}_{ii})v_i,) \tag{4.40}$$

as only the ith element of $\boldsymbol{\beta}$ (β_i) is multiplying v_i, other elements do not partake in that equation.

$$\begin{aligned} \tilde{f}(v_i|\boldsymbol{y}) \propto \exp((a_0 - 1)\ln v_i - b_0 v_i + \frac{1}{2}\ln v_i - (\widehat{\boldsymbol{\beta}\boldsymbol{\beta}^\intercal})_{ii}v_i) = \\ \exp((a_0 - 1 + \frac{1}{2})\ln v_i - (b_0 + (\widehat{\boldsymbol{\beta}\boldsymbol{\beta}^\intercal})_{ii})v_i) \end{aligned} \tag{4.41}$$

[9] Rinse and repeat, the general flow of each model will be the same, with the only change in the number of marginal distributions.

[10] A rather strong assumption would have to be made.

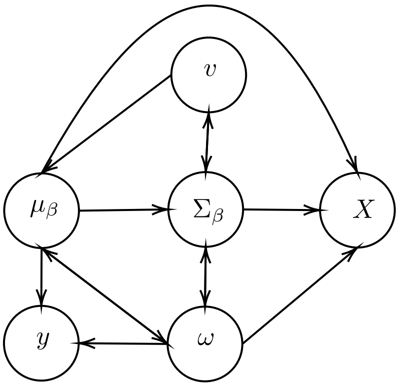


Figure 4.8 Sparse bayesian dependency graph

From that, as v_i is Gamma distributed with parameters a_i, b_i,

$$f(v_i) = \exp\left((a_i - 1)\ln(v_i) - b_i v_i + \ln(\Gamma(a_i)) + a_i \ln(b_i)\right., \tag{4.42}$$

the posterior shaping parameters are given as

$$a_i = a_0 + \frac{1}{2} \tag{4.43}$$

$$b_i = b_0 + \frac{1}{2}(\widehat{\boldsymbol{\beta}\boldsymbol{\beta}^\intercal})_{ii}). \tag{4.44}$$

The posterior distribution of ω remains unchanged (see 4.24 and 4.25). Figure 4.8 shows the addition and incorporation of parameters $\boldsymbol{v}$, and as no node is removed from the graph, it contains the previous cycles. Therefore, calculations will require initial values of (some) parameters.

4.4.3 Performance

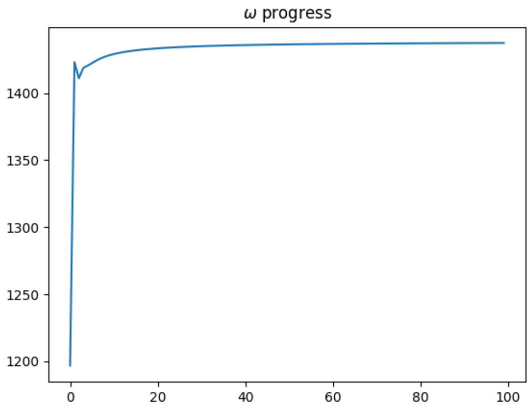


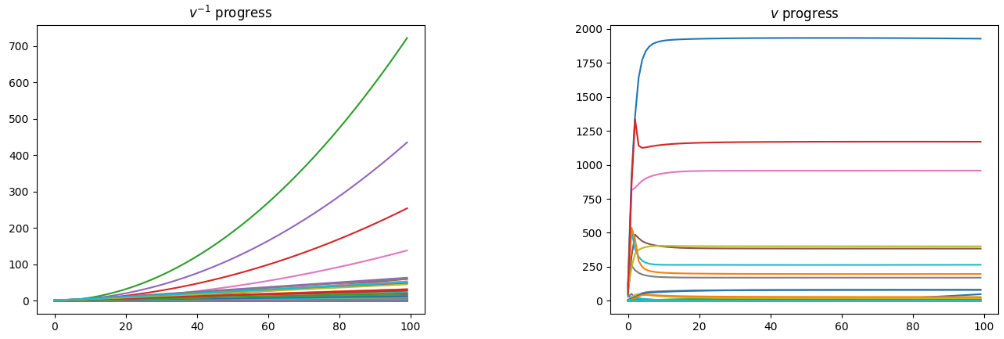
Figure 4.9 ω parameter values during 100 iterations using bayesian sparse regression on ETEX data.

Even though the estimation of ω has not changed, it acts within the calculation of others which have - namely the estimation of $\Sigma_{\boldsymbol{\beta}}$ and $\mu_{\boldsymbol{\beta}}$ (as can be also deducted from the dependency

Algorithm 2: Bayesian sparse regression algorithm

```
input  : $\boldsymbol{y} \in \mathbb{R}^{n,1}$, $\boldsymbol{X} \in \mathbb{R}^{n,p}$
output: $\mu_{\boldsymbol{\beta}} \in \mathbb{R}^{p,1}$
init
    $\omega \leftarrow \max(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$
    $\boldsymbol{V} \leftarrow \boldsymbol{I}_p$
for $i \leftarrow 1$ to 100 do
    $\Sigma \leftarrow (\omega \boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{V})^{-1}$
    $\mu \leftarrow \Sigma(\omega \boldsymbol{X}^\intercal \boldsymbol{y})$
    $\boldsymbol{\beta}, \boldsymbol{\beta\beta} \leftarrow$ Bottom truncated normal distribution's first two moments$(\mu, \operatorname{diag}(\Sigma)^{\frac{1}{2}})$
    $\operatorname{var}_{\boldsymbol{\beta\beta}} \leftarrow \boldsymbol{\beta\beta} - \boldsymbol{\beta}^2$
    $\boldsymbol{\beta\beta}^\intercal \leftarrow \boldsymbol{\beta\beta}^\intercal + \operatorname{var}_{\boldsymbol{\beta\beta}}$
    $c \leftarrow c_0 + \frac{n}{2}$
    $d \leftarrow d_0 + \frac{1}{2}((\boldsymbol{y}^\intercal \boldsymbol{y}) - 2\boldsymbol{y}^\intercal \boldsymbol{X} \mu_{\boldsymbol{\beta}} + \operatorname{Tr}(\boldsymbol{\beta\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}))$
    $\omega \leftarrow \frac{c}{d}$
    $\boldsymbol{a} \leftarrow a_0 + \frac{1}{2}$
    $\boldsymbol{b} \leftarrow b_0 + \frac{1}{2}(\boldsymbol{\beta\beta}^\intercal)_{jj}$
    $\boldsymbol{v} \leftarrow \frac{\boldsymbol{a}}{\boldsymbol{b}}$
    $\boldsymbol{V} \leftarrow \operatorname{diag} \boldsymbol{v}$
end
```

graph 4.8). It once again converges very quickly, with a minor hiccup after a steep rise during the first few iterations, ultimately seemingly converging within the first ≈ 5 iterations.



(a) Per iteration value of $\boldsymbol{v}^{-1}$in sparse bayesian regression on the ETEX dataset.

(b) Per iteration value of $\boldsymbol{v}$ in sparse bayesian regression on the ETEX dataset.

Figure 4.10 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of sparse bayesian regression on the ETEX dataset.

The progress of $\boldsymbol{v}$ is harder to depict, as it consists of n iid random variables. The progress of each v_i is depicted in a shared figure 4.10b. Most of the values stay near zero, which is exactly the behaviour expected from a sparse solution. Most $\boldsymbol{\beta}$ values are zero, or near-zero, as the activity is only in a limited window $[60, 70]$. Clearly, those values are the ones corresponding to the rising[11] curves in the figure. The important takeaway is the monotonicity of each curve, all of them are increasing.

Clearly, and especially in comparison with previous, simpler models, the sparse model has

[11] At various pace

dramatically better results. It fits the ground truth $\boldsymbol{\beta}$ very tightly, overshooting the value in a few instances, and having non-zero values outside the main activity window. This behaviour is invariant with growing number of iterations - convergence has been, apparently, achieved. The results are actually rather similar to classic ridge regression, with a visible increase of values in the activity window.

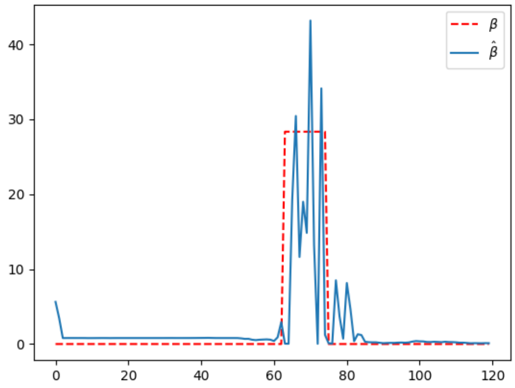


Figure 4.11 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a bayesian sparse model (blue) after 100 iterations.

As visible in figure 4.12, there are units of outliers caused by a very low estimation $\widehat{\boldsymbol{y}}$. Most of estimations are not too dissimilar of the ground truth, mainly due to the majority of $\boldsymbol{\beta}$ being 0. Nevertheless, the model marks an increase in performance and fulfills its role, introducing sparsity, allowing the model to dampen (many) values near 0.

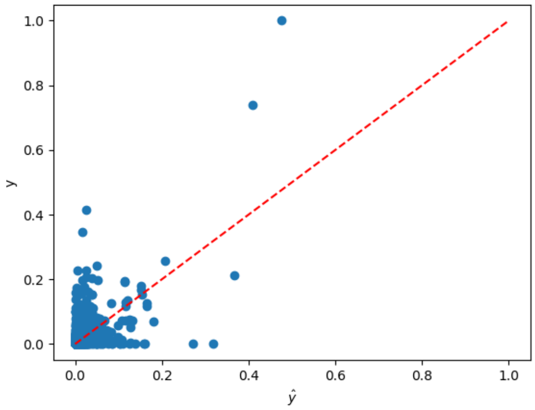


Figure 4.12 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian sparse regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

The metrics measured for this models are

- $MSE = 6.8414 \times 10^{-4}$
- $MAE = 69.3237 \times 10^{-4}$

- $RMSE = 261.5601 \times 10^{-4}$

which are lower than that of Bayesian ridge regression. A step[12] in the right direction, albeit there is discussion to be held, whether a model that is contained to the activity proper, but with lower values, is better than one that fits the activity closer, but with some values exceeding and exceeding outside the activity.

4.5 Bayesian smooth regression

The smooth regression model is a minor modification of the previously shown sparse model 4.4 by keeping the introduced sparsity, but also forcing adjacent values to have similar values.

Importantly, smoothness is a much stronger assumption about the model than sparsity. Where sparsity can be completely ignored thanks to how it is incorporated[13], and can result in a non-sparse model, smoothness can be considered as enforced, therefore performance of a smooth model is very much tied to whether the underlying ground truth data is actually smooth. The distinction will be made crystal clear when comparing the model's performance on ETEX 4.5.3 and Chernobyl 2020 fires 5.2.3 data.

4.5.1 Establishing prior distributions

Smooth beta values are modeled using a discrete first derivative, $f(\nabla\boldsymbol{\beta}|\boldsymbol{v}) \sim \mathcal{N}\left(0, \boldsymbol{V}^{-1}\right)$, where $\nabla \in \mathbb{R}^{n,n}$ is a diagonal matrix with 1 on the diagonal and -1 on the subdiagonal, i.e.:

$$\nabla = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ -1 & 1 & 0 & & 0 \\ \vdots & -1 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & -1 & 1 \end{pmatrix}. \tag{4.45}$$

▶ **Theorem 4.1** (Linear transformation of a multivariate normal distribution [63]). *Given a multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\mu_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$, a vector $\boldsymbol{a}$ and a square matrix $\boldsymbol{C}$ of appropriate dimensions, it holds that*

$$\boldsymbol{a} + \boldsymbol{C}\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a} + \boldsymbol{C}\mu_{\boldsymbol{X}}, \boldsymbol{C}\Sigma_{\boldsymbol{X}}\boldsymbol{C}^\intercal). \tag{4.46}$$

Under theorem 4.1 the final form of $\boldsymbol{\beta}$'s distribution is derived:

$$f(\nabla\boldsymbol{\beta}|\boldsymbol{v}) \sim \mathcal{N}\left(0, \boldsymbol{V}^{-1}\right) \rightarrow f(\boldsymbol{\beta}|\boldsymbol{v}) \sim \mathcal{N}\left(0, \nabla^{-1}\boldsymbol{V}^{-1}(\nabla^{-1})^\intercal\right) = \mathcal{N}\left(0, (\nabla^\intercal\boldsymbol{V}\nabla)^{-1}\right). \tag{4.47}$$

Previously, $\Sigma_{\boldsymbol{\beta}}$ was set to $\boldsymbol{V}$, allowing the variances $v_1, \dots, v_n$ to change, granting the model the ability to easily knock values down to 0. The covariance matrix $\boldsymbol{V}$ is superseded by $\nabla^\intercal\boldsymbol{V}\nabla$, keeping the sparsity capabilities and introducing smoothness. Explicitly stated, the distribution of $\boldsymbol{\beta}$ in the smooth model is

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{v}) = (2\pi)^{-\frac{n}{2}}|(\nabla^\intercal\boldsymbol{V}\nabla)^{-1}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\beta}^\intercal\nabla^\intercal\boldsymbol{V}\nabla)\boldsymbol{\beta})\right) = \\ (2\pi)^{-\frac{n}{2}}|(\nabla^\intercal\boldsymbol{V}\nabla)|^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\beta}^\intercal\nabla^\intercal\boldsymbol{V}\nabla)\boldsymbol{\beta})\right). \end{aligned} \tag{4.48}$$

▶ **Theorem 4.2** (Determinant of matrix product [99]). *Let $\boldsymbol{A_1}, \boldsymbol{A_2}, \dots, \boldsymbol{A_N} \in \mathbb{R}^{n,n}$, then* $|\prod_{i=1}^N A_i| = \prod_{i=1}^N |A_i|$,

[12] With a reasonable learning rate!

[13] Whereby the model sets the values of v_i based on the provided training sample

Using the multiplicative property of square matrices (theorem 4.2), the determinant of $\nabla^\intercal \boldsymbol{V} \nabla$ can be rewritten as noted by the following equation 4.49, as $|\nabla| = |\nabla^\intercal| = 1$ and the fact that $\boldsymbol{V}$ is a diagonal matrix:

$$|\nabla^\intercal \boldsymbol{V} \nabla| = |\nabla^\intercal| \cdot |\boldsymbol{V}||\nabla| = |\boldsymbol{V}| = \prod_{i=1}^{n} v_i. \tag{4.49}$$

The change of $\boldsymbol{\beta}$'s probability density function also naturally causes a different logarithmic form of said density function:

$$\begin{aligned} \ln f(\boldsymbol{\beta}|\boldsymbol{v}) = \frac{1}{2} \ln |\nabla^\intercal \boldsymbol{V} \nabla| - \frac{1}{2}(\boldsymbol{\beta}^\intercal \nabla^\intercal \boldsymbol{V} \nabla \boldsymbol{\beta}) = \\ \frac{1}{2} \ln (\prod_{i=1}^{n} v_i) - \frac{1}{2}(\boldsymbol{\beta}^\intercal \nabla^\intercal \boldsymbol{V} \nabla \boldsymbol{\beta}) = \\ \frac{1}{2} \sum_{i=1}^{n} \ln (v_i) - \frac{1}{2}(\boldsymbol{\beta}^\intercal \nabla^\intercal \boldsymbol{V} \nabla \boldsymbol{\beta}). \end{aligned} \tag{4.50}$$

Put together, the logarithmic joint probability model $\ln f(\boldsymbol{y}, \boldsymbol{\beta}, \omega, \boldsymbol{v})$ is

$$\begin{aligned} \ln f(\boldsymbol{y}, \boldsymbol{\beta}, \omega, \boldsymbol{v}) = -\frac{p}{2} \ln |\omega| - \frac{1}{2} \omega \boldsymbol{y} \boldsymbol{y}^\intercal + \omega \boldsymbol{y}^\intercal \boldsymbol{X} \boldsymbol{\beta} + \omega(\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta}) \\ + \frac{1}{2} \sum_{i=1}^{n} \ln (v_i) - \frac{1}{2}(\boldsymbol{\beta}^\intercal \nabla^\intercal \boldsymbol{V} \nabla \boldsymbol{\beta}) + \ln \left(\frac{c_0^{d_0}}{\Gamma(d_0)} \right) + (c_0 - 1) \ln \omega - d_0 \omega \\ + \sum_{i=1}^{n} \left(\ln \left(\frac{a_0^{b_0}}{\Gamma(b_0)} \right) + (a_0 - 1) \ln v_i - b_0 v_i \right). \end{aligned} \tag{4.51}$$

4.5.2 Deriving posterior distributions

The marginal distributions have changed.

$$\begin{aligned} \tilde{f}(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp (\widehat{\omega} \boldsymbol{y}^\intercal \boldsymbol{X} \boldsymbol{\beta} - \frac{1}{2}(\widehat{\omega} \boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{\beta}^\intercal \nabla^\intercal \boldsymbol{V} \nabla \boldsymbol{\beta}) = \\ \exp (\widehat{\omega} \boldsymbol{y}^\intercal \boldsymbol{X} \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}^\intercal (\widehat{\omega} \boldsymbol{X}^\intercal \boldsymbol{X} + \nabla^\intercal \boldsymbol{V} \nabla) \boldsymbol{\beta}). \end{aligned} \tag{4.52}$$

The two shaping parameters, $\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}$ are very similar to their previous incarnations

$$\Sigma_{\boldsymbol{\beta}} = (\widehat{\omega} \boldsymbol{X}^\intercal \boldsymbol{X} + \nabla^\intercal \widehat{\boldsymbol{V}} \nabla)^{-1} \tag{4.53}$$

$$\mu_{\boldsymbol{\beta}} = (\widehat{\omega} \boldsymbol{X}^\intercal \boldsymbol{X} + \nabla^\intercal \boldsymbol{V} \nabla)^{-1} \widehat{\omega} \boldsymbol{X}^\intercal \boldsymbol{y}. \tag{4.54}$$

Shifting focus to $\boldsymbol{v}$ by keeping only terms that include $\boldsymbol{V}$ (not specifying for v_i just yet), the following proportional form is yielded

$$\begin{aligned} \tilde{f}(\boldsymbol{V}|\boldsymbol{y}) \propto \frac{1}{2} \sum_{i=1}^{n} \ln v_i - \frac{1}{2}(\boldsymbol{\beta}^\intercal \nabla^\intercal \boldsymbol{V} \nabla \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \ln v_i - \frac{1}{2} \sum_{i=1}^{n} v_i (\beta_i - \beta_{i+1})^2 = \\ \frac{1}{2} \sum_{i=1}^{n} \ln (v_i) - \frac{1}{2} \sum_{i=1}^{n} v_i (\boldsymbol{\beta}_i^2 - 2 \boldsymbol{\beta}_i \boldsymbol{\beta}_{i+1} + \boldsymbol{\beta}_{i+1}^2). \end{aligned} \tag{4.55}$$

▶ **Definition 4.3** (Matrix trace [100]). *Given a $n \times n$ square matrix $\boldsymbol{A}$, its trace is defined as*

$$\operatorname{Tr} \boldsymbol{A} = \sum_{i=1}^{n} \boldsymbol{A}_{ii}. \tag{4.56}$$

$\boldsymbol{\beta}^\intercal\nabla^\intercal\boldsymbol{V}\nabla\boldsymbol{\beta}$ is a scalar value, and as such is equal to its trace, $\boldsymbol{\beta}^\intercal\nabla^\intercal\boldsymbol{V}\nabla\boldsymbol{\beta} = \operatorname{Tr}\boldsymbol{\beta}^\intercal\nabla^\intercal\boldsymbol{V}\nabla\boldsymbol{\beta} = \operatorname{Tr}\nabla\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\nabla^\intercal\boldsymbol{V}$.

