Aalto University
School of Science

Mikael Kuusela

Statistical Issues in Unfolding Methods for High Energy Physics

Master's thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Technology in the Degree Programme in Engineering Physics and Mathematics.

Espoo, July 26, 2012

Supervisor: Prof. Esko Valkeila

Instructors: Prof. Victor Panaretos, D.Sc. (Tech.) Mikko Voutilainen

Aalto University
School of Science

Abstract of the Master's Thesis

Author:	Mikael Kuusela		
Title:	Statistical Issues in Unfolding Methods for High Energy Physics		
Supervisor: **Instructors:**	Prof. Esko Valkeila Prof. Victor Panaretos, D.Sc. (Tech.) Mikko Voutilainen		
Degree programme:	Engineering Physics and Mathematics		
Major subject:	Mathematics	**Minor subject:**	Particle and Astrophysics
Chair (code):	Mat-1		

Abstract: Due to the finite resolution of real-world particle detectors, any measurement conducted in experimental high energy physics is contaminated by stochastic smearing. This thesis studies the problem of unfolding these measurements to estimate the true physical distribution of the observable of interest before undesired detector effects. This problem is an ill-posed statistical inverse problem in the sense that straightforward inversion of the folding operator produces in most cases highly oscillating unphysical solutions.

The first contribution of this thesis is to provide a rigorous mathematical understanding of the unfolding problem and the currently used unfolding techniques. To this end, we provide a mathematical model for the observations using indirectly observed Poisson point processes. We then explore the tools provided by both the frequentist and Bayesian paradigms of statistics for solving the problem. We show that the main issue with regularized frequentist point estimates is that the bias of these estimators makes error estimation of the unfolded solution challenging. This problem can be resolved by using Bayesian credible intervals, but then one has to make an essentially arbitrary choice for the regularization strength of the Bayesian prior.

Having gained a proper understanding about the issues involved in current unfolding methods, we proceed to propose a novel empirical Bayes unfolding technique. We solve the issue of choosing the spread of the regularizing Bayesian prior by finding a point estimate of the free hyperparameters via marginal maximum likelihood using a variant of the EM algorithm. This point estimate is then plugged into Bayes' rule to summarize our understanding of the unknowns via the Bayesian posterior. We conclude with a computational demonstration of unfolding with a particular emphasis on empirical Bayes unfolding.

Pages: vii+140	**Language:** English	**Date:** July 26, 2012

Keywords: Unfolding, inverse problems, empirical Bayes, EM algorithm, Markov chain Monte Carlo, Poisson point processes, high energy physics

Aalto-yliopisto
Perustieteiden
korkeakoulu

Diplomityön tiivistelmä

Tekijä:	Mikael Kuusela
Työn nimi:	Detektoriefektien poisto hiukkasfysiikan tilastollisessa data-analyysissä
Työn valvoja:	Prof. Esko Valkeila
Työn ohjaajat:	Prof. Victor Panaretos, TkT Mikko Voutilainen
Koulutusohjelma:	Teknillinen fysiikka ja matematiikka
Pääaine: Matematiikka	**Sivuaine:** Hiukkas- ja astrofysiikka
Opetusyksikön (ent. professuuri) koodi:	Mat-1

Tiivistelmä: Detektorien rajallisen resoluution takia jokainen kokeellisessa hiukkasfysiikassa tehtävä mittaus sisältää ei-toivottuja stokastisia efektejä. Tämä diplomityö käsittelee näiden detektoriefektien poistamista (engl. unfolding), millä tarkoitetaan kokeellisista efekteistä puhdistetun todellisen jakauman estimoimista kiinnostuksen kohteena olevalle fysikaaliselle suureelle. Koska detektoriefektejä kuvaavan operaattorin suora kääntäminen tuottaa useimmiten epäkelpoja oskilloivia ratkaisuja, kyseessä on haastava tilastollinen inversio-ongelma.

Tämän työn ensimmäinen päämäärä on muodostaa tarkka matemaattinen malli detektoriefektien poistamiselle käyttäen epäsuorasti havaittuja Poisson-pisteprosesseja. Tämän jälkeen työssä analysoidaan sekä frekventistisen että bayesilaisen tilastotieteen näkökulmasta tehtävään käytettyjä nykymenetelmiä. Analyysi osoittaa, että frekventististen piste-estimaattorien tapauksessa löydetyn ratkaisun virherajojen estimointi on hankalaa johtuen regularisoitujen estimaattorien harhaisuudesta. Ratkaisuksi ongelmaan on esitetty bayesilaisten luottamusvälien käyttöä, mutta tällöin herää kysymys siitä, kuinka regularisaatiovoimakkuutta säätelevä priorijakauma tulisi valita.

Työssä esitetään näiden ongelmien ratkaisuksi uutta detektoriefektien poistomenetelmää, joka perustuu empiiriseen Bayes-estimointiin. Menetelmässä regularisoivan priorijakauman vapaat hyperparametrit estimoidaan suurimman reunauskottavuuden menetelmällä EM-algoritmia käyttäen, minkä jälkeen tämä piste-estimaatti sijoitetaan Bayesin kaavaan. Näin saatavaa posteriorijakaumaa voidaan sitten käyttää bayesilaisten luottamusvälien muodostamiseen. Tämän uuden detektoriefektien poistomenetelmän toiminta varmennetaan simulaatiokokeita käyttäen.

Sivumäärä: vii+140	**Kieli:** englanti	**Päivämäärä:** 26.7.2012

Avainsanat: Detektoriefektien poisto, inversio-ongelma, empiirinen Bayes-estimointi, EM-algoritmi, Markovin ketju Monte Carlo, Poisson-pisteprosessi, hiukkasfysiikka

Preface

This work represents a collaboration between the Chair of Mathematical Statistics at École Polytechnique Fédérale de Lausanne (EPFL) and the CMS experiment at CERN, the European Organization for Nuclear Research, and was carried out in Switzerland during spring 2012.

I would like to express my gratitude to my instructors Victor Panaretos and Mikko Voutilainen for their guidance and continuous support during this project, for answering the numerous questions I had and for providing feedback on the manuscript. I would in addition like to thank Esko Valkeila for supervising this thesis. It was also a great pleasure to work with the CMS Statistics Committee, and I would especially like to thank to Robert Cousins, Tommaso Dorigo and Louis Lyons for encouraging and enlightening discussions. I would also like to acknowledge the numerous CMS physicists who were willing to spend their time explaining unfolding to me in great detail. Furthermore, I would like to thank Yoav Zemel for useful comments on the manuscript, András László for interesting discussions and Otto Seiskari for allowing me to use this custom-made LaTeX template.

This thesis was funded by the CMS programme at Helsinki Institute of Physics and I would like to gratefully acknowledge their contribution to this project. Partial financial support was also provided by Aalto University and Aalto University Student Union. Thanks are also due to my two host organizations, EPFL and CERN, for providing office space and access to their facilities.

Finally, I would like to thank my friends and family for all their support and encouragement in the course of this project.

Espoo, July 26, 2012

Mikael Kuusela

Contents

Notation and Abbreviations

Notation

$\perp\!\!\!\perp X_i$	the random elements X_i are independent
1_A	indicator function of the set A
$\boldsymbol{A}^\dagger$	Moore–Penrose pseudoinverse of the matrix $\boldsymbol{A}$
A^c	complement of the set A
$\mathrm{bias}(\hat{\theta})$	bias of the estimator $\hat{\theta}$ of the parameter θ
$\mathrm{Bin}(p, n)$	binomial distribution with n trials and success probability p
$\mathrm{cond}(\boldsymbol{A})$	condition number of the matrix $\boldsymbol{A}$
$\delta_{\boldsymbol{x}}$	Dirac measure at $\boldsymbol{x}$
$\det(\boldsymbol{A})$	determinant of the matrix $\boldsymbol{A}$
$\mathrm{diag}\big(a_1, \ldots, a_{\min(m,n)}\big)_{m\times n}$	$m \times n$ diagonal matrix with diagonal elements $a_1, \ldots, a_{\min(m,n)}$
$f * g$	convolution of f and g
$\Gamma(\cdot)$	gamma function
$\ker(\boldsymbol{A})$	kernel of the matrix $\boldsymbol{A}$
$\mathrm{MSE}[\hat{\boldsymbol{\theta}}]$	mean squared error of the estimator $\hat{\boldsymbol{\theta}}$
$\mathrm{Mult}(\boldsymbol{p}, n)$	multinomial distribution with probabilities $\boldsymbol{p}$ and n trials
ν	Lebesgue measure
$\mathbb{N}_0^d$	d-dimensional natural numbers including 0
$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ evaluated at $\boldsymbol{x}$
$\mathcal{P}(A)$	power set of the set A
$\mathrm{Poisson}(\boldsymbol{\lambda})$	d-variate probability distribution where the components are Poisson distributed with parameters λ_i, $i = 1, \ldots, d$
p_{T}	transverse momentum
ϱ	counting measure
$\mathbb{R}_+^d$	non-negative d-dimensional real numbers
$\mathrm{ran}(\boldsymbol{A})$	range of the matrix $\boldsymbol{A}$
$\mathrm{rank}(\boldsymbol{A})$	rank of the matrix $\boldsymbol{A}$
$\mathrm{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)$	linear span of $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$

$\mathrm{supp}(f)$	support of the function f
$\mathrm{tr}(\boldsymbol{A})$	trace of the matrix $\boldsymbol{A}$
$\hat{\theta}$	estimator of the parameter θ
$W^{\perp}$	orthogonal complement of the subspace W
$X \sim P_X$	the random element X has the distribution P_X
$X^{(k)} \overset{\text{i.i.d.}}{\sim} P_X$	the random elements $X^{(1)}, X^{(2)}, \ldots$ are independent and identically distributed with distribution P_X
$X \overset{d}{=} Y$	the random elements X and Y are equal in distribution

Abbreviations

a.e.	almost everywhere
a.s.	almost surely
cdf	cumulative distribution function
CMS	Compact Muon Solenoid
EM	expectation-maximization (algorithm)
HEP	high energy physics
i.i.d.	independent and identically distributed
LHC	Large Hadron Collider
LS	least squares
MAP	maximum a posteriori (estimator)
MC	Monte Carlo
MCEM	Monte Carlo expectation-maximization (algorithm)
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimator
MMLE	marginal maximum likelihood estimator
MoM	method of moments
MSE	mean squared error
pdf	probability density function
pmf	probability mass function
r.v.	random variable
SVD	singular value decomposition
TSVD	truncated singular value decomposition

Chapter 1

Introduction

The exponential growth of computing power during the past few decades has made 21st century science increasingly data-intensive. This is especially true for the field of high energy physics, where the scientist working at CERN, the European Organization for Nuclear Research, analyze annually some 15 petabytes of data recorded at the world's most powerful particle accelerator, the Large Hadron Collider (LHC). The study of these massive amounts of data is hoped to shed light on some the biggest mysteries of the Universe, such as the origin of mass, the nature of the mysterious dark matter or the apparent asymmetry between ordinary matter and antimatter. Apart from the obvious computational challenges related to the sheer size of the data set, the complex internal structure of the LHC data requires scientists to use complicated statistical techniques ranging from state-of-the-art multivariate classifiers to advanced statistical hypothesis testing to ensure the correct analysis and interpretation of these data. Moreover, these unprecedented statistical challenges make LHC data analysis a fertile ground for innovation on novel statistical data analysis methods.

This thesis studies a particular data analysis task, called the *unfolding problem* [13, 6, 41], encountered in the analysis of data produced by the LHC. Namely, the observations recorded with any real-world particle detector are always subject to undesired experimental effects, such as limited detector resolution, noise and detection inefficiencies. The observation of such distorted collision events instead of the desired true events is called *smearing* or *folding* of the data and often results in broadening of the physical spectra measured by the LHC experiments. Unfolding then refers to using the smeared observations to infer the true physical distribution of the events.

In high energy physics, probability distributions are, for practical reasons, often discretized using histograms. In this case, smearing of the true physical histogram $\boldsymbol{x}$, where $\boldsymbol{x}$ is a vector containing the bin counts of the histogram, can be understood as stochastic migration of events to their neighboring bins due to the noise in the detector. As a result of these migrations, we then actually observe the smeared histogram denoted by $\boldsymbol{y}$. An illustration of the effect of such smearing on the observations is shown in Figure 1.1 for a two-component Gaussian mixture model. Here, each observation constituting the true histogram $\boldsymbol{x}$ is smeared by additive Gaussian white

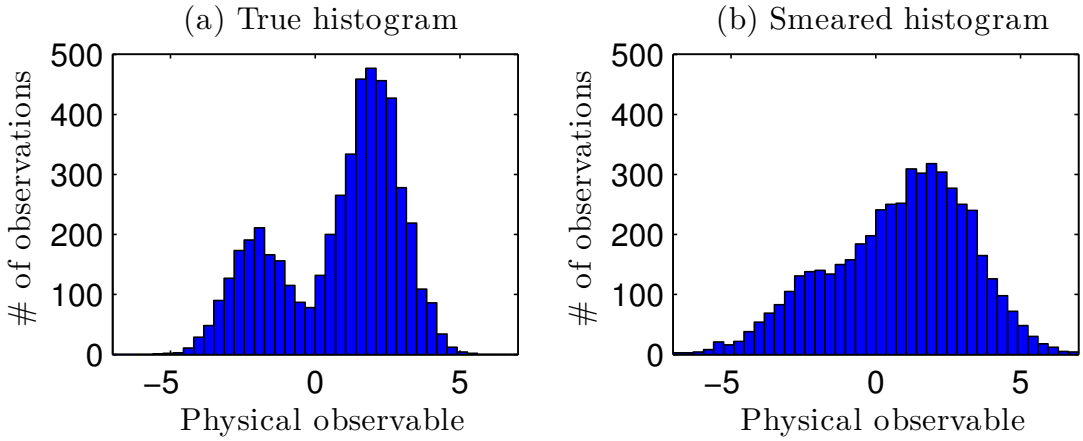


Figure 1.1: Illustration of smearing of a two-component Gaussian mixture model by a convolution operator. The peaks of the true histogram shown in Figure (a) are less prominent in the smeared histogram of Figure (b). The goal of unfolding is, roughly speaking, to recover the true histogram of Figure (a) given the observed smeared histogram of Figure (b).

noise which represents one of the simplest special cases of smearing encountered in experimental high energy physics.

In addition to smearing, there is another stochastic component in these observations. Namely, it follows from the laws of physics that the total number of observations in the histograms is Poisson distributed. As a result, all the bins of both the true histogram $\boldsymbol{x}$ and the smeared histogram $\boldsymbol{y}$ are Poisson distributed with bin means $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, respectively. We then assume that we can model the smearing by relating the bin means via $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$, where $\boldsymbol{K}$ a known *smearing matrix*. Here the (i, j)th element of $\boldsymbol{K}$ corresponds to the probability of observing an event in the ith bin of the smeared histogram given that it originates from the jth bin of the true histogram. The task in unfolding is then to use the observed smeared Poisson counts $\boldsymbol{y}$ to infer the Poisson means $\boldsymbol{\lambda}$ of the true histogram. As such, the high energy physics unfolding problem is related to deconvolution in optics and image reconstruction in medical imaging where the data are often also assumed to follow a Poisson distribution.

There are at least three reasons why it would be desirable to unfold the measurements. Firstly, publication of a non-fundamental smeared histogram is obviously intellectually dissatisfying if it was possible to publish an estimate of the true physical distribution of events. Secondly, unfolding enables comparison of measurements of two different experiments with different experimental resolutions, and thirdly, the unfolded histograms can be directly compared with theoretical predictions. This is especially valuable when a theorist comes up with a new physical theory many years from now and wants to compare his predictions with previously published measurements. Nevertheless, it should be noted that many measurements in high energy physics can be carried out using the smeared observations in which case most of the

complications discussed in this thesis can be avoided altogether.

The main problem with unfolding is that it is a challenging statistical inverse problem [31, 21, 20, 2]. In the discrete case, this means that the smearing matrix $\boldsymbol{K}$ is ill-conditioned in the sense that the solution of the linear system of equations $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ is extremely sensitive to the value of $\boldsymbol{\mu}$ which has to be inferred from the observations $\boldsymbol{y}$. Another way of putting this is to say that the (generalized) inverse of the smearing matrix corresponds to a nearly discontinuous linear map, which means that any noise in the observations might be amplified arbitrarily by the inversion. Because of the Poisson fluctuations, such noise is always present in our data and hence straightforward usage of the inverse to unfold the distribution often leads to completely unacceptable solutions. Luckily, this problem can be dealt with by injecting additional outside information into the problem, which is called *regularization* of the solution.

Due to the mathematical and statistical challenges involved, unfolding has caused a lot of confusion and controversy among the high energy physics community. Traditionally, the problem is solved by what is called *bin-by-bin correction factors*. This means that we use a Monte Carlo (MC) generator to estimate the mean $\boldsymbol{\lambda}^{\mathrm{MC}}$ of the true histograms as well as the mean $\boldsymbol{\mu}^{\mathrm{MC}}$ of the smeared histogram. The multiplicative factor between each bin of the two $C_i = \lambda_i^{\mathrm{MC}}/\mu_i^{\mathrm{MC}}$ is then used to correct the observed histogram $\boldsymbol{y}$ to the truth-level. Hence, the scaled values $\hat{\lambda}_i = C_i y_i$ are used as an estimator of $\boldsymbol{\lambda}$. The problem with this approach is that it essentially corrects for the "efficiency" of each bin instead of the migration of events between the bins. By doing so, the method has been shown to introduce a major bias for the MC model used in deriving the correction factors C_i [41].

Recently, a number of methods that perform unfolding by correcting for the bin-to-bin migrations have been proposed as an alternative to bin-by-bin corrections. The two most widely used methods are the "Bayesian" D'Agostini iteration[1] described in [16] and the SVD method of Höcker and Kartvelishvili [28]. Nevertheless, both theoretical and practical understanding of these techniques has remained somewhat limited. There have been concerns especially about the error estimation of the unfolded solutions provided by these methods. Namely, it seems that in some cases these methods can provide errors that appear to be smaller that the ones obtained in an ideal perfect detector without any smearing.

With this background, the goals of this thesis are two-fold. Firstly, we aim at mathematically rigorous understanding of the unfolding problem and the methods that are currently used for finding the unfolded solution. After gaining a solid understanding about the problem at hand and the limitations of the current unfolding methods, the second goal of this work is to determine which techniques developed by the inverse problems and statistics communities would be the most suitable for solving the high energy physics unfolding problem.

We begin by formulating in Chapter 2 a mathematical model for the smeared observations using the theory of Poisson point processes. This chapter also fixes most

[1]For reasons explained later, the term "Bayesian" in the name of the D'Agostini iteration is an unfortunate misnomer originating from the fact that D'Agostini used repeated application of Bayes' rule to derive the method.

of the notation used in the rest of this thesis. We also provide an alternative, more accessible discrete model for the observations which has traditionally been used in the relevant physics literature. Before discussing unfolding in detail, we also review in Chapter 3 the well-understood statistical inference techniques for the Poisson means $\boldsymbol{\lambda}$ in the case of direct observations without smearing.

Chapter 4 is aimed at understanding the tools provided by frequentist statistics for solving the unfolding problem. Among other things, we study the identifiability of the parameters of the model and then explain maximum likelihood and least squares estimation of the unknown means $\boldsymbol{\lambda}$. It is shown that the D'Agostini iteration in fact corresponds to the famous expectation-maximization (EM) algorithm [19] for the maximum likelihood estimator of $\boldsymbol{\lambda}$. That is, there is nothing "Bayesian" about the method. Similarly, we note that the SVD method of Höcker and Kartvelishvili is a certain generalization of Tikhonov regularization. We also explain that the error estimation of the frequentist unfolding techniques is challenging because of the bias of the regularized estimators which means that the estimated standard deviations of the estimators can no longer be used to construct approximate confidence intervals for the solution. In fact, if the bias is ignored, it is possible to make the error bars of the solution arbitrarily small by increasing the strength of the regularization.

Since the main problem with frequentist point estimates appears to be the characterization of the error associated with the solution, we move on to Bayesian analysis of the problem in Chapter 5. The use of Bayesian techniques in unfolding was recently proposed by Choudalakis in [11]. The motivation for this is that the Bayesian posterior provides a very natural way of estimating the uncertainty of the solution via Bayesian credible intervals. We show that the problem can be regularized by an appropriate choice of the prior distribution in Bayes' theorem and that Bayesian inference can then be carried out by using the Metropolis–Hastings algorithm to sample from the posterior.

The problem that remains with Bayesian unfolding is that one has to find a way of choosing the regularization strength imposed by the prior distribution, which can have a major impact on the outcome of the unfolding procedure. In Chapter 6, we propose tackling this problem using empirical Bayes techniques where the hyperparameters of the prior are fitted to the data by maximizing their marginal likelihood. To achieve this, we derive a variant of the EM algorithm for finding the marginal maximum likelihood estimator of the unknown free hyperparameters. Using such frequentist point estimator of the hyperparameters enables us to choose the optimal regularization strength objectively based on the observed data instead of performing subjective inference inherent in fully Bayesian unfolding. Even though empirical Bayes has been used earlier in solving especially geophysical inverse problems [42, 46], this is, to the best of our knowledge, the first time the technique has been applied to solving the high energy physics unfolding problem. Hence, Chapter 6 represents the main novel contribution of this thesis.

Chapter 7 is devoted to computational demonstration of unfolding with a particular emphasis on empirical Bayes unfolding. The method is first used for unfolding the Gaussian mixture model data shown in Figure 1.1 and then for unfolding a simulated data set corresponding to the inclusive jet cross section measurement [51]

at the Compact Muon Solenoid (CMS) experiment at the LHC. It is shown that by using empirical Bayes unfolding, the true histogram can be recovered with high accuracy in both of these cases, while unregularized inversion produces unsatisfactory results. We discuss ways of improving the unfolding techniques presented in this thesis in Chapter 8 before concluding with a set of general observations and recommendations on unfolding.

The presentation of this thesis assumes a good working knowledge of the main concepts of measure-theoretic probability theory, mathematical statistics, advanced linear algebra (especially the singular value decomposition and Moore–Penrose pseudoinverse) and the mathematical theory of inverse problems. A reader unfamiliar with these subjects is recommended to consult Appendix A where these topics are reviewed before proceeding with the main contents of the thesis.

Chapter 2

Formulation of the Unfolding Problem

The aim of this chapter is to establish a connection between the smeared observations and the true histogram. Statistical inference for the resulting mathematical model will form the basis of unfolding discussed in the rest of this thesis. We will provide two alternative formulations for the unfolding problem. The first formulation is based on indirectly observed Poisson point processes and is treated in Section 2.1. The model is formulated for continuous intensity functions of the Poisson processes and then discretized using histograms. We also provide a more accessible but less general alternative formulation for the unfolding problem starting from the discretized setting in Section 2.2.

2.1 Formulation as an Indirectly Observed Poisson Point Process

The mathematical theory of Poisson point processes provides a natural theoretical framework for the high energy physics unfolding problem. In this section, we formulate the unfolding problem in terms of an indirectly observed Poisson point process following the treatment presented in [49]. Other standard references for measure-theoretic introduction to Poisson point processes include [33, 17], while [14, 34] provide a more accessible treatment of the subject.

2.1.1 Introduction to Point Processes

Let us start with the definition of a point measure. Let E be a state space representing the space of our physical observables of interest. In this work, we require E to be a Borel set of the d-dimensional real space $\mathbb{R}^d$, although the theory is applicable in more general spaces as well. Let $\mathcal{B}_E$ be the Borel σ-algebra on E and $\{\boldsymbol{x}_i \in E : i \in I\}$, for some index set I, be a set of points in E. A point measure is then defined as the measure which counts the number of these points belonging to a Borel set $B \in \mathcal{B}_E$.

Definition 2.1. A *point measure* is the discrete measure

$$\xi : \mathcal{B}_E \to \mathbb{N}_0,\ B \mapsto \xi(B) = \sum_{i \in I} \delta_{\boldsymbol{x}_i}(B),$$

where $\delta_{\boldsymbol{x}}(B) = 1_B(\boldsymbol{x})$ is the Dirac measure of the set $B \in \mathcal{B}_E$ at $\boldsymbol{x} \in E$. The set of all such measures is denoted by $\Xi(E)$.

A point process is a random point measure, that is, a measurable mapping from an underlying probability space $(\Omega, \mathcal{F}, P)$ to the space of point measures $\Xi(E)$.

Definition 2.2. A *point process* $M : \Omega \to \Xi(E)$ is a point measure valued random element.

Hence, the value $M(B), B \in \mathcal{B}_E$, is a random integer counting the number of points $\boldsymbol{x}_i$ contained in B. In our case, this corresponds to the number of events seen in the particle detector with the numerical values of the observables in B. To see how to define a σ-algebra in the space of point measures, which is required to check the measurability of M, we refer the reader to [49, Section 1.1].

We are often interested in the expected number of points in a given Borel set. This information is given by the mean measure of the point process of interest.

Definition 2.3. The *mean measure* $\lambda : \mathcal{B}_E \to \mathbb{R}_+$ of a point process M is defined by the expectations

$$\lambda(B) = \mathrm{E}[M(B)], \quad B \in \mathcal{B}_E.$$

One can easily show that the mean measure is indeed a measure. In what follows, we often assume that the mean measure λ is absolutely continuous and hence, by the Radon–Nikodym theorem, we have

$$\lambda(B) = \int_B f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \quad \forall B \in \mathcal{B}_E,$$

where the almost everywhere unique density $f : E \to [0, +\infty)$ is called the *intensity function* of the mean measure λ.

The following theorem can be used to check the distributional equivalence of two point processes.

Theorem 2.4. *Let M_1 and M_2 be point processes on state space E. Then the following two are equivalent:*

(i) $M_1 \stackrel{d}{=} M_2$.

(ii) For every finite collection of pairwise disjoint sets $B_1, \ldots, B_n \in \mathcal{B}_E$:

$$\big[M_1(B_1), \ldots, M_1(B_n)\big] \stackrel{d}{=} \big[M_2(B_1), \ldots, M_2(B_n)\big].$$

Proof. See Theorem 1.1.1 and Criterion 1.1.2 in [49]. □

2.1.2 Poisson Point Processes

We now proceed to give the definition of a Poisson point process.

Definition 2.5. Let λ be a finite measure. Then the point process $M : \Omega \to \Xi(E)$ is a *Poisson point process* if

(i) $M(B) \sim \text{Poisson}(\lambda(B))$ for all $B \in \mathcal{B}_E$ and

(ii) $M(B_1), \ldots, M(B_n)$ are independent for all pairwise disjoint sets $B_1, \ldots, B_n \in \mathcal{B}_E$.

A Poisson point process, or a Poisson process for short, is hence a point process where the number of observed points $M(B)$ on any Borel set $B \in \mathcal{B}_E$ follows a Poisson distribution. Since $\mathrm{E}[M(B)] = \lambda(B)$, the measure λ in the definition is also the mean measure of the Poisson process. According to the following theorem, it uniquely determines the distribution of a Poisson process.

Theorem 2.6. *Poisson point processes with equal finite mean measures λ are equal in distribution.*

Proof. Let M_1 and M_2 be Poisson point processes with mean measure λ. Then it follows from Definition 2.5 that for disjoint sets $B_1, \ldots, B_n \in \mathcal{B}_E$, we have

$$\big[M_1(B_1), \ldots, M_1(B_n)\big] \stackrel{d}{=} \big[M_2(B_1), \ldots, M_2(B_n)\big] .$$

Hence, by Theorem 2.4, we have $M_1 \stackrel{d}{=} M_2$. □

Let us note that such a result does not hold for general point processes. Theorem 2.6 is important because it tells us that we can use the mean measure λ, or its intensity function f, to characterize a Poisson process. When the intensity is constant, i.e., $f(\boldsymbol{x}) \equiv C$, $C \geq 0$, we call the Poisson process *homogeneous*. When this is not the case, we talk about *inhomogeneous* Poisson processes.

The following theorem establishes a convenient, explicit representation for a Poisson process.

Theorem 2.7. *Let λ be a finite measure with $\lambda(E) > 0$ and let*

$$M = \sum_{i=1}^{\tau} \delta_{\boldsymbol{X}_i} \tag{2.1}$$

be a point process, where $\tau, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are independent random variables with $\tau \sim \text{Poisson}(\lambda(E))$ and the points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \in E$ are identically distributed with distribution $P_{\boldsymbol{X}_i}(B) = P_{\boldsymbol{X}}(B) = \lambda(B)/\lambda(E), B \in \mathcal{B}_E$. Then M is a Poisson point process with mean measure λ.

Proof. See Theorem 1.2.1(i) in [49]. □

Note that τ is the random total number of observed points, $M(E) = \tau$. In fact, such a representation exists not only for Poisson point processes but also for a class of more general point processes given certain regularity conditions on the measures involved [33]. Since in this work we are only interested in Poisson processes, the less general Theorem 2.7 will suffice for our needs.

Theorem 2.7 has a number of important consequences. Firstly, Equation (2.1) can be used to numerically sample from the Poisson process by first sampling τ from the Poisson distribution with parameter $\lambda(E)$ and then sampling $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau$ from the distribution $P_{\boldsymbol{X}} = \lambda/\lambda(E)$. Secondly, we have

$$\lambda(B) = \lambda(E)P_{\boldsymbol{X}}(B) = \mathrm{E}[\tau]P_{\boldsymbol{X}}(B), \quad \forall B \in \mathcal{B}_E. \tag{2.2}$$

Hence, when densities exist, we have the relation

$$f(\boldsymbol{x}) = \mathrm{E}[\tau]p_{\boldsymbol{X}}(\boldsymbol{x}) \quad \text{a.e.} \tag{2.3}$$

between the intensity function f of M and the probability density function $p_{\boldsymbol{X}}$ of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ Thus, if the points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are distributed according to $p_{\boldsymbol{X}}$ and their total number follows a Poisson distribution, we see that this is a Poisson process whose intensity function is simply the density function $p_{\boldsymbol{X}}$ scaled by the expected number of points $\mathrm{E}[\tau]$.

A number of standard, elementary operations for Poisson point processes, such as transformations and truncations, are often studied in the literature. Out of these, the concept of thinning turns out to be important for modeling the efficiency of a detector.

Definition 2.8. Let $\tau, (\boldsymbol{X}_1, Z_1), (\boldsymbol{X}_2, Z_2), \ldots$ be independent random variables with τ Poisson distributed and $(\boldsymbol{X}_i, Z_i) \in E \times \{0, 1\}$ identically distributed for all i. Furthermore, denote $\varepsilon(\boldsymbol{x}) = P(Z = 1 | \boldsymbol{X} = \boldsymbol{x})$. We then call

$$M^* = \sum_{i=1}^{\tau} Z_i \delta_{\boldsymbol{X}_i}$$

a *thinned Poisson point process* with *thinning function* $\varepsilon(\boldsymbol{x})$ and underlying Poisson point process $M = \sum_{i=1}^{\tau} \delta_{\boldsymbol{X}_i}$.

Hence, a thinned Poisson process is a Poisson process where each point $\boldsymbol{X}_i$ is observed with probability $\varepsilon(\boldsymbol{X}_i)$. The random variables Z_i are indicator variables indicating if the point $\boldsymbol{X}_i$ is observed or not. The following proposition establishes the mean measure of a thinned Poisson process.

Proposition 2.9. *Let M be a Poisson process with mean measure λ and M^* a thinning of M with thinning function $\varepsilon(\boldsymbol{x})$. The mean measure λ^* of M^* is then*

$$\lambda^*(B) = \int_B \varepsilon(\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{x}), \quad B \in \mathcal{B}_E.$$

Proof. By Definition 2.3, we have

$$\begin{aligned}
\lambda^*(B) &= \mathrm{E}[M^*(B)] \\
&= \mathrm{E}\left[\sum_{i=1}^{\tau} Z_i \delta_{\boldsymbol{X}_i}(B)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{\tau} Z_i \delta_{\boldsymbol{X}_i}(B) \middle| \tau\right]\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{\tau} \mathrm{E}[Z_i \delta_{\boldsymbol{X}_i}(B)]\right],
\end{aligned}$$

where the conditioning on τ can be dropped on the last line since (X_i, Z_i) are independent of τ. Here we have

$$\begin{aligned}
\mathrm{E}[Z_i \delta_{\boldsymbol{X}_i}(B)] &= \mathrm{E}[\mathrm{E}[Z_i \delta_{\boldsymbol{X}_i}(B) | \boldsymbol{X}_i]] \\
&= \mathrm{E}[\delta_{\boldsymbol{X}_i}(B) \mathrm{E}[Z_i | \boldsymbol{X}_i]] \\
&= \mathrm{E}[\delta_{\boldsymbol{X}_i}(B) P(Z_i = 1 | \boldsymbol{X}_i)] \\
&= \mathrm{E}[1_B(\boldsymbol{X}_i) \varepsilon(\boldsymbol{X}_i)] \\
&= \int 1_B(\boldsymbol{x}) \varepsilon(\boldsymbol{x}) \, \mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_B \varepsilon(\boldsymbol{x}) \, \mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x}),
\end{aligned}$$

where the dependence on the index i can be dropped since $\boldsymbol{X}_i$ are identically distributed. Hence

$$\lambda^*(B) = \mathrm{E}\left[\sum_{i=1}^{\tau} 1\right] \int_B \varepsilon(\boldsymbol{x}) \, \mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x}) = \mathrm{E}[\tau] \int_B \varepsilon(\boldsymbol{x}) \, \mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x}) = \int_B \varepsilon(\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{x}),$$

where the last equality follows from Equation (2.2). □

Hence, the intensity of M^* is $f^*(\boldsymbol{x}) = \varepsilon(\boldsymbol{x}) f(\boldsymbol{x})$, where $f(\boldsymbol{x})$ is the intensity function of M.

In many real-life situations, one observes a total of t i.i.d. points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t \in E$. If we know in addition that the total number of points is Poisson distributed and independent of the observations, we are then dealing with a single realization of a Poisson point process and could be interested in inferring its intensity function given the data. We call this the inference of the intensity function of a *directly observed Poisson point process*. For example, in experimental high energy physics, one usually performs the measurement of the physical quantity of interest on some interval $E = [a, b]$ and it follows from the underlying physics that the total number of observations falling on this interval is Poisson distributed. The data analysis task is then to infer the intensity function of the corresponding Poisson process, which is then used validate, reject or constrain physical theories.

2.1.3 Indirectly Observed Poisson Point Processes

Let us assume that we are interested in the Poisson process

$$M = \sum_{i=1}^{\tau} \delta_{\boldsymbol{X}_i},$$

where the points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \in E$ are independent and identically distributed with pdf $p_{\boldsymbol{X}}$. Imagine, however, that instead of M, we were to observe another Poisson process

$$N = \sum_{i=1}^{\tau} \delta_{\boldsymbol{Y}_i},$$

where the points $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots \in E$ are known to be noisy versions of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ More formally, we assume that

$$\boldsymbol{Y}_i = m(\boldsymbol{X}_i, \boldsymbol{E}_i), \quad i = 1, 2, \ldots \tag{2.4}$$

for some function m and random variables $\boldsymbol{E}_i$. In addition, we assume that the pairs $(\boldsymbol{X}_1, \boldsymbol{E}_1), (\boldsymbol{X}_2, \boldsymbol{E}_2), \ldots$ are i.i.d. and hence the resulting $\boldsymbol{Y}_i$ are also i.i.d. random variables. The pdfs $p_{\boldsymbol{X}}$ and $p_{\boldsymbol{Y}}$ of the points $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are then related by the integral equation

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \int p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int k(\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \tag{2.5}$$

where we have defined $k(\boldsymbol{x}, \boldsymbol{y}) := p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$. The function $k : E \times E \to \mathbb{R}_+$ is called the *kernel function* and, in this case, satisfies $\int k(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = 1$, $\forall \boldsymbol{x} \in E$.