Now, slicing for v_i to model its posterior distribution:

$$\tilde{f}(v_i|\boldsymbol{y}) \propto \frac{1}{2}\ln v_i - (b_0 + \frac{1}{2}((\boldsymbol{\beta\beta}^\intercal)_{i\,i} - 2(\boldsymbol{\beta\beta}^\intercal)i\,i + 1 + (\boldsymbol{\beta\beta}^\intercal)_{i+1\,i+1}), \tag{4.57}$$

giving values to the next iteration of a, b

$$a = a_0 + \frac{1}{2} \tag{4.58}$$

$$b = b_0 + (\boldsymbol{\beta\beta}^\intercal)_{i\,i} - 2(\boldsymbol{\beta\beta}^\intercal)_{i\,i+1} + (\boldsymbol{\beta\beta}^\intercal)_{i+1\,i+1}. \tag{4.59}$$

ω's posterior distribution prevails in its original form 4.24, 4.25. The flow of the algorithm does not change, nor does the algorithm grow vertically, as shown by pseudocode 3. Parameter dependency is identical to sparse model's, and as such is shown in figure 4.8, and therefore still cannot be topologically ordered.

The previously assigned, non-informative values of prior shaping parameters are kept, i.e. $a_0 = b_0 = c_0 = d_0 = 10^{-10}$.

Algorithm 3: Bayesian smooth regression algorithm

```
input  : $\boldsymbol{y} \in \mathbb{R}^{n,1}$, $\boldsymbol{X} \in \mathbb{R}^{n,p}$
output: $\mu_{\boldsymbol{\beta}} \in \mathbb{R}^{p,1}$
init
    $\omega \leftarrow \max(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$
    $\boldsymbol{V} \leftarrow \boldsymbol{I}_p$
for $i \leftarrow 1$ to 100 do
    $\Sigma \leftarrow (\omega\boldsymbol{X}^\intercal\boldsymbol{X} + \nabla\boldsymbol{V}\nabla^\intercal)^{-1}$
    $\mu \leftarrow \Sigma(\omega\boldsymbol{X}^\intercal\boldsymbol{y})$
    $\boldsymbol{\beta}, \boldsymbol{\beta\beta} \leftarrow$ Bottom truncated normal distribution's first two moments$(\mu, \operatorname{diag}(\Sigma)^{\frac{1}{2}})$
    $\operatorname{var}_{\boldsymbol{\beta\beta}} \leftarrow \boldsymbol{\beta\beta} - \boldsymbol{\beta}^2$
    $\boldsymbol{\beta\beta}^\intercal \leftarrow \boldsymbol{\beta\beta}^\intercal + \operatorname{var}_{\boldsymbol{\beta\beta}}$
    $c \leftarrow c_0 + \frac{1}{2}n$
    $d \leftarrow d_0 + \frac{1}{2}((\boldsymbol{y}^\intercal\boldsymbol{y}) - 2\boldsymbol{y}^\intercal\boldsymbol{X}\mu_{\boldsymbol{\beta}} + \operatorname{Tr}(\boldsymbol{\beta\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}))$
    $\omega \leftarrow \frac{c}{d}$
    $\operatorname{var}_{\boldsymbol{\beta\beta}} \leftarrow \mu_{\boldsymbol{\beta}}\mu_{\boldsymbol{\beta}}^\intercal - \mu_{\boldsymbol{\beta}}^\intercal\mu_{\boldsymbol{\beta}}$
    $\boldsymbol{\beta\beta}^\intercal \leftarrow \mu_{\boldsymbol{\beta}}\mu_{\boldsymbol{\beta}}^\intercal + \Sigma_{\boldsymbol{\beta}}$
    $\boldsymbol{a} \leftarrow a_0 + \frac{1}{2}$
    $\boldsymbol{b} \leftarrow b_0 + \frac{1}{2}((\boldsymbol{\beta\beta}^\intercal)_{jj} - 2(\boldsymbol{\beta\beta}^\intercal)_{jj+1} + (\boldsymbol{\beta\beta}^\intercal)_{j+1j+1})$
    $\boldsymbol{v} \leftarrow \frac{a}{b}$
end
```

4.5.3 Performance

ω has a very different value progression compared to the sparse-only variant, assumedly due to the interaction with $\boldsymbol{v}$. It rises incredibly steeply, overshooting what seems to be in hindsight the convergent value, and settles at nearly the same time as in previous models.

Sparsity parameters $v_1, \ldots, v_n$ reach their convergence incredibly quickly, as is visible in figure 4.14b - after roughly the first 15 iterations, the value stay invariant.

Figure 4.15 shows the main culprit of the previously seemingly good properties of the algorithm. Within the activity window, the behaviour seems to be closely resembling the one in

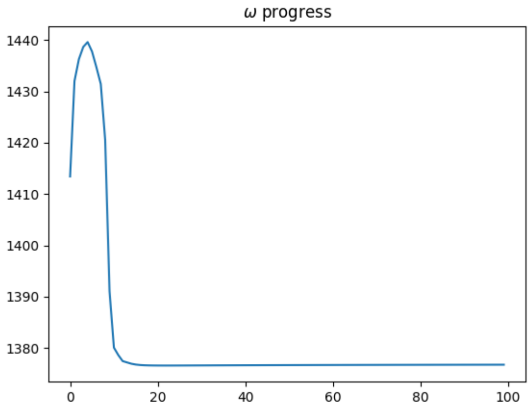
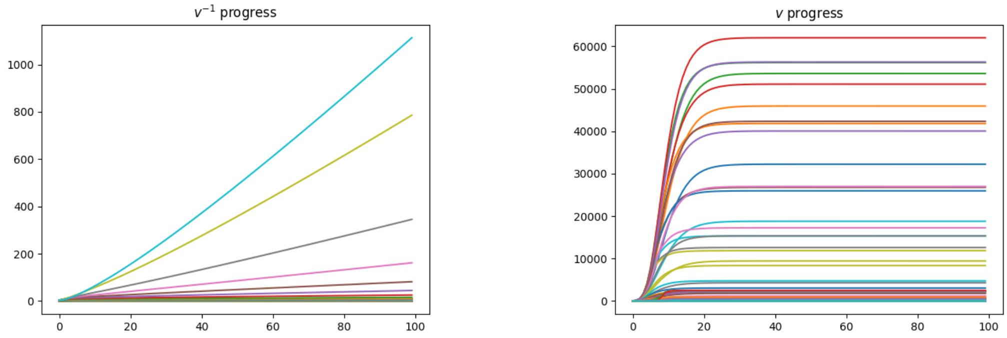


Figure 4.13 ω parameter values during 100 iterations using bayesian smooth regression on ETEX data.



(a) Per iteration value of $\boldsymbol{v}^{-1}$ in smooth bayesian regression on the ETEX dataset.

(b) Per iteration value of $\boldsymbol{v}$ in smooth bayesian regression on the ETEX dataset.

Figure 4.14 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of smooth bayesian regression on the ETEX dataset.

sparse model (4.11), including a slight deviation from 0 after the $[60, 70]$ window. The elephant in the room to address is, however, the incredibly off values preceding the activity. From the very start, the estimate $\widehat{\boldsymbol{\beta}}$ is very different from 0, almost at the same level as in the activity window. At around β_{17}, the values take off to a hitherto unprecedented extreme. The estimated values are, most likely to support smoothness in other parts, multiple times larger than the values in the narrow activity window. It is important to keep in mind that outside the activity window, β_i should cling to 0. This is indicative of an improper model for the situation. Increasing the number of does not have the effect, as can be deduced from the quick convergence of the parameters and no visible sign of deviation; this model is locked in.

Despite the grim results one could take away from $\widehat{\boldsymbol{\beta}}$ is that the actual reconstruction $\widehat{\boldsymbol{y}}$ would carry the same issues. However, that is not necessarily the case. When comparing the reconstruction and ground truth $\boldsymbol{y}$ (figure 4.16, the values do deviate in either direction - underestimation of the reconstruction $\widehat{\boldsymbol{y}}$ at some points, and overestimation at other points, with only a few outliers. This seems like a fallacy, but is actually caused by the underlying model $\boldsymbol{X}$. The values that would get multiplied ($\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{y}}$) by the extremely high values of $\widehat{\boldsymbol{\beta}}$ are, on the contrary, miniscule or 0, reducing the impact of the wildly incorrect $\widehat{\boldsymbol{\beta}}$ at those given points.

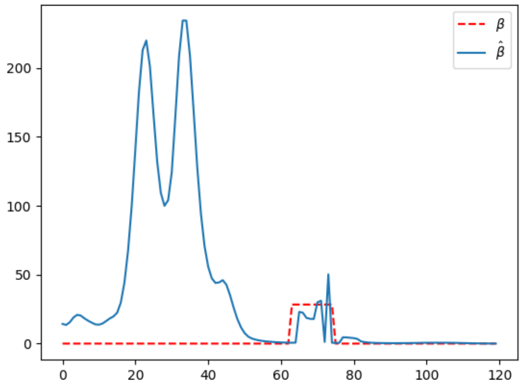


Figure 4.15 Comparison of real $\boldsymbol{\beta}$ from ETEX (red) with $\widehat{\boldsymbol{\beta}}$ as approximated by a bayesian smooth model (blue) after 100 iterations.

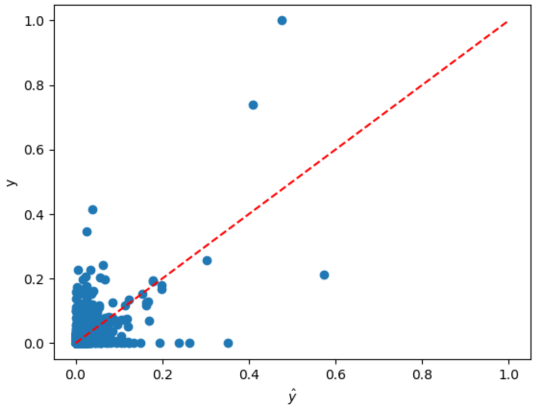


Figure 4.16 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian smooth regression attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

However, this is not a blanket behaviour and heavily relies on $\boldsymbol{X}$. This issue will be revisited on Chernobyl data in 5.2.3. The model had simply made a creative use of the task's specific formulation, yielding the following metrics for comparison:

- $MSE = 7.2237 * 10^{-4}$
- $MAE = 73.4335 * 10^{-4}$
- $RMSE = 268.7701 * 10^{-4}$.

4.6 LDL

The LDL algorithm has a promising potential to overtake any of the previous models. It expands the smooth model 4.5 much like the smooth model expanded the sparse model 4.4 [14]. In lieu of

[14] A step closer to a Rube Goldberg machine.

a constant matrix ∇, it modifies the covariance of $\boldsymbol{\beta}$ by introducing by matrix $\boldsymbol{L}$

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ l_1 & 1 & 0 & & 0 \\ \vdots & l_2 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & l_{n-1} & 1 \end{pmatrix}. \tag{4.60}$$

The name of the model being LDL stems from the covariance matrix $\boldsymbol{LDL}^\intercal$ of $\boldsymbol{\beta}$, which in this case, is $\boldsymbol{V}$.

$$\boldsymbol{D} = \boldsymbol{V} = \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & v_n \end{pmatrix} \tag{4.61}$$

LDL is a generalisation of the previous algorithms, allowing the values $\boldsymbol{l}$ influencing the smoothness of the model to change. If the values of $\boldsymbol{l} = (l_1, l_2, \dots, l_{n-1})$ were to set to fixed values 0, -1[15], respectively, then the model would degenerate into a sparse 4.4 or smooth 4.5 model, respectively. The promise of the model is reaching an equilibrium between smoothness and sparseness through a reasonable value assignment of $\boldsymbol{l}$.

4.6.1 Establishing prior distributions

LDL changes the prior covariance of $\boldsymbol{\beta}$, generalising the approach of smooth bayesian regression 4.5. To avoid confusion, due to previous use of $\boldsymbol{V}$ (sparse 4.4 and smooth 4.5 models) as the covariance matrix, going forward, it holds that $\boldsymbol{D} = \boldsymbol{V}$. The changes of parametrising matrix $\nabla \rightarrow \boldsymbol{L}$, where

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ l_1 & 1 & 0 & & 0 \\ \vdots & l_2 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & l_{n-1} & 1 \end{pmatrix}, \tag{4.62}$$

introduces a new set of parameters $l_1, \dots, l_{n-1}$ on $\boldsymbol{L}$'s subdiagonal. These parameters are iid. random variables, following a normal distribution given as follows:

$$f(l_i|\psi_i) = \mathcal{N}\left(l_0, \psi_i^{-1}\right), \tag{4.63}$$

where l_0 is empirically set as -1[16]. The newly added parameters along another set of parameters[17] $\psi_1, \dots, \psi_{n-1}$ determining the former's variances with the latter. As was the case of $l_1, \dots, l_{n-1}$, $\psi_1, \dots, \psi_{n-1}$ are identically and indepedently distributed, but following a Gamma distribution, with prior parameters e_0, f_0:

$$f(\psi_i) = \frac{e_0^{d_0}}{\Gamma(d_0)} \psi_i^{e_0 - 1} \exp\left(-d_0 \psi_i\right). \tag{4.64}$$

The model of $\boldsymbol{\beta}$ is changed to the following, denoting the vector of $l_1, \dots, l_{n-1}$ as $\boldsymbol{l}$:

$$f(\boldsymbol{\beta}|\boldsymbol{v}, \boldsymbol{l}) = (2\pi)^{-\frac{n}{2}} |(\boldsymbol{LVL}^\intercal)^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}^\intercal \boldsymbol{LVL}^\intercal \boldsymbol{\beta})\right), \tag{4.65}$$

[15] Which can be achieved by setting the mean value to 0,-1 respectively and variance to 0, transforming the normal distribution to a Dirac delta distribution [101].

[16] The value of l_0 is a subject to experiments in 5.2.4.2.

[17] Parameters within parameters within parameters.

$$\ln f(\boldsymbol{\beta}|\boldsymbol{v},\boldsymbol{l}) = -\frac{n}{2}\ln(2\pi) + \ln|\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal}| - \frac{1}{2}(\boldsymbol{\beta}^{\intercal}\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal}\boldsymbol{\beta}). \tag{4.66}$$

This equation can be further simplified by theorem 4.2. Further, as $\boldsymbol{L}$'s determinant $|\boldsymbol{L}| = |\boldsymbol{L}^{\intercal}| = 1$, the factor $|(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal})^{-1}|^{-\frac{1}{2}}$ of product 4.66 simplifies accordingly. Finally, given that $\boldsymbol{V}$ is a diagonal matrix, its determinant is the product of its diagonal elements $|\boldsymbol{V}| = \prod_{i=1}^{n} v_i$:

$$|(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal})^{-1}|^{-\frac{1}{2}} = |\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal}|^{\frac{1}{2}} = (|\boldsymbol{L}||\boldsymbol{V}||\boldsymbol{L}^{\intercal}|)^{\frac{1}{2}} = |\boldsymbol{V}|^{\frac{1}{2}} = \left(\prod_{i=1}^{n} v_i\right)^{\frac{1}{2}}. \tag{4.67}$$

Altogether the logarithm of $f(\boldsymbol{\beta}|\boldsymbol{v},\boldsymbol{l})$ has the following form

$$\ln f(\boldsymbol{\beta}|\boldsymbol{v},\boldsymbol{l}) = -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\prod_{i=1}^{n}\ln v_i - \frac{1}{2}(\boldsymbol{\beta}^{\intercal}\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal}\boldsymbol{\beta}). \tag{4.68}$$

Now, for the two newly introduced parameters that have cause this kerfuffle. Starting with the outer parameter l_i, $l_i \sim \mathcal{N}(l_0, \psi_i)$:

$$\begin{aligned}\ln f(l_i|\psi_i) = \ln(-\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\psi_i - \frac{1}{2}(l_i^2 - 2l_i l_0 + l_0^2)\psi_i,) =& \\ (2\pi)^{-\frac{1}{2}}|\psi_i^{-1}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\psi_i^{\frac{1}{2}}(l_i - l_0))^2),&\end{aligned} \tag{4.69}$$

As for the Gamma distributed ψ_i determining the normally distributed l_i's variance

$$\ln f(\psi_i) = \ln\left(\frac{e_0^{f_0}}{\Gamma(f_0)}\psi_i^{e_0-1}\exp(-f_0\psi_i)\right) = \ln\left(\frac{e_0^{f_0}}{\Gamma(e_0)}\right) + \ln\psi_i - f_0\psi_i. \tag{4.70}$$

Once again, it is assumed that all parameters are conditionally independent and thus their joint probability is equal to the product of their individual probabilities $f(\boldsymbol{y},\boldsymbol{\beta},\omega,\boldsymbol{v},\boldsymbol{l},\boldsymbol{\psi}) = f(\boldsymbol{y}|\boldsymbol{\beta},\omega)f(\boldsymbol{\beta})f(\omega)f(\boldsymbol{v})f(\boldsymbol{l})f(\boldsymbol{\psi})$ In the same vein as the previous models[18], the logarithm of joint probability function is expressed as

$$\begin{aligned}\ln f(\boldsymbol{y},\boldsymbol{\beta},\omega,\boldsymbol{v},\boldsymbol{l},\boldsymbol{\psi}) = -\frac{p}{2}\ln|\omega| - \frac{1}{2}\omega\boldsymbol{y}\boldsymbol{y}^{\intercal} + \omega\boldsymbol{y}^{\intercal}\boldsymbol{X}\boldsymbol{\beta} + \frac{1}{2}\omega(\boldsymbol{\beta}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{X}\boldsymbol{\beta})\ln\left(\frac{c_0^{d_0}}{\Gamma(d_0)}\right)& \\ +(c_0-1)\ln\omega - d_0\omega - \frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{n}\ln v_i - \frac{1}{2}\boldsymbol{\beta}^{\intercal}\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\intercal}\boldsymbol{\beta}& \\ +n\ln\left(\frac{a_0^{b_0}}{\Gamma(b_0)}\right) + \sum_{i=1}^{n}(a_0-1)\ln v_i - \sum_{i=1}^{n} b_0 v_i& \\ -\frac{1}{2}\ln(2\pi) + \sum_{i=1}^{n-1}\frac{1}{2}\ln\psi_i - \frac{1}{2}\psi_i(l_i^2 - 2l_i l_0 + l_0^2)& \\ +(n-1)\ln\left(\frac{e_0^{f_0}}{\Gamma(e_0)}\right) + \sum_{i=1}^{n-1}(e_0-1)\ln\psi_i - \sum_{i=1}^{n-1} f_0\psi_i,&\end{aligned} \tag{4.71}$$

doing the legwork and preparing the ground for posterior distributions' shape derivations.

[18] This pattern is getting rather tedious!

4.6.2 Deriving posterior distributions

The shaping parameters $\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}$ are given by the marginal distribution of 4.71 proportionally:

$$\begin{aligned}\tilde{f}(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp\left(\omega \boldsymbol{y}^\intercal \boldsymbol{X}\boldsymbol{\beta}\frac{1}{2}\omega(\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\intercal \boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal \boldsymbol{\beta}\right) = \\ \exp\left((\widehat{\omega}\boldsymbol{y}^\intercal \boldsymbol{X})\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^\intercal(\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \widehat{\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal})\boldsymbol{\beta}\right).\end{aligned} \tag{4.72}$$

Again to infer the posterior distribution, its shaping parameters are derived by setting the quadratic and linear terms of 4.72 equal to the approximated posterior distribution's (as-of-now) unknown $\Sigma_{\boldsymbol{\beta}}, \mu_{\boldsymbol{\beta}}$:

$$\Sigma_{\boldsymbol{\beta}} = (\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \widehat{\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal})^{-1} \tag{4.73}$$

$$\mu_{\boldsymbol{\beta}} = (\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{X} + \widehat{\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal})^{-1}\widehat{\omega}\boldsymbol{X}^\intercal \boldsymbol{y}. \tag{4.74}$$

The LDL matrix 4.75 is a slightly complex matter. Due to approximate nature, using various distributions, one cannot [19] set for a plain and simple matrix multiplication of matrices $\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal$. The resulting matrix, due to multiplication by $\boldsymbol{L}$ and its transposition $\boldsymbol{L}^\intercal$, is effectively multiplying the diagonal matrix $\boldsymbol{V}$ by a square of $\boldsymbol{l}$, and more precisely, by its second moments. Therefore, the moments have to be calculated separately and a formula to construct the proper matrix has to be derived.

$$\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal = \begin{pmatrix} v_1 & l_1 v_1 & 0 & 0 & \cdots & 0 \\ l_1 v_1 & l_1^2 v_1 + v_2 & l_2 v_2 & 0 & \cdots & 0 \\ 0 & l_2 v_2 & l_2^2 v_2 + v_3 & l_3 v_3 & \cdots & 0 \\ 0 & 0 & l_3 v_3 & l_3^2 v_3 + v_4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & l_{n-1}^2 v_{n-1} + v_n \end{pmatrix} \tag{4.75}$$

Fortunately, the $\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal$ matrix 4.75 is tridiagonal, and the leading diagonal, the subdiagonal and superdiagonal each follow their own, independent formulation. Further, it is of note that the matrix is symmetric, $\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal = (\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal)^\intercal$, and thus the super- and sub-diagonal (incidental diagonals to the leading diagonals) of $\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal$ are identical.

Let Λ be a vector of second non-central moments of normally distributed $l_1, \ldots, l_{n-1}$ given their particular means μ_{l_i} and variances Σ_{l_i}, i.e.

$$\Lambda = (\mu_{l_1}^2 + \Sigma_{l_1}, \mu_{l_2}^2 + \Sigma_{l_2}, \ldots, \mu_{l_{n-1}}^2 + \Sigma_{l_{n-1}})). \tag{4.76}$$

$\operatorname{diag}(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal)_i = v_{i-1} l_{i-1}^2 + v_i$. This poses an issue for the first element, which is resolved by defining $v_0 = 0, \Lambda_0 = 0$. Vector-wise, by defining vectors $\boldsymbol{v}', \Lambda'$ based on $\boldsymbol{v}$ and Λ, respectively:

$$\boldsymbol{v}' = (0, v_1, v_2, \ldots, v_{n-1}) \tag{4.77}$$

$$\Lambda' = (0, \Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1}), \tag{4.78}$$

aligning the dimensions with the dimension of $\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal$'s diagonal and not upsetting any partial assignment of the diagonal vector, resulting in a concise vector representation $\operatorname{diag}(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal) = \boldsymbol{v}'\Lambda' + \boldsymbol{v}$. As for the incidentals, the i-th element is $\text{subdiagonal}(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal)_i = \text{superdiagonal}(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal)_i = l_i v_i$. Introducing

$$\boldsymbol{v}'' = (v_1, v_2, \ldots, v_{n-1}), \tag{4.79}$$

i.e. $\boldsymbol{v}$ without the last element, the incidental diagonals of $\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal$ can be set vector-wise

$$\text{subdiagonal}(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal) = \text{superdiagonal}(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal) = \boldsymbol{v}''\boldsymbol{l}. \tag{4.80}$$

[19] Unfortunately

Altogether and with a bow on top, the proper form of $\boldsymbol{LVL}^\intercal$ is as follows 4.81:

$$\boldsymbol{LVL}^\intercal = \begin{pmatrix} v_1 & l_1v_1 & 0 & 0 & \cdots & 0 \\ l_1v_1 & \Lambda_1v_1+v_2 & l_2v_2 & 0 & \cdots & 0 \\ 0 & l_2v_2 & \Lambda_2v_2+v_3 & l_3v_3 & \cdots & 0 \\ 0 & 0 & l_3v_3 & \Lambda_3v_3+v_4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \Lambda_{n-1}v_{n-1}+v_n \end{pmatrix} \tag{4.81}$$

A similar inconvenience is caused by $\boldsymbol{\beta}$, similarly to the case in smooth regression 4.5, but turned to 11. The marginal defining the posterior distribution for $\boldsymbol{V}$ is derived from 4.71:

$$\tilde{f}(\boldsymbol{V}|\boldsymbol{y}) \propto \exp\,(-\frac{1}{2}\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta} + \frac{1}{2}\sum_{i=1}^{n}\ln v_i + \sum_{i=1}^{n}(a_0-1)\ln v_i - \sum_{i=1}^{n}b_0v_i) \tag{4.82}$$

Deriving the formulation of the first term $-\frac{1}{2}\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta}$ for v_i by plain matrix multiplication. $\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta}$ is a scalar, and therefore can equal to its trace. Utilising the invariance of trace under circular shifts[102], it holds that $\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta} = \operatorname{Tr}(\boldsymbol{L}^\intercal\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\boldsymbol{L})\boldsymbol{V}$. This gives a straight forward prescription for the coefficient of any v_i, as $\boldsymbol{V}$ is a diagonal matrix.

$$\boldsymbol{L}^\intercal\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\boldsymbol{L} = \sum_{i=1}^{n}(\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{ii} + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i}l_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i+1}l_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i+1}\Lambda_i, \tag{4.83}$$

prescribing the coefficient of v_i as $diag(\boldsymbol{L}^\intercal\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\boldsymbol{L})_i$. To match dimensions and keep calculations correct and sane, a supplementary vector Λ^v is defined as $\Lambda^v = (\Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1}, 0)$ and $(\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{1\,0} = (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{0\,1}$.

$$\tilde{f}(v_i|\boldsymbol{y}) \propto \exp\,((a_0 - 1 + \frac{1}{2})\ln v_i - (b_0 + \frac{1}{2}((\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i} + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i}l_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i+1}l_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i+1}\Lambda^v_i,))v_i), \tag{4.84}$$

yielding the follow shaping parameters a_i, b_i for random variable v_i:

$$a_i = a_0 + \frac{1}{2} \tag{4.85}$$

$$b_i = b_0 + \frac{1}{2}((\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i} + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i}l_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i+1}l_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i+1}\Lambda_i), \tag{4.86}$$

and the expected value $\mathbb{E}[v_i] = \frac{a_i}{b_i}$. Penultimate is the set of random variables $\boldsymbol{l}$. The distribution of l_i is a:

$$\tilde{f}(\boldsymbol{L}|\boldsymbol{y}) \propto \exp\,(\sum_{i=1}^{n-1}(-\frac{1}{2}\psi_i(\Lambda - 2l_0l_i)) + \operatorname{Tr}\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta}). \tag{4.87}$$

$\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta}$ is a scalar, and therefore equal to its trace, $\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta} = \operatorname{Tr}\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta}$. Given the properties of matrix traces, it holds that $\operatorname{Tr}\boldsymbol{\beta}^\intercal\boldsymbol{LVL}^\intercal\boldsymbol{\beta} = \operatorname{Tr}\boldsymbol{LVL}^\intercal\boldsymbol{\beta}\boldsymbol{\beta}^\intercal$, and it is given as

$$\operatorname{Tr}\boldsymbol{LVL}^\intercal\boldsymbol{\beta}\boldsymbol{\beta}^\intercal \propto \sum_{i=1}^{n-1} l_iv_i((\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i+1} + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i}) + \Lambda_iv_i(\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i+1}, \tag{4.88}$$

giving a cut and clear term for l_i, resulting in the posterior distribution

$$\tilde{f}(l_i|\boldsymbol{y}) \propto -\frac{1}{2}(-\psi_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i}v_i)\Lambda - \frac{1}{2}(l_0\psi_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i\,i+1}v_i + (\boldsymbol{\beta}\boldsymbol{\beta}^\intercal)_{i+1\,i})l_i. \tag{4.89}$$

The derivation of psi_i's posterior distribution is most likely the simplest and most straightforward of the bunch, the only (minor) complicatio being the presence of l_i's second non-central moment Λ_i:

$$\begin{aligned}\tilde{f}(\psi_i|\boldsymbol{y}) \propto \exp\left((e_0-1)\ln\psi_i - f_0\psi_i + \frac{1}{2}\ln\psi_i - \psi_i(\Lambda_i - 2l_0l_i + l_0^2)\right) = \\ \exp\left((e_0-1+\frac{1}{2})\ln\psi_i - (f_0+\Lambda_i-2l_0l_i+l_0^2)\psi_i\right).\end{aligned} \tag{4.90}$$

And, as ψ_i follows a Gamma distribution with unknown parameters e_i, f_i, the posterior distribution is shaped with the following

$$e_i = e_0 + \frac{1}{2} \tag{4.91}$$

$$f_i = f_0 + \Lambda_i - 2l_0l_i + l_0^2, \tag{4.92}$$

and the expected value $\mathbb{E}[\psi_i] = \frac{e_i}{f_i}$. $\tilde{f}(\omega|\boldsymbol{y})$ remains unchanged, as is the data model $\boldsymbol{y}$.

The shaping parameter of the remaining $\boldsymbol{\psi}, \boldsymbol{\beta}$'s distributions are derived from these values and do not require to be initialised. The prior shaping parameters are also set to have a non-informative nature $e_0 = f_0 = 10^{-2}$, allowing l_i to vary in range -1 ± 100[76]. Further reducing the value of these priors tightens the distribution and causes a value development closer to l_0, which is -1 in the current configuration. Having $\boldsymbol{l} = -\mathbf{1}_{n-1}$ corresponds to a smooth bayesian model 4.5.

The dependency graph of previous models is expanded by the two sets of parameters, $\boldsymbol{l}, \boldsymbol{\psi}$. No parameters have been removed from the model, therefore it retains directed cycles and thus circular dependencies. Consequently the parameter estimation cannot be topologically ordered, and requires initial values to start. The initial values are following, set to be non-informative:

- $\omega = (\max \boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$
- $\boldsymbol{v} = \mathbf{1}$[20]
- $\boldsymbol{l} = -\mathbf{1}$ per [76]

A small caveat - for correctness's sake, the pseudocode in 4 refers to the mean value μ_{l_i} and variance Σ_{l_i} in a fashion in line with expected vector behaviour, i.e. $(\mu_{\boldsymbol{l}})_i$ and $(\Sigma_{\boldsymbol{l}})_i$.

4.6.3 Performance

ω behaves wildly in the first few iterations of the algorithm (figure 4.17). It rises rapidly only to fall even more rapidly, and then the curve mirrors resembles a logarithmic function, converging to a value just shy of 1400, all that within the first twenty iterations. Long story short, wild start with a quick convergence afterwards.

Mean values $\mu_{\boldsymbol{l}}$ of $\boldsymbol{l}$ show a clear convergence within the first 50 iterations, aside from a few outliers. Most of them shift from the initial value of -1 to a slighly smaller value ≈ -0.95, deviating from the absolute smoothness that the the value -1 would enforce. The outliers converge very quickly to their final value, whilst the bulk of l_i firstly overshoots (or in a few rare cases, undershoots) it before reaching it and sticking to it. What is important is that the model clearly shows that there is "some" smoothness in the data, but is not absolutely smooth, judging by the learnt l_i values.

The approximation $\widehat{\boldsymbol{\beta}}$ as shown in 4.21 seems to have a very good fit, judging solely the similarity between the ground truth $\boldsymbol{\beta}$ and the estimate $\widehat{\boldsymbol{\beta}}$. Firstly, outside the activity window, there is but one deviation from 0, around the 80th index. That is most likely caused by the

[20] A vector of ones, $\mathbf{1} = (-1, -1, \ldots, -1)$.