A classical example of such a situation is when the points $\boldsymbol{X}_i$ are corrupted by additive noise $\boldsymbol{E}_i$, i.e., $\boldsymbol{Y}_i = \boldsymbol{X}_i + \boldsymbol{E}_i$, where $\boldsymbol{E}_i$ are independent and identically distributed with pdf $p_{\boldsymbol{E}}$ and independent of the $\boldsymbol{X}_i$. The pdf of the noisy observations $\boldsymbol{Y}_i$ is then given by the convolution

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = (p_{\boldsymbol{X}} * p_{\boldsymbol{E}})(\boldsymbol{y}) = \int p_{\boldsymbol{E}}(\boldsymbol{y} - \boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$$

and we have $k(\boldsymbol{x}, \boldsymbol{y}) = p_{\boldsymbol{E}}(\boldsymbol{y} - \boldsymbol{x})$.

Using Equation (2.2), we then know that the mean measure λ of M is $\lambda = \mathrm{E}[\tau] P_{\boldsymbol{X}}$ and the mean measure μ of N is $\mu = \mathrm{E}[\tau] P_{\boldsymbol{Y}}$. When f and h denote the intensity functions of M and N, respectively, we then have by Equation (2.3) that $f = \mathrm{E}[\tau] p_{\boldsymbol{X}}$ and $h = \mathrm{E}[\tau] p_{\boldsymbol{Y}}$. Hence, using Equation (2.5) we get

$$h(\boldsymbol{y}) = \mathrm{E}[\tau] \int k(\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int k(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.$$

From this, we see that the kernel k also relates the intensities of the two Poisson processes. In such a case, we call M an indirectly observed Poisson process.

Definition 2.10. Let M and N be Poisson point processes with state spaces E and F, mean measures λ and μ and intensity functions f and h, respectively, and assume that we observe N. Assume further that $\mu = \int K(\boldsymbol{x}, \cdot)\, \mathrm{d}\lambda(\boldsymbol{x})$ for kernel K and furthermore that $k(\boldsymbol{x}, \cdot)$ is the density of $K(\boldsymbol{x}, \cdot)$ for all $\boldsymbol{x} \in E$ so that $h(\boldsymbol{y}) = \int k(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$. We then call M an *indirectly observed Poisson point process.*

Note that this definition is more general than the treatment above since here we need not assume that the two processes share the same state space or that they always have the same number of points. The processes are also only assumed to be related on the level of intensity functions and we need not necessarily assume a relation on the point level, such as the one given by Equation (2.4).

Since the intensity function fully characterizes a Poisson process, the obvious statistical inference problem related to indirectly observed Poisson processes is to ask what can we say about the intensity function f of the process of interest M given that we have only access to the indirect observations N. In the following subsection, the unfolding problem is formulated in terms of this framework.

2.1.4 Forward Model for Unfolding

In order to formulate the unfolding problem using indirectly observed Poisson point processes, we will need to generalize the treatment of the previous subsection to include the limited efficiency of the detector. Let the Poisson process of interest M be as above

$$M = \sum_{i=1}^{\tau} \delta_{\boldsymbol{X}_i}$$

with state space E and intensity function $f(\boldsymbol{x})$. Here the points $\boldsymbol{X}_i$ correspond to the true values of the physical observable of interest and τ is the total number of events in the data sample.

Due to limitations of detector technology, some of these events might be lost in a real-world detector. Let us thus accompany each $\boldsymbol{X}_i$ by an indicator variable $Z_i \in \{0, 1\}$. Having $Z_i = 1$ indicates that $\boldsymbol{X}_i$ is observed, while $Z_i = 0$ means that $\boldsymbol{X}_i$ is lost. Let $\varepsilon(\boldsymbol{x}) = P(Z = 1 | \boldsymbol{X} = \boldsymbol{x})$ be the *efficiency function* which should be understood to account for all kinds of losses incurred in the detector. These losses can range from a simple non-detection of a particle traversing the detector without interacting with the detection medium to the smearing of $\boldsymbol{X}_i$ to a value outside of the detectable space. Removal of the lost events gives us the thinned Poisson point process

$$M^* = \sum_{i=1}^{\tau} Z_i \delta_{\boldsymbol{X}_i}$$

with efficiency $\varepsilon(\boldsymbol{x})$ as the thinning function. Let us rewrite this as

$$M^* = \sum_{i=1}^{\zeta} \delta_{\boldsymbol{X}_i^*},$$

where $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_\zeta^*$ are the observed points out of the initial points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau$ and $\zeta = \sum_{i=1}^{\tau} Z_i$. By Proposition 2.9, the intensity function f^* of the thinned process M^* is

$$f^*(\boldsymbol{x}^*) = \varepsilon(\boldsymbol{x}^*) f(\boldsymbol{x}^*).$$

Let us then assume that the points $\boldsymbol{X}_i^*$ are smeared and let us denote the smeared observations by $\boldsymbol{Y}_i$. We assume that the points $\boldsymbol{Y}_i$ lie in the space F which is not necessarily equal to the original state space E. The observed smeared Poisson point process is then

$$N = \sum_{i=1}^{\zeta} \delta_{\boldsymbol{Y}_i}$$

with state space F and by following the same line of reasoning as in the previous subsection, we find that the intensity h of N is

$$\begin{aligned} h(\boldsymbol{y}) &= \int p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}^*}(\boldsymbol{y}) f^*(\boldsymbol{x}^*) \,\mathrm{d}\boldsymbol{x}^* \\ &= \int p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}^*}(\boldsymbol{y}) \varepsilon(\boldsymbol{x}^*) f(\boldsymbol{x}^*) \,\mathrm{d}\boldsymbol{x}^* \\ &= \int k(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{2.6}$$

where we have denoted $k(\boldsymbol{x}, \boldsymbol{y}) := p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}}(\boldsymbol{y})\varepsilon(\boldsymbol{x})$. Hence, according to Definition 2.10, M is an indirectly observed Poisson point process with observations N and smearing kernel k. We see that the effect of taking into account possible losses in the detector is that the efficiency $\varepsilon(\boldsymbol{x})$ appears in the kernel and we have $\int k(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \varepsilon(\boldsymbol{x}) \in [0, 1]$, $\forall \boldsymbol{x} \in E$ instead of the kernel integrating into unity over $\boldsymbol{y}$.

It is now easy to see the relation between indirectly observed Poisson processes and the high energy physics unfolding problem. The points $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_\zeta$ of N correspond to the smeared observations seen in the particle detector and the kernel k describes the noise, efficiency and other unwanted effects induced by the imperfect measurement device. Unfolding then corresponds to the inference of the intensity function f of the true physical process M of primary interest.

Let us note that in some cases, it might be sensible to perform a second thinning for the smeared Poisson process N using a *post-smearing efficiency function* $\varepsilon_{\mathrm{PS}}$. As we will later see in Section 7.2, this is for example the case when trigger prescaling needs to be accounted for since the trigger of the experiment naturally operates with the smeared measurements. This post-smearing thinning would give us a process N^* with the intensity function

$$h^*(\boldsymbol{y}^*) = \varepsilon_{\mathrm{PS}}(\boldsymbol{y}^*) \int p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}^*}(\boldsymbol{y}^*) \varepsilon(\boldsymbol{x}^*) f(\boldsymbol{x}^*) \,\mathrm{d}\boldsymbol{x}^*.$$

Denoting

$$C(\boldsymbol{x}^*) = \left(\int \varepsilon_{\mathrm{PS}}(\boldsymbol{y}^*) p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}^*}(\boldsymbol{y}^*) \,\mathrm{d}\boldsymbol{y}^* \right)^{-1},$$

we can write this intensity in the form

$$h^*(\boldsymbol{y}^*) = \int C(\boldsymbol{x}^*)\varepsilon_{\mathrm{PS}}(\boldsymbol{y}^*)p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}^*}(\boldsymbol{y}^*)\frac{\varepsilon(\boldsymbol{x}^*)}{C(\boldsymbol{x}^*)}f(\boldsymbol{x}^*)\,\mathrm{d}\boldsymbol{x}^*.$$

Comparing this with (2.6) shows that as a result of the second thinning, we end up with the same intensity function as the one we would have in the case we simply thinned the true process M with the thinning function $\varepsilon(\boldsymbol{x}^*)/C(\boldsymbol{x}^*)$ and then used $C(\boldsymbol{x}^*)\varepsilon_{\mathrm{PS}}(\boldsymbol{y}^*)p_{\boldsymbol{Y}|\boldsymbol{X}^*=\boldsymbol{x}^*}(\boldsymbol{y}^*)$ as the conditional probability of the smeared observations. Since, by Theorem 2.6, the intensity function fully characterizes the observed Poisson process N, we see that the model (2.6) with only a single thinning is general enough to cover also the case of post-smearing thinning.

2.1.5 Discretization

We now discretize the unfolding problem using histograms to estimate the intensity functions. To this end, assume that the spaces E and F are either the one-dimensional real line $\mathbb{R}$ or some intervals of the real line. Let $\mathcal{E} = \{E_1, \ldots, E_p\}$ and $\mathcal{F} = \{F_1, \ldots, F_q\}$ be sets of intervals that form partitions of E and F, respectively. The Poisson processes M and N then correspond to the random vectors

$$\begin{aligned}\boldsymbol{x} &= \big[M(E_1), \ldots, M(E_p)\big]^{\mathrm{T}},\\ \boldsymbol{y} &= \big[N(F_1), \ldots, N(F_q)\big]^{\mathrm{T}},\end{aligned}$$

where $\boldsymbol{x} \in \mathbb{N}_0^p$ represents the unobservable true histogram for binning $\mathcal{E}$ and $\boldsymbol{y} \in \mathbb{N}_0^q$ represents the observed smeared histogram for binning $\mathcal{F}$. Similarly, for mean measures, we have

$$\boldsymbol{\lambda} = \big[\lambda(E_1), \ldots, \lambda(E_p)\big]^{\mathrm{T}} = \Big[\textstyle\int_{E_1} f(x)\,\mathrm{d}x, \ldots, \int_{E_p} f(x)\,\mathrm{d}x\Big]^{\mathrm{T}}, \tag{2.7}$$

$$\boldsymbol{\mu} = \big[\mu(F_1), \ldots, \mu(F_q)\big]^{\mathrm{T}} = \Big[\textstyle\int_{F_1} h(y)\,\mathrm{d}y, \ldots, \int_{F_q} h(y)\,\mathrm{d}y\Big]^{\mathrm{T}}, \tag{2.8}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^p$ and $\boldsymbol{\mu} \in \mathbb{R}_+^q$ represent the means of the true histogram $\boldsymbol{x}$ and the smeared histogram $\boldsymbol{y}$, respectively. Note that $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ also serve as discrete approximations of the intensity functions f and h via the relations

$$\begin{aligned} f(x) &\approx \frac{\lambda_i}{\nu(E_i)}, \quad x \in E_i, \quad i = 1, \ldots, p,\\ h(y) &\approx \frac{\mu_i}{\nu(F_i)}, \quad y \in F_i, \quad i = 1, \ldots, q,\end{aligned} \tag{2.9}$$

where ν denotes the Lebesgue measure, i.e., the length of the bin E_i or F_i.

By Definition 2.5, we know that the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent and Poisson distributed

$$\begin{aligned}\boldsymbol{x}|\boldsymbol{\lambda} &\sim \mathrm{Poisson}(\boldsymbol{\lambda}), \quad \perp\!\!\!\perp x_i|\boldsymbol{\lambda},\\ \boldsymbol{y}|\boldsymbol{\mu} &\sim \mathrm{Poisson}(\boldsymbol{\mu}), \quad \perp\!\!\!\perp y_i|\boldsymbol{\mu}.\end{aligned}$$

To see how these two Poisson distributions are related, let us use Equation (2.6) to write

$$\begin{aligned}
\mu_i &= \int_{F_i} h(y)\,\mathrm{d}y \\
&= \int_{F_i}\int_{E} k(x,y) f(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{F_i}\left(\sum_{j=1}^{p}\int_{E_j} k(x,y) f(x)\,\mathrm{d}x\right)\,\mathrm{d}y \\
&= \sum_{j=1}^{p}\int_{F_i}\int_{E_j} k(x,y) f(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \sum_{j=1}^{p}\frac{\int_{F_i}\int_{E_j} k(x,y) f(x)\,\mathrm{d}x\,\mathrm{d}y}{\int_{E_j} f(x)\,\mathrm{d}x}\lambda_j \\
&= \sum_{j=1}^{p} K_{ij}\lambda_j, \quad i = 1,\ldots,q,
\end{aligned}$$

where

$$K_{ij} = \frac{\int_{F_i}\int_{E_j} k(x,y) f(x)\,\mathrm{d}x\,\mathrm{d}y}{\int_{E_j} f(x)\,\mathrm{d}x}, \quad i = 1,\ldots,q, \quad j = 1,\ldots,p \tag{2.10}$$

are the elements of the *smearing matrix* $\boldsymbol{K}$, which can be regarded as a discretized version of the smearing kernel k. Hence, we have the relation

$$\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$$

for the Poisson means $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$.

The following proposition shows that the elements K_{ij} of the smearing matrix correspond to the probability of observing an event in smeared bin F_i when it originates from the true bin E_j. Hence, they are the *migration probabilities* from the true bin E_j to the smeared bin F_i.

Proposition 2.11. *The elements K_{ij} of the smearing matrix defined by Equation* (2.10) *satisfy*

$$K_{ij} = P(Y \in F_i | X \in E_j),$$

where Y is a point of the smeared Poisson point process N and X the corresponding point of the true process M.

Proof. Using Z to indicate if X is observed, we have

$$\begin{aligned}
&P(Y \in F_i | X \in E_j) \\
&\quad = P(Y \in F_i, Z = 1 | X \in E_j) + P(Y \in F_i, Z = 0 | X \in E_j) \\
&\quad = P(Y \in F_i, Z = 1 | X \in E_j).
\end{aligned}$$

Here we can write

$$\begin{aligned} P(Y \in F_i, Z = 1 | X \in E_j) &= \frac{P(Y \in F_i, X \in E_j, Z = 1)}{P(X \in E_j)} \\ &= \frac{P(Y \in F_i, X \in E_j | Z = 1) P(Z = 1)}{P(X \in E_j)} \\ &= \frac{\int_{F_i} \int_{E_j} p_{X,Y|Z=1}(x, y) P(Z = 1) \, \mathrm{d}x \, \mathrm{d}y}{\int_{E_j} p_X(x) \, \mathrm{d}x} \end{aligned}$$

We can rewrite the integrand in the numerator as

$$\begin{aligned} p_{X,Y|Z=1}(x, y) P(Z = 1) &= p_{Y|X=x,Z=1}(y) p_{X|Z=1}(x) P(Z = 1) \\ &= p_{Y|X^*=x}(y) P(Z = 1 | X = x) p_X(x) \\ &= p_{Y|X^*=x}(y) \varepsilon(x) p_X(x) \\ &= k(x, y) p_X(x) \end{aligned}$$

and hence we get

$$\begin{aligned} P(Y \in F_i | X \in E_j) &= \frac{\int_{F_i} \int_{E_j} k(x, y) p_X(x) \, \mathrm{d}x \, \mathrm{d}y}{\int_{E_j} p_X(x) \, \mathrm{d}x} \\ &= \frac{\int_{F_i} \int_{E_j} k(x, y) f(x) \, \mathrm{d}x \, \mathrm{d}y}{\int_{E_j} f(x) \, \mathrm{d}x} \\ &= K_{ij}, \end{aligned}$$

where the second equality follows from Equation (2.3). □

Note that due to the efficiency $\varepsilon(\boldsymbol{x})$ it is possible to have $\sum_i K_{ij} < 1$. In fact, this sum is the efficiency ε_j of the true bin E_j since

$$\sum_i K_{ij} = \sum_i P(Y \in F_i | X \in E_j) = P(Y \in F | X \in E_j) := \varepsilon_j.$$

In the following, these efficiencies are collected to the efficiency vector $\boldsymbol{\varepsilon} = \left[\varepsilon_1, \ldots, \varepsilon_p\right]^{\mathrm{T}}$.

We see from Equation (2.10) that the smearing matrix $\boldsymbol{K}$ depends on the unknown intensity f and that the significance of this dependence increases with the size of the true bins E_j. For small enough binning, we can use some approximation of $\boldsymbol{K}$ to remove this dependence. In real physics analyses, $\boldsymbol{K}$ is determined using Monte Carlo simulations, in which case its computation is based on an MC approximation f^{MC} of f. In the numerical experiments of this thesis, we simulate this by using a slightly perturbed version of the true intensity f for determining $\boldsymbol{K}$. Alternatively, we can use Equation (2.9) to approximate

$$K_{ij} = \frac{\int_{F_i} \int_{E_j} k(x, y) f(x) \, \mathrm{d}x \, \mathrm{d}y}{\int_{E_j} f(x) \, \mathrm{d}x} \approx \frac{1}{\nu(E_j)} \int_{F_i} \int_{E_j} k(x, y) \, \mathrm{d}x \, \mathrm{d}y. \tag{2.11}$$

This approximation holds as an equality if the intensity f happens to be constant over the histogram bin E_j.

To summarize, in the discrete version of the unfolding problem, we observe the smeared histogram $\boldsymbol{y}$, which follows the Poisson distribution with parameter $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$, that is,

$$\boldsymbol{y}|\boldsymbol{\lambda} \sim \text{Poisson}(\boldsymbol{K}\boldsymbol{\lambda}), \quad \perp\!\!\!\perp y_i|\boldsymbol{\lambda}, \tag{2.12}$$

and our task is to infer the unknown Poisson means $\boldsymbol{\lambda}$ of the true histogram $\boldsymbol{x}$. These can, in turn, be used to construct a piecewise constant approximation of the intensity function f of the process of interest M using Equation (2.9).

2.2 An Alternative Formulation

In this section, we give an alternative formulation for the unfolding problem without using Poisson point processes. The formulation is less general than the one presented above as it applies only in the discrete case. On the other hand, the problem can be formulated as a simple hierarchical model and there is not need to resort to measure theory or integral equations. The key element of the formulation is the following lemma:

Lemma 2.12. *Let N and $\boldsymbol{X} = \left[X_1, \ldots, X_d\right]^{\mathrm{T}}$ be random variables with $N \sim \text{Poisson}(\lambda)$ and $\boldsymbol{X}|N = n \sim \text{Mult}(\boldsymbol{p}, n)$, where $\boldsymbol{p} = [p_1, \ldots, p_d]^{\mathrm{T}}$ is a vector of probabilities that sum up to one. Then the components X_i are independent and $X_i \sim \text{Poisson}(p_i\lambda)$, $i = 1, \ldots, d$.*

Proof. We have

$$\begin{aligned} p(\boldsymbol{X} = \boldsymbol{x}) &= \sum_{n=0}^{\infty} p(\boldsymbol{X} = \boldsymbol{x}|N = n)p(N = n) \\ &= \sum_{n=0}^{\infty} \frac{n!}{x_1! \cdots x_d!} p_1^{x_1} \cdots p_d^{x_d} 1_{\{\sum_i x_i = n\}}(\boldsymbol{x}) \frac{\lambda^n}{n!} e^{-\lambda} \\ &= e^{-\lambda} \prod_{i=1}^{d} \frac{p_i^{x_i}}{x_i!} \sum_{n=0}^{\infty} 1_{\{\sum_i x_i = n\}}(\boldsymbol{x}) \lambda^n. \end{aligned}$$

Here we have

$$\sum_{n=0}^{\infty} 1_{\{\sum_i x_i = n\}}(\boldsymbol{x}) \lambda^n = \lambda^{\sum_i x_i} = \prod_{i=1}^{d} \lambda^{x_i}$$

and

$$e^{-\lambda} = e^{-\lambda \sum_i p_i} = \prod_{i=1}^{d} e^{-p_i \lambda}.$$

Hence,

$$p(\boldsymbol{X} = \boldsymbol{x}) = \prod_{i=1}^{d} \frac{(p_i\lambda)^{x_i}}{x_i!} e^{-p_i\lambda} = \prod_{i=1}^{d} p(X_i = x_i),$$

from which we can see that the X_i are independent and Poisson distributed:

$$\perp\!\!\!\perp X_i \quad \text{and} \quad X_i \sim \text{Poisson}(p_i\lambda), \quad i = 1, \ldots, d. \qquad \square$$

Now, let us assume that we arrange the true event counts before smearing in the histogram $\boldsymbol{x} = \begin{bmatrix} x_1, \ldots, x_p \end{bmatrix}^{\mathrm{T}}$ corresponding to a partition $\mathcal{E} = \{E_1, \ldots, E_p\}$ of the real line or some interval of the real line. Similarly, the event counts after smearing are recorded in the histogram $\boldsymbol{y} = \begin{bmatrix} y_1, \ldots, y_q \end{bmatrix}^{\mathrm{T}}$ for the binning $\mathcal{F} = \{F_1, \ldots, F_q\}$. Note that we do not assume the binnings $\mathcal{E}$ and $\mathcal{F}$ to be equal or for the same intervals of the real line.

Let τ denote the total number of events in the true histogram $\boldsymbol{x}$, $\tau = \sum_i x_i$. We can regard the events forming the histogram as τ independent random trials with p possible outcomes corresponding to each of the histogram bins. Hence $\boldsymbol{x}|\tau$ follows a multinomial distribution. Since we know from the underlying physics that τ is Poisson distributed, we can use Lemma 2.12 to deduce that the bins x_i are independent and Poisson distributed with some parameter $\boldsymbol{\lambda}$

$$\boldsymbol{x}|\boldsymbol{\lambda} \sim \text{Poisson}(\boldsymbol{\lambda}), \quad \perp\!\!\!\perp x_i|\boldsymbol{\lambda}.$$

Let us then assume that an event belonging to the true bin E_i is observed with an efficiency ε_i. In other words, for each true bin E_i, the efficiency vector $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1, \ldots, \varepsilon_p \end{bmatrix}^{\mathrm{T}}$ consists of the probabilities of observing an event belonging to that bin. As above, the efficiencies are assumed to take into account all sources of losses incurred in the detector. Given the true histogram $\boldsymbol{x}$, we can then think of performing a Bernoulli trial for each event in the histogram with success probabilities ε_i with the index i chosen according to the bin of the event. We then collect the successful events into a new histogram $\boldsymbol{x}^*$ corresponding to the true histogram after taking into account the inefficiency of the detector. Since the binomial distribution of the Bernoulli trials is a special case of the multinomial distribution, we can employ Lemma 2.12 to deduce that

$$x_i^*|\lambda_i \sim \text{Poisson}(\varepsilon_i\lambda_i).$$

To show the independence of these histogram bins, we need to assume that the inefficiency is independent from one bin to another

$$p(\boldsymbol{x}^*|\boldsymbol{x}) = \prod_i p(x_i^*|x_i).$$

We then have

$$
\begin{aligned}
p(\boldsymbol{x}^*|\boldsymbol{\lambda}) &= \sum_{\boldsymbol{x}} p(\boldsymbol{x}^*|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\lambda}) \\
&= \sum_{x_1} \cdots \sum_{x_p} \prod_{i=1}^{p} p(x_i^*|x_i)p(x_i|\lambda_i) \\
&= \sum_{x_1} \cdots \sum_{x_{p-1}} \prod_{i=1}^{p-1} p(x_i^*|x_i)p(x_i|\lambda_i) \sum_{x_p} p(x_p^*|x_p)p(x_p|\lambda_p) \\
&= \sum_{x_1} \cdots \sum_{x_{p-1}} \prod_{i=1}^{p-1} p(x_i^*|x_i)p(x_i|\lambda_i)p(x_p^*|\lambda_p) \\
&= p(x_p^*|\lambda_p) \sum_{x_1} \cdots \sum_{x_{p-1}} \prod_{i=1}^{p-1} p(x_i^*|x_i)p(x_i|\lambda_i) \\
&= p(x_p^*|\lambda_p)p(x_{p-1}^*|\lambda_{p-1}) \sum_{x_1} \cdots \sum_{x_{p-2}} \prod_{i=1}^{p-2} p(x_i^*|x_i)p(x_i|\lambda_i) \\
&= \ldots = \prod_{i=1}^{p} p(x_i^*|\lambda_i).
\end{aligned}
\tag{2.13}
$$

Hence, we see that the histogram bins x_i^* are conditionally independent given $\boldsymbol{\lambda}$, $\perp\!\!\!\perp x_i^*|\boldsymbol{\lambda}$.

We then proceed to form the smeared matrix $\boldsymbol{y}$. To this end, let us introduce the migration probabilities

$$p_{ij} = P(\text{event in bin } F_i \text{ of } \boldsymbol{y}|\text{event in bin } E_j \text{ of } \boldsymbol{x}^*).$$

These are the probabilities of observing the smeared event in bin F_i given that the corresponding true event was in bin E_j of the histogram $\boldsymbol{x}^*$.

Let us now denote by z_{ij} the number of events that are observed in the smeared bin F_i and originate from the true bin E_j of $\boldsymbol{x}^*$. We have

$$\boldsymbol{z}_j|x_j^* \sim \text{Mult}(\boldsymbol{p}_j, x_j^*),$$

where $\boldsymbol{z}_j = \left[z_{1j}, \ldots, z_{qj}\right]^{\mathrm{T}}$ and $\boldsymbol{p}_j = \left[p_{1j}, \ldots, p_{qj}\right]^{\mathrm{T}}$. Using Lemma 2.12, we get

$$z_{ij}|\lambda_j \sim \text{Poisson}(p_{ij}\varepsilon_j\lambda_j). \tag{2.14}$$

For the independence, we need to again assume that the smearing is independent from one bin to another

$$p(\boldsymbol{Z}|\boldsymbol{x}) = \prod_j p(\boldsymbol{z}_j|x_j),$$

where $\boldsymbol{Z} = (z_{ij})$, which gives us

$$p(\boldsymbol{Z}|\boldsymbol{\lambda}) = \prod_j p(\boldsymbol{z}_j|\lambda_j) = \prod_{ij} p(z_{ij}|\lambda_j),$$

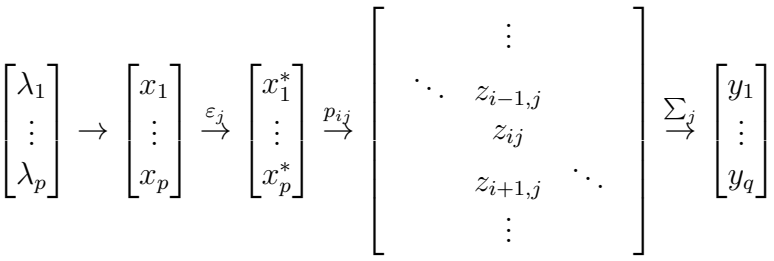


Figure 2.1: Generative forward model for unfolding. The Poisson means $\boldsymbol{\lambda}$ generate the true histogram $\boldsymbol{x}$. The histogram $\boldsymbol{x}^*$ is generated when some of these events are lost with probability $1 - \varepsilon_j$ due to the inefficiency of the detector. These events migrate with probability p_{ij} from the jth true bin to the ith smeared bin and the corresponding smeared event counts are given by z_{ij}. Finally, the row sums of the $\boldsymbol{Z}$ matrix yield the observed event counts of the smeared matrix $\boldsymbol{y}$.

where the first equality follows from a similar line of reasoning as above in Equation (2.13) and the second equality follows again from Lemma 2.12. Hence, we have $\perp\!\!\!\perp z_{ij}|\boldsymbol{\lambda}$.

The observed event counts y_i of the smeared histogram are given by the row sums of the random matrix $\boldsymbol{Z}$

$$y_i = \sum_j z_{ij}.$$

Since the z_{ij} are independent and Poisson distributed, we have

$$y_i|\boldsymbol{\lambda} \sim \text{Poisson}\Big(\sum_j p_{ij}\varepsilon_j\lambda_j\Big), \quad \perp\!\!\!\perp y_i|\boldsymbol{\lambda}.$$

When we denote here $K_{ij} := p_{ij}\varepsilon_j$, we have shown for the smeared histogram $\boldsymbol{y}$ the following:

$$\boldsymbol{y}|\boldsymbol{\lambda} \sim \text{Poisson}(\boldsymbol{K}\boldsymbol{\lambda}), \quad \perp\!\!\!\perp y_i|\boldsymbol{\lambda}, \tag{2.15}$$

where we call $\boldsymbol{K}$ the *smearing matrix*. In what follows, we will occasionally use $\boldsymbol{\mu}$ to denote the mean of the smeared histogram $\boldsymbol{y}$, i.e., $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$. This generative forward model for unfolding is illustrated Figure 2.1.

Since

$$\begin{aligned} P(Y \in F_i|X \in E_j) &= P(Y \in F_i|X^* \in E_j)P(X^* \in E_j|X \in E_j) \\ &= p_{ij}\varepsilon_j = K_{ij}, \end{aligned}$$

where X, X^* and Y denote events belonging to histograms $\boldsymbol{x}$, $\boldsymbol{x}^*$ and $\boldsymbol{y}$, respectively, we see from Proposition 2.11 that the definition of the smearing matrix $\boldsymbol{K}$ coincides with the definition given earlier in Section 2.1.5. Hence, Equation (2.10) gives an analytic expression for the elements K_{ij}. As above, some form of approximation is needed in determining $\boldsymbol{K}$ since the elements K_{ij} depend on the unknown distribution of events within the true bins E_j.

To summarize, we have shown two ways of deriving the same forward model for the discrete version of the unfolding problem. These probabilistic forward models are given be Equations (2.12) and (2.15). Given this model, the unfolding task can then be formulated as follows:

> Given the smeared observations $\boldsymbol{y}$ following the model (2.12) (or equivalently (2.15)), what can be said about the means $\boldsymbol{\lambda}$ of the true histogram $\boldsymbol{x}$?

The rest of this thesis is concerned with computational techniques for providing a solution to this statistical inference problem.

Chapter 3

Inference for Direct Observations

Before discussing unfolding, we will, in this chapter, explain the inference of the Poisson means for the case of direct observations, that is, the smearing matrix $\boldsymbol{K} = \boldsymbol{I}$ in Equation (2.12). Hence, the statistical model is

$$\boldsymbol{y}|\boldsymbol{\lambda} \sim \text{Poisson}(\boldsymbol{\lambda}), \quad \perp\!\!\!\perp y_i|\boldsymbol{\lambda} \tag{3.1}$$

and our task is to infer the mean vector $\boldsymbol{\lambda}$. This is a well-understood, routine problem in experimental high energy physics, at least as long as no underlying structure connecting the means λ_i is assumed and hence the bins can be treated separately. For an overview of the techniques used in HEP for inference under Poisson statistics, see e.g. [12].

We first provide a point estimator of $\boldsymbol{\lambda}$ via maximum likelihood in Section 3.1. We then explain the standard procedure for computing the confidence intervals of the solution in Section 3.2 followed by the corresponding Bayesian treatment in Section 3.3. In Section 3.4, we make some brief remarks about situations where $\boldsymbol{\lambda}$ is assumed to vary smoothly from one bin to another.

3.1 Maximum Likelihood Solution

The likelihood of the parameter $\boldsymbol{\lambda} \in \mathbb{R}_+^p$ in model (3.1) is

$$L(\boldsymbol{\lambda}) = p(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_i \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} = \prod_i p(y_i|\lambda_i) = \prod_i L(\lambda_i). \tag{3.2}$$

Since the full likelihood factorizes with respect to $\boldsymbol{\lambda}$, we can maximize each of the likelihoods $L(\lambda_i)$ separately. Setting the derivative to zero, we find

$$L'(\lambda_i) = \frac{y_i \lambda_i^{y_i - 1}}{y_i!} e^{-\lambda_i} - \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} = 0 \quad \Rightarrow \quad \lambda_i = y_i.$$

Hence, the MLE for $\boldsymbol{\lambda}$ is $\hat{\boldsymbol{\lambda}} = \boldsymbol{y}$. This estimator is unbiased

$$\mathrm{E}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}] = \mathrm{E}[\boldsymbol{y}|\boldsymbol{\lambda}] = \boldsymbol{\lambda}$$

and has the covariance

$$\mathrm{Cov}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}] = \mathrm{Cov}[\boldsymbol{y}|\boldsymbol{\lambda}] = \mathrm{diag}(\lambda_1, \ldots, \lambda_p) = \mathrm{diag}(\boldsymbol{\lambda}).$$

Plugging the MLE for $\boldsymbol{\lambda}$ in the equation above, we get an estimator of the covariance of $\hat{\boldsymbol{\lambda}}$

$$\widehat{\mathrm{Cov}}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}] = \mathrm{diag}(\hat{\boldsymbol{\lambda}}) = \mathrm{diag}(\boldsymbol{y}). \tag{3.3}$$

Hence, the estimated standard deviation of the MLE is

$$\widehat{\mathrm{Std}}[\hat{\lambda}_i|\lambda_i] = \sqrt{\widehat{\mathrm{Cov}}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}]_{ii}} = \sqrt{y_i}.$$

If we agree to use the estimated standard deviation to quantify the uncertainty of the inference, we could report the outcome of the measurement in the form $\hat{\lambda}_i \pm \widehat{\mathrm{Std}}[\hat{\lambda}_i|\lambda_i]$. Hence, for the MLE in the case of Poisson statistics, we would report that the mean of the ith bin is $y_i \pm \sqrt{y_i}$. In a graphical representation, this would give symmetric error bars of total length $2\sqrt{y_i}$ around the estimated mean y_i. These are often referred to as the $\sqrt{n}$ error bars, where n denotes the number of observations in the bin.

3.2 Frequentist Confidence Intervals

The rationale behind reporting $y_i \pm \sqrt{y_i}$ as the outcome of the measurement is that asymptotically the distribution of the MLE $\hat{\lambda}_i$ tends to a Gaussian with mean λ_i and standard deviation $\sqrt{\lambda_i}$ and hence for each bin E_i, this corresponds to the 68.27 % asymptotic confidence interval for the mean λ_i of the bin. The problem is that this coverage probability does not necessarily hold for finite sample sizes. Fortunately, there is a rather simple way of computing the central confidence interval for λ_i with guaranteed finite-sample coverage. While this confidence interval was first derived by Garwood [22], we follow here the more accessible modern presentation by Cowan [13, p. 126].

The central confidence interval $[a_i, b_i]$ for λ_i at confidence level $1 - \alpha$ can be constructed by solving for a_i and b_i in the following equations:

$$\frac{\alpha}{2} = \int_{\hat{\lambda}_i}^{\infty} \mathrm{d}P_{\hat{\lambda}_i|\lambda_i=a_i},$$
$$\frac{\alpha}{2} = \int_{-\infty}^{\hat{\lambda}_i} \mathrm{d}P_{\hat{\lambda}_i|\lambda_i=b_i}.$$

In other words, we are looking for a lower limit a_i (an upper limit b_i) with the property that if the true value λ_i equals a_i (b_i), then the probability of getting the observed value of the estimator $\hat{\lambda}_i$ or a value greater (smaller) than this is $\alpha/2$. For the case of Poisson observations and the estimator $\hat{\lambda}_i = y_i$, these equations become

$$\frac{\alpha}{2} = \int_{y_i}^{\infty} \mathrm{d}P_{y_i|\lambda_i=a_i} = \sum_{k=y_i}^{\infty} \frac{a_i^k}{k!} e^{-a_i},$$
$$\frac{\alpha}{2} = \int_{0}^{y_i} \mathrm{d}P_{y_i|\lambda_i=b_i} = \sum_{k=0}^{y_i} \frac{b_i^k}{k!} e^{-b_i}.$$

One can show that the solution of these equations is given by

$$a_i = \frac{1}{2} F_{\chi^2}^{-1}\left(\frac{\alpha}{2}\middle|\, 2y_i\right), \tag{3.4}$$

$$b_i = \frac{1}{2} F_{\chi^2}^{-1}\left(1-\frac{\alpha}{2}\middle|\, 2(y_i+1)\right), \tag{3.5}$$

where $F_{\chi^2}^{-1}(\cdot|k)$ denotes the inverse of the cdf of the χ^2 distribution with k degrees of freedom[1].