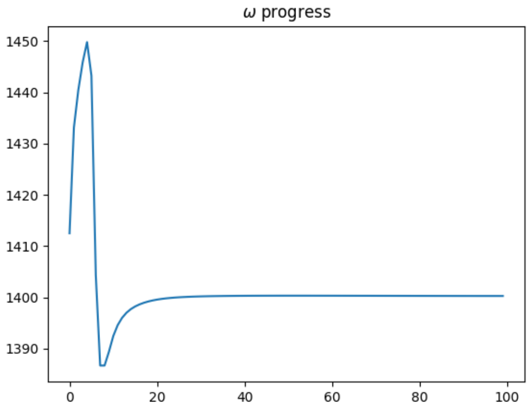
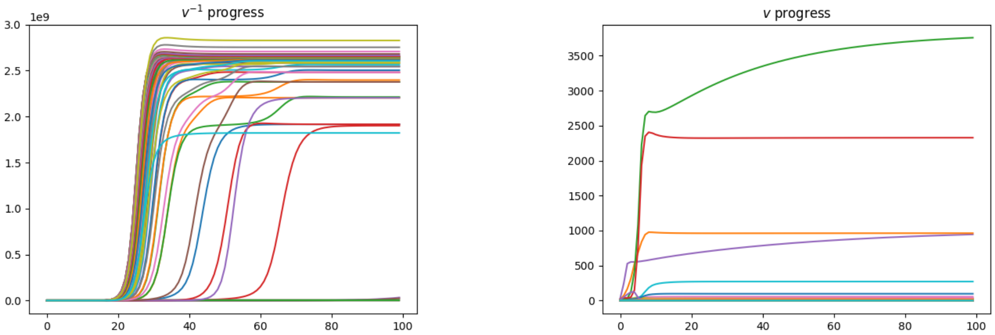


Figure 4.17 ω parameter values during 100 iterations using LDL on ETEX data.



(a) Per iteration value of $\boldsymbol{v}^{-1}$ in LDL on the ETEX 1.4 dataset.

(b) Per iteration value of $\boldsymbol{v}$ in LDL on the ETEX 1.4 dataset.

Figure 4.18 Graphs of the change in value of $\boldsymbol{v}$ during iterations of LDL on the ETEX 1.4 dataset.

exceeding peak before that and the model's attempt to keep a degree of smoothness. Within the activity window itself, aside from the narrow peak greatly exceeding the ground truth $\boldsymbol{\beta}$ at the right edge of the window and a slight drop. The interval within the window is also narrower - the steep rise lags behind and returns to 0 earlier than the real $\boldsymbol{\beta}$. This deviation might have dire consequences on the reconstruction's fit, due to how steep the window's rise is - any later calculated error will be determined by the activity and the small spike after the window.

A lot of the reconstructed $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ lies outside the regression line, deviating in both direction (from the regression line, caused by a larger $\widehat{\boldsymbol{y}}$ or $\boldsymbol{y}$, respectively). This behaviour is invariant to an increasing number of iterations and different values of l_0, as will be experimentally shown and discussed in 5.2.4.1.

The resulting model has a slighly worse performance than the sparse model 4.4, in spite of being able to attain the same form.

- $MSE = 7.1183 * 10^{-4}$
- $MAE = 67.7541 * 10^{-4}$
- $RMSE = 266.8022 * 10^{-4}$

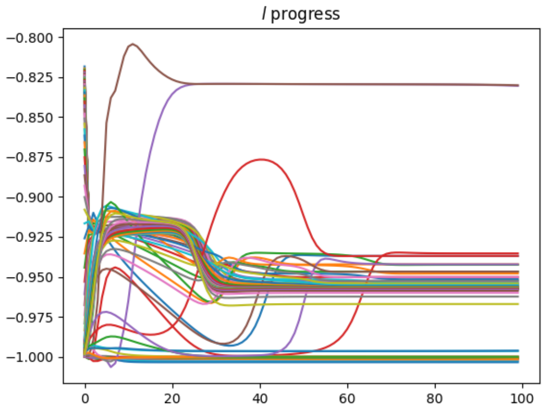
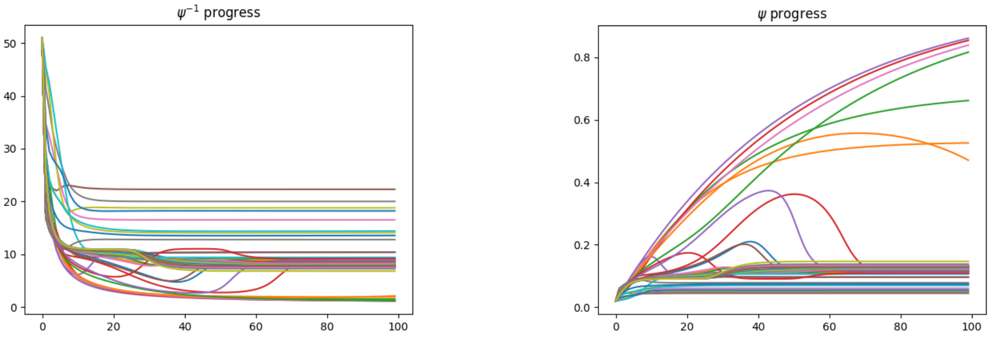


Figure 4.19 l parameter values during 100 iterations using LDL on ETEX data.



(a) Per iteration value of ψ^{-1} in LDL on the ETEX dataset.

(b) Per iteration value of ψ in LDL on the ETEX dataset.

Figure 4.20 Change in value of ψ during 100 iterations of LDL on the ETEX dataset.

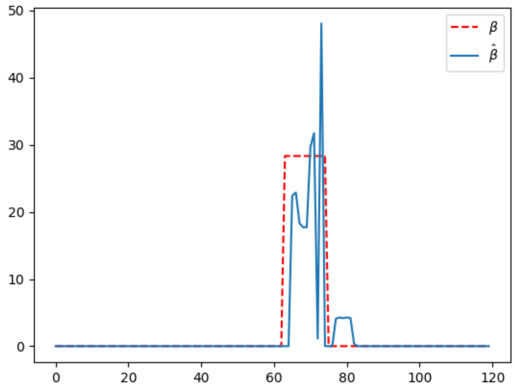


Figure 4.21 Comparison of real β from ETEX (red) with $\widehat{\beta}$ as approximated by an LDL model (blue) after 100 iterations.

Algorithm 4: LDL algorithm

```
input  : $\boldsymbol{y} \in \mathbb{R}^{n,1}$, $\boldsymbol{X} \in \mathbb{R}^{n,p}$
output: $\mu_{\boldsymbol{\beta}} \in \mathbb{R}^{p,1}$
init
    $\omega \leftarrow \max(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$
    $\boldsymbol{V} \leftarrow \boldsymbol{I}_p$
    $\mu_{\boldsymbol{l}} \leftarrow -\mathbf{1} \in \mathbb{R}^{p-1,1}$
    $\Sigma_{\boldsymbol{l}} \leftarrow \mathbf{0} \in \mathbb{R}^{p,p}$
for $i \leftarrow 1$ to 100 do
    $\Sigma \leftarrow (\omega \boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^\intercal)^{-1}$
    $\mu \leftarrow \Sigma_{\boldsymbol{\beta}}(\omega \boldsymbol{X}^\intercal \boldsymbol{y})$
    $\boldsymbol{\beta}, \boldsymbol{\beta\beta} \leftarrow$ Bottom truncated normal distribution's first two moments$(\mu, \mathrm{diag}(\Sigma)^{\frac{1}{2}})$
    $\mathrm{var}_{\boldsymbol{\beta\beta}} \leftarrow \boldsymbol{\beta\beta} - \boldsymbol{\beta}^2$
    $\boldsymbol{\beta\beta}^\intercal \leftarrow \boldsymbol{\beta\beta}^\intercal + \mathrm{var}_{\boldsymbol{\beta\beta}}$
    $c \leftarrow c_0 + \frac{n}{2}$
    $d \leftarrow d_0 + \frac{1}{2}((\boldsymbol{y}^\intercal \boldsymbol{y}) - 2\boldsymbol{y}^\intercal \boldsymbol{X} \mu_{\boldsymbol{\beta}} + \mathrm{Tr}(\boldsymbol{\beta\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}))$
    $\omega \leftarrow \frac{c}{d}$
    $\boldsymbol{v} \leftarrow$ empty vector $\in \mathbb{R}^{p-1,1}$
    for $i \leftarrow 1$ to $p$ do
        $a \leftarrow a_o + \frac{1}{2}$
        $b \leftarrow b_0 + \frac{1}{2}(\boldsymbol{\beta\beta}^\intercal_{ii} - 2\boldsymbol{\beta\beta}^\intercal ii + 1 + \boldsymbol{\beta\beta}^\intercal_{i+1i+1})$
        $v_i \leftarrow \frac{a}{b}$
    end
    $\boldsymbol{\Sigma_l} \leftarrow$ empty vector $\in \mathbb{R}^{p-1,1}$
    $\boldsymbol{\mu_l} \leftarrow$ empty vector $\in \mathbb{R}^{p-1,1}$
    for $i \leftarrow 1$ to $p-1$ do
        $(\Sigma_l)_i \leftarrow (2\psi_i - \boldsymbol{\beta\beta}^\intercal_{i+1i+1} v_i)^{-1}$
        $(\mu_l)_i \leftarrow \Sigma_{l_i}(l_0 \psi_i - \boldsymbol{\beta\beta}^\intercal_{ii+1} v_i)$
    end
    $\boldsymbol{\psi} \leftarrow$ empty vector $\in \mathbb{R}^{p-1,1}$
    for $i \leftarrow 1$ to $p-1$ do
        $f \leftarrow f_0 + \frac{1}{2}(l_i^2 - 2l_0 + l_0^2)$
        $e \leftarrow e_0 + \frac{1}{2}$
        $\psi_i \leftarrow \frac{e_i}{f_i}$
    end
end
```

Algorithm	MSE (10^4)	MAE (10^4)	RMSE (10^4)
Linear regression	NaN	NaN	NaN
Ridge regression	7.1020	65.2597	266.4950
Bayesian ridge regression	9.2399	71.2809	303.9722
Bayesian sparse regression	6.8414	69.3237	261.5601
Bayesian smooth regression	7.2237	73.4335	268.7701
LDL	7.1183	67.7540	266.8022

Table 4.1 Metrics of different algorithms on the ETEX 1.4 dataset.

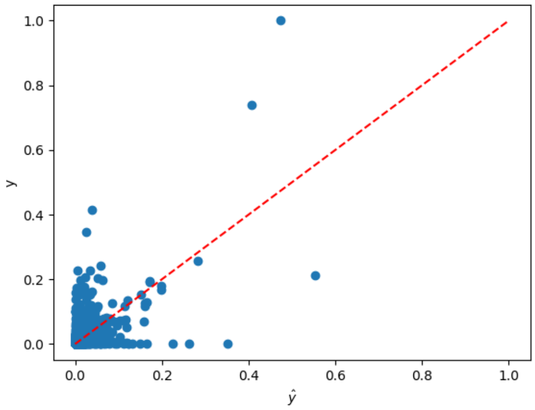


Figure 4.22 Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL-attained reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

Chapter 5
Uncontrolled emission modeling

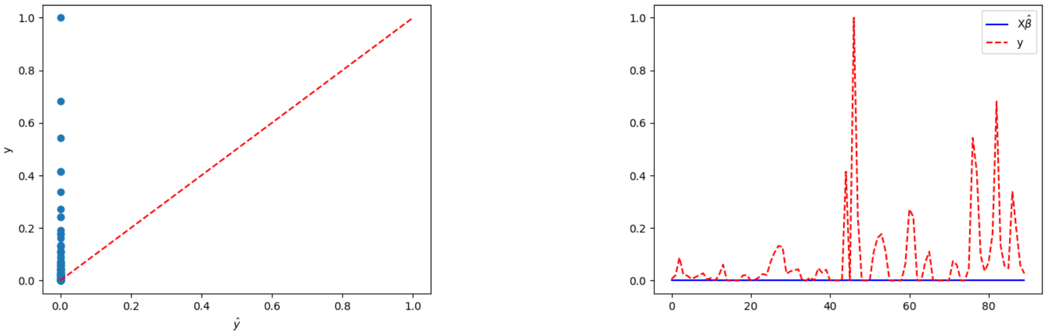
The dataset picked for modeling of an uncontrolled emission is a subset of the 2020 Chernobyl fires. Following steps were taken to simplify the convoluted model:

1. Single height of particle emission was considered.
2. Single particle size was considered.
3. Data was narrowed down to measurements between April 3rd 2020 and April 27th 2020.

5.1 Blind run

Firstly, the models developed in 4 were used on the new dataset, with the same prior constants and initial values to get a picture of general performance. Afterwards, experiments through a grid search of initial values were carried out to determine the best-performing, task-specific set model.

5.2 Classic ridge regression



(a) Chernobyl classic ridge y reconstruction scatter

(b) Chernobyl classic ridge y reconstruction

Once again, $\boldsymbol{X}^\intercal \boldsymbol{X}$ is singular and does not have the needed inverse to apply the OLS method. Attempting to generalise it to a ridge regression results in an ill-fitting model, as can deduced from comparative figures 5.1b and 5.1a, as well as the performance measured by the three metrics used prior (for comparison with other methods, refer to table 5.5. The L-curve method was used to determine the most suitable λ penalisation coefficient, but given the nature of the data, the sought after L-curve does not have the needed curve to determine such a value.

- $MSE = 3.0122 * 10^{-2}$
- $MAE = 7.9480 * 10^{-2}$
- $RMSE = 17.3557 * 10^{-2}$

5.2.1 Bayesian ridge regression

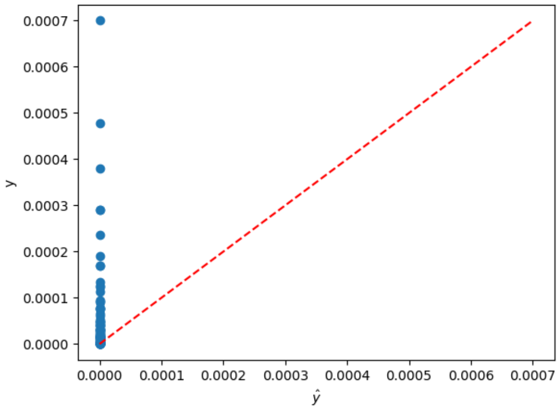


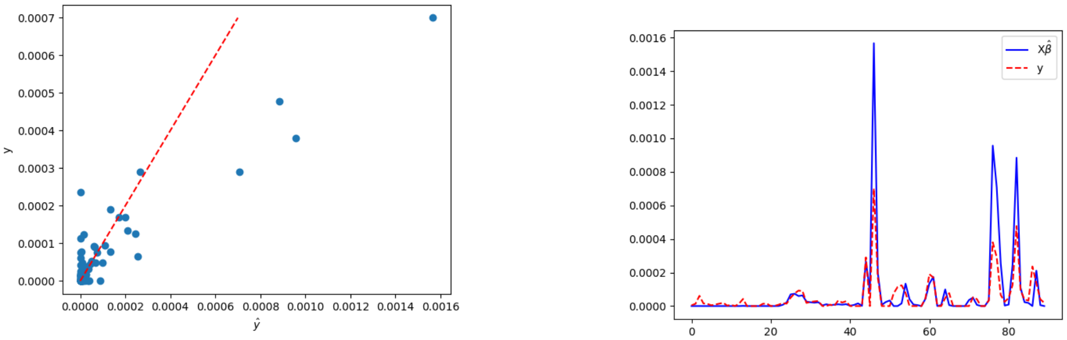
Figure 5.2 Comparison of measured data $\boldsymbol{y}$ (y-axis) and bayesian ridge regression approximated retrieved $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

Bayesian ridge regression, initialised with the same values as in the ETEX case, performs incredibly poorly on the Chernobyl dataset. Aside from a handful of values at $(0, 0)$, none of the measurements $\boldsymbol{y}$ are anywhere near well approximated by the reconstruction $\widehat{y}$, as shown by both the figure 5.2 and the metrics.

- $MSE = 3.0122 * 10^{-2}$
- $MAE = 7.9485 * 10^{-2}$
- $RMSE = 17.3557 * 10^{-2}$

5.2.2 Bayesian sparse regression

Figure 5.3a shows that the sparse regression fits the data much better than the simple bayesian ridge regression 5.2.1. A large number of points (which should be the vast majority, given the nature of the task) is still correctly sitting at, or near $(0, 0)$, and a plenty of other points near or on the regression line. There are new outliers in the direction of $\widehat{\boldsymbol{y}}$ (the x-axis), showing overestimation of some values by the reconstruction. In this case, it might be easier to see the quality of the fit by plotting both $\boldsymbol{y}$ on top of $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. The three peaks are overshot by the reconstruction, but the areas of main activity are detected by the model, aside from a few minor ones at the start. Overall, it seems only as a matter of scale.



(a) Comparison of measured data from Chernobyl fires $\boldsymbol{y}$ (y-axis) and bayesian smooth regression retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

(b) Side by side visualisation of Chernobyl fires' $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ estimated by a bayesian sparse model after 100 iterations.

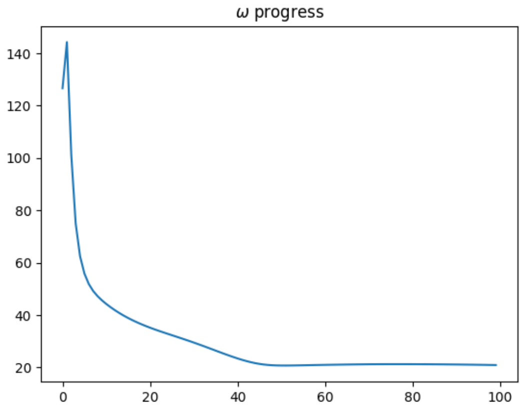


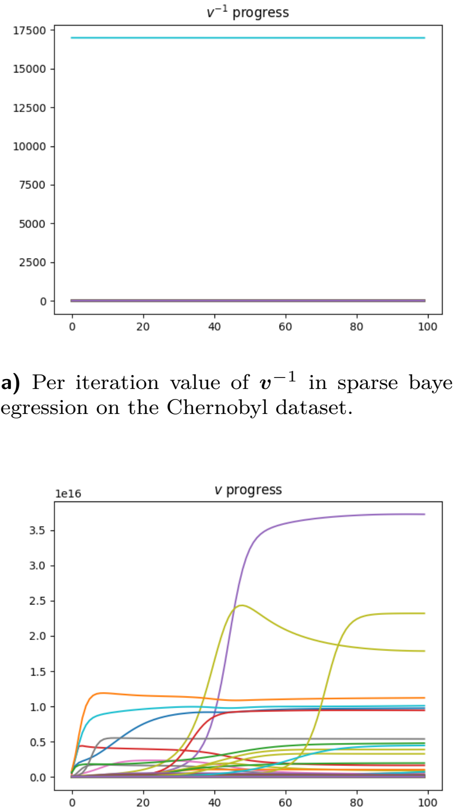
Figure 5.4 ω parameter values during 100 iterations using bayesian sparse regression on Chernobyl data.

ω shares a similar progress of its value with the ETEX case 4.9 - a slight deviation at the start, and then a rapid descent within the first few iterations, seemingly reaching convergence at $\approx$ 50th iteration.

The progress of V^{-1} (or rather, of $v_1, v_2, \ldots, v_n$) in figure 5.5a, which was selected as an inverse for convenience in model construction, is impossible to depict in a representative, informative manner. Even if the massive outlier (nearing 17000) would be removed, the difference between the next highest value and all the remaining plot lines is so drastic (on a smaller scale), that one might not even notice any outlier has been removed at all[1]. As such, it might be more convenient to observe the inversion's inversion 5.5b, which shows a convergence of majority of v_i in the first half of the run. There seems to be one or two delayed, which started changing value in the second half, but that is an exception. Overall, the values seem to have converged within the 100 iterations runs.

- $MSE = 3.6516 * 10^{-2}$
- $MAE = 7.0076 * 10^{-2}$

[1] Only the y-axis scale visibly changes!



(a) Per iteration value of $\boldsymbol{v}^{-1}$ in sparse bayesian regression on the Chernobyl dataset.

(b) Per iteration value of $\boldsymbol{v}$ in sparse bayesian regression on the Chernobyl dataset.

Figure 5.5 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of sparse bayesian regression on the Chernobyl fires data.

- $RMSE = 19.1092 * 10^{-2}$

5.2.3 Bayesian smooth regression

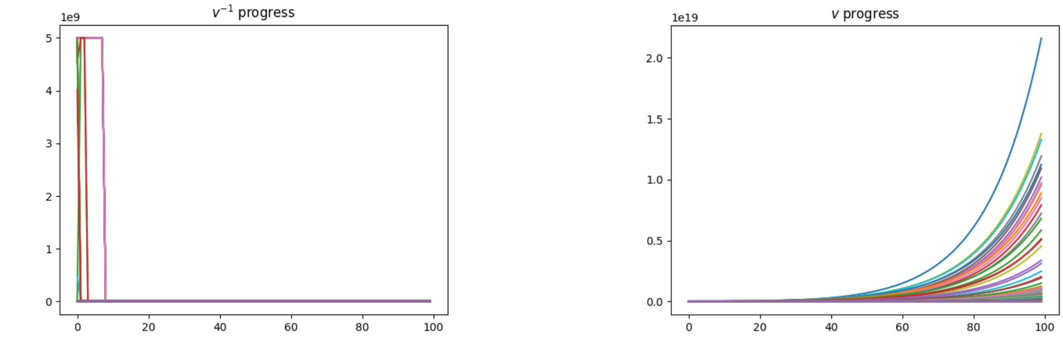
The smooth approach applied to Chernobyl data is a great example of things going very, very wrong, which is visible rightaway from the value progression of its parameters 5.6a, 5.6b, 5.8 alone.

and whilst having features of a picturesque wall decoration, in terms of bayesian modeling, it is of no practical use. Clearly, adding absolute smoothness is not the way

While the model correctly detects the intervals of emission, it wildly overestimates the actual values, and that is not even consistently proportional to the ground truth $\boldsymbol{y}$ - the location of the large peaks of the reconstruction do not match the peaks of the ground truth. What is peculiar is the troughs. One would expect them to be few and far in between, but there seems to be isolated activity with a rapid drop in value in between, something that is rather surprising given that a smooth model was used to model the data. The takeaway is that the assumption of smoothness of the model is not met by this particular dataset.

ω adheres to the behaviour of sparse model's ω 5.4 with a slight up at the start before quickly dropping and converging within the first 50 iterations.

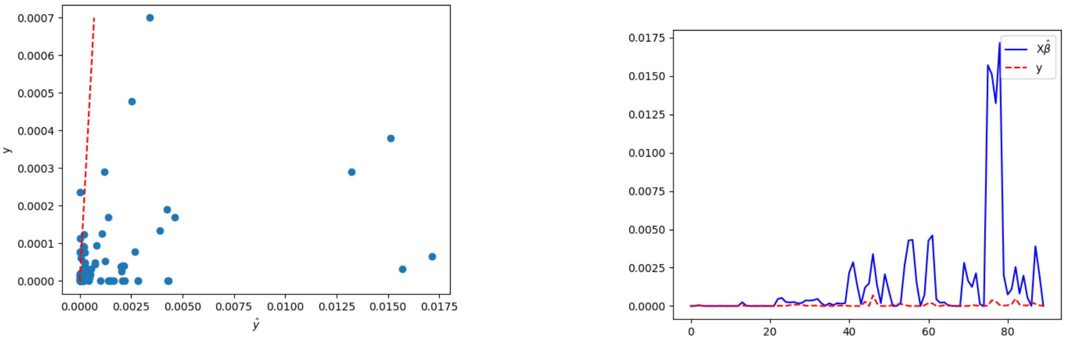
The performance-tracking metrics are, unsurprisingly, the worst by a long shot. Thanks to



(a) Per iteration value of $\boldsymbol{v}^{-1}$ in smooth bayesian regression on the Chernobyl dataset.

(b) Per iteration value of $\boldsymbol{v}$ in smooth bayesian regression on the Chernobyl dataset.

Figure 5.6 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of smooth bayesian regression on the Chernobyl fires data.



(a) Comparison of measured data from Chernobyl fires $\boldsymbol{y}$ (y-axis) and bayesian smooth regression retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

(b) Side by side visualisation of Chernobyl fires' $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ estimated by a bayesian smooth model after 100 iterations.

the incredibly large deviation of the reconstruction $\widehat{\boldsymbol{y}}$ (refer to figure 5.7b for a clear visual), the squared error is also incredinly large. Do keep in mind that the metrics are calculated on the same data as approximated on, in terms of scaling and units. This should clear some confusion as to why 100 values, with individual error of at most 0.0175 (highest value of $\widehat{\boldsymbol{y}}$ could sum up to a mean squared error of 24.65. Nevertheless, the error-rate would be the same on any form of the data, only scaled. The metrics measured are, as follows (and also included in the summary table 5.5:

- $MSE = 2465.2423 * 10^{-2}$
- $MAE = 201.3704 * 10^{-2}$
- $RMSE = 496.5121 * 10^{-2}$

5.2.4 LDL

LDL was firstly used to model with configuration 5.1, with values set to be ideally non-informative.

Once again, the initial value of ω was a far cry from its final value, immediately experiencing a sharp drop. Peculiarly, before dropping further, it rises back, nearly to the same value, before

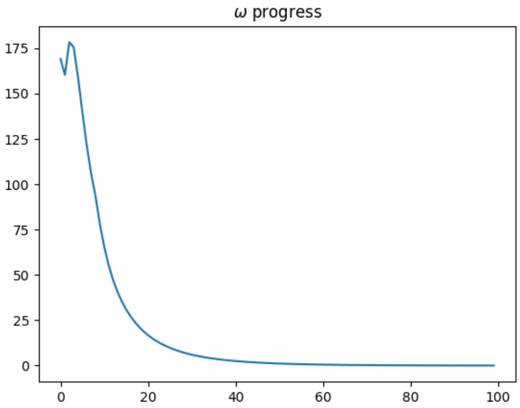


Figure 5.8 ω parameter values during 100 iterations using bayesian smooth regression on the Chernobyl fires dataset.

Parameter	Initial value
ω	$(\max \boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$
μ_{l_i}	-1
Σ_{l_i}	0
v_i	1
# of iterations	100
Prior constant	Value
a_0	10^{-10}
b_0	10^{-10}
c_0	10^{-10}
d_0	10^{-10}
e_0	10^{-2}
f_0	10^{-2}
l_0	-1

Table 5.1 Initial values and prior constants set for the first run of LDL on Chernobyl data, set to be non-informative.

slowly falling near the convergent, just to lightly meander on the upper bound of 150. A few more iterations would most likely cause the value to ultimately settle, but the difference would be in the domain of digits.

$\boldsymbol{v}$ is a prime example of the need to experiment with the number of iterations. The sparsity-controlling random variables vary in the speed of their convergence (figure 5.10a). Most of them quickly adapt a value different from the initial non-informative value 1 (see table 5.1, which is a good sign. However, the issue arises in the roughly 50th iteration, where the major bulk of v_i suddenly rises to a different value and sticks to it. Another set of v_i follows this behaviour near the end of the run. All of this is telling of the need to investigate behaviour on a larger number of iterations. This task is carried out in 5.2.4.1.

The intermediate values of $l_1, \ldots, l_{n-1}$ behave rather erratically. Their range spans covers $[-1, -0.65]$, where -1 corresponds to smoothness. The resulting interval is narrower, $[-1, -0.85]$, depicting a degree of smoothness in all values. Where the behaviour differs is the various convergence. Some l_i converge almost immediately (whether sticking to -1, from whence they started, or reaching their value swiftly), while the others stabilise in vastly different iterations - and by

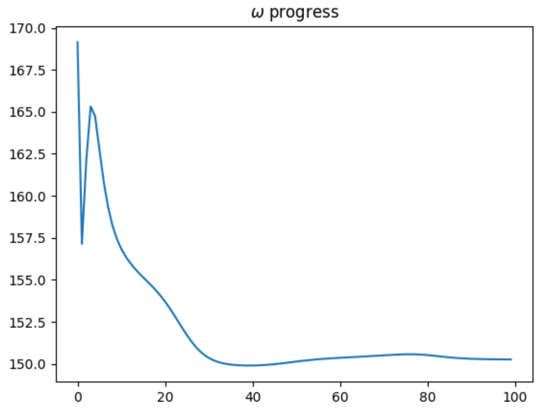
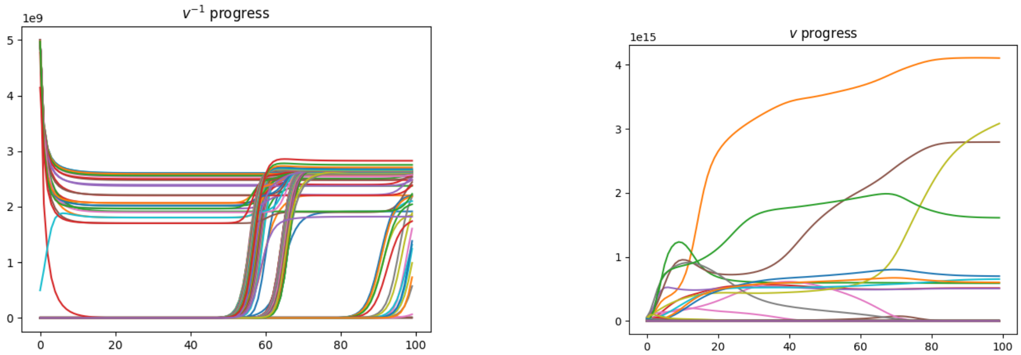


Figure 5.9 ω parameter values during 100 iterations using LDL on Chernobyl data.



(a) Per iteration value of $\boldsymbol{v}^{-1}$ in LDL on the Chernobyl dataset.

(b) Per iteration value of $\boldsymbol{v}$ in LDL on the Chernobyl dataset.

Figure 5.10 Approximate intermediate values of $\boldsymbol{v}$ during 100 iterations of LDL on the Chernobyl fires data.

the end, some do not seem to be converged yet. This observation sparked a need to experiment with the number of iterations, described in 5.2.4.1.

$\boldsymbol{\psi}$, modeling the variance of $\boldsymbol{l}$, shown in 5.12a has very quick convergence for most ψ_i; within the first ≈ 20 iterations. Some values oscillate in the final few iterations, with a few more iterations potentially allowing them to settle. However, this is not such a strong case as was for $\boldsymbol{v}$ or $\boldsymbol{l}$, shown in 5.10a,5.9, respectively.

LDL immediately outperforms the other models. While it does not approximate many of the smaller activities, it near perfectly fits the activity's peaks, as visible in figure 5.13b, while the ignored activity is also clearly visible in the comparison plot of $\widehat{\boldsymbol{y}}$ and $\boldsymbol{y}$ in figure 5.13a.

- $MSE = 0.6470 * 10^{-2}$
- $MAE = 4.6179 * 10^{-2}$
- $RMSE = 8.0436 * 10^{-2}$

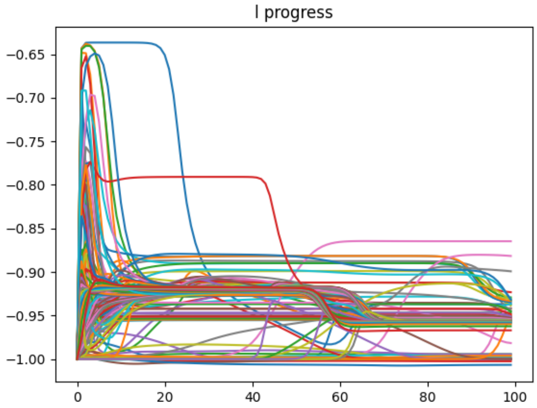
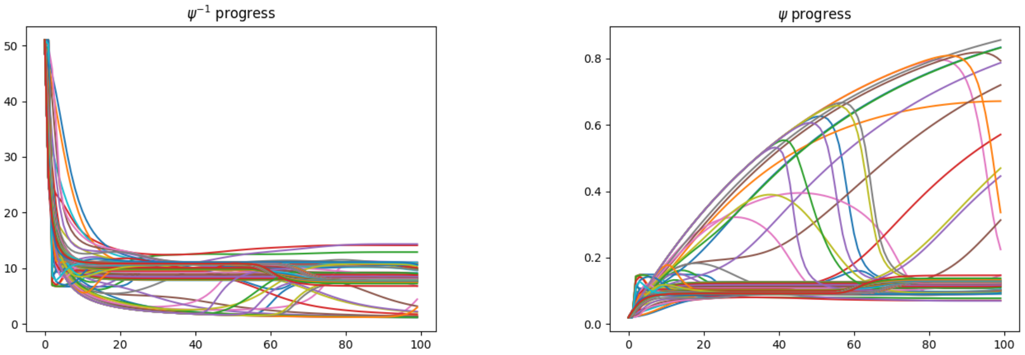


Figure 5.11 $\boldsymbol{l}$ parameter values during 100 iterations using LDL on Chernobyl data.



(a) Per iteration value of $\boldsymbol{\psi}^{-1}$ in LDL on the Chernobyl dataset.

(b) Per iteration value of $\boldsymbol{\psi}$ in LDL on the Chernobyl dataset.

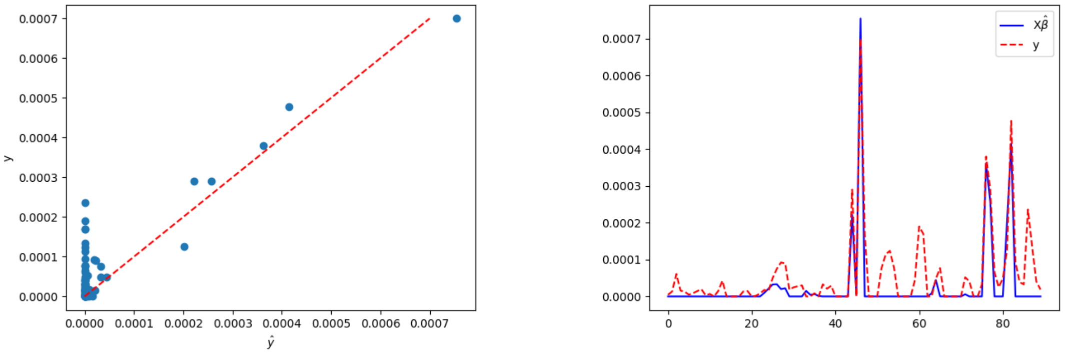
Figure 5.12 Approximate intermediate values of $\boldsymbol{\psi}$ during 100 iterations of LDL on the Chernobyl fires data.