It is known that the resulting random interval $[a_i, b_i] = [a_i(\hat{\lambda}_i), b_i(\hat{\lambda}_i)]$ obtained by this construction satisfies the coverage property

$$P\left(a_i(\hat{\lambda}_i) \leq \lambda_i \leq b_i(\hat{\lambda}_i)\middle|\, \lambda_i\right) \geq 1-\alpha, \quad \forall \lambda_i > 0.$$

This means that the confidence interval $[a_i, b_i]$ is guaranteed to satisfy the minimum coverage of $1-\alpha$ with possible overcoverage for some values of λ_i. In fact, one can further show, that the minimum coverage is attained in the asymptotic limit $\lambda_i \to \infty$ and that the interval $[a_i, b_i]$ is conservative (i.e. it overcovers λ_i) for any finite true value λ_i. Due to the discrete nature of the Poisson distribution, it is not possible to construct confidence intervals for λ_i with exact coverage. If one requires a minimum coverage of $1-\alpha$, there will always be overcoverage for some true values of λ_i, while the alternative requirement for mean coverage of $1-\alpha$ would result in undercoverage for some values of λ_i. For a coverage plot of the central confidence interval $[a_i, b_i]$, see [27, p. 13].

When confidence intervals are used to report the outcome of a HEP experiment, the standard convention is to report the 68.27 % confidence intervals[2] obtained by setting $\alpha = 1 - 0.6827 = 0.3173$. The result of the experiment is then usually expressed in the form $y_{i-c_i}^{+d_i}$, where $c_i = y_i - a_i$ and $d_i = b_i - y_i$ and y_i is the MLE of λ_i. In graphical form, the outcome would be expressed as asymmetric error bars ranging from a_i to b_i with the point estimate at y_i. By a simple computational experiment, it is easy to verify that with small y_i the these error bars are significantly distinct from the naïve $\pm\sqrt{y_i}$ symmetric error bars, but when the number of observations y_i tends to infinity, the error bars become increasingly symmetric and converge to the $\pm\sqrt{y_i}$ errors, as one would expect based on the discussion above on asymptotics.

3.3 Bayesian Credible Intervals

We now proceed to find Bayesian credible intervals for the means $\boldsymbol{\lambda}$ in model (3.1). Using Bayes' theorem (A.8), we can write the posterior of $\boldsymbol{\lambda}$ as

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}}, \tag{3.6}$$

[1] Note that Equation (3.4) cannot be used for setting the lower limit when we have zero observed counts, $y_i = 0$. In this case, the lower limit is set to $a_i = 0$ and analogously with Equation (3.5), we can obtain the upper bound b_i at confidence level $1-\alpha$ from $b_i = \frac{1}{2} F_{\chi^2}^{-1}\left(1-\alpha|\, 2(y_i+1)\right)$.

[2] These are also called 1σ confidence intervals since 0.6827 corresponds to the probability mass contained within $\mu \pm 1\sigma$ of a Gaussian pdf with mean μ and standard deviation σ.

where the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$ is given by Equation (3.2). Assuming prior independence of the means, i.e., $p(\boldsymbol{\lambda}) = \prod_i p(\lambda_i)$, the posterior factorizes with respect to $\boldsymbol{\lambda}$

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \prod_i \frac{p(y_i|\lambda_i)p(\lambda_i)}{\int p(y_i|\lambda_i)p(\lambda_i)\,\mathrm{d}\lambda_i} = \prod_i p(\lambda_i|y_i).$$

Hence, the λ_i are independent in the posterior and we can perform the inference individually for each of them. We see that the posterior of λ_i is proportional to the likelihood times the prior

$$p(\lambda_i|y_i) \propto p(y_i|\lambda_i)p(\lambda_i).$$

Assuming the uniform non-negativity prior, $p(\lambda_i) \propto 1_{[0,\infty)}(\lambda_i)$, we can write this as

$$p(\lambda_i|y_i) \propto p(y_i|\lambda_i)1_{[0,\infty)}(\lambda_i).$$

Writing out the normalization coefficient, we have

$$\begin{aligned} p(\lambda_i|y_i) &= \frac{p(y_i|\lambda_i)}{\int_0^\infty p(y_i|\lambda_i)\,\mathrm{d}\lambda_i} 1_{[0,\infty)}(\lambda_i) \\ &= \frac{\lambda_i^{y_i} e^{-\lambda_i}}{\int_0^\infty \lambda_i^{y_i} e^{-\lambda_i}\,\mathrm{d}\lambda_i} 1_{[0,\infty)}(\lambda_i) \\ &= \frac{\lambda_i^{y_i} e^{-\lambda_i}}{\Gamma(y_i+1)} 1_{[0,\infty)}(\lambda_i), \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. Via a change of variables $z_i = 2\lambda_i$, we find

$$p(z_i|y_i) = \frac{1}{2^{y_i+1}\Gamma(y_i+1)} z_i^{y_i} e^{-z_i/2} 1_{[0,\infty)}(z_i),$$

which is the χ^2 distribution with $2(y_i+1)$ degrees of freedom

$$z_i|y_i = 2\lambda_i|y_i \sim \chi^2(2(y_i+1)). \tag{3.7}$$

The $100(1-\alpha)\%$ Bayesian central credible interval $[a_i, b_i]$ can then be found as a solution of the following two equations:

$$\begin{aligned} \frac{\alpha}{2} &= \int_0^{a_i} p(\lambda_i|y_i)\,\mathrm{d}\lambda_i = \int_0^{2a_i} p(z_i|y_i)\,\mathrm{d}\lambda_i = F_{\chi^2}(2a_i|2(y_i+1)), \\ \frac{\alpha}{2} &= \int_{b_i}^{\infty} p(\lambda_i|y_i)\,\mathrm{d}\lambda_i = \int_{2b_i}^{\infty} p(z_i|y_i)\,\mathrm{d}\lambda_i = 1 - F_{\chi^2}(2b_i|2(y_i+1)), \end{aligned}$$

where $F_{\chi^2}(\cdot|k)$ is the cdf of the χ^2 distribution with k degrees of freedom. Hence, we have

$$a_i = \frac{1}{2} F_{\chi^2}^{-1}\left(\frac{\alpha}{2}\middle|\, 2(y_i+1)\right), \tag{3.8}$$

$$b_i = \frac{1}{2} F_{\chi^2}^{-1}\left(1-\frac{\alpha}{2}\middle|\, 2(y_i+1)\right). \tag{3.9}$$

Comparison of these Bayesian limits to the corresponding frequentist limits given in Equations (3.4) and (3.5), shows that the resulting upper limits b_i are equal, but the lower limits are different. The frequentist lower limit is computed with $2y_i$ degrees of freedom, while the Bayesian limit uses $2(y_i+1)$ degrees of freedom corresponding to the frequentist limit with one more observation. Hence, the Bayesian credible intervals are are always shorter than the corresponding frequentist confidence intervals.

The Bayesian credible intervals of Equations (3.8) and (3.9) were obtained using the prior $p(\lambda_i) \propto 1_{[0,\infty)}(\lambda_i)$, which is the uniform prior on the non-negative real axis and represents the choice of an *uninformative* prior for λ_i. The first disconcerting feature of this choice is that $p(\lambda_i)$ cannot be normalized to be a pdf. However, the posterior (3.7) turns out to be a valid density function, which is often the case with such *improper* priors, and hence this is not a major concern. A more significant issue with the uniform prior is the fact that it is not invariant under nonlinear changes of variables. That is, the distribution of $g(\lambda_i)$, where g is some nonlinear function, is not in general the uniform distribution. Hence, $p(\lambda_i)$ is uninformative for λ_i but informative for $g(\lambda_i)$. Because of this complication, the choice of an uninformative prior is not unambiguous and one can come up with various "uninformative" priors depending on which metric one decides to be uninformative. For example, $p(\lambda_i) \propto \frac{1}{\lambda_i^2} 1_{[0,\infty)}(\lambda_i)$ would be uniform for $\frac{1}{\lambda_i}$, while $p(\lambda_i) \propto \frac{1}{\lambda_i} 1_{[0,\infty)}(\lambda_i)$ would be uniform for $\log \lambda_i$. Yet another widely-used prior is $p(\lambda_i) \propto \frac{1}{\sqrt{\lambda_i}} 1_{[0,\infty)}(\lambda_i)$, which is the so-called Jeffreys prior [30] for the case of Poisson observations. Out of the various possible options, we will mostly be using $p(\lambda_i) \propto 1_{[0,\infty)}(\lambda_i)$ as the uninformative prior, but it is important to keep in mind that this is essentially just a convenient arbitrary choice.

When comparing the frequentist confidence intervals (Equations (3.4) and (3.5)) and the Bayesian credible intervals (Equations (3.8) and (3.9)), it is also important to keep in mind that these results describe two fundamentally different things. In the frequentist paradigm, the parameter λ_i is a non-negative real number with some fixed true value and not a random variable. The $100(1-\alpha)\%$ frequentist confidence interval is a random interval which covers this true value in at least $100(1-\alpha)\%$ of the cases when the experiment is repeated infinitely many times. On the other hand, in the Bayesian paradigm, λ_i is a random variable and the posterior describes our degree of belief about its true value encoded in the form of a probability density. The $100(1-\alpha)\%$ Bayesian credible interval then represents the interval where we expect to find the true value with a probability of $1-\alpha$ given the observation y_i and our prior beliefs $p(\lambda_i)$.

3.4 Smoothing

We have so far considered inference of the means λ_i in cases where absolutely nothing is known about them in advance and we have shown that in such a case one can perform the inference individually for each of the bins. It is, however, often the case that one does have more information at hand about the structure of the intensity function $f(x)$ of the underlying Poisson process. In some cases, there are theoretical

justifications for a certain parametric family of intensity functions $f(x) = f(x|\boldsymbol{\theta})$ and the task is then to infer the parameters $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$ by, e.g., maximum likelihood estimation. One could also make a more general nonparametric statement about the expected characteristics of the intensity. We could for example argue that physically plausible intensities should be smooth functions. In the discrete case, the corresponding statement is that the finite differences for the vector $\boldsymbol{\lambda}$ should be small. We now proceed to show how to incorporate such a criterion in both the frequentist maximum likelihood paradigm and the Bayesian framework.

We showed in Section 3.1 that the maximum likelihood solution to the direct inference problem is $\hat{\boldsymbol{\lambda}} = \boldsymbol{y}$, which can alternatively also be regarded as the solution of the least squares problem

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \|\boldsymbol{\lambda} - \boldsymbol{y}\|^2. \tag{3.10}$$

To impose smoothness of the solution, we penalize for the finite differences of the solution and hence consider the following optimization problem:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left(\|\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \delta \|\boldsymbol{L}\boldsymbol{\lambda}\|^2 \right), \tag{3.11}$$

where $\boldsymbol{L}$ is the finite-difference matrix of desired order and $\delta > 0$ a suitably chosen *smoothing parameter*. The larger the value of δ, the smoother solutions we obtain, while $\delta \to 0$ yields the maximum likelihood solution.

When we move from Equation (3.10) to (3.11), we are no longer able to consider each λ_i separately since the neighboring values are connected via the matrix $\boldsymbol{L}$. The important consequence of this is that the variance of each estimated mean $\hat{\lambda}_i$ is reduced from the nominal value λ_i at the expense of having a biased estimator. The reason for this is that one is able to use observations from the neighboring bins in addition to the observation y_i from the current bin to estimate the mean λ_i which reduces the uncertainty of the estimate. Since Equation (3.11) is a special case of Tikhonov regularization for unfolding, we postpone a more thorough analysis of the solution to Section 4.2.2.

In the Bayesian paradigm, one can impose the smoothness of the solution by choosing the prior distribution $p(\boldsymbol{\lambda})$ in (3.6) appropriately. One possibility is to use the truncated multivariate Gaussian distribution

$$p(\boldsymbol{\lambda}) \propto 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \exp\left(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2\right),$$

where $\boldsymbol{L}$ is again a finite-difference matrix and the hyperparameter α controls the strength of this *Gaussian smoothness prior*. The resulting posterior is

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) = 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \left(\prod_i \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} \right) \exp\left(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2\right), \tag{3.12}$$

which does not factorize with respect to $\boldsymbol{\lambda}$ because of the matrix $\boldsymbol{L}$. Hence, the λ_i are not independent in the posterior and one needs to resort to more advanced techniques than above in order to find the Bayesian credible intervals for λ_i. We discuss methods for exploring posteriors of the type of (3.12) as well as techniques for selecting the hyperparameter α later in Chapters 5 and 6 in the context of Bayesian and empirical Bayes unfolding.

Chapter 4

Frequentist Unfolding Techniques

This chapter provides an overview of the unfolding techniques provided by the frequentist paradigm of statistics. We tackle the problem of estimating $\boldsymbol{\lambda}$ in model (2.12) using two different approaches. In Section 4.1, we study maximum likelihood estimation of $\boldsymbol{\lambda}$. It turns out that there is no closed form expression for the MLE and we will have to resort to expectation-maximization algorithm with early stopping in order to solve and regularize the problem. In Section 4.2, the unfolding problem is formulated as a least squares optimization problem. In this case, the problem can be regularized using the truncated singular value decomposition or Tikhonov regularization. Although these point estimates for $\boldsymbol{\lambda}$ are both conceptually and computationally simple, their major limitation is that there is no straightforward way of estimating the uncertainty of the solutions in a non-asymptotic way. We conclude this chapter by making brief remarks about the optimal choice of the regularization strength of the frequentist unfolding algorithms in Section 4.3.

4.1 Maximum Likelihood Estimation

The likelihood of the mean of the true histogram $\boldsymbol{\lambda}$ in model (2.12) is

$$L(\boldsymbol{\lambda}) = p(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_i \frac{\left(\sum_j K_{ij}\lambda_j\right)^{y_i}}{y_i!} e^{-\sum_j K_{ij}\lambda_j}, \quad \boldsymbol{\lambda} \in \mathbb{R}^p_+. \tag{4.1}$$

We can equivalently parametrize the model using the mean of the smeared histogram $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ in which case the likelihood is

$$L(\boldsymbol{\mu}) = p(\boldsymbol{y}|\boldsymbol{\mu}) = \prod_i \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}, \quad \boldsymbol{\mu} \in \mathbb{R}^q_+.$$

Let us first study the identifiability of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. This is established in the following theorem:

Theorem 4.1. *In model* (2.12)*,* $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ *is always identifiable and* $\boldsymbol{\lambda}$ *is identifiable if and only if* $\boldsymbol{K} \in \mathbb{R}^{q\times p}$ *has full column rank (i.e.* $\operatorname{rank}(\boldsymbol{K}) = p$*).*

Proof. Let us first consider the situation with $q = 1$, i.e., the smeared histogram has only a single bin. Then for $\mu_1, \mu_1' > 0$ we have:

$$\begin{aligned}
& p(y_1|\mu_1) = p(y_1|\mu_1'), \quad \forall y_1 \in \mathbb{N}_0 \\
\Leftrightarrow \quad & \frac{\mu_1^{y_1}}{y_1!} e^{-\mu_1} = \frac{{\mu_1'}^{y_1}}{y_1!} e^{-\mu_1'} \\
\Leftrightarrow \quad & \mu_1^{y_1} e^{-\mu_1} = {\mu_1'}^{y_1} e^{-\mu_1'} \\
\Rightarrow \quad & \begin{cases} \mu_1 e^{-\mu_1} = \mu_1' e^{-\mu_1'} \\ \mu_1^2 e^{-\mu_1} = \mu_1'^2 e^{-\mu_1'} \end{cases} \\
\Rightarrow \quad & \frac{\mu_1^2 e^{-\mu_1}}{\mu_1 e^{-\mu_1}} = \frac{\mu_1'^2 e^{-\mu_1'}}{\mu_1' e^{-\mu_1'}} \\
\Leftrightarrow \quad & \mu_1 = \mu_1'
\end{aligned}$$

Clearly the implication also holds if $\mu_1 = 0$ or $\mu_1' = 0$. Hence μ_1 is identifiable. When $q > 1$, the ith marginal $p(y_i|\mu_i)$ only depends on μ_i. Since $p(\boldsymbol{y}|\boldsymbol{\mu}) = p(\boldsymbol{y}|\boldsymbol{\mu}')$, $\forall \boldsymbol{y}$ implies that the marginals have to be equal, i.e., $p(y_i|\mu_i) = p(y_i|\mu_i')$, $\forall y_i \in \mathbb{N}_0$, $i = 1, \ldots, q$, we can use the argument above to deduce that $\boldsymbol{\mu} = \boldsymbol{\mu}'$ and thus $\boldsymbol{\mu}$ is identifiable.

Let us then study the identifiability of $\boldsymbol{\lambda}$. Using the identifiability of $\boldsymbol{\mu}$ we have:

$$\begin{aligned}
& p(\boldsymbol{y}|\boldsymbol{\lambda}) = p(\boldsymbol{y}|\boldsymbol{\lambda}') \\
\Leftrightarrow \quad & \boldsymbol{K}\boldsymbol{\lambda} = \boldsymbol{K}\boldsymbol{\lambda}' \\
\Leftrightarrow \quad & \boldsymbol{K}(\boldsymbol{\lambda} - \boldsymbol{\lambda}') = \boldsymbol{0} \\
\Leftrightarrow \quad & \boldsymbol{\lambda} - \boldsymbol{\lambda}' \in \ker(\boldsymbol{K}) \\
\Leftrightarrow \quad & \boldsymbol{\lambda} = \boldsymbol{\lambda}' + \boldsymbol{w}, \ \boldsymbol{w} \in \ker(\boldsymbol{K})
\end{aligned} \tag{4.2}$$

Hence we see that $\boldsymbol{\lambda}$ is identifiable if and only if $\ker(\boldsymbol{K}) = \{\boldsymbol{0}\}$ or equivalently $\boldsymbol{K}$ has full column rank. □

We see from Equation (4.2) that if $\boldsymbol{K}$ does not have full column rank, all the elements of the set

$$S_{\boldsymbol{\lambda}} = \{\boldsymbol{\lambda} + \boldsymbol{w} : \boldsymbol{w} \in \ker(\boldsymbol{K})\} \cap \mathbb{R}_+^p \tag{4.3}$$

will generate smeared histograms $\boldsymbol{y}$ following the same distribution as the histograms generated by the true parameter $\boldsymbol{\lambda}$. This means that using only the data $\boldsymbol{y}$, there is no way to discriminate between the elements of this set and without additional information about plausible solutions, the best we can hope to achieve in unfolding is to identify the correct set $S_{\boldsymbol{\lambda}}$. However, using additional constraints related for instance to the smoothness of the solution, it is possible to remove such an ambiguity as will be demonstrated in several examples throughout this chapter.

Let us then consider the problem of finding the maximum likelihood estimator of $\boldsymbol{\lambda}$. Hence, our task is to maximize (4.1) subject to the non-negativity constraint

$\lambda_j \geq 0,\ j = 1, \ldots, p$. Equivalently, we could try to find the maximum of the log-likelihood function

$$\begin{aligned} l(\boldsymbol{\lambda}) = \log p(\boldsymbol{y}|\boldsymbol{\lambda}) &= \sum_i \left[y_i \log \left(\sum_j K_{ij}\lambda_j \right) - \sum_j K_{ij}\lambda_j \right] + \text{const} \\ &= \sum_i \Big(y_i \log \mu_i - \mu_i \Big) + \text{const}, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$ is the mean of the smeared histogram. Hence, the maximum likelihood estimator $\hat{\boldsymbol{\lambda}}_{\text{MLE}}$ is the solution to the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & l(\boldsymbol{\lambda}) = \sum_i \left[y_i \log \left(\sum_j K_{ij}\lambda_j \right) - \sum_j K_{ij}\lambda_j \right] + \text{const} \\ \text{subject to:} \quad & \boldsymbol{\lambda} \geq \mathbf{0} \end{aligned} \tag{4.5}$$

It turns out that this optimization problem is not as innocuous as it would seem on the first glance. We start our analysis with the following theorem which shows that we will not have to worry about local maxima when maximizing (4.4). Furthermore, when $\boldsymbol{\lambda}$ is identifiable, the maximum is unique. These results were first shown in [56].

Theorem 4.2. *Assume $K_{ij} > 0$ and $\boldsymbol{y} \neq \mathbf{0}$. Then the following hold for the log-likelihood $l : \mathbb{R}^p_+ \to \mathbb{R} \cup \{-\infty\}$ defined by Equation (4.4):*

(i) The log-likelihood has a maximum.

(ii) The log-likelihood is concave and hence all the maxima are global maxima.

(iii) The maximum is unique if and only if $\boldsymbol{K}$ has full column rank.

Proof.

(i) $l(\boldsymbol{\lambda})$ is continuous in $\mathbb{R}^p_+$ except for the singularity at $\boldsymbol{\lambda} = \mathbf{0}$. But $\lim_{\boldsymbol{\lambda}\to\mathbf{0}} l(\boldsymbol{\lambda}) = -\infty$ and hence the singularity will not cause problems for the existence of a maximum. Furthermore, $l(\boldsymbol{\lambda}) \to -\infty$ when we take any number of λ_j's to infinity. Hence, we conclude that a maximum exists.

(ii) The first and second derivative of the log-likelihood are

$$\begin{aligned} \frac{\partial l}{\partial \lambda_j} &= \sum_i \frac{\partial l}{\partial \mu_i}\frac{\partial \mu_i}{\partial \lambda_j} = \sum_i K_{ij}\left(\frac{y_i}{\mu_i} - 1\right) \\ \frac{\partial^2 l}{\partial \lambda_j \partial \lambda_k} &= \sum_l \left(\frac{\partial}{\partial \mu_l}\frac{\partial l}{\partial \lambda_j} \right)\frac{\partial \mu_l}{\partial \lambda_k} = -\sum_l K_{lj}K_{lk}\frac{y_l}{\mu_l^2}. \end{aligned}$$

From this, we find that the quadratic form of the Hessian matrix is negative semidefinite

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w} &= \sum_j \sum_k \frac{\partial^2 l}{\partial\lambda_j \partial\lambda_k} w_j w_k \\ &= -\sum_j \sum_k \sum_l K_{lj} K_{lk} \frac{y_l}{\mu_l^2} w_j w_k \\ &= -\sum_l \frac{y_l}{\mu_l^2} \left(\sum_j K_{lj} w_j\right)\left(\sum_k K_{lk} w_k\right) \\ &= -\sum_l \frac{y_l}{\mu_l^2} \widetilde{w}_l^2 \leq 0, \end{aligned} \tag{4.6}$$

where $\boldsymbol{w} \in \mathbb{R}^p \backslash \{\boldsymbol{0}\}$ is arbitrary and we have denoted $\widetilde{w}_l = \sum_j K_{lj} w_j$. It follows that $l(\boldsymbol{\lambda})$ is concave.

(iii) $l(\boldsymbol{\lambda})$ has a unique maximum if it is strictly concave or equivalently its Hessian is negative definite. We have an equality in (4.6) for some $\boldsymbol{w} \in \mathbb{R}^p \backslash \{\boldsymbol{0}\}$ iff $\sum_j K_{lj} w_j = 0,\ \forall l$ or equivalently $\boldsymbol{k}_l \perp \boldsymbol{w},\ \forall l$, where $\boldsymbol{k}_l = \left[K_{l1}, \ldots, K_{lp}\right]^{\mathrm{T}}$ is the transpose of the lth row of $\boldsymbol{K}$. This is equivalent to $\operatorname{span}(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_q) \neq \mathbb{R}^p$ and consequently to $\operatorname{rank}(\boldsymbol{K}) < p$. Hence, the quadratic form is negative definite iff $\operatorname{rank}(\boldsymbol{K}) = p$. □

Theorem 4.2 shows that the MLE always exists but need not be unique. However, in same cases, it should be possible to circumvent this issue by reducing the number of bins p of the true histogram until $\boldsymbol{K}$ has full column rank. A more serious practical issue is that it is not possible to express the MLE in a closed form. The reason for this is that we are dealing with a nonlinear constrained optimization problem. Such problems often have to be solved using various numerical techniques. To see the extent of the problem, assume that $\boldsymbol{K}$ is an invertible square matrix. By $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$, we then have a one-to-one correspondence between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$. From Section 3.1, we know that the MLE of $\boldsymbol{\mu}$ is $\boldsymbol{y}$ and hence it would seem that we could find the MLE of $\boldsymbol{\lambda}$ as the solution of

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{\lambda}. \tag{4.7}$$

The problem is that there is no guarantee that $\boldsymbol{y} \in \operatorname{ran}(\boldsymbol{K})_+$, where $\operatorname{ran}(\boldsymbol{K})_+ = \{\boldsymbol{K}\boldsymbol{\lambda} : \boldsymbol{\lambda} \geq \boldsymbol{0}\}$. Hence, inversion of (4.7) could give us a solution which does not satisfy the non-negativity constraint of (4.5). Luckily efficient numerical techniques have been developed for finding the MLE and we will discuss below in Sections 4.1.1 and 4.1.2 one possible solution based on the expectation-maximization algorithm with guaranteed convergence to the non-negative MLE.

The last and most fundamental problem with the MLE is nevertheless due to the ill-posedness of the unfolding problem. Namely, it turns out that in practice the maximum likelihood solutions are highly oscillating and often unusable. This is due to the very high variance of the MLE which can be reduced using some form of regularization. When iterative algorithms are used it is customary to impose

regularization by stopping the iteration before oscillations start to appear. Another option would be to add a penalty term $G(\boldsymbol{\lambda})$ with regularization parameter $\delta > 0$ to the likelihood

$$\widetilde{L}(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda}) - \delta G(\boldsymbol{\lambda})$$

and then maximize $\widetilde{L}(\boldsymbol{\lambda})$ subject to $\boldsymbol{\lambda} \geq 0$. This is called *penalized maximum likelihood estimation*.

4.1.1 The Expectation-Maximization Algorithm

We now describe the *expectation-maximization* (EM) *algorithm* for finding the MLE of the parameters $\boldsymbol{\theta}$ of a parametric model $p(\boldsymbol{y}|\boldsymbol{\theta})$ for the data $\boldsymbol{y}$. We then show in the next subsection how the algorithm can be used for finding the MLE of the unfolding problem. The EM algorithm is a versatile iterative technique for finding the MLE in problems where the observations $\boldsymbol{y}$ can be regarded as incomplete in a sense that will be made precise below. The algorithm has been discovered independently in various forms on several different fields of science but it was the famous paper by Dempster, Laird and Rubin [19] in 1977 that made the EM algorithm popular by presenting the algorithm in its general, widely applicable form and by establishing the main theoretical properties of the algorithm. We will follow in our treatment the book by McLachnan and Krishnan [43] which is a comprehensive, modern reference on the EM algorithm and various related computational methods.

Assume that we have observed the random variable $\boldsymbol{y}$ and we know that its distribution depends on some parameters $\boldsymbol{\theta}$. We would then like to find the MLE of $\boldsymbol{\theta}$ by maximizing the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ but find that for numerical or algorithmic reasons this is difficult to accomplish. Assume then that we can regard $\boldsymbol{y}$ as an incomplete version of some *complete data* random variable $\boldsymbol{x}$ with density $p(\boldsymbol{x}|\boldsymbol{\theta})$. That is, $\boldsymbol{y} = g(\boldsymbol{x})$ for some many-to-one function g and the incomplete-data likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ is given by

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\{\boldsymbol{x}:\boldsymbol{y}=g(\boldsymbol{x})\}} p(\boldsymbol{x}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} = \int_{g^{-1}(\boldsymbol{y})} p(\boldsymbol{x}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} = \int_{g^{-1}(\boldsymbol{y})} L(\boldsymbol{\theta}; \boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

where $g^{-1}(\boldsymbol{y})$ is the preimage of $\boldsymbol{y}$ and $L(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})$ is the complete-data likelihood. It is often the case that we have $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$ with g the projection to the $\boldsymbol{y}$-component of $\boldsymbol{x}$ and $\boldsymbol{z}$ some unobserved *latent variables*. In this case, the incomplete-data likelihood is simply given by

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z} = \int L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{4.8}$$

The power of the EM algorithm lies on the fact that in many cases the complete data $\boldsymbol{x}$ can be chosen in such a way that the complete-data likelihood $L(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})$, or equivalently the complete-data log-likelihood $l(\boldsymbol{\theta}; \boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{\theta})$, can be easily maximized. The EM algorithm then exploits this to indirectly find the maximum of the original incomplete-data likelihood $L(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$.

It follows that we would like to maximize $l(\boldsymbol{\theta}; \boldsymbol{x})$, but since the complete data $\boldsymbol{x}$ cannot be observed, we compute the expectation of $l(\boldsymbol{\theta}; \boldsymbol{x})$ over the unobservable parts of $\boldsymbol{x}$ given the observations $\boldsymbol{y}$ and the current value of $\boldsymbol{\theta}$. That is, in the *expectation step* (or E-step) of the algorithm, we compute

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \mathrm{E}\big[l(\boldsymbol{\theta}; \boldsymbol{x}) | \boldsymbol{y}, \boldsymbol{\theta}^{(k)}\big] = \mathrm{E}\big[\log p(\boldsymbol{x}|\boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta}^{(k)}\big],$$

where $\boldsymbol{\theta}^{(k)}$ is the current value of the parameters $\boldsymbol{\theta}^{(k)}$. In the special case of $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$, this expectation is given by

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) &= \mathrm{E}\big[l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) | \boldsymbol{y}, \boldsymbol{\theta}^{(k)}\big] \\ &= \mathrm{E}\big[\log p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta}^{(k)}\big] \\ &= \int p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}^{(k)}) \log p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{z} \\ &= \int \frac{p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}^{(k)})}{p(\boldsymbol{y}|\boldsymbol{\theta}^{(k)})} \log p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{z}. \end{aligned}$$

On the subsequent *maximization step* (or M-step) of the algorithm, the parameters for the next iteration are found as the maximizer of this expected complete-data log-likelihood with respect to $\boldsymbol{\theta}$, that is

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}).$$

In essence, this algorithm uses the current values of the parameters $\boldsymbol{\theta}$ to fix the values of the unknown parts of the complete data $\boldsymbol{x}$ and then uses this estimate to update the values of the parameters $\boldsymbol{\theta}$.

To summarize, the EM algorithm for finding the maximum of the incomplete-data likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ and hence the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ of $\boldsymbol{\theta}$ is given by the following iteration:

1. Pick some initial guess $\boldsymbol{\theta}^{(0)}$ and set $k = 0$.
2. E-step: Compute $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \mathrm{E}\big[\log p(\boldsymbol{x}|\boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta}^{(k)}\big]$, where $\boldsymbol{x}$ is the complete data.
3. M-step: Set $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.
4. Set $k \leftarrow k + 1$.
5. If some stopping rule $\mathcal{C}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k-1)}, \ldots, \boldsymbol{\theta}^{(0)})$ is satisfied, set $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \boldsymbol{\theta}^{(k)}$ and terminate the iteration, else go to step 2.

If on some iteration the maximum of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ in the M-step of the algorithm is not unique, it suffices to choose any of the $\boldsymbol{\theta}$'s that maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ as the next iterate. The stopping rule is often selected to be based on how much the parameter $\boldsymbol{\theta}$ or the incomplete-data log-likelihood $l(\boldsymbol{\theta}; \boldsymbol{y})$ changed on the last iteration. One could

for example say that the algorithm has converged when $l(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}) - l(\boldsymbol{\theta}^{(k-1)}; \boldsymbol{y}) < \varepsilon$ for some constant $\varepsilon > 0$.

It was proved by Dempster, Laird and Rubin [19] that after each EM iteration, the original incomplete-data likelihood $L(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ is either increased or remains at its current value.

Theorem 4.3. *The EM sequence of likelihoods* $\{L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})\}$ *increases monotonically, that is,* $L(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{y}) \geq L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ *for all* $k = 0, 1, 2, \ldots$

Proof. See [19, Theorem 1] or alternatively [43, Section 3.2]. □

Corollary 4.4. *If the EM sequence of likelihoods* $\{L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})\}$ *is bounded, it converges monotonically to some value* L^**, that is* $L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}) \uparrow L^* \in \mathbb{R}_+$.

Proof. The claim follows directly from the monotonicity of the EM sequence of likelihoods $\{L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})\}$. □

Hence we know that for bounded likelihoods $L(\boldsymbol{\theta}; \boldsymbol{y})$, the likelihood sequence $\{L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})\}$ converges. However, this does not imply the existence of a point $\boldsymbol{\theta}^*$ such that $L^* = L(\boldsymbol{\theta}^*)$, nor does it imply that L^* would be the maximum of $L(\boldsymbol{\theta}; \boldsymbol{y})$. However, under weak regularity conditions, it can be shown that $L(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}) \uparrow L(\boldsymbol{\theta}^*; \boldsymbol{y})$, where $\boldsymbol{\theta}^*$ is a stationary point of the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$. Furthermore, in many cases, it can also be shown that the iterates $\boldsymbol{\theta}^{(k)}$ converge to a stationary point $\boldsymbol{\theta}^*$ of the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$. [61]

In most practical applications, the more serious convergence issue with the EM algorithm is that there are often no guarantees for the stationary point $\boldsymbol{\theta}^*$ to be the global maximum of the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$. In particular, convergence to a local maximum of the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ is often a serious issue with the EM iteration. Fortunately, Theorem 4.2 tells us that all the maxima of the likelihood of the unfolding problem are global and hence we need not worry about local maxima when using the EM algorithm for unfolding.

4.1.2 Unfolding with the EM Algorithm

We now describe the use of the EM algorithm for finding the maximum of the log-likelihood $l(\boldsymbol{\lambda}; \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\lambda})$ given by Equation (4.4). These results were discovered independently by Shepp and Vardi [53] and Lange and Carson [37] in the beginning of the 1980s in the context of image reconstruction in positron emission tomography. This work was later extended by Vardi, Shepp and Kaufman in the subsequent paper [56]. Our treatment here follows the presentation of [43, Section 2.5].