5.2.4.1 Iteration impact

As mentioned in the previous section 5.2.4, the progress of values of certain parameters is indicative of the need for more iterations of the algorithm. The algorithm was run with 100 (the original, empirically set), 200, 300 and 400 iterations to study the impact of the number of iterations on the convergence of parameters. At these points, $\widehat{\boldsymbol{y}}$ was calculated and based on it the metrics, in the same vein as the previous cases. The progress of parameters is coalesced on the parameter basis due to overlap of the runs - a run with 200 iterations shares the first 100 iterations with the run of 100 iterations, etc. This progress is depicted in figures 5.14a,5.14b, 5.15b, 5.15a, with vertical lines at 100, 200, 300 and 400 iterations to paint a better picture of the outcome of each iterations count. Reconstructions were plotted against the ground truth separately for each checkpoint, shown in figures 5.16a,5.16b,5.16c, 5.16d.

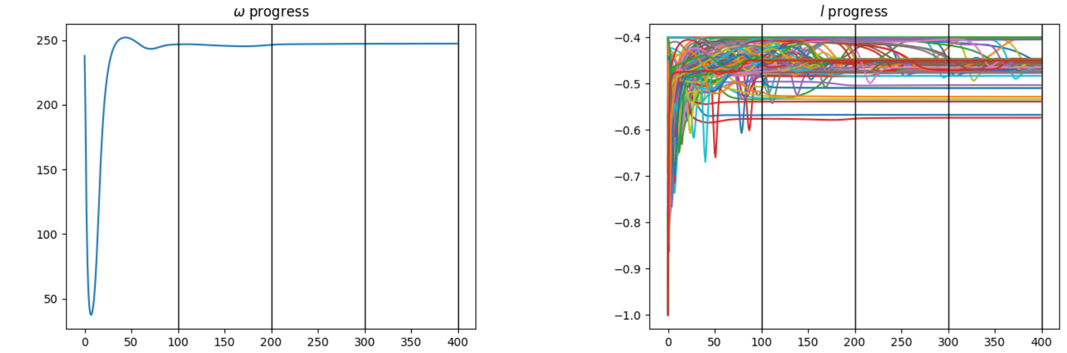
The convergence of ω occurs very early on in the iterative algorithm (figure 5.14a), within the first $\approx$ 20 iterations the algorithm reaches its vicinity and then within another 20 or so iterations it stabilises - the difference is in terms of units. Absolute convergence is achieved after 200 iterations.

$\boldsymbol{v}$ follows a similar behaviour in latter iterations as it did in the first 100 - a lot of the values



(a) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations.

(b) Side by side visualisation of Chernobyl fires' $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ estimated by an LDL model after 100 iterations.



(a) Per iteration value of ω during 400 iterations with markers at 100, 200, 300 and 400 iterations.

(b) Per iteration value of $\boldsymbol{l}$ during 400 iterations with markers at 100, 200, 300 and 400 iterations.

are 0 until a certain point, where they sharply rise. This happens primarily on the interval $[60, 200]$, and continues on a smaller scale onwards. Out of the 175 values, 143 are greater than 0 at the 400th iteration, as compared to 100 at the 100th iteration.

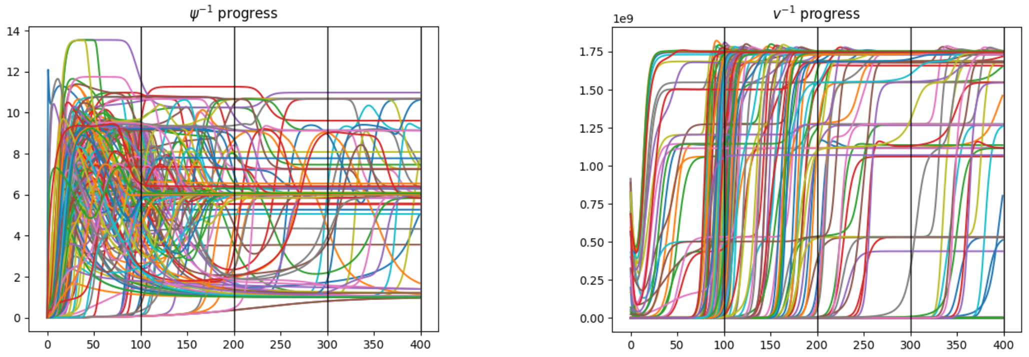
Omega converges quickly, whilst v takes a while to take off and then converges quickly, l oscillates a bit and psi does not converge within the original 100 iterations. 300 seems to capture the convergence, also leads to better results.

# of iterations	MSE ($\times 10^4$)	MAE ($\times 10^4$)	RMSE ($\times 10^4$)
100	37.9331	358.4933	615.8982
200	38.0000	358.1226	616.4373
300	37.8754	357.0150	615.4299
400	37.8561	356.8722	615.2730

Table 5.2 Metrics for the LDL algorithm given various number of iterations.

Increasing the number of iterations from 100 to 200 and higher had a positive impact on parameters' convergence and results in a lower error up to a point. It seems that increasing the number of iterations beyond 300 does not improve the results, only (albeit slightly) increases the algorithm's runtime in a linear fashion[2]. For each checkpoint, metrics are recorded in table 5.2,

[2] With the given data, number of iterations and a mid-range personal computer, the runtime was under 10



(a) Per iteration value of $\boldsymbol{\psi}^{-1}$ during 400 iterations with markers at 100, 200, 300 and 400 iterations.

(b) Per iteration value of $\boldsymbol{v}$ during 400 iterations with markers at 100, 200, 300 and 400 iterations.

showing that the performance improve as the number of iterations increases. However, there are diminishing returns, and as such it is not worth chasing this avenue further.

Given these results, the major takeaway is to increase the number of iterations to at least 200, preferably 300, strengthening the belief in the model's peformance with convergence in parameters.

5.2.4.2 Configuration consideration

To determine whether the task-agnostic setting of initial values and prior constants 5.1 is optimal for the task at hand, a grid search was informed with the following values tried in combination:

- $\omega \in \{(\max(\boldsymbol{X}^\intercal \boldsymbol{X}))^{-1} * 10^i \mid i \in -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$
- $v_i \in \{10^i \mid i \in -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$
- $l_i \in \{-1 * 10^i \mid i \in -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$
- $l_0 \in \{-1, -\frac{1}{2}, -\frac{1}{4}, -\frac{1}{8}, 0, 1\}$

This leads to a large number of runs $(6 * 10^3)$. The primary focus was to first find the best-performing combination of initial values for a fixed $l_0 = -1$. Originally, a smaller range of initial values was considered to get an idea as to which direction is more suitable, was later expanded to continue in that way. Further increasing the examinedp range does not yield substantially better results.

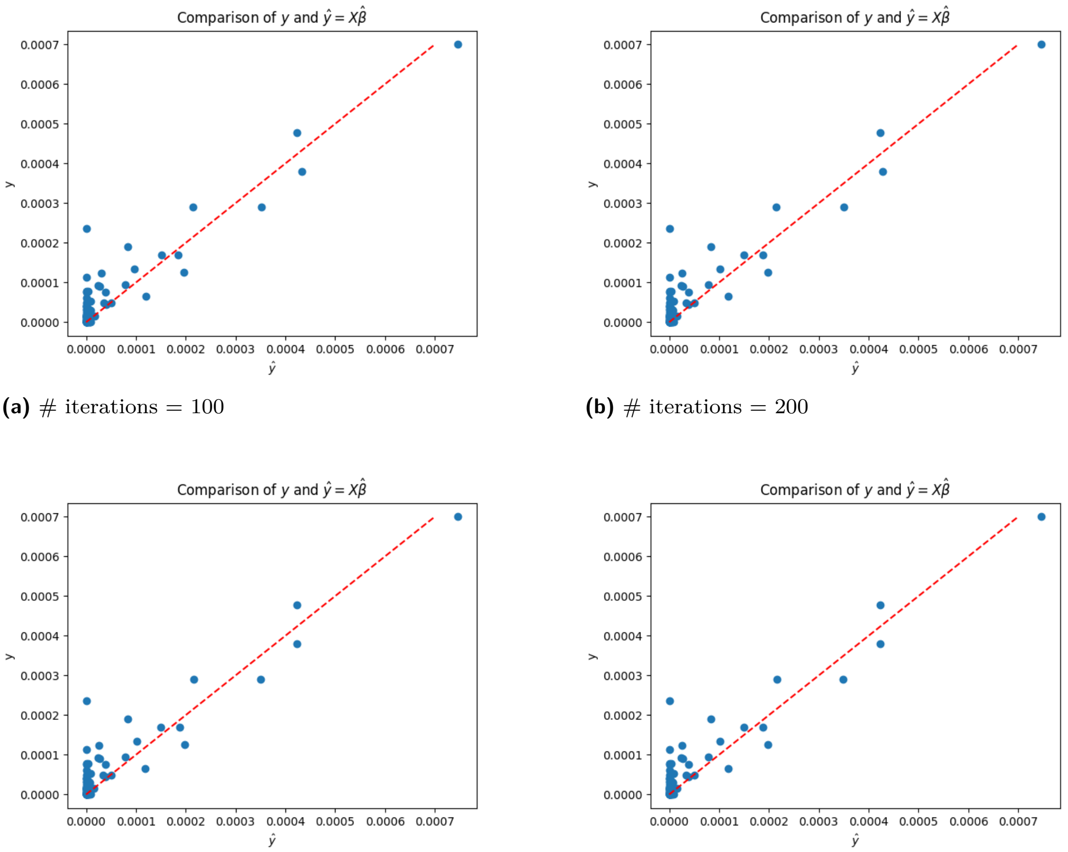
From this experimentation and examination, it was deduced that the best initial values are $\omega = (\max(\boldsymbol{X}^\intercal \boldsymbol{X}))^{-1} * 10^{-3}, v_i = 10^{-3}, l_i = -10$. Further, to observe whether softening the pressure of smoothing by the prior constant $l_0 = -1$, various landmarks points were examined as its potential replacement, resulting in the multiple figures in 5.17, using the previously set initial values[3] .

Table 5.3 holds the result of LDL runs with different values of l_0 with same initial values and 300 iterations. Landmark values, from the interval $[-1, 0]$ are extended by 1 just to dissuade any use of positive values of l_0. The goal of LDL (and smooth regression, as an extreme case of LDL), is to incorporate some degree of smoothness into the model, where $l_0 = -1$ creates a discrete derivative of the values and $l_0 = 0$ ignores smoothness altogether. Best results in terms of the mean squared error are achieved for $l_0 = -\frac{1}{2}$[4].

seconds for 400 iterations.

[3] One might consider returning to the tuning of initial values after determining the optimal l_0, which would lead to an iterative approach over an iterative approach . . . a metaiteration?

[4] This is the way.



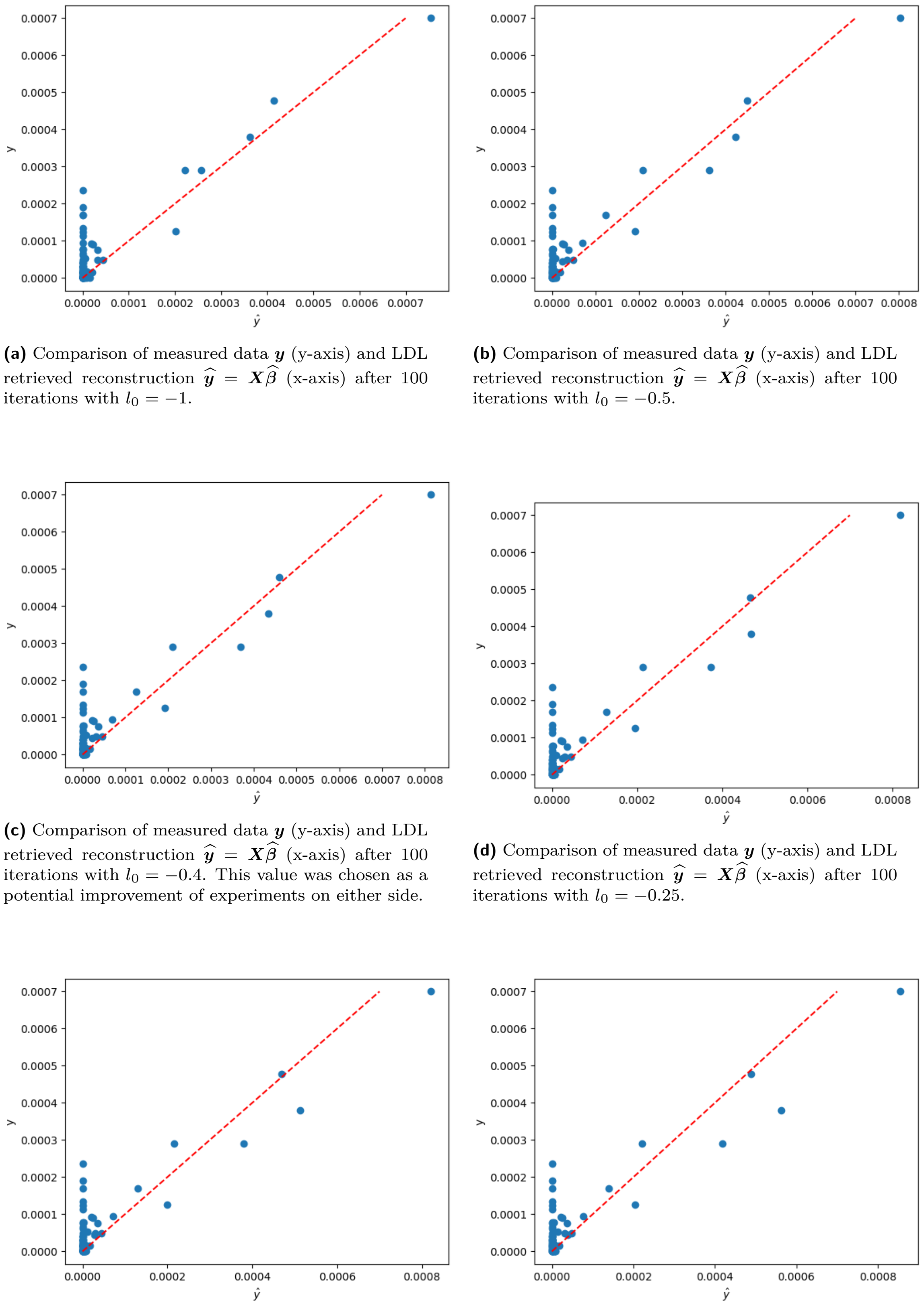
(a) # iterations = 100

(b) # iterations = 200

(c) # iterations = 300

(d) # iterations = 400

Figure 5.16 Comparison of $\boldsymbol{y}$ and $\widehat{y}$ after different number of iterations.



(a) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations with $l_0 = -1$.

(b) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations with $l_0 = -0.5$.

(c) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations with $l_0 = -0.4$. This value was chosen as a potential improvement of experiments on either side.

(d) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations with $l_0 = -0.25$.

(e) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations with $l_0 = -0.125$.

(f) Comparison of measured data $\boldsymbol{y}$ (y-axis) and LDL retrieved reconstruction $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ (x-axis) after 100 iterations with $l_0 = 0$.

Figure 5.17 Comparison of $\boldsymbol{y}$ and $\widehat{y}$ given different l_0 with 300 iterations and empirically set initial parameters.

l_0	MSE ($\times 10^4$)	MAE ($\times 10^4$)	RMSE ($\times 10^4$)
-1	0.647	4.618	8.044
-0.5	0.582	4.366	7.631
-0.25	0.606	4.462	7.788
-0.125	0.632	4.533	7.950
0	0.703	4.673	8.384
1	143205962973783	359801	1196687

Table 5.3 Metrics for the LDL algorithm given various prior constant l_0.

5.2.4.3 Master model

Given the previous experiments and investigations, a conclusion is drawn as to the best selection of prior constants and initial values for the LDL algorithm.

Parameter	Initial value
ω	$(\max \boldsymbol{X}^\intercal \boldsymbol{X})^{-1} * 10^{-3}$
μ_{l_i}	-10
Σ_{l_i}	0
v_i	10^{-3}
# of iterations	300
Prior constant	Value
a_0	10^{-10}
b_0	10^{-10}
c_0	10^{-10}
d_0	10^{-10}
e_0	10^{-2}
f_0	10^{-2}
l_0	-0.5

Table 5.4 Initial values and prior constants set for the best fit of LDL on Chernobyl data.

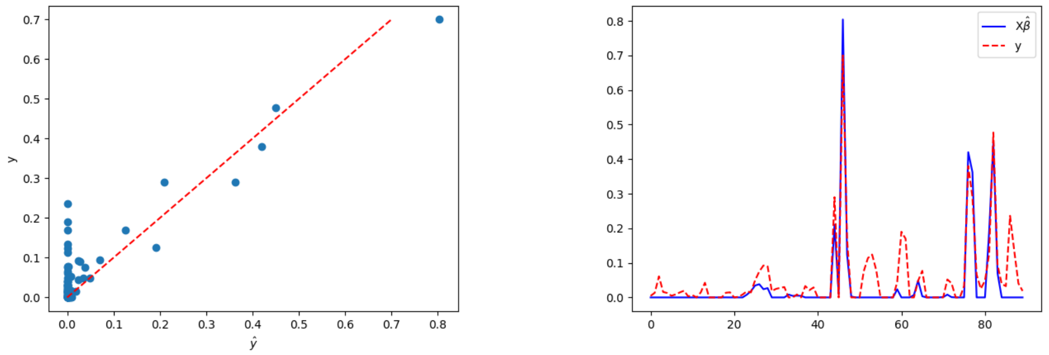
Compared to the results of the original run 5.2.4, there is an improvement in all regards. Do keep in mind that the metrics are calculated on the data used for fitting the model, which includes scaling and preprocessing. These values are not comparable with results on different data, such as ETEX 4.1.