The natural choice for the complete data of this problem are the counts z_{ij} of Equation (2.14). Recall from Section 2.2 that the random variable z_{ij} represents the number of events originating from the jth true bin and observed in the ith smeared bin. Hence, the incomplete data $\boldsymbol{y}$ are related to the complete data $\boldsymbol{Z} = (z_{ij})$ via the row sums

$$\boldsymbol{y} = g(\boldsymbol{Z}) = \left[\textstyle\sum_j z_{1j}, \ldots, \sum_j z_{pj}\right]^{\mathrm{T}}.$$

We also know from Section 2.2 that the random variable z_{ij} are Poisson distributed and conditionally independent

$$z_{ij}|\lambda_j \sim \text{Poisson}(K_{ij}\lambda_j), \quad \perp\!\!\!\perp z_{ij}|\boldsymbol{\lambda}.$$

Hence, the complete-data log-likelihood $l(\boldsymbol{\lambda}; \boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{\lambda})$ is given by

$$l(\boldsymbol{\lambda}; \boldsymbol{Z}) = \sum_{ij} \Big(z_{ij} \log(K_{ij}\lambda_j) - \log z_{ij}! - K_{ij}\lambda_j \Big).$$

On the E-step of the EM algorithm, we compute the conditional expectation of the complete-data log-likelihood $l(\boldsymbol{\lambda}; \boldsymbol{Z})$ given the observations $\boldsymbol{y}$ and the current value $\boldsymbol{\lambda}^{(k)}$ of the unknown parameter $\boldsymbol{\lambda}$

$$Q(\boldsymbol{\lambda}; \boldsymbol{\lambda}^{(k)}) = \mathrm{E}\big[l(\boldsymbol{\lambda}; \boldsymbol{Z})|\boldsymbol{y}, \boldsymbol{\lambda}^{(k)}\big] \propto \sum_{ij} \Big(\log(K_{ij}\lambda_j)\mathrm{E}\big[z_{ij}|y_i, \boldsymbol{\lambda}^{(k)}\big] - K_{ij}\lambda_j \Big). \quad (4.9)$$

To find the conditional distribution $p(z_{ij}|y_i, \boldsymbol{\lambda}^{(k)})$ of z_{ij}, we note that we have y_i observations for the ith smeared bin and these observations can either originate from the jth true bin or some other true bin besides the jth bin. Hence, we can regard z_{ij} as indicating the number of successes in y_i Bernoulli trials with success probability $K_{ij}\lambda_j^{(k)} / \left(\sum_l K_{il}\lambda_l^{(k)}\right)$. It follows that

$$z_{ij}|y_i, \boldsymbol{\lambda}^{(k)} \sim \text{Bin}\left(\frac{K_{ij}\lambda_j^{(k)}}{\sum_l K_{il}\lambda_l^{(k)}}, y_i \right).$$

Hence, in Equation (4.9),

$$\mathrm{E}\big[z_{ij}|y_i, \boldsymbol{\lambda}^{(k)}\big] = \frac{K_{ij}\lambda_j^{(k)}}{\sum_l K_{il}\lambda_l^{(k)}} y_i$$

and we need to find the maximum of

$$\widetilde{Q}(\boldsymbol{\lambda}; \boldsymbol{\lambda}^{(k)}) = \sum_{ij} \left(\frac{K_{ij}\lambda_j^{(k)}}{\sum_l K_{il}\lambda_l^{(k)}} \log(K_{ij}\lambda_j) y_i - K_{ij}\lambda_j \right)$$

with respect to $\boldsymbol{\lambda}$ on the M-step of the algorithm.

To find the maximum of $\widetilde{Q}$, we set the derivative with respect to the mth component to zero

$$\frac{\partial}{\partial \lambda_m} \widetilde{Q}(\boldsymbol{\lambda}; \boldsymbol{\lambda}^{(k)}) = \sum_i \left(\frac{K_{im}\lambda_m^{(k)}}{\sum_l K_{il}\lambda_l^{(k)}} \frac{y_i}{\lambda_m} - K_{im} \right) = 0.$$

Solving this for λ_m and replacing m with j leads us to update λ_j on the M-step using

$$\lambda_j^{(k+1)} = \frac{\lambda_j^{(k)}}{\sum_i K_{ij}} \sum_i \frac{K_{ij} y_i}{\sum_l K_{il}\lambda_l^{(k)}}.$$

Hence, the EM iteration for finding the MLE of $\boldsymbol{\lambda}$ in model (2.15) is simply given by the following iteration:

1. Pick some initial guess $\boldsymbol{\lambda}^{(0)} > \mathbf{0}$ and set $k = 0$.

2. Compute:
$$\lambda_j^{(k+1)} = \frac{\lambda_j^{(k)}}{\sum_i K_{ij}} \sum_i \frac{K_{ij} y_i}{\sum_l K_{il} \lambda_l^{(k)}}, \quad j = 1, \ldots, p. \tag{4.10}$$

3. Set $k \leftarrow k + 1$.

4. If some stopping rule $\mathcal{C}(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^{(k-1)}, \ldots, \boldsymbol{\lambda}^{(0)})$ is satisfied, set $\hat{\boldsymbol{\lambda}}_{\text{MLE}} = \boldsymbol{\lambda}^{(k)}$ and terminate the iteration, else go to step 2.

If we now compare the iteration outlined above to the one derived by D'Agostini in [16], we find that the D'Agostini iteration is equal to the EM iteration. This is a quite remarkable coincidence given that D'Agostini derives his iteration by repeatedly applying Bayes' theorem to infer the values of the bin means λ_j. This leads him to call the iteration "Bayesian" unfolding, although we now see that it is merely a frequentist iteration for the MLE. Some authors have previously identified the equivalence of the EM and D'Agostini iterations (see, e.g., [55, Section 1]), but it appears that this finding has not received much attention in the HEP community. In fact, as early as in the 1980s right after the publication of the seminal paper [56] by Vardi, Shepp and Kaufman, some preliminary work was carried out at CERN on applying the EM iteration to the HEP unfolding problem [48], but for some reason the iteration only came to be used in real physics analyses as a result of D'Agostini's work. In addition, it turns out that the same iteration is also used in optics and astronomy for removing optical distortions from images [50, 40]. In these fields, the algorithm goes by the name of *Lucy–Richardson deconvolution*.

The EM iteration for unfolding has a number of convenient properties. Firstly, we see that given that each component of the first iterate $\boldsymbol{\lambda}^{(0)}$ is strictly positive, the iteration (4.10) will always produce strictly positive solutions. Hence, the non-negativity constraint of the maximum likelihood problem (4.5) is satisfied for the maximum likelihood estimator $\hat{\boldsymbol{\lambda}}_{\text{MLE}}$ produced by the iteration. When it comes to the convergence of the iteration, it is quite interesting to note that the standard EM convergence theorems of Wu [61] are not applicable here. Nevertheless, the convergence of the resulting EM sequence of iterates $\{\boldsymbol{\lambda}^{(k)}\}$ to a maximum of the log-likelihood $l(\boldsymbol{\lambda}; \boldsymbol{y})$ was shown by Vardi, Shepp and Kaufman in [56, Theorem A.1].

As noted earlier, the ill-posedness of the problem means that we are, in fact, in most cases, not interested in the maximum likelihood solution of the unfolding problem since it often exhibits large oscillations due to the large variance of the estimator. However, when we are using an iterative method for finding the MLE, it is straightforward to regularize the solution by stopping the iteration prematurely before oscillations start to appear. Hence, the number of EM iterations performed controls the strength of the regularization and the sooner the iteration is terminated,

the more we regularize the problem. For a discussion on various alternative stopping rules for the iteration, see Section 4.3.

We conclude this section by noting that there is one major drawback with the EM algorithm when used in HEP applications. Namely, since the iteration (4.10) is nonlinear in $\boldsymbol{\lambda}$, we cannot use the methods outlined in Section 4.2.3 to estimate the standard deviations of the resulting estimator and hence to quantify the uncertainty of the solution. Nevertheless, when the EM iteration is allowed to converge to the MLE, standard, well-understood methods exist for estimating the standard deviations of the solution, see [43, Chapter 4]. The problem is that it is not immediately clear how these methods should be adapted to the case where the iteration is stopped prematurely to avoid oscillating solutions. On the other hand, in the HEP literature, D'Agostini [16] provides a way of estimating the errors of the solution, but his calculations have been criticized by Adye [1]. For the time being, it appears that the error estimation of the EM iteration remains an issue that that has not yet been completely settled.

4.2 Least Squares Estimation

Let us study method of moments (MoM) estimation of $\boldsymbol{\lambda}$. The expectation of the smeared data $\boldsymbol{y}$ is given by

$$\mathrm{E}[\boldsymbol{y}|\boldsymbol{\lambda}] = \boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}.$$

Since we have only one observation of the smeared histogram, we use it as an estimator of the mean. Equating the sample mean $\boldsymbol{y}$ with the theoretical mean, we have

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{\lambda}. \tag{4.11}$$

The solution of this equation would then give us the method of moments estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{MoM}}$. Unfortunately, there are no guarantees about the existence or uniqueness of the solution of this linear system of equation. When $\boldsymbol{K}$ is row-rank deficient (i.e. not surjective), it could be that $\boldsymbol{y} \notin \mathrm{ran}(\boldsymbol{K})$ in which case the MoM estimator does not exist. On the other hand, when $\boldsymbol{K}$ is column-rank deficient (i.e. not injective), when it exists, the solution would not be unique.

To find an estimator that always exists, let us consider the least squares solution $\boldsymbol{\lambda}_{\mathrm{LS}}$ to Equation (4.11) defined by

$$\|\boldsymbol{K}\boldsymbol{\lambda}_{\mathrm{LS}} - \boldsymbol{y}\|^2 = \min_{\boldsymbol{\lambda}\in\mathbb{R}^p} \|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2. \tag{4.12}$$

The following theorem establishes the existence of the least squares solution $\boldsymbol{\lambda}_{\mathrm{LS}}$ and gives an explicit formula for the solution using the Moore–Penrose pseudoinverse.

Theorem 4.5. *The least squares solution $\boldsymbol{\lambda}_{\mathrm{LS}}$ in (4.12) always exists, but is not necessarily unique, and all the solutions are given by*

$$\boldsymbol{\lambda}_{\mathrm{LS}} = \boldsymbol{K}^\dagger \boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{K}^\dagger \boldsymbol{K})\boldsymbol{w}, \quad \boldsymbol{w} \in \mathbb{R}^p, \tag{4.13}$$

where $\boldsymbol{K}^\dagger$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{K}$.

Proof. The existence and nonuniqueness of the least squares solution $\boldsymbol{\lambda}_{\text{LS}}$ is a well-known property of the least squares problem (4.12). For proof, see e.g. [38, Theorem 2.3]. For proof[1] of Equation (4.13), see [4, Corollary 3.1]. □

The following corollary shows that when $\boldsymbol{\lambda}$ is identifiable, the least squares solution is unique.

Corollary 4.6. *If $\boldsymbol{K}$ has full column rank, the least squares solution $\boldsymbol{\lambda}_{\text{LS}}$ is unique and given by*

$$\boldsymbol{\lambda}_{\text{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y} = (\boldsymbol{K}^{\text{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\text{T}}\boldsymbol{y}.$$

Proof. When $\boldsymbol{K}$ has full column rank, the pseudoinverse is given by

$$\boldsymbol{K}^{\dagger} = (\boldsymbol{K}^{\text{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\text{T}}.$$

Substituting this in (4.13), gives us

$$\boldsymbol{\lambda}_{\text{LS}} = (\boldsymbol{K}^{\text{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\text{T}}\boldsymbol{y} + (\boldsymbol{I} - (\boldsymbol{K}^{\text{T}}\boldsymbol{K})^{-1}(\boldsymbol{K}^{\text{T}}\boldsymbol{K}))\boldsymbol{w} = (\boldsymbol{K}^{\text{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\text{T}}\boldsymbol{y}. \qquad \square$$

In cases where $\boldsymbol{K}$ is column-rank deficient, we can still get around the nonuniqueness of the least squares solution by picking the solution with the smallest norm.

Theorem 4.7. *Let S be the set of all solutions to the least squares problem* (4.12)*. Then the problem*

$$\min_{\boldsymbol{\lambda}_{\text{LS}} \in S} \|\boldsymbol{\lambda}_{\text{LS}}\| \tag{4.14}$$

has a unique solution $\boldsymbol{\lambda}^{*}_{\text{LS}}$ given by $\boldsymbol{\lambda}^{*}_{\text{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y}$.

Proof. To prove this, we use the *closest-point theorem* from convex analysis [3, Theorem 2.4.1] which, when applied to our case, states that if S is a closed convex set, then there exists a unique point $\boldsymbol{\lambda}^{*}_{\text{LS}} \in S$ with minimum norm and $\boldsymbol{\lambda}^{*}_{\text{LS}} \in S$ is this minimizing point if and only if

$$(\boldsymbol{\lambda}^{*}_{\text{LS}})^{\text{T}}(\boldsymbol{\lambda}_{\text{LS}} - \boldsymbol{\lambda}^{*}_{\text{LS}}) \geq 0, \quad \forall \boldsymbol{\lambda}_{\text{LS}} \in S. \tag{4.15}$$

Using Equation (4.13), we see that

$$S = \{\boldsymbol{K}^{\dagger}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{K}^{\dagger}\boldsymbol{K})\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^{p}\}.$$

Since $\boldsymbol{I} - \boldsymbol{K}^{\dagger}\boldsymbol{K}$ is the orthogonal projection onto $\ker(\boldsymbol{K})$, we can equivalently write this as

$$S = \{\boldsymbol{K}^{\dagger}\boldsymbol{y} + \boldsymbol{v} : \boldsymbol{v} \in \ker(\boldsymbol{K})\}.$$

Since $\ker(\boldsymbol{K})$ is a linear subspace of $\mathbb{R}^{p}$, is it closed and convex. It follows that also S is closed and convex and problem (4.14) has a unique solution. Clearly,

[1] The formula shown in [4, Corollary 3.1] is in fact for a more general inverse of $\boldsymbol{K}$ but includes the Moore–Penrose pseudoinverse $\boldsymbol{K}^{\dagger}$ as a special case.

$\boldsymbol{\lambda}^*_{\mathrm{LS}} = \boldsymbol{K}^\dagger \boldsymbol{y} \in S$ and hence is the solution to (4.14) if it satisfies the condition (4.15). We have:

$$\begin{aligned} & (\boldsymbol{\lambda}^*_{\mathrm{LS}})^{\mathrm{T}}(\boldsymbol{\lambda}_{\mathrm{LS}} - \boldsymbol{\lambda}^*_{\mathrm{LS}}) \geq 0, \quad \forall \boldsymbol{\lambda}_{\mathrm{LS}} \in S \\ \Leftrightarrow \quad & (\boldsymbol{K}^\dagger \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{K}^\dagger \boldsymbol{y} + \boldsymbol{v} - \boldsymbol{K}^\dagger \boldsymbol{y}) \geq 0, \quad \forall \boldsymbol{v} \in \ker(\boldsymbol{K}) \\ \Leftrightarrow \quad & (\boldsymbol{K}^\dagger \boldsymbol{y})^{\mathrm{T}} \boldsymbol{v} \geq 0, \quad \forall \boldsymbol{v} \in \ker(\boldsymbol{K}) \end{aligned} \tag{4.16}$$

Since $\boldsymbol{K}^\dagger \boldsymbol{y} \in \mathrm{ran}(\boldsymbol{K}^\dagger) = \mathrm{ran}(\boldsymbol{K}^{\mathrm{T}}) = \ker(\boldsymbol{K})^\perp$, we have $(\boldsymbol{K}^\dagger \boldsymbol{y})^{\mathrm{T}} \boldsymbol{v} = 0$, $\forall \boldsymbol{v} \in \ker(\boldsymbol{K})$. Hence, (4.16) is a true statement and it follows that the condition (4.15) is satisfied. □

We call this minimum norm least squares solution $\boldsymbol{\lambda}^*_{\mathrm{LS}}$ the *least squares estimator* $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ of $\boldsymbol{\lambda}$. Hence, the least squares estimator is given by

$$\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} = \boldsymbol{\lambda}^*_{\mathrm{LS}} = \boldsymbol{K}^\dagger \boldsymbol{y}.$$

The least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ has several attractive features. Firstly, it is simple to implement provided that the computational platform in use provides a ready-made implementation for the pseudoinverse. Secondly, when $\boldsymbol{\lambda}$ is identifiable, $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ is unbiased.

Proposition 4.8. *When $\boldsymbol{K}$ has full column rank, the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} = \boldsymbol{K}^\dagger \boldsymbol{y}$ is unbiased.*

Proof. Using a similar line of reasoning as in the proof of Corollary 4.6, we find

$$\mathrm{E}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} | \, \boldsymbol{\lambda}] = \boldsymbol{K}^\dagger \mathrm{E}[\boldsymbol{y} | \boldsymbol{\lambda}] = \boldsymbol{K}^\dagger \boldsymbol{K} \boldsymbol{\lambda} = (\boldsymbol{K}^{\mathrm{T}} \boldsymbol{K})^{-1} (\boldsymbol{K}^{\mathrm{T}} \boldsymbol{K}) \boldsymbol{\lambda} = \boldsymbol{\lambda}$$

and hence $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ is unbiased. □

On the other hand, when $\boldsymbol{K}$ is column-rank deficient, we have

$$\mathrm{E}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} | \, \boldsymbol{\lambda}] = \boldsymbol{K}^\dagger \boldsymbol{K} \boldsymbol{\lambda} = \boldsymbol{\lambda} - (\boldsymbol{I} - \boldsymbol{K}^\dagger \boldsymbol{K}) \boldsymbol{\lambda} = \boldsymbol{\lambda} - \boldsymbol{P} \boldsymbol{\lambda}, \tag{4.17}$$

where $\boldsymbol{P} : \mathbb{R}^p \to \ker(\boldsymbol{K})$ is the orthogonal projection onto the kernel of $\boldsymbol{K}$. Comparison of this with the discussion on identifiability of $\boldsymbol{\lambda}$ in Section 4.1 shows that the potential bias in $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ is an inevitable consequence of the unidentifiability of $\boldsymbol{\lambda}$. Nevertheless, it could happen that $\mathrm{E}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} | \, \boldsymbol{\lambda}] \notin S_{\boldsymbol{\lambda}}$ as defined by Equation (4.3) since there are no guarantees about the non-negativity of the expectation. This reveals the first major problem with the least squares estimator. Namely, it could, and in practice often does, give solutions with negative values.

As with all naïve estimators of $\boldsymbol{\lambda}$, the second major problem with $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ is the extremely large variance of the estimator caused by the ill-posedness of the unfolding problem. Using Equation (A.6), the covariance of $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ can be written as

$$\mathrm{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} | \, \boldsymbol{\lambda}] = \boldsymbol{K}^\dagger \, \mathrm{Cov}[\boldsymbol{y} | \boldsymbol{\lambda}] \, (\boldsymbol{K}^\dagger)^{\mathrm{T}}. \tag{4.18}$$

Here $\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{\lambda}] = \mathrm{diag}(\boldsymbol{\mu}) = \mathrm{diag}(\boldsymbol{K}\boldsymbol{\lambda})$. Especially in cases, where the condition number $\mathrm{cond}(\boldsymbol{K})$ is large, the diagonal elements of $\mathrm{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}|\,\boldsymbol{\lambda}]$ can become significantly larger than the diagonal elements of $\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{\lambda}]$. Hence, unfolding of data $\boldsymbol{y}$ with reasonable errors can lead to huge variations in the unfolded spectrum if the plain least squares estimator is used. Note also that in general $\mathrm{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}|\,\boldsymbol{\lambda}]$ is not diagonal meaning that the components of the least squares estimator are correlated even though the bins of both the true histogram $\boldsymbol{x}$ and the smeared histogram $\boldsymbol{y}$ are independent.

4.2.1 Truncated Singular Value Decomposition

We now turn our attention into showing how the variance of the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ can be reduced by using regularization. Using Theorem A.28, the singular value decomposition of the smearing matrix $\boldsymbol{K} \in \mathbb{R}^{q\times p}$ can be written as

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{4.19}$$

where $\boldsymbol{U} \in \mathbb{R}^{q\times q}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times p}$ is a diagonal matrix with the non-negative singular values σ_i on the diagonal. Let $r = \mathrm{rank}(\boldsymbol{K})$, $1 \leq r \leq \min(p, q)$. Consequently, by Proposition A.29, r coincides with the number of strictly positive singular values of $\boldsymbol{K}$, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min(p,q)} = 0$. By Proposition A.32, we can use the singular value decomposition (4.19) to write the pseudoinverse of $\boldsymbol{K}$ as

$$\boldsymbol{K}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{U}^{\mathrm{T}}, \tag{4.20}$$

where $\boldsymbol{\Sigma}^{\dagger} \in \mathbb{R}^{p\times q}$ is the pseudoinverse of $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma}^{\dagger} = \mathrm{diag}(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)_{p\times q}.$$

Writing out the matrix product in Equation (4.20), the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ can be written as

$$\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y} = \sum_{i=1}^{r} \frac{1}{\sigma_i}(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y})\boldsymbol{v}_i, \tag{4.21}$$

where $\boldsymbol{u}_i$ is the ith column of $\boldsymbol{U}$ and $\boldsymbol{v}_i$ the ith column of $\boldsymbol{V}$. From this form, it is intuitively clear what is the source of the large oscillations observed with the least squares estimator. Namely, the factor $1/\sigma_i$ significantly amplifies fluctuations in $\boldsymbol{y}$ if the corresponding singular value σ_i is very small. Hence, we can expect trouble in cases where the smallest singular values of $\boldsymbol{K}$ are close to zero which is, more often than not, the case with real-world smearing matrices $\boldsymbol{K}$.

Equation (4.21) immediately suggest a possible way of taming these oscillations by simply ignoring the problematic terms in the sum. To this end, let us introduce a *truncation index* t, $t \leq r$, and truncate the sum in (4.21) above this index. We call the resulting estimator

$$\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}} = \sum_{i=1}^{t} \frac{1}{\sigma_i}(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y})\boldsymbol{v}_i \tag{4.22}$$

the *truncated singular value decomposition* (TSVD) *estimator* of $\boldsymbol{\lambda}$ [31]. Letting

$$\boldsymbol{K}_t^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}_t^\dagger \boldsymbol{U}^{\mathrm{T}},$$

where $\boldsymbol{\Sigma}_t^\dagger = \mathrm{diag}(1/\sigma_1, \ldots, 1/\sigma_t, 0, \ldots, 0)_{p\times q}$ is a truncation of $\boldsymbol{\Sigma}^\dagger$, we can equivalently write (4.22) in the matrix form

$$\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}} = \boldsymbol{K}_t^\dagger \boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Sigma}_t^\dagger \boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}. \tag{4.23}$$

The truncation index t represents the regularization parameter of the TSVD estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}$. The smaller the value of t, the smoother the resulting estimates, while setting $t = r$ takes us back to the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$. Methods for choosing the value of t are discussed below in Section 4.3.

Using Equation (A.6), the covariance of $\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}$ can be written as

$$\mathrm{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}} |\, \boldsymbol{\lambda}] = \boldsymbol{K}_t^\dagger \,\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{\lambda}]\, (\boldsymbol{K}_t^\dagger)^{\mathrm{T}},$$

which is better behaved than the covariance of $\mathrm{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} |\, \boldsymbol{\lambda}]$ given in Equation (4.18) because of the truncation in $\boldsymbol{K}_t^\dagger$. However, this reduction of variance comes with a price. Namely, the TSVD estimator is biased.

Proposition 4.9. *The bias of the TSVD estimator* (4.23) *is given by*

$$\mathrm{bias}\Big(\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}\Big) = (\boldsymbol{K}_t^\dagger - \boldsymbol{K}^\dagger)\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{P}\boldsymbol{\lambda},$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{K}^\dagger\boldsymbol{K}$ *is the orthogonal projection onto* $\ker(\boldsymbol{K})$.

Proof. By definition

$$\mathrm{bias}\Big(\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}\Big) = \mathrm{E}[\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}} |\, \boldsymbol{\lambda}] - \boldsymbol{\lambda}.$$

Here we can write

$$\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}} = \boldsymbol{K}_t^\dagger \boldsymbol{y} = \boldsymbol{K}^\dagger \boldsymbol{y} + (\boldsymbol{K}_t^\dagger - \boldsymbol{K}^\dagger)\boldsymbol{y}.$$

Using the linearity of the expectation and Equation (4.17), we find

$$\mathrm{bias}\Big(\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}\Big) = \boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{\lambda} + (\boldsymbol{K}_t^\dagger - \boldsymbol{K}^\dagger)\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{\lambda} = (\boldsymbol{K}_t^\dagger - \boldsymbol{K}^\dagger)\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{P}\boldsymbol{\lambda},$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{K}^\dagger\boldsymbol{K}$. □

When $\boldsymbol{K}$ has full column rank, $\ker(\boldsymbol{K}) = \{\boldsymbol{0}\}$ and hence $\boldsymbol{P}\boldsymbol{\lambda} = \boldsymbol{0}$, $\forall\boldsymbol{\lambda}$ giving us the bias

$$\mathrm{bias}\Big(\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}\Big) = (\boldsymbol{K}_t^\dagger - \boldsymbol{K}^\dagger)\boldsymbol{K}\boldsymbol{\lambda}.$$

Naturally, the bias vanishes by setting $t = r$.

4.2.2 Tikhonov Regularization

Instead of taming the oscillations of the solution by truncating the singular value spectrum of the smearing matrix, we could try to explicitly enforce some desired properties of the solution. This is the general idea of *Tikhonov regularization* [31]. In its simplest from, the Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\text{Tik}}$ is defined as the solution to the following penalized least squares problem, where the penalty term is equal to the squared 2-norm of the solution.

Definition 4.10. Let $\delta > 0$ be a constant. Then the *Tikhonov regularized estimator* $\hat{\boldsymbol{\lambda}}_{\text{Tik}}$ of $\boldsymbol{\lambda}$ is the solution to the optimization problem

$$\min_{\boldsymbol{\lambda}} \ \|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \delta\|\boldsymbol{\lambda}\|^2. \tag{4.24}$$

The constant δ is called the *regularization parameter*.

This means that we are trying to strike a balance between fitting the data, i.e., making $\|\boldsymbol{K}\boldsymbol{\lambda}-\boldsymbol{y}\|^2$ small, and finding a solution with a small norm $\|\boldsymbol{\lambda}\|^2$. The degree of this compromise is controlled with the regularization parameter δ. The larger the parameter δ, the smaller the norm of the solution, while taking $\delta \to 0$ gives us the least squares solution. The choice of δ is discussed in Section 4.3.

The following theorem shows that the Tikhonov regularized estimator exists and is unique. Furthermore, it establishes a convenient formula for computing the solution.

Theorem 4.11. *The Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\text{Tik}}$ exists and is unique for all smearing matrices $\boldsymbol{K} \in \mathbb{R}^{q\times p}$. Furthermore, the estimator is given by*

$$\hat{\boldsymbol{\lambda}}_{\text{Tik}} = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{I})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y}. \tag{4.25}$$

Proof. We can write

$$\begin{aligned}
\|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \delta\|\boldsymbol{\lambda}\|^2 &= \|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \|\sqrt{\delta}\boldsymbol{\lambda}\|^2 \\
&= \left\| \begin{bmatrix} \boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y} \\ \sqrt{\delta}\boldsymbol{\lambda} \end{bmatrix} \right\|^2 \\
&= \left\| \begin{bmatrix} \boldsymbol{K} \\ \sqrt{\delta}\boldsymbol{I} \end{bmatrix} \boldsymbol{\lambda} - \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \right\|^2 \\
&= \|\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{b}\|^2,
\end{aligned} \tag{4.26}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{K} \\ \sqrt{\delta}\boldsymbol{I} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}.$$

Hence, we see that the Tikhonov optimization problem (4.24) is equal to the least squares problem for matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ and hence, by Theorem 4.5, the Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\text{Tik}}$ exists. Furthermore, since the augmented matrix $\boldsymbol{A}$ has

linearly independent columns and hence full column rank, we can use Corollary 4.6 to deduce that the solution us unique and given by

$$\hat{\boldsymbol{\lambda}}_{\text{Tik}} = (\boldsymbol{A}^{\text{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\text{T}}\boldsymbol{b}. \tag{4.27}$$

Here

$$\boldsymbol{A}^{\text{T}}\boldsymbol{A} = \begin{bmatrix}\boldsymbol{K}^{\text{T}}\sqrt{\delta}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{K}\\ \sqrt{\delta}\boldsymbol{I}\end{bmatrix} = \boldsymbol{K}^{\text{T}}\boldsymbol{K} + \delta\boldsymbol{I}$$

and

$$\boldsymbol{A}^{\text{T}}\boldsymbol{b} = \begin{bmatrix}\boldsymbol{K}^{\text{T}}\sqrt{\delta}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix} = \boldsymbol{K}^{\text{T}}\boldsymbol{y}.$$

Substituting to (4.27), we find

$$\hat{\boldsymbol{\lambda}}_{\text{Tik}} = (\boldsymbol{K}^{\text{T}}\boldsymbol{K} + \delta\boldsymbol{I})^{-1}\boldsymbol{K}^{\text{T}}\boldsymbol{y}. \tag{4.28}$$

□

Using the singular value decomposition of $\boldsymbol{K}$, we gain another perspective on the workings of Tikhonov regularization.

Corollary 4.12. *Let $\{\boldsymbol{u}_i\}_{i=1}^{p}$ be the left singular vector, $\{\boldsymbol{v}_i\}_{i=1}^{q}$ be the right singular vectors and $\{\sigma_i\}_{i=1}^{\min(p,q)}$ be the singular values of the smearing matrix $\boldsymbol{K}$. Then the Tikhonov regularized estimator is given by*

$$\hat{\boldsymbol{\lambda}}_{\text{Tik}} = \sum_{i=1}^{r}\frac{\sigma_i}{\sigma_i^2+\delta}(\boldsymbol{u}_i^{\text{T}}\boldsymbol{y})\boldsymbol{v}_i, \tag{4.29}$$

where $r = \text{rank}(\boldsymbol{K})$.

Proof. Writing out the SVD of $\boldsymbol{K}$ given by Equation (4.19), we get

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\text{T}} = \sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^{\text{T}},$$

where $r = \text{rank}(\boldsymbol{K})$. Using this, we can write

$$\begin{aligned}\boldsymbol{K}^{\text{T}}\boldsymbol{K} + \delta\boldsymbol{I} &= \left(\sum_{i=1}^{r}\sigma_i\boldsymbol{v}_i\boldsymbol{u}_i^{\text{T}}\right)\left(\sum_{j=1}^{r}\sigma_j\boldsymbol{u}_j\boldsymbol{v}_j^{\text{T}}\right) + \delta\boldsymbol{I}\\ &= \sum_{i=1}^{r}\sum_{j=1}^{r}\sigma_i\sigma_j\boldsymbol{v}_i\boldsymbol{u}_i^{\text{T}}\boldsymbol{u}_j\boldsymbol{v}_j^{\text{T}} + \delta\boldsymbol{I}.\end{aligned}$$

Since $\boldsymbol{U}$ is orthogonal, its columns are orthonormal. Hence, $\boldsymbol{u}_i^{\text{T}}\boldsymbol{u}_j = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases}1, & \text{if } i = j\\ 0, & \text{if } i \neq j\end{cases}$$

is the Kronecker delta and we have

$$\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{I} = \sum_{i=1}^{r} \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}} + \delta\boldsymbol{I}.$$

Letting $\boldsymbol{\Sigma}^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0)_{p\times p}$ and using the orthogonality of $\boldsymbol{V}$, we can write this in matrix form

$$\begin{aligned}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{I} &= \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathrm{T}} + \delta\boldsymbol{I} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathrm{T}} + \delta\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \\ &= \boldsymbol{V}(\boldsymbol{\Sigma}^2 + \delta\boldsymbol{I})\boldsymbol{V}^{\mathrm{T}}.\end{aligned}$$

Since $\boldsymbol{\Sigma}^2 + \delta\boldsymbol{I}$ is a diagonal $p \times p$ matrix with strictly positive diagonal elements and $\boldsymbol{V}$ is orthogonal, we have

$$(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{I})^{-1} = \boldsymbol{V}(\boldsymbol{\Sigma}^2 + \delta\boldsymbol{I})^{-1}\boldsymbol{V}^{\mathrm{T}} = \sum_{i=1}^{p} \frac{1}{\sigma_i^2 + \delta} \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}. \tag{4.30}$$

Similarly,

$$\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y} = \sum_{i=1}^{r} \sigma_i (\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y}) \boldsymbol{v}_i. \tag{4.31}$$

Substituting (4.30) and (4.31) in (4.25), we obtain

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\mathrm{Tik}} &= (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{I})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{r} \frac{\sigma_j}{\sigma_i^2 + \delta} (\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{y}) \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_j \\ &= \sum_{i=1}^{p}\sum_{j=1}^{r} \frac{\sigma_j}{\sigma_i^2 + \delta} (\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{y}) \delta_{ij} \boldsymbol{v}_i \\ &= \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \delta} (\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y}) \boldsymbol{v}_i. \end{aligned}$$

□

Comparing Equations (4.21) and (4.29), we see that Tikhonov regularization stabilizes the least squares solution by replacing the multiplicative factor $1/\sigma_i$ with $\sigma_i/(\sigma_i^2 + \delta)$. When $\sigma_i \gg \delta$, $\sigma_i/(\sigma_i^2 + \delta) \approx 1/\sigma_i$ showing that the terms related to large singular values are left unaffected by the regularization. On the other hand, when $\sigma_i \to 0$, $\sigma_i/(\sigma_i^2 + \delta) \to 0$ which avoids the instabilities related to the small singular values. Furthermore, comparing Equations (4.22) and (4.29) allows us to easily see how Tikhonov regularization differs from TSVD. Namely, in TSVD, we imposed a hard cut-off for small singular values above some truncation index t, while in Tikhonov regularization, these singular values are truncated using a soft cut-off dictated by the regularization parameter δ.

Tikhonov regularization can be generalized by changing the original optimization problem (4.24). The most commonly encountered variation of the problem replaces the original penalty term by the squared norm of some linear mapping of $\boldsymbol{\lambda}$.

Definition 4.13. Let $\delta > 0$ be a constant regularization parameter. Then the *generalized Tikhonov regularized estimator* $\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}}$ of $\boldsymbol{\lambda}$ is the solution to the optimization problem

$$\min_{\boldsymbol{\lambda}} \ \|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \delta\|\boldsymbol{L}\boldsymbol{\lambda}\|^2, \tag{4.32}$$

where $\boldsymbol{L}$ is a matrix with p columns.

The matrix $\boldsymbol{L}$ is usually chosen to be the discretized version of either the first-order or the second-order derivative operator. In this case, the generalized penalty term $\|\boldsymbol{L}\boldsymbol{\lambda}\|^2$ penalizes, respectively, either for the slope or the curvature of the solution. Obviously, setting $\boldsymbol{L} = \boldsymbol{I}$ takes us back to the original Tikhonov regularized solution.

The following theorem shows that the matrix $\boldsymbol{L}$ need not have full column rank to guarantee the uniqueness of the generalized Tikhonov regularized estimator.

Theorem 4.14. *The generalized Tikhonov regularized estimator* $\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}}$ *always exists. Furthermore, when the intersection of the null spaces of* $\boldsymbol{K}$ *and* $\boldsymbol{L}$ *is trivial, i.e.,* $\ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}) = \{\boldsymbol{0}\}$*, the solution is unique and given by*

$$\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}} = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y}. \tag{4.33}$$

Proof. Using the same line of reasoning as in Equation (4.26), we find that the generalized problem (4.32) is equivalent to the least squares problem

$$\min_{\boldsymbol{\lambda}} \|\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{b}\|^2,$$

where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{K} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}.$$

Hence, Theorem 4.5 guarantees the existence of the generalized Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}}$. Similarly, Corollary 4.6 guarantees the uniqueness of the solution if the augmented matrix $\boldsymbol{A}$ has full column rank. This is equivalent to saying that the null space of $\boldsymbol{A}$ is trivial. Furthermore,

$$\begin{aligned}
& \ker(\boldsymbol{A}) = \{\boldsymbol{0}\} \\
\Leftrightarrow \quad & (\boldsymbol{A}\boldsymbol{\lambda} = \boldsymbol{0} \Leftrightarrow \boldsymbol{\lambda} = \boldsymbol{0}) \\
\Leftrightarrow \quad & \left(\begin{bmatrix} \boldsymbol{K}\boldsymbol{\lambda} \\ \sqrt{\delta}\boldsymbol{L}\boldsymbol{\lambda} \end{bmatrix} = \boldsymbol{0} \Leftrightarrow \boldsymbol{\lambda} = \boldsymbol{0} \right) \\
\Leftrightarrow \quad & (\boldsymbol{K}\boldsymbol{\lambda} = \boldsymbol{0} \ \& \ \boldsymbol{L}\boldsymbol{\lambda} = \boldsymbol{0} \Leftrightarrow \boldsymbol{\lambda} = \boldsymbol{0}) \\
\Leftrightarrow \quad & (\boldsymbol{\lambda} \in \ker(\boldsymbol{K}) \ \& \ \boldsymbol{\lambda} \in \ker(\boldsymbol{L}) \Leftrightarrow \boldsymbol{\lambda} = \boldsymbol{0}) \\
\Leftrightarrow \quad & (\boldsymbol{\lambda} \in \ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}) \Leftrightarrow \boldsymbol{\lambda} = \boldsymbol{0}) \\
\Leftrightarrow \quad & \ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}) = \{\boldsymbol{0}\}
\end{aligned}$$

showing that the solution is unique if the intersection of $\ker(\boldsymbol{K})$ and $\ker(\boldsymbol{L})$ is trivial. Equation (4.33) follows from a similar computation as the one in Equations (4.27)–(4.28). □

If $\ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}) \neq \{\boldsymbol{0}\}$, $\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}}$ is not unique but we can still compute one solution using

$$\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}} = \boldsymbol{A}^{\dagger}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{K} \\ \sqrt{\delta}\boldsymbol{L} \end{bmatrix}^{\dagger} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}$$

and Theorem 4.7 guarantees that this is the the minimum-norm solution.

Using a generalization of the singular value decomposition, the generalized Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\text{Tik},\boldsymbol{L}}$ given by Equation (4.33) can be written in a form analogous to the SVD form (4.29) of $\hat{\boldsymbol{\lambda}}_{\text{Tik}}$. For more details, we refer the interested reader to [2, Section 5.4].

We then turn our attention to the explicit form of the matrix $\boldsymbol{L}$ appearing in the generalized penalty term $\|\boldsymbol{L}\boldsymbol{\lambda}\|^2$. Firstly, if we wish to penalize for the magnitude of the intensity function f of the true Poisson process, we may set $\boldsymbol{L} = \boldsymbol{I} := \boldsymbol{L}^0$ which takes us back to standard Tikhonov regularization. Let us then assume that we would instead like to penalize for the first derivative of the intensity function f and that we are using bins of uniform size $\nu(E)$ in the discretization of f, i.e., $\nu(E_i) = \nu(E)$, $\forall i$. We can use the forward finite-difference approximation of f' and Equation (2.9) to write

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{\nu(E)} \approx \frac{\lambda_{i+1} - \lambda_i}{(\nu(E))^2},$$

where x_i denotes the center of the ith bin. Since the factor $1/(\nu(E))^2$ can be absorbed into the regularization parameter δ, we can choose $\boldsymbol{L}$ to be

$$\boldsymbol{L}_1^1 = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(p-1)\times p}. \tag{4.34}$$

One can easily check that $\operatorname{rank}(\boldsymbol{L}_1^1) = p - 1$. This choice for $\boldsymbol{L}$ does not enforce any boundary conditions, but on the other hand $\boldsymbol{L}_1^1$ does not have full column rank and we need to check for the condition $\ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}_1^1) = \{\boldsymbol{0}\}$ in Theorem 4.14. We can have an $\boldsymbol{L}$ matrix with full column rank, in which case the solution is unique for all smearing matrices $\boldsymbol{K}$, if we use the Dirichlet boundary condition on either boundary of the histogram. Often the spectra studied in high energy physics are such that we know that the intensity $f(x) = 0$ outside the right boundary of the true histogram. In such a case, the appropriate choice for $\boldsymbol{L}$ would be

$$\boldsymbol{L}_2^1 = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & -1 \end{bmatrix} \in \mathbb{R}^{p\times p}. \tag{4.35}$$

This matrix has full rank, $\operatorname{rank}(\boldsymbol{L}_2^1) = p$.

Alternatively, we could penalize for the curvature of the intensity function f. Assuming again uniform binning for the true histogram, the central finite-difference approximation for the second derivative of f is given by

$$f''(x_i) \approx \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\nu(E))^2} \approx \frac{\lambda_{i+1} - 2\lambda_i + \lambda_{i-1}}{(\nu(E))^3}.$$

When no boundary conditions are enforced, the matrix $\boldsymbol{L}$ becomes

$$\boldsymbol{L}_1^2 = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(p-2)\times p}. \tag{4.36}$$

This matrix has $\operatorname{rank}(\boldsymbol{L}_1^2) = p - 2$, i.e., full row rank but deficient column rank. To obtain a second-order finite-difference matrix with full column rank, one has to impose boundary conditions on both boundaries of the histogram. Assuming the Dirichlet boundary condition for both boundaries, i.e., $f(x) = 0$ outside of the true histogram, we obtain

$$\boldsymbol{L}_2^2 = \begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{p\times p}. \tag{4.37}$$

This is again a full-rank matrix, $\operatorname{rank}(\boldsymbol{L}_2^2) = p$, and hence guarantees the uniqueness of the corresponding generalized Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}_2^2}$.

Before discussing the statistical properties of generalized Tikhonov regularization, let us note that several other generalizations of the standard procedure (4.24) have been proposed in the literature. If one has a good idea about a likely unfolded solution $\boldsymbol{\lambda}_0$, it is possible to change the penalty term to reflect this and consider the problem

$$\min_{\boldsymbol{\lambda}} \ \|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \delta\|\boldsymbol{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\|^2.$$

In the context of HEP unfolding, $\boldsymbol{\lambda}_0$ could be a Monte Carlo simulated theoretical prediction for the true histogram. The obvious problem with this approach is that this introduces a bias towards $\boldsymbol{\lambda}_0$.

Another commonly encountered generalization is of the form

$$\min_{\boldsymbol{\lambda}} \ (\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}) + \delta\|\boldsymbol{L}\boldsymbol{\lambda}\|^2, \tag{4.38}$$

where $\boldsymbol{C} = \widehat{\operatorname{Cov}}[\boldsymbol{y}|\boldsymbol{\lambda}]$. For example, the popular SVD unfolding technique by Höcker and Kartvelishvili [28] is based on using the singular value decomposition to minimize an objective function which has the form (4.38), but in their approach, $\boldsymbol{\lambda}$ is replaced

by the ratio of the unfolded histogram to some expected theoretical histogram. The rationale for this scaling is that the ratios should be close to unity throughout the histogram and hence one can confidently enforce the smoothness of the resulting vector.

As in the case of TSVD, Tikhonov regularization achieves a reduction in the variance of the estimator by making the estimator biased. The following proposition establishes the bias of the generalized Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}$.

Proposition 4.15. *Assume that* $\ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}) = \{\boldsymbol{0}\}$. *Then the bias of the generalized Tikhonov regularized estimator* $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}$ *of Equation* (4.33) *is given by*

$$\mathrm{bias}\Big(\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}\Big) = -\left(\frac{1}{\delta}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\lambda}.$$

Proof. The Tikhonov estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}$ is equivalently given by

$$\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}} = \boldsymbol{A}^{\dagger}\boldsymbol{b},$$

where

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{K}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix}.$$

Hence

$$\mathrm{E}[\,\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}|\,\boldsymbol{\lambda}] = \boldsymbol{A}^{\dagger}\mathrm{E}[\boldsymbol{b}|\boldsymbol{\lambda}]. \tag{4.39}$$

Here

$$\begin{aligned}\mathrm{E}[\boldsymbol{b}|\boldsymbol{\lambda}] &= \mathrm{E}\left[\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix}\middle|\,\boldsymbol{\lambda}\right]\\ &= \begin{bmatrix}\boldsymbol{K}\boldsymbol{\lambda}\\ \boldsymbol{0}\end{bmatrix}\\ &= \left(\begin{bmatrix}\boldsymbol{K}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix} - \begin{bmatrix}\boldsymbol{0}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix}\right)\boldsymbol{\lambda}\\ &= \left(\boldsymbol{A} - \begin{bmatrix}\boldsymbol{0}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix}\right)\boldsymbol{\lambda}.\end{aligned}$$

Substituting in (4.39), we get

$$\mathrm{E}\Big[\,\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}\Big|\,\boldsymbol{\lambda}\Big] = \boldsymbol{A}^{\dagger}\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{A}^{\dagger}\begin{bmatrix}\boldsymbol{0}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix}\boldsymbol{\lambda} = \boldsymbol{\lambda} - \boldsymbol{A}^{\dagger}\begin{bmatrix}\boldsymbol{0}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix}\boldsymbol{\lambda},$$

where we have used $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}$. This follows from the full column rank of $\boldsymbol{A}$ which is, in turn, guaranteed the assumption $\ker(\boldsymbol{K}) \cap \ker(\boldsymbol{L}) = \{\boldsymbol{0}\}$ as shown in the proof of

Theorem 4.14. Using Corollary 4.6 and a computation similar to Equations (4.27)-(4.28), it follows that

$$\begin{aligned}\operatorname{bias}\left(\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}\right) &= -\boldsymbol{A}^{\dagger}\begin{bmatrix}\mathbf{0}\\ \sqrt{\delta}\boldsymbol{L}\end{bmatrix}\boldsymbol{\lambda}\\ &= -(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\begin{bmatrix}\mathbf{0}\\ \sqrt{\delta}\boldsymbol{L}\boldsymbol{\lambda}\end{bmatrix}\\ &= -(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}+\delta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L})^{-1}\delta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\lambda}\\ &= -\left(\frac{1}{\delta}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}+\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\lambda}.\end{aligned}$$

□

The covariance of the Tikhonov estimator can again be expressed using Equations (A.6) and (4.33) as follows:

$$\operatorname{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}|\,\boldsymbol{\lambda}] = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}+\delta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L})^{-1}\boldsymbol{K}^{\mathrm{T}}\operatorname{Cov}[\boldsymbol{y}|\boldsymbol{\lambda}]\,\boldsymbol{K}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}+\delta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L})^{-1}.$$

One might then be tempted to think that the square roots of the diagonal elements of the covariance matrix $\operatorname{Cov}[\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}|\,\boldsymbol{\lambda}]$ could be used to construct approximate confidence intervals for the true means λ_i. Unfortunately, this is not the case as will be explained in the next subsection.

4.2.3 Error Estimation

We have studied above the least squares estimator and two related regularized estimators based on TSVD and Tikhonov regularization. All these estimators are linear in the sense that they can be expressed in the form

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{+}\boldsymbol{y},$$

where in the case of the least squares estimator, $\boldsymbol{K}^{+}$ is the pseudoinverse of $\boldsymbol{K}$ and in the case of TSVD and Tikhonov regularization, a regularized version of the pseudoinverse. As we have shown, the covariance of such an estimator is given by

$$\operatorname{Cov}[\hat{\boldsymbol{\lambda}}|\,\boldsymbol{\lambda}] = \boldsymbol{K}^{+}\operatorname{Cov}[\boldsymbol{y}|\boldsymbol{\lambda}]\,(\boldsymbol{K}^{+})^{\mathrm{T}} = \boldsymbol{K}^{+}\operatorname{diag}(\boldsymbol{\mu})(\boldsymbol{K}^{+})^{\mathrm{T}}.$$

If we use the MLE, which is simply given by the observations $\boldsymbol{y}$, to estimate $\boldsymbol{\mu}$, the covariance can be estimated using

$$\widehat{\operatorname{Cov}}[\hat{\boldsymbol{\lambda}}|\,\boldsymbol{\lambda}] = \boldsymbol{K}^{+}\operatorname{diag}(\hat{\boldsymbol{\mu}})(\boldsymbol{K}^{+})^{\mathrm{T}} = \boldsymbol{K}^{+}\operatorname{diag}(\boldsymbol{y})(\boldsymbol{K}^{+})^{\mathrm{T}}$$

and we could then estimate the standard deviation of each component of the estimator $\hat{\boldsymbol{\lambda}}$ with

$$\widehat{\operatorname{Std}}[\hat{\lambda}_i|\,\boldsymbol{\lambda}] = \sqrt{\widehat{\operatorname{Cov}}[\hat{\boldsymbol{\lambda}}|\,\boldsymbol{\lambda}]_{ii}}. \tag{4.40}$$

We could then proceed as we did in Section 3.1 and report the outcome of the measurement using $\hat{\lambda}_i \pm \widehat{\operatorname{Std}}[\hat{\lambda}_i|\,\boldsymbol{\lambda}]$. However, when regularization is involved, this does not serve as an approximate 68.27 % confidence interval for the true mean λ_i.

To see why, let us remind ourselves that the diagonal elements of $\widehat{\mathrm{Cov}}[\hat{\boldsymbol{\lambda}}|\,\boldsymbol{\lambda}]$ represent the spread of the estimator $\hat{\boldsymbol{\lambda}}$ around its expectation $\mathrm{E}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}]$. However, only when the estimator is unbiased, i.e., $\mathrm{E}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}] = \boldsymbol{\lambda}$, does this coincide with the true value of the parameter. When the estimator is biased, $\mathrm{E}[\hat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}] \neq \boldsymbol{\lambda}$ and hence the diagonal elements of $\widehat{\mathrm{Cov}}[\hat{\boldsymbol{\lambda}}|\,\boldsymbol{\lambda}]$ do not represent the variability of the estimator around the true parameter value $\boldsymbol{\lambda}$.

As we discussed above, regularization achieves a reduction in the variance of the estimator by making it biased. In particular, the TSVD estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}$ and the Tikhonov estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}$ are in general biased as shown by Propositions 4.9 and 4.15. Hence, when these estimators are used and $\hat{\lambda}_i \pm \widehat{\mathrm{Std}}[\hat{\lambda}_i|\,\boldsymbol{\lambda}]$ is reported as the outcome of the measurement, this should not be thought of as a confidence interval for λ_i. Instead, the interval $\hat{\lambda}_i \pm \widehat{\mathrm{Std}}[\hat{\lambda}_i|\,\boldsymbol{\lambda}]$ represents the spread of the estimators $\hat{\lambda}_i$ if the measurement were to be repeated several times, but due to the bias, this interval is systematically off with respect to the true parameter λ_i. The same is true even for the least squares estimator when the smearing matrix $\boldsymbol{K}$ does not have full column rank as shown by Equation (4.17).

Note also that as the regularization strength is increased, i.e., we make the regularization parameter δ in Tikhonov regularization larger or the truncation index t in TSVD smaller, the variance of the regularized estimators is decreased. With strong enough regularization, the variance can be made arbitrarily small. For example, in standard Tikhonov regularization (i.e. $\boldsymbol{L} = \boldsymbol{I}$), taking the limit $\delta \to \infty$ gives us $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik}} = \boldsymbol{0}$ with zero variance. This resolves the long-standing paradox in the high energy physics community about obtaining errors smaller than $\sqrt{\hat{\lambda}_i}$ for the unfolded histogram seemingly providing better resolution than that obtained from a perfect detector [41]. Namely, there is no contradiction in having the estimated standard deviations $\widehat{\mathrm{Std}}[\hat{\lambda}_i|\,\boldsymbol{\lambda}]$ smaller than $\sqrt{\hat{\lambda}_i}$ simply because they do not represent the resolution of the experiment when one does not account for the bias of the regularized estimator.

Unfortunately, the frequentist paradigm of statistics does not provide a straightforward way of constructing confidence intervals for λ_i starting from such complicated estimators as the TSVD estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{TSVD}}$ or the Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik}}$. For instance, one could naïvely try to estimate the bias of $\hat{\boldsymbol{\lambda}}$ and consider a new estimator of the form $\hat{\boldsymbol{\lambda}}^* = \hat{\boldsymbol{\lambda}} - \widehat{\mathrm{bias}}(\hat{\boldsymbol{\lambda}})$ which should be nearly unbiased and would hence solve most of our problems. However, since the estimator $\widehat{\mathrm{bias}}(\hat{\boldsymbol{\lambda}})$ is a random variable, subtracting it from $\hat{\boldsymbol{\lambda}}$ would again cause the variance of the resulting estimator $\hat{\boldsymbol{\lambda}}^*$ to blow up and would just take us back to square one.

4.3 Choice of the Regularization Strength

A common theme with all the unfolding techniques discussed in this chapter is that one has to find a way to choose the regularization strength of the method. That is, we need to strike a balance between the bias and the variance of the regularized estimator $\hat{\boldsymbol{\lambda}}$. In the EM algorithm, the regularization strength is controlled by the length of the iteration, in TSVD, it is the truncation index t that controls this

balance, while in Tikhonov regularization, we need to choose the regularization parameter δ. Several methods for choosing these parameters have been proposed in the literature. In this discussion, we first focus on Tikhonov regularization and then make brief remarks about the EM algorithm and TSVD at the end of this section. It should be noted that while all methods for choosing the regularization strength are more or less heuristic in their nature, a detailed analysis of their statistical properties can still be carried out. In this treatment, we will merely focus on introducing some of the most commonly encountered techniques for choosing the regularization strength and we refer the reader to the literature for a more detailed analysis of the methods. Good references to consider are, e.g., Chapter 4 in [20] and Chapter 7 in [58].

Recall from Section 4.2.2 that, in generalized Tikhonov regularization, we solve the optimization problem

$$\min_{\boldsymbol{\lambda}} \ \|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 + \delta\|\boldsymbol{L}\boldsymbol{\lambda}\|^2 \tag{4.41}$$

and the question we are trying to address here is how should the regularization parameter δ be selected. To emphasize its dependence on δ, let us for the rest of this section denote the solution of (4.41) by $\hat{\boldsymbol{\lambda}}_\delta$ instead of the earlier used notation $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}$.

The optimization problem (4.41) gives an immediate motivation for the so-called *L-curve method* [25] for choosing δ. In this method, one plots on a log-log scale the norm of the residual $\|\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta - \boldsymbol{y}\|$ versus $\|\boldsymbol{L}\hat{\boldsymbol{\lambda}}_\delta\|$ for different values of δ. Because $\|\boldsymbol{L}\hat{\boldsymbol{\lambda}}_\delta\|$ decreases as a function of δ and $\|\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta - \boldsymbol{y}\|$ increases as a function of δ, the resulting curve often takes a distinctive shape resembling the letter L. One then picks the parameter δ that corresponds to the corner of the curve as the optimal regularization parameter. The rationale for this is that the point at the corner offers, in some sense, the optimal compromise between $\|\boldsymbol{K}\boldsymbol{\lambda} - \boldsymbol{y}\|$ and $\|\boldsymbol{L}\boldsymbol{\lambda}\|$.

Another commonly encountered method for choosing δ is the *Morozov discrepancy principle* [47] (see also [31, Section 2.3]). In this method, one chooses the largest value of δ which satisfies[2]

$$\|\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta - \boldsymbol{y}\| \leq \varepsilon, \tag{4.42}$$

where ε characterizes the noise level of the data. The rationale behind this approach is that we can safely expect to be able to fit the data up to the noise level, but any solution that matches the observations with an accuracy higher than that is potentially *overfitting* the data.

Unfortunately, the choice of the noise level ε in (4.42) is not unambiguous and depends on the particular problem under consideration. For example, in the case of unfolding, we could set

$$\begin{aligned}\varepsilon^2 &= \mathrm{E}\big[\|\boldsymbol{\mu} - \boldsymbol{y}\|^2|\boldsymbol{\mu}\big] = \sum_{i=1}^{q} \mathrm{E}\big[(\mu_i - y_i)^2|\mu_i\big] \\ &= \sum_{i=1}^{q} \mathrm{Var}[y_i|\mu_i] = \sum_{i=1}^{q} \mu_i = \|\boldsymbol{\mu}\|_1,\end{aligned}$$

[2]Because in Tikhonov regularization the regularization parameter δ is continuous, the inequality in (4.42) can actually be replaced by an equality. The inequality is required for cases where the regularization parameter is discrete, such as TSVD and iterative methods.

where $\boldsymbol{y}|\boldsymbol{\mu} \sim \text{Poisson}(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \boldsymbol{K}\boldsymbol{\lambda}$. This gives us $\varepsilon = \sqrt{\|\boldsymbol{\mu}\|_1}$, which is unknown because $\boldsymbol{\mu}$ is unknown. However, we can still estimate this by estimating $\boldsymbol{\mu}$ using the observations $\boldsymbol{y}$. Hence, the estimated noise level is $\hat{\varepsilon} = \sqrt{\|\hat{\boldsymbol{\mu}}\|_1} = \sqrt{\|\boldsymbol{y}\|_1}$ and we would choose the regularization parameter δ so that it satisfies

$$\|\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta - \boldsymbol{y}\| = \sqrt{\|\boldsymbol{y}\|_1}.$$

A third method for choosing δ is *cross validation*, which is a general statistical technique often used in solving parameter and model selection problems. First, consider the Tikhonov optimization problem (4.41) where we have left out the kth observation from the objective function, that is,

$$\min_{\boldsymbol{\lambda}} \sum_{i \neq k} ((\boldsymbol{K}\boldsymbol{\lambda})_i - y_i)^2 + \delta \|\boldsymbol{L}\boldsymbol{\lambda}\|^2. \tag{4.43}$$

Let us denote the solution of this problem by $\hat{\boldsymbol{\lambda}}_\delta^{[k]}$. We would then expect that for each k, the kth component of $\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta^{[k]}$ should be close to y_k even though y_k was not used for finding $\hat{\boldsymbol{\lambda}}_\delta^{[k]}$. That is, for an appropriate choice of the regularization parameter δ, the solution should *generalize* for new observations. If this is not the case, then we could be overfitting or underfitting the data. We hence choose δ by minimizing the cross validation function

$$\text{CV}(\delta) = \frac{1}{q} \sum_{k=1}^{q} ((\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta^{[k]})_k - y_k)^2.$$

We call this approach for choosing δ the *leave-one-out cross validation* method.

From a computational point of view, the main issue with leave-one-out cross validation is that, for each value of δ, one has to solve the problem (4.43) q times in order to evaluate $\text{CV}(\delta)$. Luckily, it can be shown (see, e.g., [2, Section 5.7]) that $\text{CV}(\delta)$ can be approximated using a *generalized cross validation* function $\text{GCV}(\delta)$ that depends only on the solution $\hat{\boldsymbol{\lambda}}_\delta$ of the full problem (4.41). Namely,

$$\text{CV}(\delta) \approx \text{GCV}(\delta) = \frac{q\|\boldsymbol{K}\hat{\boldsymbol{\lambda}}_\delta - \boldsymbol{y}\|^2}{\text{tr}(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{K}^+)},$$

where $\boldsymbol{K}^+ = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \delta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L})^{-1}\boldsymbol{K}^{\mathrm{T}}$. The regularization parameter δ is then chosen by minimizing $\text{GCV}(\delta)$. The formal justification to using generalized cross validation lies on an analysis of its asymptotic properties. For details, we refer the reader to [15, 59].

Out of the three methods considered for choosing the regularization strength, generalized cross validation has the best theoretical foundation but it can be easily used only in the case of TSVD and Tikhonov regularization. On the other hand, the L-curve technique and Morozov discrepancy principle can be used with all the unfolding techniques discussed in this chapter, including the iterative EM algorithm. The main problem with the discrepancy principle is the heuristic choice of the noise level ε, while the L-curve technique suffers from arbitrariness in determining the

corner of the L-curve. Not surprisingly, a number of other criteria for stopping the EM iteration have also been proposed. Veklerov and Llacer [57], for example, perform on each EM iteration a statistical hypothesis test to see if $\boldsymbol{y}$ can be regarded as a realization of a Poisson($\boldsymbol{\mu}^{(k)}$) distributed random variable, where $\boldsymbol{\mu}^{(k)} = \boldsymbol{K}\boldsymbol{\lambda}^{(k)}$, while in high energy physics applications, one usually uses a χ^2 comparison of two consecutive iterates $\boldsymbol{\lambda}^{(k)}$ and $\boldsymbol{\lambda}^{(k-1)}$ [16].

Chapter 5

Bayesian Unfolding

Bayesian techniques are widely regarded as a versatile alternative to frequentist techniques for solving inverse problems [31]. The basic idea is to use Bayes' theorem to compute the distribution of the unknown parameters given the observed data and then use this posterior distribution to describe our understanding about the unknowns. Efficient sampling techniques based on Markov chains make this approach computationally feasible for a wide range of models. The Bayesian perspective is especially well suited for HEP data analysis since it provides a natural way of quantifying the uncertainty of the solution via Bayesian credible intervals. The basic concepts of Bayesian inference for unfolding are first explained in Section 5.1. We then show in Section 5.2 how Markov chain Monte Carlo samplers can be used to generate a sample from the posterior. The important question about the choice of the prior distribution, which regularizes the otherwise ill-posed problem, is addressed in Section 5.3.

5.1 Bayesian Inference for Unfolding

The key idea in Bayesian unfolding is to use Bayes' theorem (A.8) to make inferences about the unknown means $\boldsymbol{\lambda}$ given the observed histogram $\boldsymbol{y}$ distributed according to model (2.12). This was recently proposed in the context of HEP unfolding by Choudalakis [11]. For unfolding, Bayes' theorem reads

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{y})}, \tag{5.1}$$

where $p(\boldsymbol{\lambda}|\boldsymbol{y})$ is the posterior of the means $\boldsymbol{\lambda}$, $p(\boldsymbol{y}|\boldsymbol{\lambda})$ is the likelihood of $\boldsymbol{\lambda}$ given by Equation (4.1) and $p(\boldsymbol{\lambda})$ is the prior density. The marginal $p(\boldsymbol{y})$ does not depend on the unknown $\boldsymbol{\lambda}$ and can be regarded as a normalization constant for the posterior density.

In the strict context of the Bayesian paradigm of statistics, Bayes' formula (5.1) should be interpreted as follows: The prior $p(\boldsymbol{\lambda})$ represents our a priori subjective degree of belief about the true histogram $\boldsymbol{\lambda}$ before looking at the data $\boldsymbol{y}$. When the data come in, the combination of this prior information and our information

about the process that generated the data reflected by the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$ is used to update our degree of belief about the true histogram. This a posteriori understanding of $\boldsymbol{\lambda}$ is then captured by the posterior density $p(\boldsymbol{\lambda}|\boldsymbol{y})$. In essence, the posterior contains *all* the information we have about $\boldsymbol{\lambda}$ given the observations $\boldsymbol{y}$ and our prior beliefs $p(\boldsymbol{\lambda})$.

In the case of inverse problems, the prior $p(\boldsymbol{\lambda})$ has an additional, important role. Namely, it is the prior that regularizes the otherwise ill-posed problem. In Bayesian thinking, the ill-posedness of the problem manifests itself as a likelihood function which is almost flat for a large number of very different solutions. Since the posterior is proportional to the product of the likelihood and the prior density, choosing, for instance, the flat uniform prior would result in a posterior suffering from the same flatness problem and lead to largely arbitrary inferences about $\boldsymbol{\lambda}$. On the other hand, if the prior places the majority of its probability mass on physically plausible solutions, this will make the posterior to peak near such solutions and thus drive the inference towards physically acceptable histograms.

Often the prior in (5.1) is chosen among some parametric family of models in which case it depends on some hyperparameters $\boldsymbol{\alpha}$. In this case, Bayes' rule can be written as follows

$$p(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha})}. \tag{5.2}$$

This form emphasizes the fact that the posterior depends on the choice of the prior via the parameters $\boldsymbol{\alpha}$. Here the denominator is given by the integral

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\lambda}. \tag{5.3}$$

From the Bayesian point of view, the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})$ is the complete solution to the unfolding problem. However, when there are p bins in the true histogram, this is a p-dimensional probability density function. Since usually $p \gg 1$, it is not practical to provide such a density as the outcome of the unfolding procedure. Hence, we need some more accessible ways of summarizing the information contained in the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})$. Some frequently used measures for the location of the posterior are its expectation $\mathrm{E}[\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha}]$ or its mode $\arg\max_{\boldsymbol{\lambda}} p(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})$ assuming that these are well-defined quantities. The estimator corresponding to the mode of the posterior is called the *maximum a posteriori* (MAP) *estimator* $\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}}$. Similarly, the covariance of the posterior $\mathrm{Cov}[\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha}]$ can be used to measure the spread of the posterior around its mean.

Recall that our main motivation for using Bayesian techniques was the possibility of using the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})$ to construct credible intervals which are analogous to confidence intervals in frequentist statistics. From the point of view of computation, interpretation and presentation of the results, the most straightforward way of doing this is to consider the one-dimensional marginals of the posterior

$$p(\lambda_i|\boldsymbol{y},\boldsymbol{\alpha}) = \int p(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\lambda}_{-i}, \tag{5.4}$$

where $\mathrm{d}\boldsymbol{\lambda}_{-i} = \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{i-1}\mathrm{d}\lambda_{i+1} \cdots \mathrm{d}\lambda_p$. The marginal posterior $p(\lambda_i|\boldsymbol{y},\boldsymbol{\alpha})$ captures all the information we have about the ith bin when we want to summarize its

inference without making reference to the rest of the bins[1]. The $100(1-\alpha)\%$ marginal credible interval $[a_i, b_i]$ for λ_i is then defined as the solution to

$$\frac{\alpha}{2} = \int_0^{a_i} p(\lambda_i|\boldsymbol{y}, \boldsymbol{\alpha}) \, \mathrm{d}\lambda_i,$$
$$\frac{\alpha}{2} = \int_{b_i}^{\infty} p(\lambda_i|\boldsymbol{y}, \boldsymbol{\alpha}) \, \mathrm{d}\lambda_i.$$

In practice, we shall not be working with the closed-form marginal posterior $p(\lambda_i|\boldsymbol{y}, \boldsymbol{\alpha})$ but instead with a sample $\{\lambda_i^{(k)}\}_{k=1}^N$ from $p(\lambda_i|\boldsymbol{y}, \boldsymbol{\alpha})$. In this case, the lower limit a_i can be estimated using the $(100 \cdot \alpha/2)$th percentile of the sample and the upper limit using the $100(1 - \alpha/2)$th percentile of the sample.

It is often desirable to complement the credible intervals $[a_i, b_i]$ with other tools from descriptive statistics to gain a maximum amount of information about the marginal posteriors. When reporting the results of the computational experiments of Chapter 7, we accompany the credible intervals $[a_i, b_i]$ with the medians $\hat{\lambda}_{\mathrm{med},i}$ of the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \boldsymbol{\alpha})$ to summarize our understanding about their locations. As above, $\hat{\lambda}_{\mathrm{med},i}$ is obtained as the 50th percentile of the sample $\{\lambda_i^{(k)}\}_{k=1}^N$. In addition, we use box plots of these samples to provide an alternative way of summarizing our understanding of the marginals $p(\lambda_i|\boldsymbol{y}, \boldsymbol{\alpha})$.

We conclude this section by showing that there is a simple relationship between the Bayesian maximum a posteriori point estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}}$ and the frequentist maximum likelihood estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}$ of the true means $\boldsymbol{\lambda}$. Namely, these two estimators coincide when we use the uniform non-negativity prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) = p(\boldsymbol{\lambda}) \propto 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda})$. This is easily seen as follows:

$$\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}} = \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} p(\boldsymbol{\lambda}|\boldsymbol{y}) = \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} p(\boldsymbol{y}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) = \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}_+^p} p(\boldsymbol{y}|\boldsymbol{\lambda}) = \hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}.$$

Using different priors, various other MAP estimators can be constructed. In cases where the parameters can take any real values, it is often possible to find a similar correspondence between Tikhonov regularized estimators discussed in Section 4.2.2 and the MAP estimator for a suitably chosen prior. However, in our case, such a correspondence does not exist which can be seen by noting that the MAP estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}}$ is always non-negative[2] while the components of the (generalized) Tikhonov regularized estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{Tik},\boldsymbol{L}}$ can also take negative values.

[1] If needed, the full posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ can also be used to construct confidence envelopes for the whole histogram $\boldsymbol{\lambda}$ using Bayesian credible sets (see Definition A.24).

[2] Note that it is ultimately the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$ that enforces the non-negativity of the solution in Bayesian unfolding and not the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$. This is because the likelihood, which is given in Equation (4.1), is only defined for $\boldsymbol{\lambda} \in \mathbb{R}_+^p$. As a result, all the integrals and densities over $\boldsymbol{\lambda}$ should only be considered over $\mathbb{R}_+^p$ (see Sections A.1 and A.2). We could thus also write the uniform prior in the form $p(\boldsymbol{\lambda}) \propto 1$, $\boldsymbol{\lambda} \in \mathbb{R}_+^p$. However, for the clarity of the presentation and in order to avoid any confusion, we prefer adopting the convention of writing the priors with the indicator function $1_{\mathbb{R}_+^p}(\boldsymbol{\lambda})$ and to call such priors non-negativity priors even though this would not be absolutely necessary.

5.2 Markov Chain Monte Carlo Sampling

From a computational point of view, the main problem with Bayesian unfolding is that the denominator $p(\boldsymbol{y}|\boldsymbol{\alpha})$ of Bayes' rule (5.2) is given by the p-dimensional integral of Equation (5.3) which often cannot be evaluated analytically. In addition, when the number of true bins $p \gg 1$, this becomes extremely difficult to compute numerically using standard quadrature rules. This means that we are in most cases unable to normalize the posterior density $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$. Similarly, integrals of the posterior, such as the one in Equation (5.4), cannot be evaluated using standard integration techniques.

Fortunately, a class of algorithms called *Markov chain Monte Carlo* (MCMC) *sampling* algorithms allows one to draw a sample from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ without the need to know the normalization constant $p(\boldsymbol{y}|\boldsymbol{\alpha})$ [31, 23]. It is largely due to these algorithms that Bayesian methods have become commonplace in contemporary statistics and have been successfully applied to a wide class of problems that could not have been solved using traditional frequentist methods.

The main idea of MCMC sampling is to construct a time-homogeneous Markov chain having the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ as its equilibrium distribution. The terminology here is defined as follows:

Definition 5.1. A *time-homogeneous Markov chain* is an ordered set of random variables $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$, $\boldsymbol{X}_j \in \mathbb{R}^d$, which satisfy

$$P_{\boldsymbol{X}_{j+1}}(B|\boldsymbol{X}_j = \boldsymbol{x}_j, \ldots, \boldsymbol{X}_1 = \boldsymbol{x}_1) = P_{\boldsymbol{X}_{j+1}}(B|\boldsymbol{X}_j = \boldsymbol{x}_j) = P(\boldsymbol{x}_j, B), \quad \forall B \in \mathcal{B}^d,$$

where $\mathcal{B}^d$ is the Borel σ-algebra on $\mathbb{R}^d$. Here, the probability measure $P(\boldsymbol{x}, \cdot)$ is called the *transition kernel* and does not depend on the time index j.

This means that the distribution of the state of the chain at time index $j+1$ only depends on the history of the chain via the state of the chain at the previous time index j. If we know the distribution of the chain at time step j, we can use the transition kernel P to write the distribution of the chain at the subsequent time step $j+1$

$$P_{\boldsymbol{X}_{j+1}}(B) = \int P_{\boldsymbol{X}_{j+1}}(B|\boldsymbol{X}_j = \boldsymbol{x}_j) P_{\boldsymbol{X}_j}(\mathrm{d}\boldsymbol{x}_j) = \int P(\boldsymbol{x}_j, B) P_{\boldsymbol{X}_j}(\mathrm{d}\boldsymbol{x}_j).$$

A distribution of states which is invariant under this mapping is called the invariant distribution of the kernel P.