- $MSE = 0.5815 * 10^{-2}$
- $MAE = 4.3589 * 10^{-2}$
- $RMSE = 7.6259 * 10^{-2}$

The resulting emissions, after reversing preprocessing, standardising and conversion from Bq to GBq, are shown for April 2020 in figure 5.19a. There are three estimated emissions of ^{137}Cs altogether, on the 7th, 9th and 10th of April, resulting in ≈ 450 GBq.

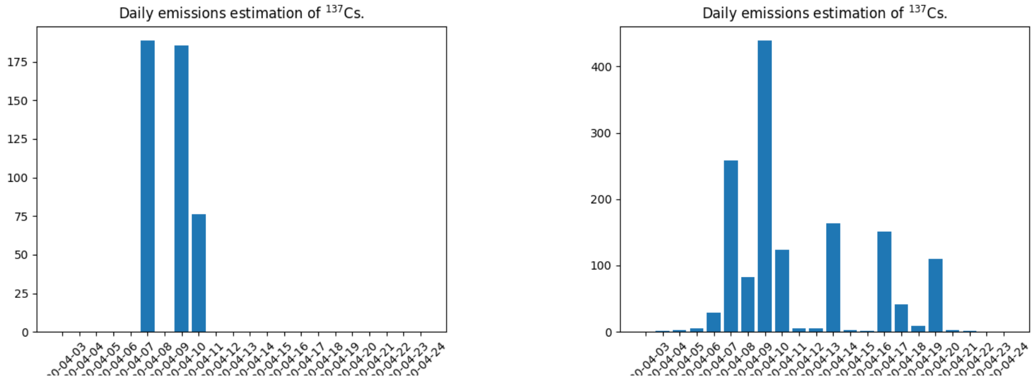
Even though LDL can model the degree of smoothness enforced, it is still enforced. This can be a potential issue in very sparse models, such as the modeled emissions, as the most of the values will be 0, forcing other values down near 0 to satisfy the smoothness criteria. To compare with a model that does not entertain the idea of smoothness, the sparse model is picked. The resulting emission profile for the same time period (April 2020) is plotted in 5.19b, resulting in a threefold total emission of ≈ 1434 GBq.

This best performing model approximated the total emission of 450 GBq. The daily profile is shown in figure 5.19a. This result falls near the results reported by other models on the same



(a) Best performing model's comparison of ground truth $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

(b) Best performing model's comparison of ground truth $\boldsymbol{y}$ and estimate $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.



(a) Fine-tuned LDL's emission estimate of ^{137}Cs.

(b) Sparse model's emission estimate of ^{137}Cs.

Figure 5.19 Daily emissions estimate of ^{137}Cs between 2.4.2020 and 25.4.2020.

dataset, such as [29], [25] and [103], which have estimated the total emission to be 574 GBq, 341 GBq and 650 GBq, respectively. Despite differing wildly, just as the aforementioned models do, in the day-to-day estimated emissions, the general profile (points of activity) is similar.

Their modeled emissions differ wildly in the day-to-day estimates as well, but capturing the activity in similar time steps. Within the domain of atmospheric dispersion modeling, due to the sheer magnitude of variables and the simplifying steps used in modeling, landing the result within a similar magnitude (±100%) is a good-enough result, but yielding a result not dissimilar from the others in either direction is a solid indicator of the model's quality.

Algorithm	MSE ($\times 10^2$)	MAE ($\times 10^2$)	RMSE ($\times 10^2$)
Linear regression	NaN	NaN	NaN
Ridge regression	3.0122	7.9480	17.3557
Bayesian ridge regression	3.0122	7.9485	17.3557
Bayesian sparse regression	3.6516	7.0076	19.1092
Bayesian smooth regression	2465.2423	201.3704	96.5121
LDL	0.6470	4.6179	8.0436
Fine-tuned LDL	0.5815	4.3589	7.6259

Table 5.5 Metrics of different algorithms on the 2020 Chernobyl fires 1.3.1 dataset.

Chapter 6
Conclusion

The goal of the thesis was to examine various methods of inference of an emission based on a set of measurements and to use a select few models from the Variational Bayes family to construct and compare the models.

The ETEX dataset 1.4 was used to develop and compare the models on a meticulously constructed dataset and to identify potential both strengths and weaknesses of individual implemented models. The best performing model on this well-documented dataset, based on the metrics recorded in table 4.1, is the sparse model 4.4, followed by the LDL model 4.6.

The developed models were then experimentally evaluated on a subset of a dataset with uncontrolled dispersion from the 2020 Chernobyl fires 1.3.1. The dataset was simplified for the purposes of modeling, taking into account only a single height of dispersion, as well as a single emitted particle size. These simplifications can cause a major discourse and difference in results with other established models, such as [29],[25] or [103], which aside from different modeling techniques have utilised different simplifications.

The best performing model on this dataset was a fine-tuned LDL 5.2.4.3, resulting in a total emission of 450 GBq, which is comparable to other models. Just as those models differ from each other in their day-to-day estimates, so does this model, estimating that the emission occurred within three separate days. Due to the sheer score of variables and the various simplifying steps used in atmospheric dispersion modeling, the models can differ drastically but still land within a similar resulting domain, all caused by simplifications of the underlying reality. Without a ground truth to compare against, it is difficult to determine which model performs the best. Altogether, this thesis produces a model that is sound, stable and with results on par with previously developed models.

However, even the best performing models, which have been able to capture the most important emissions, have their obvious shortcomings. Further fine-tuning of hyperparameters, such as prior constants, initial values or even different distributions which the parameters are assumed to follow might be necessary. In the case of the models presented and developed in this thesis, they could be further improved in particular by expanding the dataset used, removing a few of the simplifications utilised, which would require further modeling to encompass these realities. These would require taking into account other heights of initial emission, rather than the single one, as well as considering different particle sizes, as well as modeling the relationships between these different heights, particle sizes and other tangible parameters of the emission.

Bibliography

1. HAVER (HTTPS://TEX.STACKEXCHANGE.COM/USERS/190460/HAVER). *Drawing gamma function in LaTeX* [LaTeX Stack Exchange]. 2019. Available also from: `https://tex.stackexchange.com/a/520121`.
2. ZEVIANI, Walmes. *dist gamma*. 2020. Available also from: `https://github.com/walmes/Tikz/blob/master/src/dist%5C_gamma.pgf`.
3. SEIBERT, P.; FRANK, A. Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode. *Atmospheric Chemistry and Physics*. 2004, vol. 4, no. 1, pp. 51–63. Available from DOI: `10.5194/acp-4-51-2004`.
4. VAN DOP, H; ADDIS, R; FRASER, G; GIRARDI, F; GRAZIANI, G; INOUE, Y; KELLY, N; KLUG, W; KULMALA, A; NODOP, K; PRETEL, J. ETEX: A EUROPEAN TRACER EXPERIMENT; OBSERVATIONS, DISPERSION MODELLING AND EMERGENCY RESPONSE. *Atmospheric Environment*. 1998, vol. 32, no. 24, pp. 4089–4094. Available from DOI: `https://doi.org/10.1007/978-94-011-4570-1_11`.
5. HANNA, S R; BRIGGS, G A; HOSKER Jr, R P. Handbook on atmospheric diffusion. 1982. Available from DOI: `10.2172/5591108`.
6. INTERNATIONAL NUCLEAR SAFETY ADVISORY GROUP. Summary Report on the Post-accident Review Meeting on the Chernobyl Accident. *Summary Report on the Post-accident Review Meeting on the Chernobyl Accident*. 1986, vol. 1. Available also from: `https://www.iaea.org/publications/3598/summary-report-on-the-post-accident-review-meeting-on-the-chernobyl-accident`.
7. ASSOCIATION, World Nuclear. *RBMK reactors — reactor bolshoy moshchnosty kanalny — positive void coefficient - world nuclear association*. World Nuclear Association, 2022. Available also from: `https://www.world-nuclear.org/information-library/nuclear-fuel-cycle/nuclear-power-reactors/appendices/rbmk-reactors.aspx`.
8. WORLD NUCLEAR ASSOCIATION. *Chernobyl — Chernobyl Accident — Chernobyl Disaster - World Nuclear Association*. World Nuclear Association, 2021. Available also from: `https://world-nuclear.org/information-library/safety-and-security/safety-of-plants/chernobyl-accident.aspx`.
9. TOURAN, Nick. *The Chernobyl Disaster*. [N.d.]. Available also from: `https://whatisnuclear.com/chernobyl-main.html`.
10. NUCLEAR ENERGY INSTITUTE. *Comparing Fukushima and Chernobyl*. 2019. Available also from: `https://www.nei.org/resources/fact-sheets/comparing-fukushima-and-chernobyl`.

11. BEZPIATCHUK, Zhanna; LOMAKIN, Andrey. *The people who moved to Chernobyl - BBC News*. 2018. Available also from: https://www.bbc.co.uk/news/resources/idt-sh/moving_to_Chernobyl.

12. MERRIAM-WEBSTER. Die-off. In: *Merriam-Webster.com dictionary*. 2023. Available also from: https://www.merriam-webster.com/dictionary/die-off.

13. L'INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE. *Ukraine: Report on fires in the Chernobyl exclusion zone*. L'Institut de Radioprotection et de Sûreté Nucléaire, 2022. Available also from: https://en.irsn.fr/EN/newsroom/News/Pages/20220325_Ukraine-Report-on-fires-in-the-Chernobyl-exclusion-zone.aspx.

14. MILMAN, Oliver. *Forest fires erupt around Chernobyl nuclear plant in Ukraine*. Guardian Media Group, 2022. Available also from: https://www.theguardian.com/world/2022/mar/22/chernobyl-forest-fires-ukraine-nuclear-plant.

15. EVANGELIOU, N.; BALKANSKI, Y.; COZIC, A.; HAO, W. M.; MOUILLOT, F.; THONICKE, K.; PAUGAM, R.; ZIBTSEV, S.; MOUSSEAU, T. A.; WANG, R.; POULTER, B.; PETKOV, A.; YUE, C.; CADULE, P.; KOFFI, B.; KAISER, J. W.; MØLLER, A. P. Fire evolution in the radioactive forests of Ukraine and Belarus: future risks for the population and the environment. *Ecological Monographs*. 2015, vol. 85, no. 1, pp. 49–72. Available from DOI: https://doi.org/10.1890/14-1227.1.

16. C2ES. *Wildfires and Climate Change*. 2018. Available also from: https://www.c2es.org/content/wildfires-and-climate-change/.

17. NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. *Wildfire climate connection*. 2022. Available also from: https://www.noaa.gov/noaa-wildfire/wildfire-climate-connection.

18. IQAIR. *IQAir — First in Air Quality*. Fordos Andras, [n.d.]. Available also from: https://www.iqair.com/newsroom/wildfires-increasing-or-decreasing.

19. WORLD METEOROLOGICAL ORGANIZATION. *Number of wildfires forecast to rise by 50 % by 2100*. World Meteorological Organization, 2022. Available also from: https://public.wmo.int/en/media/news/number-of-wildfires-forecast-rise-50-2100.

20. SULLIVAN, Andrew; BAKER, Elaine; KURVITS, Tiina. *Spreading Like Wildfire*. UNEP, 2022. Available also from: https://www.researchgate.net/publication/358989662_Spreading_like_wildfire_the_rising_threat_of_extraordinary_landscape_fires.

21. BRAXTON LITTLE, Jane. Forest Fires Are Setting Chernobyl's Radiation Free. *The Atlantic*. 2020. Available also from: https://www.theatlantic.com/science/archive/2020/08/chernobyl-fires/615067/.

22. NUWER, Rachel. Forest Fires Threaten New Fallout From Chernobyl. *The New York Times*. 2015. Available also from: https://www.nytimes.com/2015/04/07/science/forest-fires-threaten-new-fallout-from-chernobyl.html.

23. MARA, Darren; BOLSOVER, Catherine. *Chernobyl threat – DW – 08/11/2010*. Ed. by KUEBLER, Martin. Deutsche Welle, 2020. Available also from: https://www.dw.com/en/russian-wildfires-spread-to-nuclear-contaminated-chernobyl-area/a-5889953.

24. EVANGELIOU, Nikolaos; BALKANSKI, Yves; COZIC, Anne; HAO, Wei Min; MØLLER, Anders Pape. Wildfires in Chernobyl-contaminated forests and risks to the population and the environment: A new nuclear disaster about to happen? *Environment International*. 2014, vol. 73, pp. 346–358. Available from DOI: https://doi.org/10.1016/j.envint.2014.08.012.

25. EVANGELIOU, Nikolaos; ECKHARDT, Sabine. Uncovering transport, deposition and impact of radionuclides released after the early spring 2020 wildfires in the Chernobyl Exclusion Zone. *Scientific Reports*. 2020, vol. 10, no. 1. Available from DOI: `https://doi.org/10.1038/s41598-020-67620-3`.

26. DE MEUTTER, Pieter; GUEIBE, Christophe; TOMAS, Jasper; OUTER, Peter den; APITULEY, Arnoud; BRUGGEMAN, Michel; CAMPS, Johan; DELCLOO, Andy; KNETSCH, Gert-Jan; ROOBOL, Lars; VERHEYEN, Leen. The assessment of the April 2020 chernobyl wildfires and their impact on Cs-137 levels in Belgium and The Netherlands. *Journal of Environmental Radioactivity*. 2021, vol. 237, no. 237, p. 106688. Available from DOI: `https://doi.org/10.1016/j.jenvrad.2021.106688`.

27. CDC. *CDC Radiation Emergencies*. 2018. Available also from: `https://www.cdc.gov/nceh/radiation/emergencies/isotopes/cesium.htm`.

28. CDC. *CDC Radiation Emergencies — Radioisotope Brief: Strontium-90 (Sr-90)*. 2019. Available also from: `https://www.cdc.gov/nceh/radiation/emergencies/isotopes/strontium.htm`.

29. TALERKO, Mykola; KOVALETS, Ivan; LEV, Tatiana; IGARASHU, Yasunori; ROMANENKO, Olexandr. Simulation study of radionuclide atmospheric transport after wildland fires in the Chernobyl Exclusion Zone in April 2020. *Atmospheric Pollution Research*. 2021, vol. 12, no. 3, pp. 193–204. ISSN 1309-1042. Available from DOI: `https://doi.org/10.1016/j.apr.2021.01.010`.

30. WOOD, Michael D.; BERESFORD, Nicholas A.; BARNETT, Catherine L.; BURGESS, Peter H.; MOBBS, Shelly. *Chornobyl radiation spikes are not due to military vehicles disturbing soil*. 2022. Available from arXiv: `2204.03157 [physics.med-ph]`.

31. KISELEV, A N. How much nuclear fuel is present in the lavalike fuel-containing mass in the fourth power-generating unit of the Chernobyl nuclear power plant? *Atomic Energy (New York)*. 1995, vol. 78, no. 4. Available from DOI: `10.1007/BF02416427`.

32. FOWLER, David. A chronology of global air quality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2020, vol. 378, no. 2183, p. 20190314. Available from DOI: `https://doi.org/10.1098/rsta.2019.0314`.

33. AGENCY, United States Environmental Protection. *Evolution of the Clean Air Act — US EPA*. 2018. Available also from: `https://www.epa.gov/clean-air-act-overview/evolution-clean-air-act`.

34. EVOTECH AIR QUALITY. 2022. Available also from: `https://www.evotechairquality.co.uk/articles/history-of-air-quality`.

35. MOORE, G.E. Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE*. 1998, vol. 86, no. 1, pp. 82–85. Available from DOI: `https://doi.org/10.1109/jproc.1998.658762`.

36. EUROPEAN UNION, Publications Office of the. *What is open data — data.europa.eu*. European Union, 2011. Available also from: `https://data.europa.eu/en/trening/what-open-data`.

37. HENRY, Christophe; MINIER, Jean-Pierre; BRAMBILLA, Sara. Particle resuspension: Challenges and perspectives for future models. *Physics Reports*. 2023, vol. 1007, pp. 1–98. ISSN 0370-1573. Available from DOI: `https://doi.org/10.1016/j.physrep.2022.12.005`. Particle resuspension: challenges and perspectives for future models.

38. NODOP, K.; CONNOLLY, R.; GIRARDI, F. The field campaigns of the European Tracer Experiment (ETEX). *Atmospheric Environment*. 1998, vol. 32, no. 24, pp. 4095–4108. Available from DOI: `https://doi.org/10.1016/s1352-2310(98)00190-3`.

39. MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. *Introduction to Linear Regression Analysis*. New York, NY: John Wiley & Sons, 2015. ISBN 9781119180173.

40. PROVOST, Foster. *Glossary of Terms Journal of Machine Learning*. Ed. by KOHAVI, Ron. Leland Stanford Junior University, 1998. Available also from: `https://ai.stanford.edu/~ronnyk/glossary.html`.

41. ALVES, Rodrigo. *Linear regression* [`https://courses.fit.cvut.cz/BIE-VZD/lectures/files/2022/BI-VZD-04-WS2223-en-slides.pdf`]. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022.