Definition 5.2. A probability measure μ is the *invariant distribution* of the transition kernel $P(\boldsymbol{x}, \cdot)$ if

$$\mu(B) = \int P(\boldsymbol{x}, B) \mu(\mathrm{d}\boldsymbol{x}), \quad \forall B \in \mathcal{B}^d.$$

Let us also define the k-step transition kernel $P^{(k)}(\boldsymbol{x}, \cdot)$ using

$$P^{(k)}(\boldsymbol{x}_i, B) = P_{\boldsymbol{X}_{j+k}}(B|\boldsymbol{X}_j = \boldsymbol{x}_j), \quad B \in \mathcal{B}^d.$$

Clearly, $P^{(1)} = P$ and the k-step transition kernel can be constructed recursively using the one-step kernel P

$$\begin{aligned} P^{(k)}(\boldsymbol{x}_j, B) &= \int P_{\boldsymbol{X}_{j+k}}(B|\boldsymbol{X}_{j+k-1} = \boldsymbol{x}_{j+k-1}) P_{\boldsymbol{X}_{j+k-1}}(\mathrm{d}\boldsymbol{x}_{j+k-1}|\boldsymbol{X}_j = \boldsymbol{x}_j) \\ &= \int P(\boldsymbol{x}_{j+k-1}, B) P^{(k-1)}(\boldsymbol{x}_j, \mathrm{d}\boldsymbol{x}_{j+k-1}), \quad k > 2. \end{aligned}$$

The distribution of the chain at time index k is then given by

$$P_{\boldsymbol{X}_k}(B) = \int P_{\boldsymbol{X}_k}(B|\boldsymbol{X}_1 = \boldsymbol{x}_1) P_{\boldsymbol{X}_1}(\mathrm{d}\boldsymbol{x}_1) = \int P^{(k-1)}(\boldsymbol{x}_1, B) P_{\boldsymbol{X}_1}(\mathrm{d}\boldsymbol{x}_1). \tag{5.5}$$

It can be shown [31, Proposition 3.11] that if the transition kernel P satisfies certain regularity conditions, the asymptotic limit of the k-step transition kernel $P^{(k)}$ is given by the invariant distribution μ of P. That is, for μ-almost all $\boldsymbol{x}_1 \in \mathbb{R}^d$

$$\lim_{k\to\infty} P^{(k)}(\boldsymbol{x}_1, B) = \mu(B), \quad \forall B \in \mathcal{B}^d. \tag{5.6}$$

If this holds, then the invariant distribution μ is called the *equilibrium distribution* of the Markov chain $\{\boldsymbol{X}_k\}_{k=1}^{\infty}$. If the distribution of the starting point $\boldsymbol{X}_1$ is degenerate at $\boldsymbol{x}_1$, $P_{\boldsymbol{X}_1} = \delta_{\boldsymbol{x}_1}$, and μ is the equilibrium distribution of the chain, Equation (5.5) gives us

$$P_{\boldsymbol{X}_k}(B) = P^{(k-1)}(\boldsymbol{x}_1, B) \to \mu(B)$$

when $k \to \infty$. This means that for large time indices k, the states $\boldsymbol{X}_k$ are approximately distributed according to μ irrespective of the starting point $\boldsymbol{x}_1$ of the chain. It follows that if we are able to construct such a Markov chain that its equilibrium distribution is the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$, then for large indices k, the realizations of the random variables $\boldsymbol{X}_k$ form a sample from the posterior.

Under the same regularity conditions required for the convergence result (5.6), one can also show the following ergodicity property for the invariant distribution μ:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\boldsymbol{X}_k) = \int f(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x}) \quad \text{a.s.}$$

for all μ-integrable functions f. Assuming that μ is absolutely continuous with density $p(\boldsymbol{x})$, this result shows that when $\{\boldsymbol{x}_k\}_{k=1}^{N}$ is a large enough realization of the Markov chain, we can perform MC integration with respect to the density $p(\boldsymbol{x})$ using

$$\mathrm{E}[f(\boldsymbol{X})] = \int f(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \approx \frac{1}{N} \sum_{k=1}^{N} f(\boldsymbol{x}_k). \tag{5.7}$$

Several algorithms for constructing a Markov chain having the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ as its equilibrium distribution have been proposed in the literature. Out of these MCMC samplers, the most well-known is the Metropolis–Hastings algorithm. Assume that at time step k, the chain is at point $\boldsymbol{\lambda}^{(k)}$. Then, the basic idea of the

Metropolis–Hastings algorithm is to draw a proposal for the next point $\boldsymbol{\lambda}^*$ from some proposal density $p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)})$. This proposal is then either accepted or rejected. Letting $\pi(\boldsymbol{\lambda}) = p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ and $q(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*) = p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)})$, the probability of accepting $\boldsymbol{\lambda}^*$ is

$$a(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*) = \min\left(1, \frac{\pi(\boldsymbol{\lambda}^*)q(\boldsymbol{\lambda}^*, \boldsymbol{\lambda}^{(k)})}{\pi(\boldsymbol{\lambda}^{(k)})q(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*)}\right). \tag{5.8}$$

If the proposal $\boldsymbol{\lambda}^*$ is accepted, this becomes the next point of the chain $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^*$. Otherwise, the chain remains at its current location, $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}$. More specifically, the Metropolis–Hastings algorithm for sampling N observations from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ is given by the iteration:

1. Pick the initial point $\boldsymbol{\lambda}^{(1)} \in \mathbb{R}_+^p$ and let $k = 1$.
2. Sample the proposal $\boldsymbol{\lambda}^*$ from $p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)})$.
3. Compute $a(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*)$ using Equation (5.8).
4. Sample $r \sim U(0, 1)$, where $U(0, 1)$ is the uniform distribution on the interval $[0, 1]$.
5. If $r \leq a(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*)$, accept $\boldsymbol{\lambda}^*$ and set $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^*$, else set $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}$.
6. Set $k \leftarrow k + 1$.
7. If $k = N$, terminate, else go to step 2.

It is easy to show that the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ is the invariant distribution of the Markov chain given by the Metropolis–Hastings algorithm. To show that the posterior is also the equilibrium distribution, some mild assumptions about the proposal density $p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)})$ are required. For details, we refer the reader to [54].

The reason the Metropolis–Hastings algorithm is so useful in Bayesian computations is the fact that the posterior only appears in the algorithm in the ratio $\pi(\boldsymbol{\lambda}^*)/\pi(\boldsymbol{\lambda}^{(k)})$ in Equation (5.8). Hence, the normalization factor $p(\boldsymbol{y}|\boldsymbol{\alpha})$ cancels out and we are able to sample from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ without having to evaluate the problematic integral (5.3).

Naturally, different choices for the proposal density $p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)}) = q(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*)$ will lead to different sampling schemes. A common choice, is the random-walk proposal density, which satisfies

$$q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{x} - \boldsymbol{y}), \quad q(-\boldsymbol{x}) = q(\boldsymbol{x}). \tag{5.9}$$

It follows that $q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y}, \boldsymbol{x})$ and the acceptance probability (5.8) simplifies to

$$a(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}^*) = \min\left(1, \frac{\pi(\boldsymbol{\lambda}^*)}{\pi(\boldsymbol{\lambda}^{(k)})}\right).$$

This is the original version of the algorithm as proposed by Metropolis et al. in [44]. The simplified acceptance probability has a straightforward interpretation: If the proposed move $\boldsymbol{\lambda}^*$ is in the direction of higher posterior density, it is always

accepted, while a move to the direction of lower posterior density is accepted with the probability given by the ratio of the posterior densities $\pi(\boldsymbol{\lambda}^*)/\pi(\boldsymbol{\lambda}^{(k)})$. Hence, as expected, the algorithm tends to create more observations for regions with high posterior density.

When the posterior is the equilibrium distribution of the Metropolis–Hastings chain, Equation (5.6) guarantees the asymptotic convergence of the chain. The obvious question that arises is when is the time index k large enough so that we can consider the observations $\boldsymbol{\lambda}^{(k)}$ to follow the posterior distribution. In practice, this is often dealt with by plotting each component $\lambda_i^{(k)}$ as a function of k and looking for an index b after which the time series seems stationary. Then the b first observations, which are often called the *burn-in* of the chain, are discarded and the remaining $N - b$ observations can be considered as a sample from the posterior.

Another important issue with MCMC sampling is the speed of the *mixing* of the chain. Because the Metropolis–Hastings chain is, in most practical cases, constructed in such a way that the two consecutive observations $\boldsymbol{\lambda}^{(k)}$ and $\boldsymbol{\lambda}^{(k+1)}$ are correlated, one is likely to find the next observation close to the previous one. Due to this correlation, the chain explores one local part of the state space at a time and slowly moves around the global support of the posterior. If the speed at which it explores the posterior is very low, the chain is said to be mixing slowly. In such a case, one needs to have a very large number of observations to get a good idea about the global structure of the posterior. One way to monitor the mixing of the chain is to compute the acceptance rate of the Metropolis–Hastings jumps, that is the percentage of the proposals $\boldsymbol{\lambda}^*$ that are accepted in the algorithm. If the acceptance rate is very low, the proposed jumps are too large and the chain mostly stays still at its current position. On the other hand, if the acceptance rate is very large, the proposed jumps are often too small and again the chain moves around very slowly. There are theoretical results saying that the acceptance rate should be of the order of 20–40 % [23, Section 11.9]. Such optimal acceptance rates can be obtained by an appropriate choice of the proposal density $p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)})$. The general recommendation is to try to set the shape of the proposal density close to the shape of the posterior and then tune the scale of the proposals to obtain the recommended acceptance rates.

It is clear from the discussion above that the issues related to the convergence and mixing of the Metropolis–Hastings algorithm often make its use more of an art than exact science. Nevertheless, it is a well-understood, widely applicable sampling algorithm which often gives satisfactory results given that the number of samples N is large enough and at least some basic convergence checks are performed. For an overview of more advanced convergence analysis tools for MCMC sampling, see [8].

The issues related to the suboptimal mixing of the Metropolis–Hastings chain are partially resolved by using another popular MCMC algorithm called *Gibbs sampling*. Here the basic idea is to update each component of the Markov chain by sampling from the full posterior conditionals $p(\lambda_i|\boldsymbol{\lambda}_{-i}, \boldsymbol{y}, \boldsymbol{\alpha})$, where $\boldsymbol{\lambda}_{-i}$ is the vector $\boldsymbol{\lambda}$ without the ith component, $\boldsymbol{\lambda}_{-i} = \left[\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_p\right]^{\mathrm{T}}$. More precisely, the ith component of the new observation $\boldsymbol{\lambda}^{(k+1)}$ is sampled conditioning on the new values

of dimensions $1, \ldots, i-1$ and the previous values of dimensions $i+1, \ldots, p$, that is,

$$\lambda_i^{(k+1)} \sim p(\lambda_i | \lambda_1 = \lambda_1^{(k+1)}, \ldots, \lambda_{i-1} = \lambda_{i-1}^{(k+1)}, \lambda_{i+1} = \lambda_{i+1}^{(k)}, \ldots, \lambda_p = \lambda_p^{(k)}, \boldsymbol{y}, \boldsymbol{\alpha}).$$

The sampling is started by sampling $\lambda_1^{(k+1)}$ conditional on the old values of dimensions $2, \ldots, p$ followed by sampling $\lambda_2^{(k+1)}$ conditional on the sampled value of $\lambda_1^{(k+1)}$ and the old values of dimensions $3, \ldots, p$ and so forth. After p samplings, we have obtained the next full observation $\boldsymbol{\lambda}^{(k+1)}$. Note that in Gibbs sampling, all the jumps of the chain are always accepted.

When compared to Metropolis–Hastings sampling, Gibbs sampling has several attractive features. Firstly, there are no free parameters to tune making the algorithm easy to use for the end user. Secondly, the length of the jumps in each dimension are automatically adapted to the shape of the posterior while in Metropolis–Hastings sampling one needs to choose the shape of the proposal density $p(\boldsymbol{\lambda}^* | \boldsymbol{\lambda}^{(k)})$ to match the shape of the posterior. Theoretically, the price one has to pay for this flexibility is that one has to impose some mild regularity conditions on the posterior $p(\boldsymbol{\lambda} | \boldsymbol{y}, \boldsymbol{\alpha})$ in order to guarantee that it is the equilibrium distribution of the Gibbs sampler (see [31, Proposition 3.12]). In terms of computation, the applicability of Gibbs sampling is very much dependent on the full posterior conditionals $p(\lambda_i | \boldsymbol{\lambda}_{-i}, \boldsymbol{y}, \boldsymbol{\alpha})$. If these are available in closed form and happen to belong to some family of distributions for which efficient sampling algorithms are available, then Gibbs sampling is usually computationally fast and explores the posterior very effectively. On the other hand, if the conditionals are not available in closed form, one will have to resort to expensive numerical techniques in order to sample from them, which makes Gibbs sampling very slow.

Unfortunately, the likelihood of the unfolding problem given by Equation (4.1) is such that no choice of the prior will lead to posterior conditionals belonging to any of the well-known, standard families of probability densities. Hence, sampling from the conditionals is computationally challenging. Because of this, the computational experiments of Chapter 7 are conducted using the Metropolis–Hastings algorithm.

5.3 Prior Models

The choice of the prior $p(\boldsymbol{\lambda} | \boldsymbol{\alpha})$ in Bayes' formula (5.2) is crucial for successful unfolding. The reason for this is that due to the ill-posedness of the problem, the information contained in the likelihood $p(\boldsymbol{y} | \boldsymbol{\lambda})$ alone is too vague for us to obtain physically plausible solutions with reasonable uncertainties. Therefore, additional information about plausible solutions needs to be injected into the problem using the prior. For an overview of the most commonly used prior models in Bayesian inference for inverse problems, see [31, Section 3.3]. In the case of unfolding, the main challenge is to pick the prior in such a way that it is restrictive enough to stabilize the problem but uninformative enough not to bias any particular solution. In this section, we will concentrate on the choice of the family of prior densities $\{p(\boldsymbol{\lambda} | \boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$ parametrized by some hyperparameters $\boldsymbol{\alpha}$. In fully Bayesian thinking,

the free hyperparameters $\boldsymbol{\alpha}$ should then be chosen by the analyst to reflect their subjective degree of belief in the regularization imposed by the prior.

It is often the case that the prior is in fact not a density in the sense that it cannot be normalized to integrate into unity. Such priors are called *improper priors*. For example, the uniform non-negativity prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) = p(\boldsymbol{\lambda}) \propto 1_{\mathbb{R}^p_+}(\boldsymbol{\lambda})$ is clearly improper, since $\int_{\mathbb{R}^p_+} 1 \,\mathrm{d}\boldsymbol{\lambda} = \nu(\mathbb{R}^p_+) = \infty$. In Bayesian inference, improper priors are not considered a problem as long as they define a proper posterior density. Let us use $q(\boldsymbol{x})$ to denote an unnormalized pdf of the random variable $\boldsymbol{x}$. Hence, the possibly improper prior $q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ should be chosen in such a way that the unnormalized posterior $q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) = p(\boldsymbol{y}|\boldsymbol{\lambda})q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ is normalizable, i.e., $\int q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda} \in (0, \infty)$, and hence defines a proper posterior density. The following proposition says that in the case of unfolding this is true for most reasonable priors.

Proposition 5.3. *Consider the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$ given by* (4.1) *and assume that $K_{ij} > 0$. Let $q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \geq 0$ be an unnormalized, possibly improper prior pdf for $\boldsymbol{\lambda}$ depending on hyperparameters $\boldsymbol{\alpha}$. Then the unnormalized posterior $q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) = p(\boldsymbol{y}|\boldsymbol{\lambda})q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ is normalizable, that is $\int_{\mathbb{R}^p_+} q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda} \in (0, \infty)$, if the prior pdf is bounded and its support is a subset of $\mathbb{R}^p_+$ with a strictly positive Lebesgue measure.*

Proof. Let us first show the lower bound $\int_{\mathbb{R}^p_+} q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda} > 0$. Under the assumption $K_{ij} > 0$, the likelihood is strictly positive on $\mathbb{R}^p_+$, that is $p(\boldsymbol{y}|\boldsymbol{\lambda}) > 0$, except for the origin $\boldsymbol{\lambda} = \mathbf{0}$ where the likelihood is either undefined or vanishes depending on the value of $\boldsymbol{y}$. We have

$$\begin{aligned}\int_{\mathbb{R}^p_+} q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda} &= \int_{\mathbb{R}^p_+} p(\boldsymbol{y}|\boldsymbol{\lambda})q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda} \\ &= \int\limits_{\mathrm{supp}(q(\boldsymbol{\lambda}|\boldsymbol{\alpha}))\setminus\{\mathbf{0}\}} p(\boldsymbol{y}|\boldsymbol{\lambda})q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda}.\end{aligned}$$

Since $p(\boldsymbol{y}|\boldsymbol{\lambda})q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) > 0$ on $\mathrm{supp}(q(\boldsymbol{\lambda}|\boldsymbol{\alpha}))\setminus\{\mathbf{0}\}$ and $\nu(\mathrm{supp}(q(\boldsymbol{\lambda}|\boldsymbol{\alpha}))) > 0$, we conclude that

$$\int_{\mathbb{R}^p_+} q(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\lambda} > 0.$$

To show that this integral is finite, we first show that

$$\int_{\mathbb{R}^p_+} p(\boldsymbol{y}|\boldsymbol{\lambda}) \,\mathrm{d}\boldsymbol{\lambda} < \infty. \tag{5.10}$$

To this end, consider the set $\{\boldsymbol{\lambda} \in \mathbb{R}^p_+ : \|\boldsymbol{\lambda}\|_\infty > M\}$, where $M \geq 0$ is a finite constant. By increasing M, we can make $\sum_j K_{ij}\lambda_j$ arbitrarily large on this set, given that $K_{ij} > 0$. Hence, for large enough M, we have the upper bound $\left(\sum_j K_{ij}\lambda_j\right)^{y_i} \leq e^{\frac{1}{2}\sum_j K_{ij}\lambda_j}$. We can then write

$$\int_{\mathbb{R}^p_+} p(\boldsymbol{y}|\boldsymbol{\lambda}) \,\mathrm{d}\boldsymbol{\lambda} = \int\limits_{\{\boldsymbol{\lambda}\in\mathbb{R}^p_+ : \|\boldsymbol{\lambda}\|_\infty \leq M\}} p(\boldsymbol{y}|\boldsymbol{\lambda}) \,\mathrm{d}\boldsymbol{\lambda} \; + \int\limits_{\{\boldsymbol{\lambda}\in\mathbb{R}^p_+ : \|\boldsymbol{\lambda}\|_\infty > M\}} p(\boldsymbol{y}|\boldsymbol{\lambda}) \,\mathrm{d}\boldsymbol{\lambda}.$$

Here the first term is finite since both $p(\boldsymbol{y}|\boldsymbol{\lambda})$ and the domain of integration are bounded. For the second term, we have the upper bound

$$\begin{aligned}
\int\limits_{\{\boldsymbol{\lambda}\in\mathbb{R}_+^p:\|\boldsymbol{\lambda}\|_\infty>M\}} p(\boldsymbol{y}|\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} &= \int\limits_{\{\boldsymbol{\lambda}\in\mathbb{R}_+^p:\|\boldsymbol{\lambda}\|_\infty>M\}} \prod_i \frac{\left(\sum_j K_{ij}\lambda_j\right)^{y_i}}{y_i!} e^{-\sum_j K_{ij}\lambda_j}\,\mathrm{d}\boldsymbol{\lambda} \\
&\leq \int\limits_{\{\boldsymbol{\lambda}\in\mathbb{R}_+^p:\|\boldsymbol{\lambda}\|_\infty>M\}} \prod_i \frac{1}{y_i!} e^{-\frac{1}{2}\sum_j K_{ij}\lambda_j}\,\mathrm{d}\boldsymbol{\lambda} \\
&\leq \left(\prod_i \frac{1}{y_i!}\right) \int_{\mathbb{R}_+^p} \prod_i e^{-\frac{1}{2}\sum_j K_{ij}\lambda_j}\,\mathrm{d}\boldsymbol{\lambda} \\
&= \left(\prod_i \frac{1}{y_i!}\right) \int_{\mathbb{R}_+^p} e^{-\frac{1}{2}\sum_{ij} K_{ij}\lambda_j}\,\mathrm{d}\boldsymbol{\lambda}.
\end{aligned}$$

Let us denote $\varepsilon_j = \sum_i K_{ij}$. By the assumption $K_{ij} > 0$, we have $\varepsilon_j > 0$. Furthermore, we can write $\sum_{ij} K_{ij}\lambda_j = \sum_j \left(\sum_i K_{ij}\right)\lambda_j = \sum_j \varepsilon_j\lambda_j$. Hence

$$\begin{aligned}
\int\limits_{\{\boldsymbol{\lambda}\in\mathbb{R}_+^p:\|\boldsymbol{\lambda}\|_\infty>M\}} p(\boldsymbol{y}|\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} &\leq \left(\prod_i \frac{1}{y_i!}\right) \int_{\mathbb{R}_+^p} \prod_j e^{-\frac{1}{2}\varepsilon_j\lambda_j}\,\mathrm{d}\boldsymbol{\lambda} \\
&= \left(\prod_i \frac{1}{y_i!}\right) \left(\prod_j \int_0^\infty e^{-\frac{1}{2}\varepsilon_j\lambda_j}\,\mathrm{d}\lambda_j\right) \\
&= \left(\prod_i \frac{1}{y_i!}\right) \left(\prod_j \frac{2}{\varepsilon_j}\right) < \infty.
\end{aligned}$$

Hence, we have shown (5.10). Using the boundedness of $q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$, it then follows for the unnormalized posterior $q(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})$ that

$$\begin{aligned}
\int_{\mathbb{R}_+^p} q(\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\lambda} &= \int_{\mathbb{R}_+^p} p(\boldsymbol{y}|\boldsymbol{\lambda}) q(\boldsymbol{\lambda}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\lambda} \\
&\leq \sup_{\boldsymbol{\lambda}\in\mathbb{R}_+^p} q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \int_{\mathbb{R}_+^p} p(\boldsymbol{y}|\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} < \infty
\end{aligned}$$

since both factors in the product are finite. □

Proposition 5.3 gives us the freedom of choosing any bounded function supported on $\mathbb{R}_+^p$ as the prior. It is often desirable to choose the prior in such a way that it is as objective about the unknown as possible. Such priors are called *uninformative priors*. The uniform non-negativity prior

$$p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \propto 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \tag{5.11}$$

is a prototypical example of an uninformative prior since this prior regards all non-negative solutions as equally likely. However, as noted in Section 3.3, the uniform

prior is uninformative only in the current metric and becomes informative after a change of variables. Hence, care should be taken when interpreting such a prior and the resulting posterior.

Analogously with the case of frequentist maximum likelihood estimation, the uniform prior is not sufficient to regularize the ill-posed unfolding problem. In the frequentist framework, this resulted in oscillating point estimates, while in the Bayesian analysis this is evident in the large posterior uncertainty about the solution. Since we expect in most cases the intensity function of the true Poisson process to be a smooth function, the corresponding true histogram should also be smooth in the sense that we expect the values of adjacent histogram bins to be close to one another. Hence, we should choose a *smoothness prior* which incorporates this a priori smoothness information into the unfolding problem.

Among the priors enforcing the smoothness of the solution, the most commonly encountered example is the Gaussian smoothness prior of which we use the non-negativity constrained version given by

$$\begin{aligned} p(\boldsymbol{\lambda}|\alpha) &\propto 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \exp\left(-\alpha \|\boldsymbol{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\|^2\right) \\ &= 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \exp\left(-\alpha (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\right), \quad \alpha > 0. \end{aligned} \tag{5.12}$$

This type of a prior penalizes for some discrete differential operator $\boldsymbol{L}$ of the solution $\boldsymbol{\lambda}$. When desired, $\boldsymbol{\lambda}_0$ can be used to bias the inference towards some specific solution. The hyperparameter α controls the scale and hence the strength of the prior and, in the fully Bayesian paradigm, has to be hand-picked by the analyst. The larger the value of α, the stronger the regularization imposed by the prior. Using Proposition 5.3, we know that (5.12) defines a proper posterior density for all choices of $\boldsymbol{L}$. Hence, standard Bayesian inference can be carried out for any choice of the matrix $\boldsymbol{L}$ among which we consider the cases where $\boldsymbol{L}$ is the identity matrix, $\boldsymbol{L} = \boldsymbol{I}$, or one of the finite-difference matrices defined in Equations (4.34)-(4.37) of Section 4.2.2. In the standard Bayesian framework, it is a matter of taste if $\boldsymbol{L}$ and $\boldsymbol{\lambda}_0$ are regarded as hyperparameters of the prior in addition to α. However, in the subsequent empirical Bayes formulation, we consider $\boldsymbol{L}$ and $\boldsymbol{\lambda}_0$ to have fixed, known values but fit the scale α to the data $\boldsymbol{y}$. Because of this, we will already here think of (5.12) as being parametrized only by α.

It could also be the case that instead of smooth solutions, we expect to have an unfolded histogram consisting of blocks of small variation and possibly discontinuous jumps between the blocks. In that case, the more appropriate prior would be the *total variation prior*

$$p(\boldsymbol{\lambda}|\alpha) \propto 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \exp\left(-\alpha \|\boldsymbol{L}_1^1 \boldsymbol{\lambda}\|_1\right), \quad \alpha > 0,$$

where $\boldsymbol{L}_1^1$ is given by (4.34). If on the other hand, we expect to see a small number of counts in most bins of the true histogram with a few outstanding bins, we could use the l^1 *prior*

$$p(\boldsymbol{\lambda}|\alpha) \propto 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \exp\left(-\alpha \|\boldsymbol{\lambda}\|_1\right), \quad \alpha > 0.$$

Nevertheless, in most cases in high energy physics, the Gaussian smoothness prior (5.12) is the appropriate physical choice to represent our a priori knowledge about the solution.

Chapter 6

Empirical Bayes Unfolding

The main problem with the fully Bayesian approach to unfolding discussed in the previous chapter is that the solution depends on the subjective choice of the hyperparameter $\boldsymbol{\alpha}$ of the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$. In this chapter, we explain how to use empirical Bayes techniques for choosing the prior objectively based on the data. The basic idea of empirical Bayes, that is, marginal maximum likelihood estimation of the hyperparameter $\boldsymbol{\alpha}$, is first explained in Section 6.1. We then derive a variant of the EM algorithm for estimating $\boldsymbol{\alpha}$ in Section 6.2 and finally explain how to implement the algorithm for the Gaussian smoothness prior in Section 6.3.

6.1 Parametric Empirical Bayes for Unfolding

Up to now we have defined the Bayesian solution of the unfolding problem via a sample from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha})$ and we have discussed how to select the family of priors $\{p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$ parametrized by some hyperparameters $\boldsymbol{\alpha}$. The problem that remains is to find a way to select an appropriate value for $\boldsymbol{\alpha}$. The hyperparameter $\boldsymbol{\alpha}$ is analogous to the regularization parameter δ in Tikhonov regularization and the truncation index t in TSVD in the frequentist paradigm and can have a significant effect on the solution of the unfolding problem.

In strict Bayesian thinking, the value of $\boldsymbol{\alpha}$ should be selected by the analyst to reflect their subjective a priori knowledge about the solution. The problem is that when such abstract priors as for example the Gaussian smoothness prior of Equation (5.12) are used, it is very difficult to make an informed decision about the hyperparameters. When we are uncertain about the value of the hyperparameters $\boldsymbol{\alpha}$, the correct Bayesian procedure is to consider them part of the inference problem. Hence, we would define a hyperprior $p(\boldsymbol{\alpha}|\boldsymbol{\beta})$ for $\boldsymbol{\alpha}$ and consider the posterior

$$p(\boldsymbol{\lambda}, \boldsymbol{\alpha}|\boldsymbol{y}, \boldsymbol{\beta}) \propto p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{\beta}).$$

Since, at the end of the day, we are not interested in the value of the nuisance parameter $\boldsymbol{\alpha}$, we integrate it out giving us the posterior

$$p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\beta}) = \int p(\boldsymbol{\lambda}, \boldsymbol{\alpha}|\boldsymbol{y}, \boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\alpha} \propto \int p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\alpha}.$$

As we see, the problem here is that the hyperprior itself often depends on some additional hyperparameters $\boldsymbol{\beta}$ which are often even more difficult to choose than $\boldsymbol{\alpha}$. Hence, by doing this, we have done nothing more than moved the problem one level higher up. There are arguments saying that the outcome of the inference is less sensitive to the choice of the hyperprior $p(\boldsymbol{\alpha}|\boldsymbol{\beta})$ than to the choice of the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ but nevertheless the conceptual problem remains. What is more, the subjectivity of the results, which is inherent in standard Bayesian inference, is not too well suited to experimental natural science aiming at reporting the outcome of the measurement in as objective way as possible.

Empirical Bayes methods provide an essentially non-Bayesian way of selecting the hyperparameters $\boldsymbol{\alpha}$ objectively based on the data $\boldsymbol{y}$. To a large extent, this solves the issue of subjectivity of Bayesian unfolding as the only subjective choice made in the procedure is the choice of the family of priors $\{p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$. In addition, this results in an automatic unfolding machinery with no free parameters to fine-tune except for the choice of the family of priors[1].

The key idea of *parametric empirical Bayes* [10, Chapter 5] is to regard the marginal $p(\boldsymbol{y}|\boldsymbol{\alpha})$ in Bayes' rule (5.2) as a parametric model for the observations $\boldsymbol{y}$ and use any of the standard tools in frequentist statistics to find a point estimate $\hat{\boldsymbol{\alpha}}$ for the hyperparameters $\boldsymbol{\alpha}$. This point estimate is then plugged into Bayes' rule (5.2) to obtain the posterior

$$p(\boldsymbol{\lambda}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}})}{p(\boldsymbol{y}|\hat{\boldsymbol{\alpha}})}.$$

The inferences based on this posterior are then regarded as the solution to the unfolding problem. The term "parametric" appears in the name of the method since we assume that the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ can be parametrized using the hyperparameters $\boldsymbol{\alpha}$. The more general alternative is *nonparametric empirical Bayes* where no parametric form is assumed for the prior. Since, in our case, we actually take advantage of the restrictive parametric form of the prior by using it to regularize the problem, we will not pursue the nonparametric form of empirical Bayes further in here.

In terms of its interpretation, empirical Bayes unfolding can be seen as a combination of frequentist and Bayesian inference: the unknown hyperparameter $\boldsymbol{\alpha}$ is fitted to the data $\boldsymbol{y}$ using a frequentist point estimator and the estimated value $\hat{\boldsymbol{\alpha}}$ is then used to obtain the Bayesian posterior. In fact, if we summarize the information contained in the posterior using its mean, the resulting point estimator $\hat{\boldsymbol{\lambda}} = \mathrm{E}[\boldsymbol{\lambda}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}]$ admits a frequentist interpretation via statistical decision theory. Namely, the posterior mean $\mathrm{E}[\boldsymbol{\lambda}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}]$ is the decision rule that minimizes the Bayes risk for the prior $p(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}})$ and the squared error loss function [62, Section 3.2]. Since in empirical Bayes the prior $p(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}})$ is inferred from the data $\boldsymbol{y}$, this can be seen as a fully frequentist procedure. When credible intervals of the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \hat{\boldsymbol{\alpha}})$ are used, interpretation in terms of decision theory is no longer possible. Nevertheless, we still retain the advantage of selecting the prior $p(\boldsymbol{\lambda}|\hat{\boldsymbol{\alpha}})$ objectively

[1] In practice, there could still be parameters related to the numerical algorithms involved but their fine-tuning is more of a technicality than a conceptual issue. An example of such a parameter is the scale of the proposal density $p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)})$ in Metropolis–Hastings sampling.

based on the data $\boldsymbol{y}$. Hence, the only Bayesian element present in the procedure is the use of the posterior probability density over the unknown $\boldsymbol{\lambda}$ to characterize our degree of belief about its true value. This is in stark contrast with fully Bayesian inference where the essential defining characteristics are both the use of densities over $\boldsymbol{\lambda}$ *and* the use of the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ to quantify our subjective knowledge about the unknown before the measurement took place.

6.2 Marginal Maximum Likelihood Estimation with the MCEM Algorithm

While in principle any frequentist point estimator could be used when estimating the hyperparameters $\boldsymbol{\alpha}$ in the marginal $p(\boldsymbol{y}|\boldsymbol{\alpha})$, the most natural way to proceed is to regard $L(\boldsymbol{\alpha};\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\alpha})$ as the likelihood of the hyperparameters $\boldsymbol{\alpha}$ and then find the point estimator $\hat{\boldsymbol{\alpha}}$ as the maximizer of this marginal likelihood. We will hence use the *marginal maximum likelihood estimator* (MMLE) defined by

$$\hat{\boldsymbol{\alpha}}_{\text{MMLE}} = \arg\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha};\boldsymbol{y}) = \arg\max_{\boldsymbol{\alpha}} p(\boldsymbol{y}|\boldsymbol{\alpha})$$

to estimate the hyperparameters $\boldsymbol{\alpha}$. Using Equation (5.3), the marginal likelihood is given by

$$L(\boldsymbol{\alpha};\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\lambda}.$$

Since this is an intractable integral, we could use the law of large numbers (Theorem A.13) to find its MC approximation based on a sample from the prior

$$L(\boldsymbol{\alpha};\boldsymbol{y}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{y}|\boldsymbol{\lambda}^{(i)}), \quad \boldsymbol{\lambda}^{(i)} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{\lambda}|\boldsymbol{\alpha}). \tag{6.1}$$

Unfortunately, this approach does not work well in practice. The reason for this is that in the high-dimensional sample space, most of the sampled values of $\boldsymbol{\lambda}$ fall on regions of the space where the likelihood is numerically zero. This is true even for very reasonable priors $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$. Hence, an extremely large sample from the prior would be required to get even a rough idea about the value of the likelihood $L(\boldsymbol{\alpha};\boldsymbol{y})$.

To solve this problem, we use the EM algorithm described in Section 4.1.1 to find the maximum of the marginal likelihood $L(\boldsymbol{\alpha};\boldsymbol{y})$. To do this, we regard the means of the true histogram $\boldsymbol{\lambda}$ as unobserved latent variables. Hence, the complete data are $(\boldsymbol{y},\boldsymbol{\lambda})$ with the likelihood function $L(\boldsymbol{\alpha};\boldsymbol{y},\boldsymbol{\lambda}) = p(\boldsymbol{y},\boldsymbol{\lambda}|\boldsymbol{\alpha})$ for the hyperparameters $\boldsymbol{\alpha}$. The two likelihood functions are related by

$$L(\boldsymbol{\alpha};\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y},\boldsymbol{\lambda}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\lambda} = \int L(\boldsymbol{\alpha};\boldsymbol{y},\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda},$$

which corresponds to Equation (4.8) in our general formulation of the EM algorithm. Denoting the complete-data log-likelihood by $l(\boldsymbol{\alpha};\boldsymbol{y},\boldsymbol{\lambda}) = \log p(\boldsymbol{y},\boldsymbol{\lambda}|\boldsymbol{\alpha})$, it follows that on the kth E-step of the algorithm, we compute the conditional expectation

$$Q(\boldsymbol{\alpha};\boldsymbol{\alpha}^{(k)}) = \mathrm{E}\big[l(\boldsymbol{\alpha};\boldsymbol{y},\boldsymbol{\lambda})|\boldsymbol{y},\boldsymbol{\alpha}^{(k)}\big] = \mathrm{E}\big[\log p(\boldsymbol{y},\boldsymbol{\lambda}|\boldsymbol{\alpha})|\boldsymbol{y},\boldsymbol{\alpha}^{(k)}\big].$$

Since $p(\boldsymbol{y}, \boldsymbol{\lambda}|\boldsymbol{\alpha}) = p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ and we are only interested in how Q depends on $\boldsymbol{\alpha}$, we can write

$$\begin{aligned} Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)}) &= \mathrm{E}\big[\log p(\boldsymbol{\lambda}|\boldsymbol{\alpha})|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)}\big] + \mathrm{E}\big[\log p(\boldsymbol{y}|\boldsymbol{\lambda})|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)}\big] \\ &= \mathrm{E}\big[\log p(\boldsymbol{\lambda}|\boldsymbol{\alpha})|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)}\big] + \text{const} \\ &= \int p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)}) \log p(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\lambda} + \text{const}. \end{aligned}$$

This is again an intractable integral, so we use the MC approximation of Equation (5.7) to find

$$Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)}) \approx \frac{1}{N} \sum_{i=1}^{N} \log p(\boldsymbol{\lambda}^{(i)}|\boldsymbol{\alpha}) + \text{const}, \quad \boldsymbol{\lambda}^{(i)} \sim p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)}),$$

where the sample $\{\boldsymbol{\lambda}^{(i)}\}_{i=1}^{N}$ is produced using the Metropolis–Hastings algorithm described in Section 5.2. Thus, on the E-step of the algorithm, we sample N observations from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)})$ computed for the current iterate of the hyperparameters. The arithmetic mean of the values of the log-prior corresponding to this sample is then used to approximate the value of $Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)})$ up to a constant which does not depend on $\boldsymbol{\alpha}$. On the subsequent M-step of the algorithm, the approximate value of $Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)})$ is maximized with respect to $\boldsymbol{\alpha}$. The difficulty of this maximization depends on the choice of the family of priors $\{p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$. We show below that for the Gaussian smoothness prior (5.12), the maximization can be carried out analytically, while for more complicated choices of $\{p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$, it might be necessary to find the maximum numerically using standard nonlinear optimization algorithms.

Since we need to resort to MC integration when computing the conditional expectation, the iteration outlined above is not exactly the EM iteration described in Section 4.1.1 but instead its stochastic Monte Carlo version. This extension of the original EM algorithm was first proposed by Wei and Tanner in [60], who called it the *Monte Carlo EM* (MCEM) *algorithm*. See also [43, Section 6.3] for a review of the literature on the MCEM algorithm.

To summarize the discussion above, the MCEM algorithm for finding the marginal maximum likelihood estimator $\hat{\boldsymbol{\alpha}}_{\mathrm{MMLE}}$ of the hyperparameters $\boldsymbol{\alpha}$ consists of the following iteration:

1. Pick some initial guess $\boldsymbol{\alpha}^{(0)}$ and set $k = 0$.
2. E-step:
   (a) Sample $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \ldots, \boldsymbol{\lambda}^{(N)}$ from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}^{(k)})$.
   (b) Compute:
   $$\widetilde{Q}(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)}) = \frac{1}{N} \sum_{i=1}^{N} \log p(\boldsymbol{\lambda}^{(i)}|\boldsymbol{\alpha}). \tag{6.2}$$

3. M-step: Set $\boldsymbol{\alpha}^{(k+1)} = \arg\max_{\boldsymbol{\alpha}} \widetilde{Q}(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)})$.

4. Set $k \leftarrow k + 1$.

5. If some stopping rule $\mathcal{C}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\alpha}^{(k-1)}, \ldots, \boldsymbol{\alpha}^{(0)})$ is satisfied, set $\hat{\boldsymbol{\alpha}}_{\text{MMLE}} = \boldsymbol{\alpha}^{(k)}$ and terminate the iteration, else go to step 2.

Replacing the E-step with its MC approximation complicates both the theoretical and practical convergence analysis of the EM algorithm. Firstly, the random fluctuations of the MC estimator invalidate the monotonicity of the original EM algorithm (see Theorem 4.3). Secondly, the stopping rule $\mathcal{C}(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\alpha}^{(k-1)}, \ldots, \boldsymbol{\alpha}^{(0)})$ should consider more than just the latest iteration of the algorithm to see if the iterates appear to fluctuate around some central value before claiming convergence. Nevertheless, despite these complications, the MCEM algorithm has been successfully applied to various problems of practical interest, see e.g. [7, 39].

The MCEM algorithm has a rather intuitive interpretation. First, on the E-step, we use the current iterate $\boldsymbol{\alpha}^{(k)}$ to produce a sample of $\boldsymbol{\lambda}$'s from the posterior. Since this sample summarizes our current understanding of $\boldsymbol{\lambda}$, we then tune the prior by changing $\boldsymbol{\alpha}$ on the M-step to match this sample as well as possible and the $\boldsymbol{\alpha}$ that matches the posterior sample the best will then become the next iterate $\boldsymbol{\alpha}^{(k+1)}$.

There are two reasons why the MCEM algorithm is numerically a lot more stable than the first MC approximation (6.1) that we tried to use to find the MMLE. Firstly, in MCEM, the $\boldsymbol{\lambda}$'s are sampled from the posterior and hence most of them are reasonable true histograms. This means that they should also lie within the bulk of the prior probability density making $\widetilde{Q}$ in Equation (6.2) well behaved. On the contrary, in Equation (6.1), the sample is generated from the prior resulting mostly in very unlikely true histograms. Secondly, the sum in (6.1) is over plain densities instead of log-densities which is the case in (6.2). This makes the computations in MCEM a lot more robust against small pdf values.

6.3 Empirical Bayes Unfolding with the Gaussian Smoothness Prior

In the treatment above, we implicitly assumed that the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ used in empirical Bayes unfolding is proper. The reason for this is that if an improper prior $q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ was used, then instead of the nominal version of Bayes' theorem

$$p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha})}$$

the posterior is given by

$$p(\boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{\alpha}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})q(\boldsymbol{\lambda}|\boldsymbol{\alpha})}{q(\boldsymbol{y}|\boldsymbol{\alpha})}.$$

Since the posterior has to be normalized, the denominator is given by

$$q(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y}|\boldsymbol{\lambda}) q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\lambda}.$$

The problem is that this can no longer be interpreted as the marginal of $\boldsymbol{y}$ given $\boldsymbol{\alpha}$. In fact, when $q(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ is improper, $q(\boldsymbol{y}|\boldsymbol{\alpha})$ does not even define a proper probability density, that is,

$$\begin{aligned}
\sum_{\boldsymbol{y}} q(\boldsymbol{y}|\boldsymbol{\alpha}) &= \sum_{\boldsymbol{y}} \int p(\boldsymbol{y}|\boldsymbol{\lambda}) q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\lambda} \\
&= \int \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{\lambda}) q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\lambda} \\
&= \int q(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\lambda} = \infty.
\end{aligned}$$

Here the interchange of summation and integration is allowed by the monotone convergence theorem. As a result, it is not clear if the value of $\boldsymbol{\alpha}$ that maximizes $q(\boldsymbol{y}|\boldsymbol{\alpha})$ for the observed data $\boldsymbol{y}$ is a reasonable estimator of the hyperparameters and, even if it was, it is not clear if the EM algorithm can, in this case, be used for finding the maximum. Because of these complications, we will only consider proper priors in empirical Bayes unfolding.

In the computational experiments of this thesis, we use the Gaussian smoothness prior with the non-negativity constraint defined by Equation (5.12). The following proposition establishes a sufficient condition for this prior to be proper.

Proposition 6.1. *The Gaussian smoothness prior with the non-negativity constraint defined by Equation* (5.12) *is proper if the $k \times p$ matrix $\boldsymbol{L}$ with $k \geq p$ has full column rank.*

Proof. Consider the unnormalized Gaussian prior

$$q(\boldsymbol{\lambda}|\alpha) = \exp\left(-\alpha \|\boldsymbol{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\|^2\right) = \exp\left(-\alpha (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\right).$$

This is a proper density if and only if the $p \times p$ matrix $\boldsymbol{\Lambda} = \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L}$ is positive definite. Since $\boldsymbol{\Lambda}$ is the Gram matrix of $\boldsymbol{L}$, it is positive semidefinite and $\operatorname{rank}(\boldsymbol{\Lambda}) = \operatorname{rank}(\boldsymbol{L})$. Hence, the statement that $\boldsymbol{\Lambda}$ is positive definite is equivalent to having $\operatorname{rank}(\boldsymbol{\Lambda}) = p$ and it follows that full column rank of $\boldsymbol{L}$ is equivalent to $q(\boldsymbol{\lambda}|\alpha)$ being a proper density. Since the prior (5.12) is the truncation of $q(\boldsymbol{\lambda}|\alpha)$ to $\mathbb{R}^p_+$, having $\operatorname{rank}(\boldsymbol{L}) = p$ then implies that (5.12) is proper. □

This proposition hence guarantees that the smoothness prior is proper when we set the $\boldsymbol{L}$ to be the identity $\boldsymbol{I}$, the full-rank first-order finite-difference matrix $\boldsymbol{L}_2^1$ defined by Equation (4.35) or the full-rank second-order finite-difference matrix $\boldsymbol{L}_2^2$ of Equation (4.37). In cases where $\boldsymbol{L}$ is column-rank deficient, it depends on the orientation of the kernel of $\boldsymbol{L}$ whether or not (5.12) can be normalized. In particular,

due to the truncation in (5.12), it is possible to have a proper prior even in cases where $q(\boldsymbol{\lambda}|\alpha)$ in the proof above does not define a proper Gaussian density.

Unfortunately, for the two remaining prototypical choices of $\boldsymbol{L}$, namely $\boldsymbol{L}_1^1$ of Equation (4.34) and $\boldsymbol{L}_1^2$ of Equation (4.36), the Gaussian smoothness prior is improper. To see this, let us find the kernel of $\boldsymbol{L}_1^1$. The condition $\boldsymbol{L}_1^1\boldsymbol{\lambda} = \mathbf{0}$ is equivalent to

$$\begin{cases} -\lambda_1 + \lambda_2 = 0 \\ -\lambda_2 + \lambda_3 = 0 \\ \quad\vdots \\ -\lambda_{p-1} + \lambda_p = 0 \end{cases} \Leftrightarrow \begin{cases} \lambda_1 = \lambda_2 \\ \lambda_2 = \lambda_3 \\ \quad\vdots \\ \lambda_{p-1} = \lambda_p \end{cases} \Leftrightarrow \quad \lambda_1 = \lambda_2 = \ldots = \lambda_p.$$

Hence, $\ker(\boldsymbol{L}_1^1) = \{\boldsymbol{\lambda} \in \mathbb{R}^p : \lambda_1 = \lambda_2 = \ldots = \lambda_p\}$. Now, let us consider the line $\boldsymbol{\lambda} = \boldsymbol{\lambda}^* + t\boldsymbol{v}$, $t \in \mathbb{R}$, where $\boldsymbol{v} \in \ker(\boldsymbol{L}_1^1)$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^p$. It follows that for any t, $\boldsymbol{L}_1^1\boldsymbol{\lambda} = \boldsymbol{L}_1^1\boldsymbol{\lambda}^*$ meaning that $q(\boldsymbol{\lambda}|\alpha)$ is constant on this line. Since $\boldsymbol{\lambda}^*$ was arbitrary, $q(\boldsymbol{\lambda}|\alpha)$ has a constant value on any line with direction vector $\boldsymbol{v} \in \ker(\boldsymbol{L}_1^1)\backslash\{\mathbf{0}\}$. Since $\ker(\boldsymbol{L}_1^1) \cap \mathbb{R}_+^p \neq \{\mathbf{0}\}$, we can choose a non-zero element of $\ker(\boldsymbol{L}_1^1) \cap \mathbb{R}_+^p$ as the direction vector $\boldsymbol{v}$ and hence deduce that $\int_{\mathbb{R}_+^p} q(\boldsymbol{\lambda}|\alpha)\,\mathrm{d}\boldsymbol{\lambda} = \infty$, when $\boldsymbol{L} = \boldsymbol{L}_1^1$.

Similarly, for $\boldsymbol{L}_1^2$, we find

$$\boldsymbol{L}_1^2\boldsymbol{\lambda} = \mathbf{0} \Leftrightarrow \begin{cases} \lambda_1 - 2\lambda_2 + \lambda_3 = 0 \\ \lambda_2 - 2\lambda_3 + \lambda_4 = 0 \\ \quad\vdots \\ \lambda_{p-2} - 2\lambda_{p-1} + \lambda_p = 0 \end{cases}.$$

Setting $\boldsymbol{\lambda} = t\mathbf{1}, t \in \mathbb{R}$, where $\mathbf{1}$ is the $p \times 1$ vector of ones, satisfies these equations for all t. Hence, $\ker(\boldsymbol{L}_1^2) \supset \ker(\boldsymbol{L}_1^1)$ and we can use the argument above to deduce that $\int_{\mathbb{R}_+^p} q(\boldsymbol{\lambda}|\alpha)\,\mathrm{d}\boldsymbol{\lambda} = \infty$, when $\boldsymbol{L} = \boldsymbol{L}_1^2$.

In contrast to standard Bayesian inference, in empirical Bayes unfolding, we need to be able to compute the normalization constant of the Gaussian smoothness prior, at least up to its dependence on α. The reason for this is that on the M-step of the MCEM algorithm we need to find the maximum of

$$\widetilde{Q}(\alpha; \alpha^{(k)}) = \frac{1}{N}\sum_{i=1}^{N} \log p(\boldsymbol{\lambda}^{(i)}|\alpha)$$

with respect to α. Since α appears on the normalization coefficient of $p(\boldsymbol{\lambda}|\alpha)$ we can no longer consider the unnormalized prior. When $p(\boldsymbol{\lambda}|\alpha)$ is the standard multivariate Gaussian supported on $\mathbb{R}^p$, finding the α-dependence of the normalization coefficient is trivial. However, when we impose the non-negativity constraint, we are in fact dealing with the truncation of this multivariate Gaussian to $\mathbb{R}_+^p$. In the general case, the normalization coefficient of a truncated multivariate Gaussian cannot be computed in closed form. Also, its numerical computation is challenging when the number of dimensions p is large.

Fortunately, if we set $\boldsymbol{\lambda}_0 = \mathbf{0}$, i.e., we are not interested in biasing any particular solution in the prior, we can find the α dependence of the normalization coefficient analytically. To this end, consider the corresponding truncated Gaussian

$$p(\boldsymbol{\lambda}|\alpha) = 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) C(\alpha) \exp(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2).$$

Here the normalization coefficient $C(\alpha)$ is given by

$$C(\alpha) = \frac{1}{\int_{\mathbb{R}_+^p} \exp(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2)\, \mathrm{d}\boldsymbol{\lambda}}.$$

When $\boldsymbol{L}$ has full column rank, we know that the integral in the denominator is finite. By making the change of variables $\boldsymbol{\lambda}^* = \sqrt{\alpha}\boldsymbol{\lambda}$, we can write

$$\int_{\mathbb{R}_+^p} \exp(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2)\, \mathrm{d}\boldsymbol{\lambda} = \alpha^{-p/2} \int_{\mathbb{R}_+^p} \exp(-\|\boldsymbol{L}\boldsymbol{\lambda}^*\|^2)\, \mathrm{d}\boldsymbol{\lambda}^*.$$

Hence, we can write the prior $p(\boldsymbol{\lambda}|\alpha)$ as follows

$$p(\boldsymbol{\lambda}|\alpha) = 1_{\mathbb{R}_+^p}(\boldsymbol{\lambda}) \alpha^{p/2} \frac{\exp(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2)}{\int_{\mathbb{R}_+^p} \exp(-\|\boldsymbol{L}\boldsymbol{\lambda}^*\|^2)\, \mathrm{d}\boldsymbol{\lambda}^*}.$$

Here the difficult integral in the denominator does not depend on $\boldsymbol{\alpha}$. Hence, when this pdf is considered as a function of α, we have

$$p(\boldsymbol{\lambda}|\alpha) \propto \alpha^{p/2} \exp(-\alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2)$$

or equivalently for the log-density

$$\log p(\boldsymbol{\lambda}|\alpha) = \frac{p}{2} \log \alpha - \alpha \|\boldsymbol{L}\boldsymbol{\lambda}\|^2 + \text{const},$$

where the constant does not depend on α. This allows us to find analytically the maximum of $\widetilde{Q}(\alpha; \alpha^{(k)})$. Namely, we can write

$$\begin{aligned}
\widetilde{Q}(\alpha; \alpha^{(k)}) &= \frac{1}{N} \sum_{i=1}^{N} \log p(\boldsymbol{\lambda}^{(i)}|\alpha) \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{p}{2} \log \alpha - \alpha \|\boldsymbol{L}\boldsymbol{\lambda}^{(i)}\|^2 \right) + \text{const} \\
&= \frac{p}{2} \log \alpha - \frac{\alpha}{N} \sum_{i=1}^{N} \|\boldsymbol{L}\boldsymbol{\lambda}^{(i)}\|^2 + \text{const}.
\end{aligned}$$

Setting the derivative to zero

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \widetilde{Q}(\alpha; \alpha^{(k)}) = \frac{p}{2\alpha} - \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{L}\boldsymbol{\lambda}^{(i)}\|^2 = 0,$$

we find

$$\alpha = \frac{1}{\frac{2}{pN}\sum_{i=1}^{N} \|\boldsymbol{L}\boldsymbol{\lambda}^{(i)}\|^2} > 0. \tag{6.3}$$

Since the second derivative is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\widetilde{Q}(\alpha;\alpha^{(k)}) = -\frac{p}{2\alpha^2} < 0,$$

we see that the value of α given by Equation (6.3) is the unique global maximum of $\widetilde{Q}(\alpha;\alpha^{(k)})$.

To conclude, when the Gaussian smoothness prior (5.12) is used in empirical Bayes unfolding, we set $\boldsymbol{\lambda}_0 = \mathbf{0}$ and require $\boldsymbol{L}$ to have full column rank. Then the MCEM algorithm for finding the marginal maximum likelihood estimator $\hat{\alpha}_{\mathrm{MMLE}}$ of the hyperparameter α is given by the following iteration:

1. Pick some initial guess $\alpha^{(0)} > 0$ and set $k = 0$.
2. E-step: Sample $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \ldots, \boldsymbol{\lambda}^{(N)}$ from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \alpha^{(k)})$.
3. M-step: Set
$$\alpha^{(k+1)} = \frac{1}{\frac{2}{pN}\sum_{i=1}^{N} \|\boldsymbol{L}\boldsymbol{\lambda}^{(i)}\|^2}. \tag{6.4}$$
4. Set $k \leftarrow k + 1$.
5. If some stopping rule $\mathcal{C}(\alpha^{(k)}, \alpha^{(k-1)}, \ldots, \alpha^{(0)})$ is satisfied, set $\hat{\alpha}_{\mathrm{MMLE}} = \alpha^{(k)}$ and terminate the iteration, else go to step 2.

Once the iteration has converged, we sample M observations from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$ and use this sample to draw inferences about $\boldsymbol{\lambda}$ as explained in Section 5.1.

Chapter 7

Computational Experiments

In this chapter, unfolding is demonstrated in practice using two different simulated data sets. The emphasis of the computational experiments is on empirical Bayes unfolding as described in Chapter 6. In Section 7.1, a Gaussian mixture model which is smeared using a convolution with a Gaussian is unfolded using various levels of regularization. In Section 7.2, a more realistic simulation study is performed by emulating the inclusive jet cross section measurement at the CMS experiment. In addition to the results of the computational experiments, this chapter addresses a number of practical issues with unfolding such as the details of the MCMC sampling scheme, the use of non-uniform binning and the choice of the spaces E and F in cases where is it possible to have smeared events which in reality originate from outside the observable space F. All the experiments reported in here were conducted using custom-made implementations of the unfolding algorithms in Matlab R2011a.

7.1 Gaussian Mixture Model

7.1.1 Description of the Data

We first study the computational performance of unfolding techniques using artificial toy data where the true observations follow a Gaussian mixture model and the smearing operator is a simple Gaussian convolution operator. We generated the data by sampling the number of true observations τ from a Poisson distribution with mean T, that is $\mathrm{E}[\tau] = T$. We then sampled τ observations from the Gaussian mixture model with two components

$$p_X(x) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x|\mu_2, \sigma_2^2).$$

Using Equation (2.3), the intensity function of the true Poisson process is thus

$$f(x) = \mathrm{E}[\tau] p_X(x) = \pi_1 T \mathcal{N}(x|\mu_1, \sigma_1^2) + \pi_2 T \mathcal{N}(x|\mu_2, \sigma_2^2).$$

The parameters of the first Gaussian component were set to $\pi_1 = 0.3, \mu_1 = -2, \sigma_1 = 1$ and the parameters of the second component to $\pi_2 = 0.7, \mu_2 = 2, \sigma_2 = 1$. The expected number of observations was $T = 5000$.

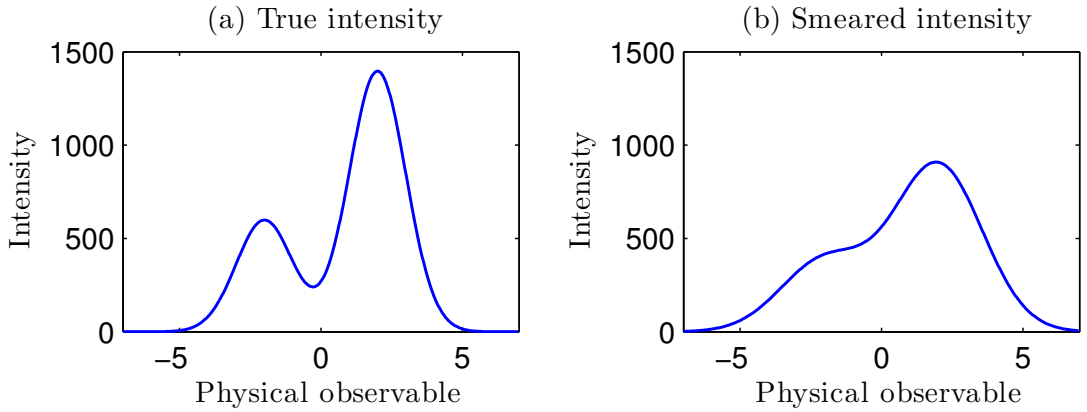


Figure 7.1: Intensities of the Poisson processes for the Gaussian mixture model data. Figure (a) shows the intensity function f of the true Poisson process and Figure (b) the intensity function h of the smeared Poisson process obtained by convolving the true intensity with a Gaussian of standard deviation $\sigma_n = 1.2$.

The true observations were then corrupted with additive Gaussian white noise with variance σ_n^2. The magnitude of the noise was set to $\sigma_n = 1.2$. As discussed in Section 2.1.3, the intensity function of the smeared Poisson process is then given by

$$h(y) = \int k(x, y) f(x) \, \mathrm{d}x$$

with the smearing kernel $k(x, y) = \mathcal{N}(y - x|0, \sigma_n^2)$. It follows that we can write the smeared intensity using convolutions of Gaussian pdfs

$$\begin{aligned} h(y) &= (\mathcal{N}(0, \sigma_n^2) * f)(y) \\ &= \pi_1 T(\mathcal{N}(\mu_1, \sigma_1^2) * \mathcal{N}(0, \sigma_n^2))(y) + \pi_2 T(\mathcal{N}(\mu_2, \sigma_2^2) * \mathcal{N}(0, \sigma_n^2))(y). \end{aligned}$$

Using the property that the convolution of two Gaussians with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is a Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$, we can write this intensity as another mixture of two Gaussians

$$h(y) = \pi_1 T \mathcal{N}(\mu_1, \sigma_1^2 + \sigma_n^2) + \pi_2 T \mathcal{N}(\mu_2, \sigma_2^2 + \sigma_n^2).$$

The intensities f and h are shown in Figure 7.1.

We then discretize the problem using histograms as described in Section 2.1.5. When doing this we consider both the true and the smeared Poisson process on the interval $[-7, 7]$, that is $E = F = [-7, 7]$. We then discretize both of these intervals using $p = q = 40$ histogram bins of uniform size. The observations of the two processes are then recorded into the histograms corresponding to this binning resulting in the true histogram $\boldsymbol{x}$ and the smeared histogram $\boldsymbol{y}$. The bin means $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ of these histograms are obtained from the intensity function f and h using Equations (2.7) and (2.8). Figure 7.2 shows one realization of the these histograms along with their means.

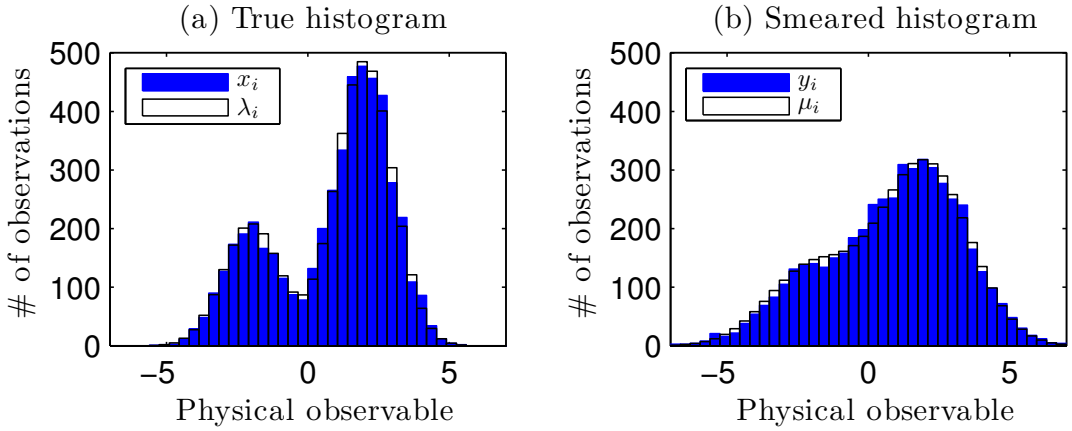


Figure 7.2: Discretization of the Gaussian mixture model data on the interval $[-7, 7]$ using histograms with 40 bins. Figure (a) shows one realization of the true histogram $\boldsymbol{x}$ and its mean $\boldsymbol{\lambda}$ and Figure (b) shows the corresponding realization of the smeared histogram $\boldsymbol{y}$ along with its mean $\boldsymbol{\mu}$. The task in unfolding is to infer $\boldsymbol{\lambda}$ using the observations $\boldsymbol{y}$.

The smearing matrix $\boldsymbol{K} \in \mathbb{R}^{40\times 40}$ was computed using the approximation (2.11). The resulting approximate smearing matrix has a (numerical) rank of 34. This was computed using the Matlab command `rank` which uses SVD for determining the rank. Hence, $\boldsymbol{K}$ is (numerically) singular and we know by Theorem 4.1 that $\boldsymbol{\lambda}$ is in fact non-identifiable. Nevertheless, as we shall see, we will be able to estimate $\boldsymbol{\lambda}$ rather well by injecting more information into the problem via regularization. The condition number of $\boldsymbol{K}$ is $\text{cond}(\boldsymbol{K}) = 1.6 \cdot 10^{17}$ which is consistent with a singular or badly ill-posed matrix.

7.1.2 Sampling Scheme

In the Bayesian and empirical Bayes unfolding experiments that follow, we use the Metropolis–Hastings algorithm for sampling from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \alpha)$ as described on Section 5.2. The proposal density in the Metropolis–Hastings sampler is chosen to be a multivariate Gaussian

$$p(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)}) = \mathcal{N}(\boldsymbol{\lambda}^*|\boldsymbol{\lambda}^{(k)}, \boldsymbol{\Sigma}).$$

Since this satisfies (5.9), the resulting algorithm is an example of random-walk Metropolis–Hastings sampling. As noted in Section 5.2, the shape of the proposal density should be close to the shape of the posterior. If this is not the case, it could happen that we propose large jumps for bins with only a few events and small jumps for bins with a huge number of events which naturally results in slow convergence and mixing of the chain. Hence, in our case, it is especially important to have the relative sizes of the jumps on different dimensions to be somewhat similar to the expected relative heights of the corresponding true histogram bins. To this end, we

take the proposal covariance to be diagonal and set the proposal variance on each dimension either to

$$\sigma_{1,i}^2 = \gamma \max(1, y_i) \tag{7.1}$$

or

$$\sigma_{2,i}^2 = (\gamma \max(1, y_i))^2. \tag{7.2}$$

The idea with the first choice $\sigma_{1,i}^2$ is to first set the variance of each component to the value of the corresponding smeared observation y_i since this would be the estimated variance of the MLE in the case with no smearing (see Equation (3.3)). All the variances are then scaled using the parameter γ to achieve steps of optimal size. The second choice $\sigma_{2,i}^2$ is based on first setting the standard deviations of each dimension to y_i which fixes the relative sizes of the dimensions to reasonable values. After fixing the shape of the density, we then scale the standard deviations by γ to optimize the spread of the proposal density. The maximum is required in the definitions to deal with bins with no observations.

We start the Metropolis–Hastings chain from the observations $\boldsymbol{y}$, that is $\boldsymbol{\lambda}^{(1)} = \boldsymbol{y}$. This allows us to start sampling from a reasonable region of the state space which should not be too far away from the bulk of the posterior probability mass and hence facilitates convergence of the chain. To deal with burn-in, we produce a chain of length $\frac{3}{2}N$, when N observations are needed, and simply discard the first third of the chain. The convergence and mixing of the chain are then verified by plotting the time series of each dimension λ_i and computing the acceptance rate for the sample with the burn-in removed.

7.1.3 Unfolding Results

As expected, unfolding of $\boldsymbol{y}$ to find $\boldsymbol{\lambda}$ without any regularization results in a completely unacceptable solution. We demonstrate this by using the least squares estimator $\hat{\boldsymbol{\lambda}}_{\text{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y}$ of $\boldsymbol{\lambda}$. The resulting estimator is shown in Figure 7.3. The figure also shows $\boldsymbol{\lambda}$, but the oscillations of the LS estimator $\hat{\boldsymbol{\lambda}}_{\text{LS}}$ have such a high amplitude that the histogram for $\boldsymbol{\lambda}$ is indistinguishable from the horizontal axis of the plot.

From Figure 7.3, it is clear that some sort of regularization is needed to find a reasonable estimate of $\boldsymbol{\lambda}$. In what follows, we demonstrate Bayesian and empirical Bayes unfolding with various regularization schemes in increasing order of regularization strength. We start with Bayesian unfolding with the uniform non-negativity prior (5.11). We then demonstrate empirical Bayes unfolding using the truncated Gaussian smoothness prior defined by Equation (5.12). As discussed in Section 6.3, we set $\boldsymbol{\lambda}_0 = \boldsymbol{0}$ and set the matrix $\boldsymbol{L}$ either to the identity $\boldsymbol{I}$, the first-order finite-difference matrix $\boldsymbol{L}_2^1$ with the Dirichlet boundary condition for the right boundary (4.35) or the second-order finite-difference matrix $\boldsymbol{L}_2^2$ with the Dirichlet boundary condition for both boundaries (4.37). We see from Figure 7.1(a) that the desired solution to the problem at hand satisfies these boundary conditions. We found empirically that when less regularization is applied, the Metropolis–Hastings sampler mixes better with the proposal variance $\sigma_{2,i}^2$ of Equation (7.2), while in the case of stronger regularization, $\sigma_{1,i}^2$ gives better performance. Based on this analysis, we

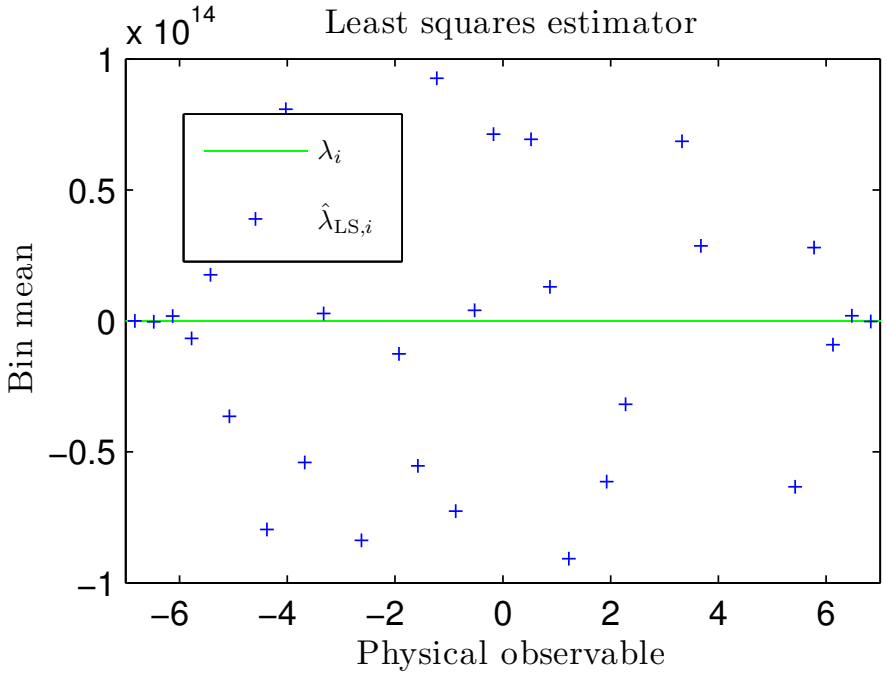


Figure 7.3: Unfolding of the Gaussian mixture model data using the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y}$ of $\boldsymbol{\lambda}$. Due to the scale of the least squares estimator, the true histogram $\boldsymbol{\lambda}$ is indistinguishable from the horizontal axis. Because of the ill-posedness of the problem, the resulting solution is completely unacceptable.

used $\sigma^2_{2,i}$ with the uniform prior and with the Gaussian smoothness prior for $\boldsymbol{L} = \boldsymbol{I}$ and $\sigma^2_{1,i}$ with the Gaussian smoothness prior for $\boldsymbol{L} = \boldsymbol{L}^1_2$ and $\boldsymbol{L} = \boldsymbol{L}^2_2$.

We performed Bayesian unfolding with the uniform non-negativity prior by sampling $N = 400\,000$ observations from the corresponding posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$[1]. The size of the steps γ was adjusted until the acceptance rate after burn-in was roughly 30 %. We settled on the choice $\gamma = 0.035$ which gave an acceptance rate of 32 %. Figure 7.4 shows the time series of the Metropolis–Hastings chain for each component λ_i. We see that the mixing of the chain in rather good in the central bins of the histogram but the chain has some trouble sampling the bins near the boundaries of the histogram. The behavior of the chain in acceptable but we will later see examples of chains that mix a lot better than the one in here.

Each dimension of the Metropolis–Hastings sample forms a sample from the marginal posterior $p(\lambda_i|\boldsymbol{y})$. Figure 7.5(a) shows the central 68.27 % credible intervals for each mean λ_i computed using these marginal posterior samples. Although the solution has a high uncertainty, we see that we are already doing a lot better than with the least squares estimator. The most likely reason for this is that the Bayesian approach forces the solution to be non-negative which can be seen as a first step towards regularizing the problem. The red error bars in Figure 7.5(a) represent the

[1]Note that here the prior does not depend on any hyperparameters $\boldsymbol{\alpha}$ and hence we have omitted the dependence on $\boldsymbol{\alpha}$ in the posterior too.