42. ROTH, Dan; ZHOU, Ben; CERVANTES, C.; CHENG, C. *Introduction to Machine Learning*. 3330 Walnut Street , Levine Hall, Philadelphia, PA 19104-6309, 2016. Available also from: `https://www.cis.upenn.edu/~danroth/Teaching/CS446-17/LectureNotesNew/intro/main.pdf`.

43. KURAMA, Vihar. *Regression in Machine Learning: What it is and Examples of Different Models*. 2019. Available also from: `https://builtin.com/data-science/regression-machine-learning`.

44. ALVES, Rodrigo. *Data Mining Supervised Learning Concepts*. Faculty of Information Technology, Czech Technical University, 2022. Available also from: `https://courses.fit.cvut.cz/BIE-VZD/lectures/files/2022/BI-VZD-02-WS2223-en-slides_v2.pdf`.

45. KALVODA, Tomáš; VAŠATA, Daniel. *Základy matematické analýzy*. Thákurova 9, 160 00 Prague 6: Faculty of Information Technology, Czech Technical University, Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022. Available also from: `https://courses.fit.cvut.cz/BI-ZMA/textbook/bi-zma-textbook.pdf`.

46. JAKKANWAR, Atharva. *A Machine Learning primer: Almost without the math — Part 1*. A Medium Corporation, 2018. Available also from: `https://towardsdatascience.com/a-machine-learning-primer-almost-without-the-math-part-1-19ed04e352c0`.

47. BROWNLEE, Jason. *Train-Test Split for Evaluating Machine Learning Algorithms*. 2020. Available also from: `https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/`.

48. JAMES, Gareth Michael; WITTEN, Daniela; HASTIE, Trevor J; TIBSHIRANI, Robert. *An introduction to statistical learning : with applications in R*. New York: Springer, 2013. ISBN 9781461471387.

49. KLOUDA, Karel; VAŠATA, Daniel. *Lineární regrese* [`https://courses.fit.cvut.cz/BI-VZD/lectures/files/BI-VZD-05-cs-handout.pdf`]. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022.

50. STEPHENS, David A. *Regression Modelling And Least-Squares*. Imperial College London, 2005. Available also from: `https://www.ma.imperial.ac.uk/~das01/GSACourse/Regression.pdf`.

51. WEISSTEIN, Eric W. *Metric*. Wolfram, [n.d.]. Available also from: `https://mathworld.wolfram.com/Metric.html`.

52. STAROSTA, Štěpán; JAN, Spěvák. *Vícerozměrný prostor a funkce* [`https://courses.fit.cvut.cz/NI-MPI/latex/lectures/czech/mi-mpi-prednaska-11-analyza-I-handout.pdf`]. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022.

53. LI, Chi-Kwong. Norms, Isometries, and Isometry Groups. *The American Mathematical Monthly*. 2000, vol. 107, no. 4, pp. 334–340. Available from DOI: `https://doi.org/10.1080/00029890.2000.12005201`.

54. WEISSTEIN, Eric W. *Norm.* Wolfram Research, [n.d.]. Available also from: https://mathworld.wolfram.com/Norm.html.

55. HTTPS://WWW.FACEBOOK.COM/JASON.BROWNLEE.39. *Gradient Descent for Machine Learning.* Guiding Tech Media, 2016. Available also from: https://machinelearningmastery.com/gradient-descent-for-machine-learning/.

56. KWIATKOWSKI, Robert. *Gradient Descent Algorithm — a deep dive.* A Medium Corporation, 2021. Available also from: https://towardsdatascience.com/gradient-descent-algorithm-a-deep-dive-cf04e8115f21.

57. ANWAR, Aqeel. *Types of Regularization in Machine Learning.* A Medium Corporation, 2021. Available also from: https://towardsdatascience.com/types-of-regularization-in-machine-learning-eb5ce5f9bf50.

58. TEAM, CFI. *Elastic Net.* CFI Education Inc., 2022. Available also from: https://corporatefinanceinstitute.com/resources/data-science/elastic-net/.

59. *Andrey Nikolayevich Tikhonov.* 2023. Available also from: https://en.wikipedia.org/wiki/Andrey_Nikolayevich_Tikhonov.

60. HOERL, Arthur E.'; KENNARD, Robert W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics.* 2000, vol. 42, no. 1, pp. 80–86. Available from DOI: https://doi.org/10.1080/00401706.2000.10485983.

61. FORNACON-WOOD, Isabella; MISTRY, Hitesh; JOHNSON-HART, Corinne; FAIVRE-FINN, Corinne; O'CONNOR, James P.B.; PRICE, Gareth J. Understanding the Differences Between Bayesian and Frequentist Statistics. *International Journal of Radiation Oncology*Biology*Physics.* 2022, vol. 112, no. 5, pp. 1076–1082. Available from DOI: https://doi.org/10.1016/j.ijrobp.2021.12.011.

62. BLAŽEK, Rudolf B.; KOTECKÝ, Roman; VAŠATA, Daniel; HRABÁKOVÁ, Jitka; NOVÁK, Petr. *BI-PST – Pravděpodobnost a statistika.* Thákurova 9, 160 00 Prague 6: Faculty of Information Technology, Czech Technical University, Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022. Available also from: https://courses.fit.cvut.cz/BI-PST/media/lectures/BI-PST-Textbook.pdf.

63. BLAŽEK, Rudolf B.; HRABÁKOVÁ, Jitka; HRABÁK, Pavel; KOTECKÝ, Roman; NOVÁK, Petr; VAŠATA, Daniel. *NI-VSM – Vybrané statistické metody.* Thákurova 9, 160 00 Prague 6: Faculty of Information Technology, Czech Technical University, Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2023. Available also from: https://courses.fit.cvut.cz/NI-VSM/lectures/files/NI-VSM-TextBook-Handout.pdf.

64. DEDECIUS, Kamil. *Základy a specifika bayesovské teorie.* Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022. Available also from: https://gitlab.fit.cvut.cz/koristo1/ni-bml-lectures/-/blob/master/prednasky/Uvod_do_modelovani-KD.ipynb.

65. KLOUDA, Karel. *Logistická regrese* [https://courses.fit.cvut.cz/BI-VZD/lectures/files/BI-VZD-09-cs-handout.pdf]. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2022.

66. FISHER, R. A.; RUSSELL, Edward John. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character.* 1922, vol. 222, no. 594-604, pp. 309–368. Available from DOI: 10.1098/rsta.1922.0009.

67. GOYAL, CHIRAG. *Moments - A Must Known Statistical Concept for Data Science*. 2022. Available also from: https://www.analyticsvidhya.com/blog/2022/01/moments-a-must-known-statistical-concept-for-data-science/.

68. GUNDERSEN, Gregory. *Understanding Moments*. 2020. Available also from: https://gregorygundersen.com/blog/2020/04/11/moments/.

69. LANE, David M; SCOTT, David; HEBL, Mikki; GUERRA, Rudy; OSHERSON, Dan; ZIMMER, Heidi. *Online Statistics Education*. Ed. by LANE, David M. Independently published, 2014. ISBN 9781687894250. Available also from: https://onlinestatbook.com/.

70. LOPES, AA. Statistical Inference (part II): the Normal and Related Distributions. *The Brazilian Journal of Infectious Diseases: an Official Publication of the Brazilian Society of Infectious Diseases*. 1998, vol. 2, no. 4, pp. 170–174. Available also from: https://pubmed.ncbi.nlm.nih.gov/11103005/.

71. ELLIOTT, Peter. *Are all the moments of the normal distribution finite?* 2017. Available also from: https://www.quora.com/Are-all-the-moments-of-the-normal-distribution-finite.

72. LABARBERA, Michael. *It's Alive!* University of Chicago Press, 2013. ISBN 9780226094885. Available also from: https://fathom.lib.uchicago.edu/2/21701757/.

73. ALLEN, David H. *How Mechanics Shaped the Modern World*. Cham Springer International Publishing, 2014. ISBN 9783319017013.

74. BOX, George E. P. Science and Statistics. *Journal of the American Statistical Association*. 1976, vol. 71, no. 356, pp. 791–799. Available from DOI: https://doi.org/10.1080/01621459.1976.10480949.

75. PINNOW, Stefan. *How can I draw the probability density function a truncated normal distribution?* [LaTeX Stack Exchange]. 2016. Available also from: https://tex.stackexchange.com/a/341886.

76. TICHÝ, O.; ŠMÍDL, V.; HOFMAN, R.; STOHL, A. LS-APC v1.0: a tuning-free method for the linear inverse problem and its application to source-term determination. *Geoscientific Model Development*. 2016, vol. 9, no. 11, pp. 4297–4311. Available from DOI: 10.5194/gmd-9-4297-2016.

77. TICHÝ, Ondřej. *Přednáška 12: Bilineární model a bayesovská maticová dekompozice - 2. část*. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2020. Available also from: https://gitlab.fit.cvut.cz/dedeckam/ni-bml-lectures/-/blob/master/prednasky/Bilinearni_model_pokr_OT.ipynb.

78. VÁCLAV, Šmídl; TICHÝ, Ondřej. Sparsity in Bayesian Blind Source Separation and Deconvolution. In: BLOCKEEL, Hendrik; KERSTING, Kristian; NIJSSEN, Siegfried; ŽELEZNÝ, Filip (eds.). *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–563. ISBN 978-3-642-40991-2.

79. TABOGA, Marco. *Gamma function*. Kindle Direct Publishing, 2021. Available also from: https://www.statlect.com/mathematical-tools/gamma-function.

80. FINK, Daniel. A Compendium of Conjugate Priors. *Technical Report*. 1997.

81. DEDECIUS, Kamil. *Sekvenční odhad lineárních modelů, predikce*. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology, Czech Technical University in Prague, 2023. Available also from: https://gitlab.fit.cvut.cz/dedeckam/ni-bml-lectures/-/blob/master/prednasky/02_Linearni_modely_a_regrese_KD.ipynb.

82. 2023. Available also from: `https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions`.

83. JORDAN, Michael I. *The exponential family: Conjugate priors.* 2010. Available also from: `https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter9.pdf`.

84. ŠMÍDL, Václav. *The Variational Bayes Approach in Signal Processing.* 2004.

85. FRAGOSO, Tiago M.; BERTOLI, Wesley; LOUZADA, Francisco. Bayesian Model Averaging: A Systematic Review and Conceptual Classification. *International Statistical Review.* 2017, vol. 86, no. 1, pp. 1–28. Available from DOI: `10.1111/insr.12243`.

86. JOYCE, James M. Kullback-Leibler Divergence. In: *International Encyclopedia of Statistical Science.* Ed. by LOVRIC, Miodrag. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 720–722. ISBN 978-3-642-04898-2. Available from DOI: `10.1007/978-3-642-04898-2_327`.

87. HAN, Jiawei. *2.4.8 Kullback-Leibler Divergence.* [N.d.]. Available also from: `http://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf`.

88. ATTIAS, Hagai. A Variational Baysian Framework for Graphical Models. In: SOLLA, S.; LEEN, T.; MÜLLER, K. (eds.). *Advances in Neural Information Processing Systems.* MIT Press, 1999, vol. 12. Available also from: `https://proceedings.neurips.cc/paper_files/paper/1999/file/74563ba21a90da13dacf2a73e3ddefa7-Paper.pdf`.

89. ŠMÍDL, Václav; QUINN, Anthony. *The Variational Bayes Method in Signal Processing.* Berlin/Heidelberg: Springer-Verlag, 2006. ISBN 3540288198. Available from DOI: `https://doi.org/10.1007/3-540-28820-1`.

90. TICHÝ, O.; ULRYCH, L.; ŠMÍDL, V.; EVANGELIOU, N.; STOHL, A. On the tuning of atmospheric inverse methods: comparisons with the European Tracer Experiment (ETEX) and Chernobyl datasets using the atmospheric transport model FLEXPART. *Geoscientific Model Development.* 2020, vol. 13, no. 12, pp. 5917–5934. Available from DOI: `10.5194/gmd-13-5917-2020`.

91. FOUNDATION, Python Software. *Python Documentation.* Python Software Foundation, 2021. Available also from: `https://docs.python.org/3/index.html`.

92. KLUYVER, Thomas; RAGAN-KELLEY, Benjamin; PÉREZ, Fernando; GRANGER, Brian; BUSSONNIER, Matthias; FREDERIC, Jonathan; KELLEY, Kyle; HAMRICK, Jessica; GROUT, Jason; CORLAY, Sylvain; IVANOV, Paul; AVILA, Damián; ABDALLA, Safia; WILLING, Carol. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: LOIZIDES, F.; SCHMIDT, B. (eds.). *Positioning and Power in Academic Publishing: Players, Agents and Agendas.* IOS Press, 2016, pp. 87–90.

93. VIRTANEN, Pauli; GOMMERS, Ralf; OLIPHANT, Travis E.; HABERLAND, Matt; REDDY, Tyler; COURNAPEAU, David; BUROVSKI, Evgeni; PETERSON, Pearu; WECKESSER, Warren; BRIGHT, Jonathan; VAN DER WALT, Stéfan J.; BRETT, Matthew; WILSON, Joshua; MILLMAN, K. Jarrod; MAYOROV, Nikolay; NELSON, Andrew R. J.; JONES, Eric; KERN, Robert; LARSON, Eric; CAREY, C J; POLAT, İlhan; FENG, Yu; MOORE, Eric W.; VANDERPLAS, Jake; LAXALDE, Denis; PERKTOLD, Josef; CIMRMAN, Robert; HENRIKSEN, Ian; QUINTERO, E. A.; HARRIS, Charles R.; ARCHIBALD, Anne M.; RIBEIRO, Antônio H.; PEDREGOSA, Fabian; VAN MULBREGT, Paul; SCIPY 1.0 CONTRIBUTORS. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods.* 2020, vol. 17, pp. 261–272. Available from DOI: `10.1038/s41592-019-0686-2`.

94. PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011, vol. 12, pp. 2825–2830.

95. HARRIS, Charles R.; MILLMAN, K. Jarrod; WALT, Stéfan J. van der; GOMMERS, Ralf; VIRTANEN, Pauli; COURNAPEAU, David; WIESER, Eric; TAYLOR, Julian; BERG, Sebastian; SMITH, Nathaniel J.; KERN, Robert; PICUS, Matti; HOYER, Stephan; KERKWIJK, Marten H. van; BRETT, Matthew; HALDANE, Allan; RÍO, Jaime Fernández del; WIEBE, Mark; PETERSON, Pearu; GÉRARD-MARCHANT, Pierre; SHEPPARD, Kevin; REDDY, Tyler; WECKESSER, Warren; ABBASI, Hameer; GOHLKE, Christoph; OLIPHANT, Travis E. Array programming with NumPy. *Nature*. 2020, vol. 585, no. 7825, pp. 357–362. Available from DOI: `10.1038/s41586-020-2649-2`.

96. MEURER, Aaron; SMITH, Christopher P.; PAPROCKI, Mateusz; ČERTÍK, Ondřej; KIRPICHEV, Sergey B.; ROCKLIN, Matthew; KUMAR, AMiT; IVANOV, Sergiu; MOORE, Jason K.; SINGH, Sartaj; RATHNAYAKE, Thilina; VIG, Sean; GRANGER, Brian E.; MULLER, Richard P.; BONAZZI, Francesco; GUPTA, Harsh; VATS, Shivam; JOHANSSON, Fredrik; PEDREGOSA, Fabian; CURRY, Matthew J.; TERREL, Andy R.; ROUČKA, Štěpán; SABOO, Ashutosh; FERNANDO, Isuru; KULAL, Sumith; CIMRMAN, Robert; SCOPATZ, Anthony. SymPy: symbolic computing in Python. *PeerJ Computer Science*. 2017, vol. 3, e103. ISSN 2376-5992. Available from DOI: `10.7717/peerj-cs.103`.

97. MCKINNEY, Wes. Data Structures for Statistical Computing in Python. In: WALT, Stéfan van der; MILLMAN, Jarrod (eds.). *Proceedings of the 9th Python in Science Conference*. 2010, pp. 56–61. Available from DOI: `10.25080/Majora-92bf1922-00a`.

98. HUNTER, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. 2007, vol. 9, no. 3, pp. 90–95. Available from DOI: `10.1109/MCSE.2007.55`.

99. DOMBEK, Daniel; KALVODA, Tomáš; KLEPRLÍK, Luděk; KLOUDA, Karel. *Lineární algebra Studijní text*. Thákurova 9, 160 00 Prague 6: Department of applied mathematics, Faculty of Information Technology Czech Technical University in Prague, 2019. Available also from: `https://kam.fit.cvut.cz/deploy/bi-lin//lin-text.pdf`.

100. WEISSTEIN, Eric W. *The CRC concise encyclopedia of mathematics*. London: Chapman & Hall, 2011. ISBN 9781420072174.

101. WEISSTEIN, Eric W. *Delta Function*. Wolfram Research, Inc., [n.d.]. Available also from: `https://mathworld.wolfram.com/DeltaFunction.html`.

102. WEISSTEIN, Eric W. *Matrix Trace*. Wolfram Research, Inc., 2002. Available also from: `https://mathworld.wolfram.com/MatrixTrace.html`.

103. LAPTEV, Gennady; VOITSEKHOVITCH, Oleg; PROTSAK, Valentyn. Estimation of Radioactive Source Term Dynamics for Atmospheric Transport during Wildfires in Chernobyl Zone in Spring 2020. 2020.