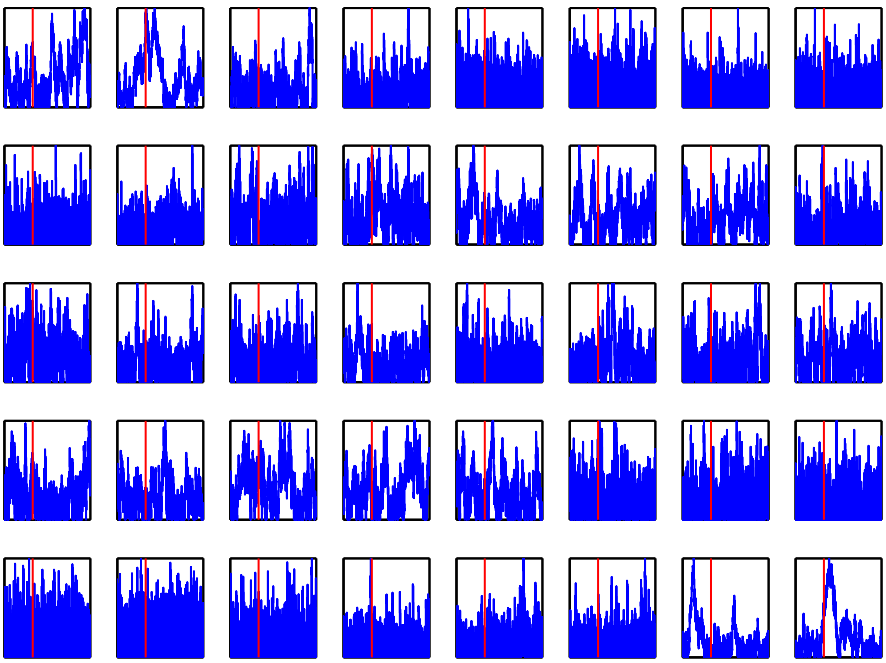


Figure 7.4: Time series of the components λ_i of the Metropolis–Hastings chain with the uniform non-negativity prior. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.

so-called $\sqrt{n}$ errors. They show how large the errors would have been if the medians of the marginal posteriors had been our frequentist MLEs for each λ_i and there was no smearing. They are used here and in further experiments to compare the errors of the unfolded histogram to the errors we would have gotten had we used a hypothetical perfect detector without any smearing.

Furthermore, Figure 7.5(b) shows a box plot of the marginal posterior samples. The horizontal lines of the plot are the medians of the samples and the boxes show their interquartile ranges (IQRs). The whiskers extend to the smallest (largest) data point still within 1.5 IQR from the lower (upper) quartile. The whiskers constructed this way can be interpreted as the range of the data excluding outliers. When plotted this way, we see that there is indeed a large uncertainty about the solution. For example, all the lower whiskers extend all the way to the limiting horizontal axis.

We now move on to describe the empirical Bayes experiments. In all these experiments, we ran the MCEM algorithm as described in Section 6.3 for 30 iterations starting with $\alpha^{(0)} = 1 \cdot 10^{-4}$ and verified the convergence of the algorithm by plotting the time series of the iterates $\alpha^{(k)}$. On each iteration of the algorithm $N = 100\,000$ observations were sampled from the current posterior using the Metropolis–Hastings algorithm. The step size γ was chosen to have approximately 30 % acceptance rate at the convergence of the algorithm. The last iterate from MCEM was then chosen as our MMLE of $\boldsymbol{\alpha}$ and was used in another Metropolis–Hastings chain of length $M = 400\,000$ to produce a sample from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. This sample

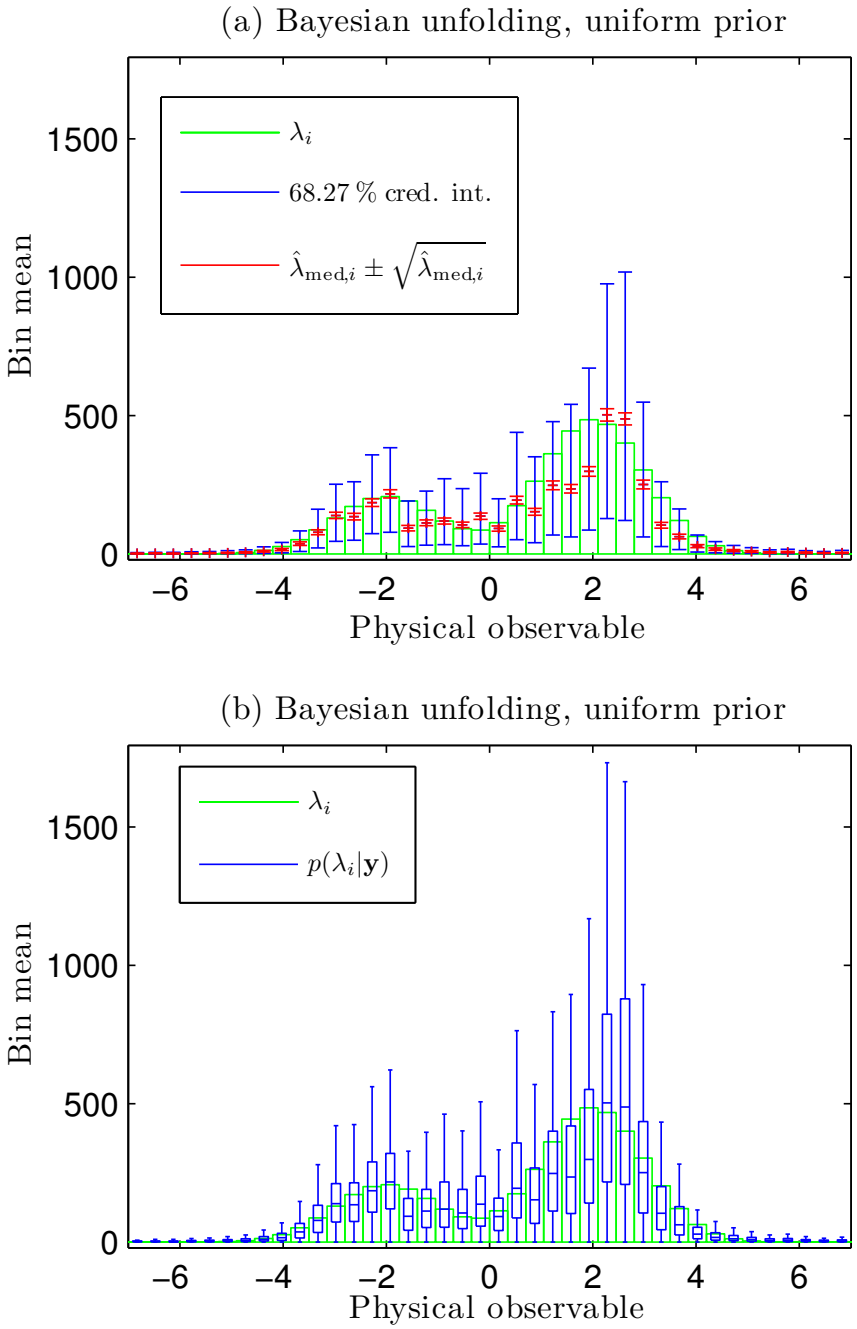


Figure 7.5: Comparison of the results of Bayesian unfolding of the Gaussian mixture model data to the correct value of $\boldsymbol{\lambda}$ (green histograms) when using the uniform non-negativity prior. Figure (a) shows the central 68.27 % credible intervals for each marginal posterior $p(\lambda_i|\boldsymbol{y})$. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of a perfect detector without any smearing. Figure (b) shows box plots which are computed for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y})$. The horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

was then used to make inferences about $\boldsymbol{\lambda}$ in the same way we described above for Bayesian unfolding.

Let us first set $\boldsymbol{L} = \boldsymbol{I}$ in the prior and investigate the convergence of the MCEM algorithm. Figure 7.6 shows that the iteration converges to a value of $\hat{\alpha}_{\text{MMLE}} = 1.2 \cdot 10^{-5}$ in just a couple of iterations and that there is little MC variation in the sequence. Figure 7.7 shows the time series of the Metropolis–Hastings chain for the MMLE and indicates that the chain converges and mixes nicely except for the boundary bins. The step size was $\gamma = 0.04$ in all these experiments which resulted in an acceptance rate of 29 % in the final Metropolis–Hastings chain. The unfolded histograms are then shown in Figure 7.8. We see that we have managed to reduce the uncertainty of the solution slightly in comparison to the mere uniform non-negativity prior but the variance of the posterior is still considerably large.

We then penalize for the first derivatives by setting $\boldsymbol{L} = \boldsymbol{L}_2^1$ in the prior. Figure 7.9 shows that now the MCEM algorithm converges in roughly 10 iterations to $\hat{\alpha}_{\text{MMLE}} = 2.2 \cdot 10^{-4}$ with only slight MC variation in the sequence. The time series of the Metropolis–Hastings chain for the MMLE are shown in Figure 7.10. This time around, it seems that all the components of the chain converge and explore the posterior very well. The step size was set to $\gamma = 0.06$ giving an acceptance rate of 32 % in the final Metropolis–Hastings sampling. The unfolded histograms are shown in Figure 7.11. Regularization with the first derivatives substantially reduces the uncertainty of the solution while still mostly maintaining the correct value of λ_i within the 68.27 % error bars. In addition, the whiskers of the box plot of Figure 7.11(b) nicely cover the correct histogram.

To conclude the empirical Bayes experiments, we penalize for the second derivative by setting $\boldsymbol{L} = \boldsymbol{L}_2^2$ in the prior. Figure 7.12 shows that the MCEM algorithm converges nicely to the hyperparameter $\hat{\alpha}_{\text{MMLE}} = 1.0 \cdot 10^{-3}$ and Figure 7.13 shows that the Metropolis–Hastings chain corresponding to the MMLE mixes very well for all the components λ_i. These results were obtained with the step size $\gamma = 0.025$ which gave an acceptance rate of 34 % for the final sample. The unfolded histograms are shown in Figure 7.14. We see that we have managed to further reduce the uncertainty of the solution. We also see that with such a strong regularization we are approaching the ideal-detector $\sqrt{n}$ errors on some bins near the right-hand peak. The reason for this is that especially in this region of high curvature the regularization makes the solution biased while at the same time decreasing its variance and hence the credible intervals are short and slightly off the desired values. It is a matter of taste whether one prefers the results obtained in Figure 7.11 for $\boldsymbol{L} = \boldsymbol{L}_2^1$ to the ones obtained in here for $\boldsymbol{L} = \boldsymbol{L}_2^2$. On one hand, one could be inclined to say that $\boldsymbol{L} = \boldsymbol{L}_2^2$ overregularizes the problem, while on the other hand, with $\boldsymbol{L} = \boldsymbol{L}_2^1$, there still remains considerable uncertainty about the solution across the whole histogram.

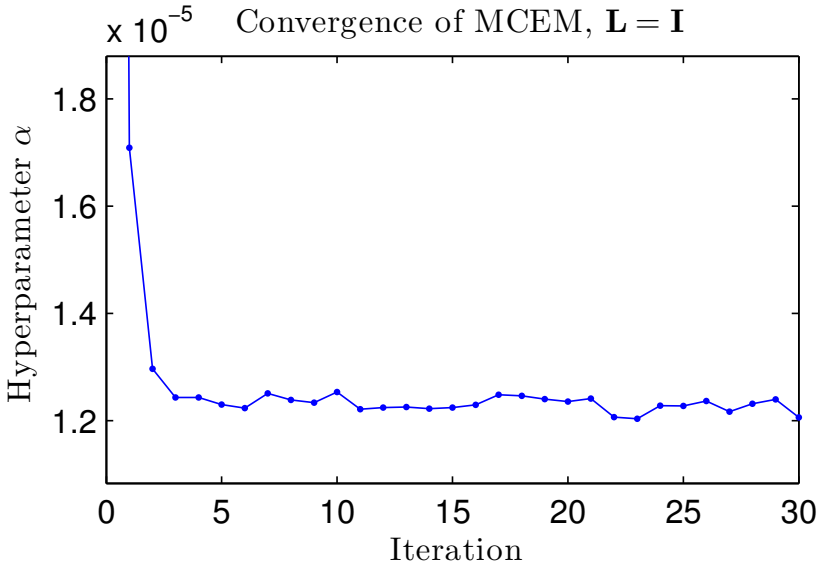


Figure 7.6: Convergence of the MCEM algorithm in empirical Bayes unfolding for the truncated Gaussian smoothness prior penalizing for the norm of the solution with $\boldsymbol{L} = \boldsymbol{I}$.

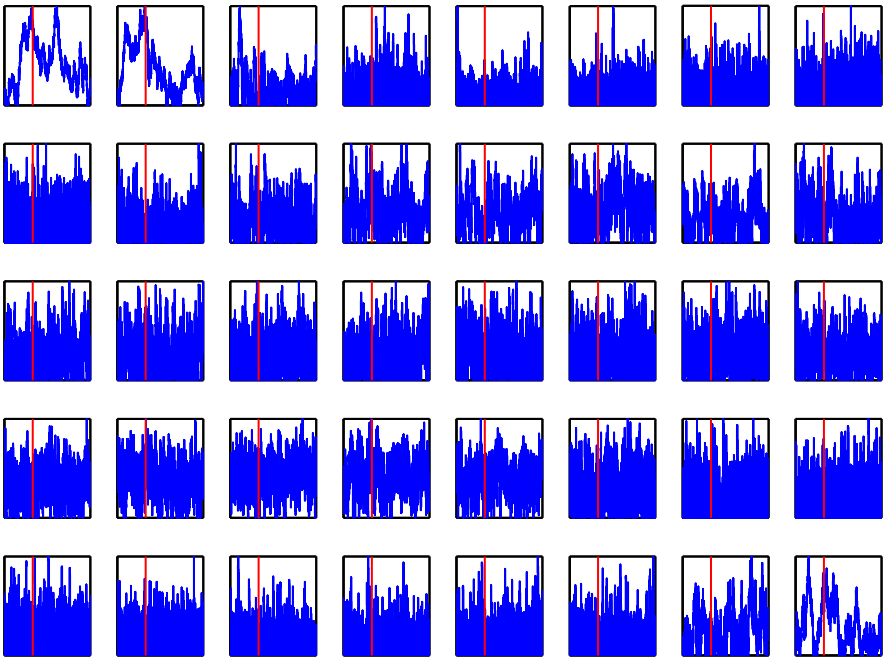


Figure 7.7: Time series of the components λ_i of the Metropolis–Hastings chain for the MMLE obtained from the MCEM algorithm in empirical Bayes unfolding for the truncated Gaussian smoothness prior penalizing for the norm of the solution with $\boldsymbol{L} = \boldsymbol{I}$. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.

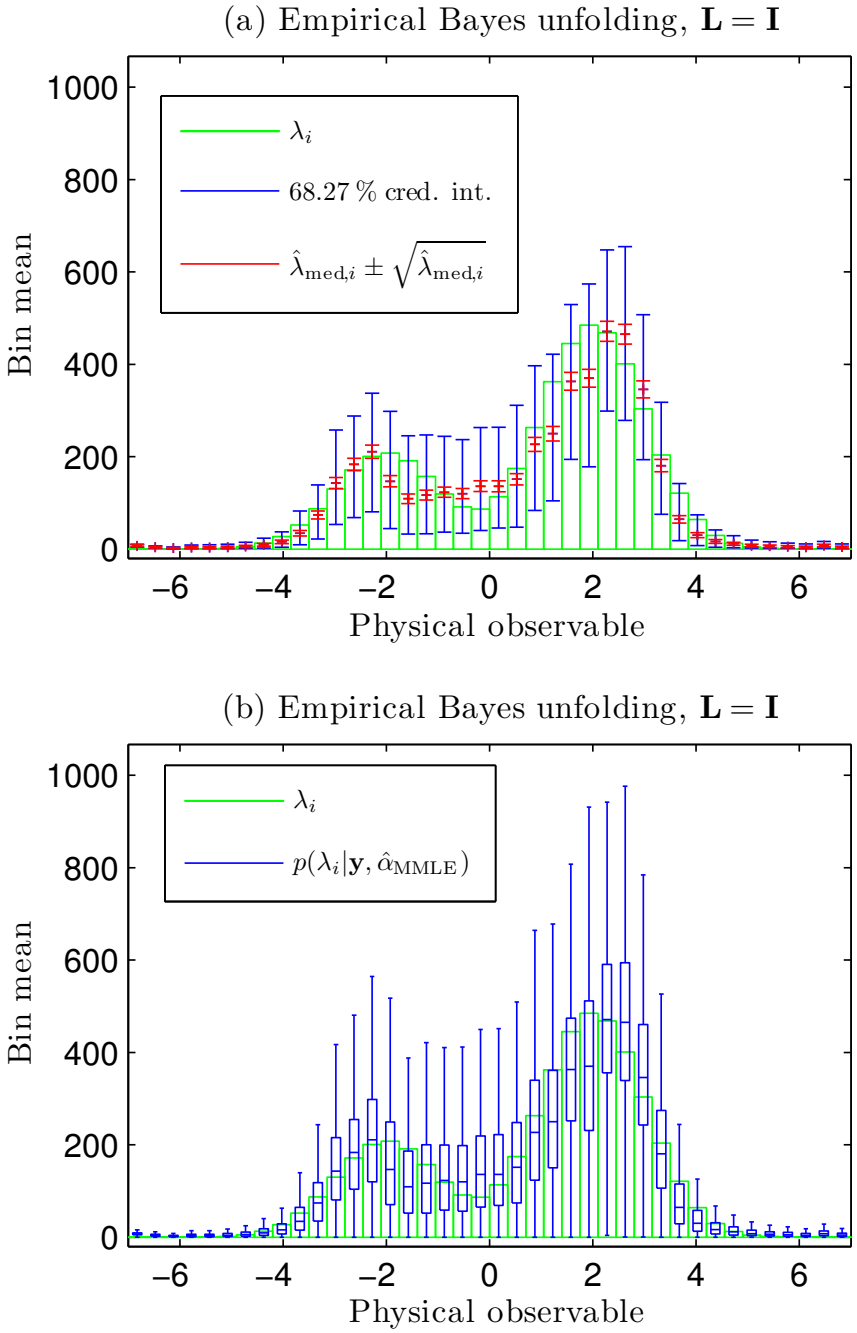


Figure 7.8: Comparison of the results of empirical Bayes unfolding of the Gaussian mixture model data to the correct value of $\boldsymbol{\lambda}$ (green histograms) when using the truncated Gaussian smoothness prior penalizing for the norm of the solution with $\boldsymbol{L} = \boldsymbol{I}$. Figure (a) shows the central 68.27 % credible intervals for each marginal posterior $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of a perfect detector without any smearing. Figure (b) shows box plots which are computed for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. The horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

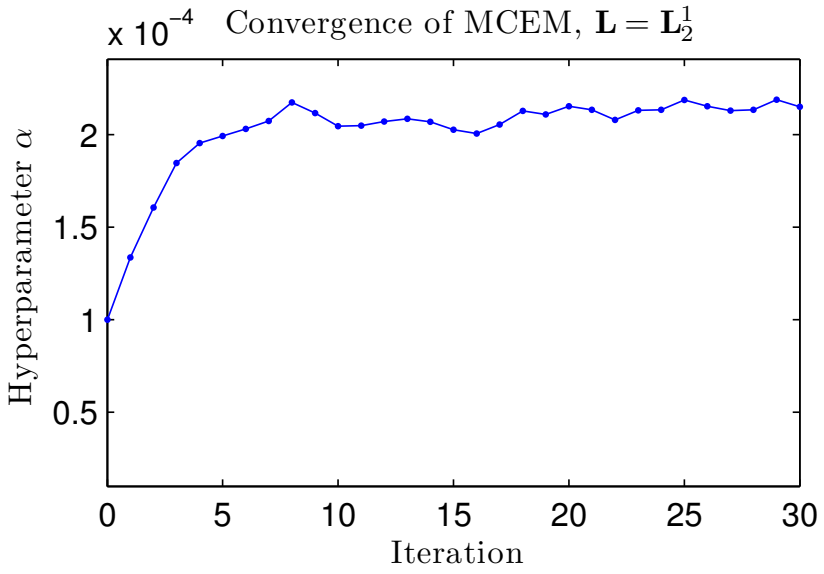


Figure 7.9: Convergence of the MCEM algorithm in empirical Bayes unfolding for the truncated Gaussian smoothness prior penalizing for the first derivatives with $\boldsymbol{L} = \boldsymbol{L}_2^1$.

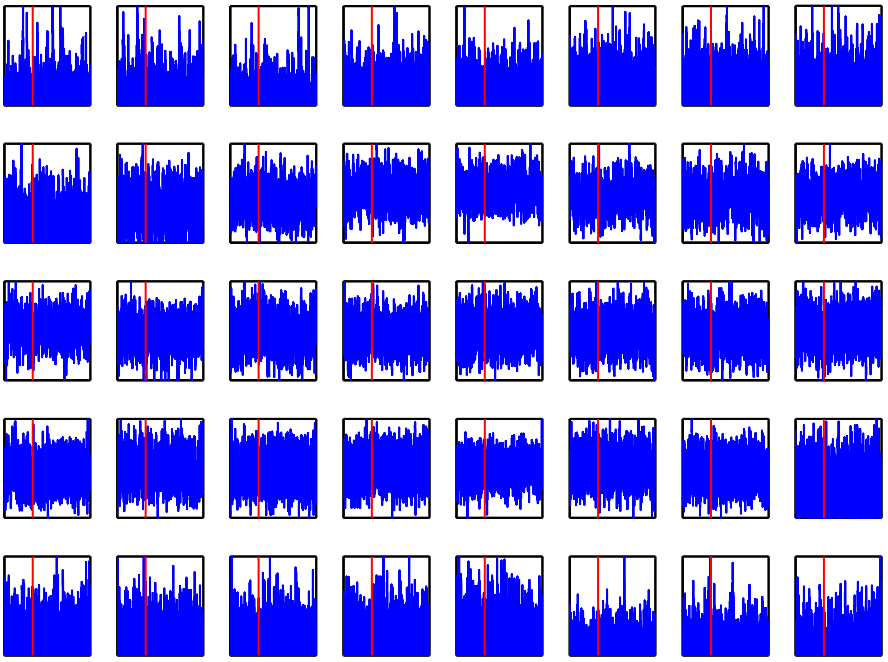


Figure 7.10: Time series of the components λ_i of the Metropolis–Hastings chain for the MMLE obtained from the MCEM algorithm in empirical Bayes unfolding for the truncated Gaussian smoothness prior penalizing for the first derivatives with $\boldsymbol{L} = \boldsymbol{L}_2^1$. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.

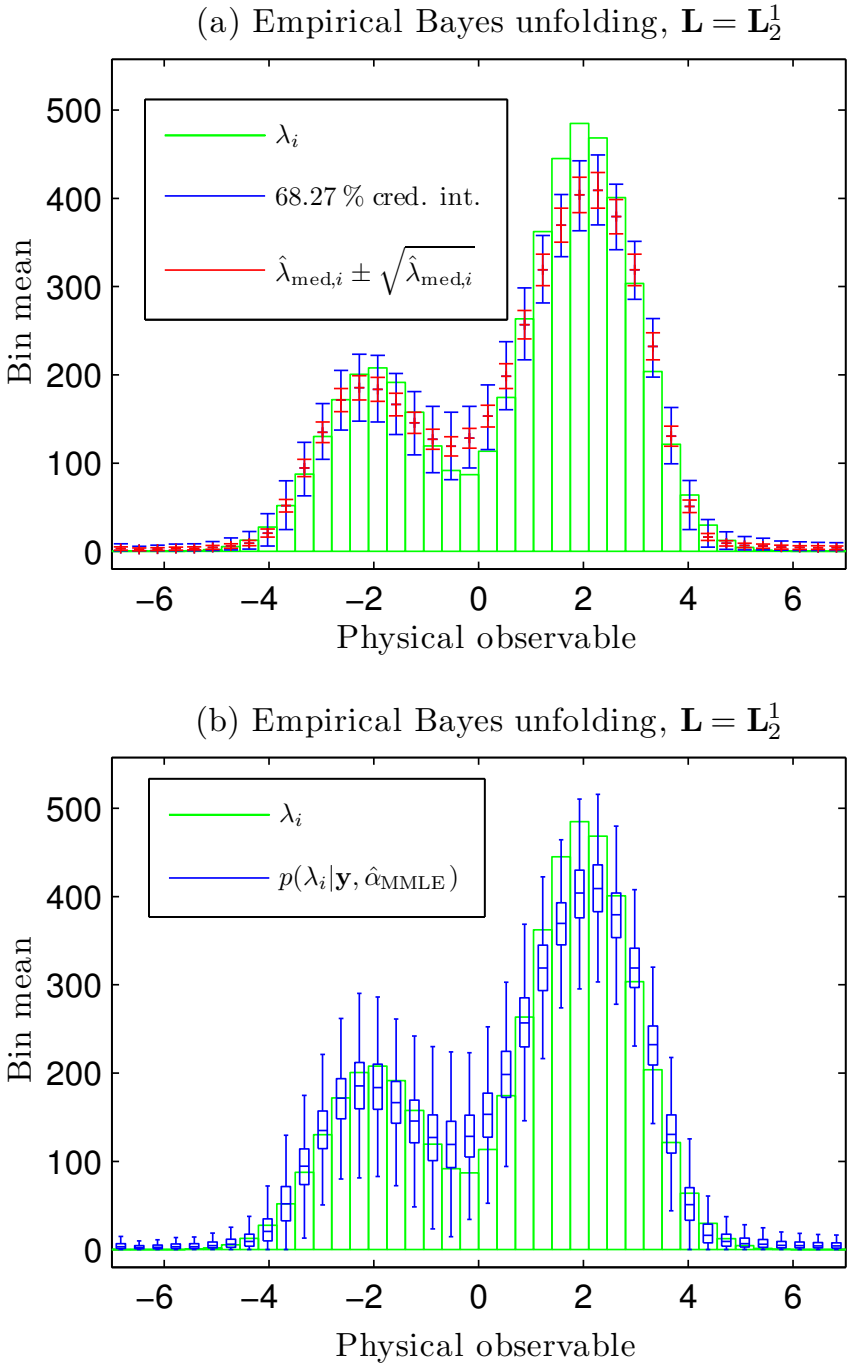


Figure 7.11: Comparison of the results of empirical Bayes unfolding of the Gaussian mixture model data to the correct value of $\boldsymbol{\lambda}$ (green histograms) when using the truncated Gaussian smoothness prior penalizing for the first derivatives with $\boldsymbol{L} = \boldsymbol{L}_2^1$. Figure (a) shows the central 68.27 % credible intervals for each marginal posterior $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of a perfect detector without any smearing. Figure (b) shows box plots which are computed for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. The horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

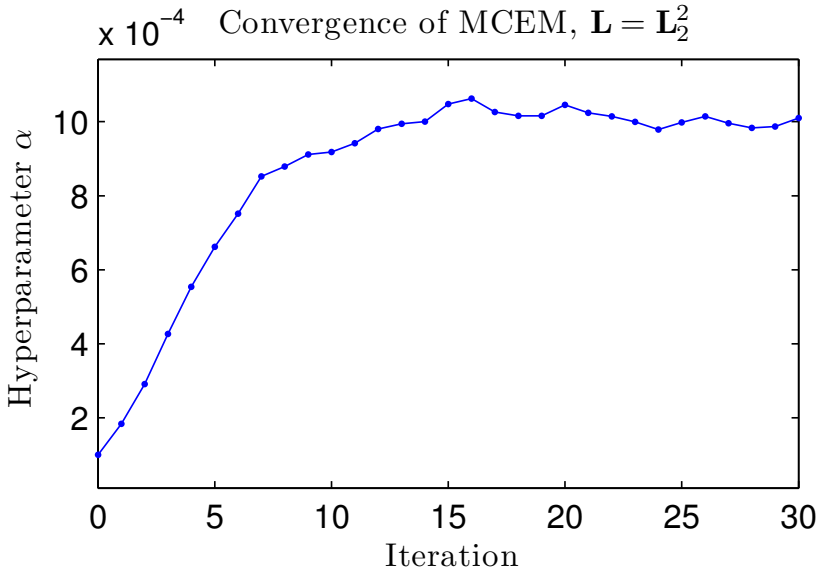


Figure 7.12: Convergence of the MCEM algorithm in empirical Bayes unfolding for the truncated Gaussian smoothness prior penalizing for the second derivatives with $\boldsymbol{L} = \boldsymbol{L}_2^2$.

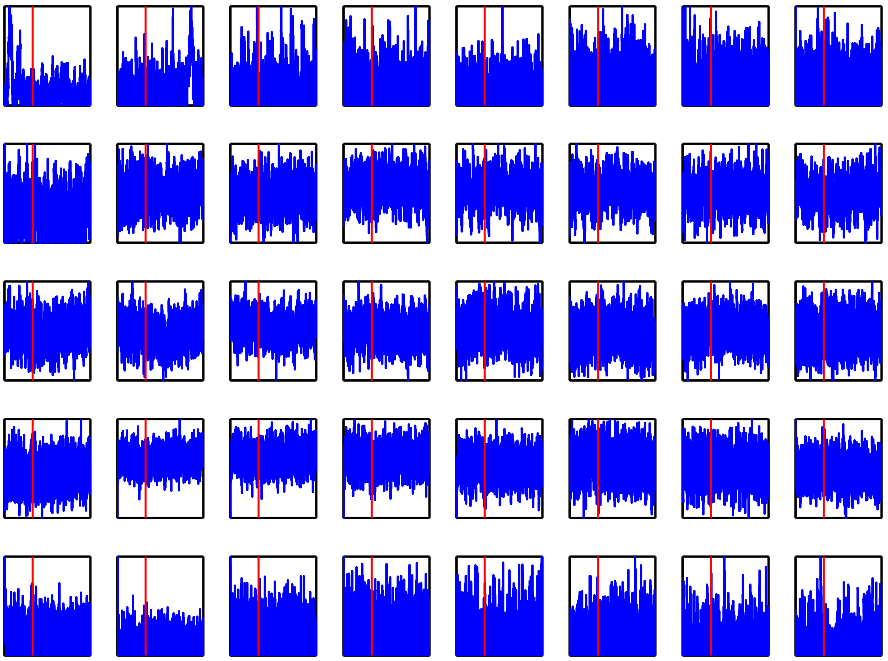


Figure 7.13: Time series of the components λ_i of the Metropolis–Hastings chain for the MMLE obtained from the MCEM algorithm in empirical Bayes unfolding for the truncated Gaussian smoothness prior penalizing for the second derivatives with $\boldsymbol{L} = \boldsymbol{L}_2^2$. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.

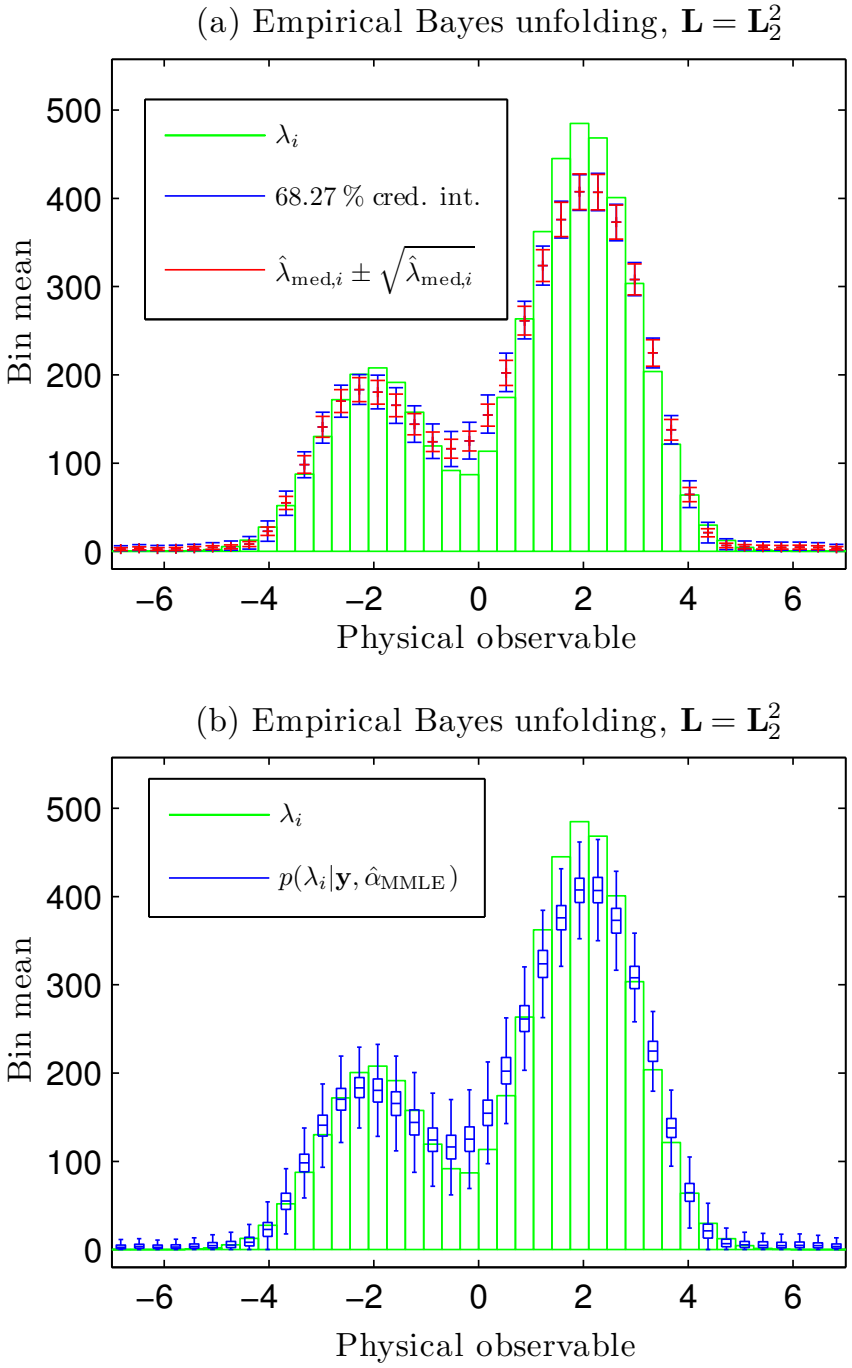


Figure 7.14: Comparison of the results of empirical Bayes unfolding of the Gaussian mixture model data to the correct value of $\boldsymbol{\lambda}$ (green histograms) when using the truncated Gaussian smoothness prior penalizing for the second derivatives with $\boldsymbol{L} = \boldsymbol{L}_2^2$. Figure (a) shows the central 68.27 % credible intervals for each marginal posterior $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of a perfect detector without any smearing. Figure (b) shows box plots which are computed for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$. The horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

7.2 Inclusive Jet Cross Section

7.2.1 Description of the Data

We now demonstrate unfolding of a steeply falling spectrum of events. The goal of this analysis is to mimic with a simulation study the analysis of the inclusive jet cross section measurement at the Compact Muon Solenoid (CMS) experiment of the Large Hadron Collider at CERN. The latest version of this analysis is presented in [51] for a center-of-mass energy of $\sqrt{s} = 7$ TeV and an integrated luminosity[2] of $L = 34\ \text{pb}^{-1}$.

A jet is a collimated stream of energetic particles seen in a particle detector. The detection of a jet indicates that either a quark or a gluon was created in a proton–proton collision at the heart of the detector. The goal of the inclusive jet analysis is to measure the probability of jet production as a function of the transverse momentum[3] p_T and the rapidity[4] y of the jet. To be more precise, the goal is to measure the differential cross section of the jets

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\text{T}\mathrm{d}y} = \frac{1}{L}\frac{\mathrm{d}^2N}{\mathrm{d}p_\text{T}\mathrm{d}y}, \tag{7.3}$$

where N is the expected number of jets produced. Mathematically, the differential cross section is the function which integrated over the desired values of p_T and y and multiplied by the integrated luminosity L gives the Poisson mean for the number of jets. Hence, up to the multiplicative factor $1/L$, this is the intensity function of the underlying Poisson point process. In what follows, we consider the differential cross section integrated over the rapidity range $|y| < 0.5$ corresponding to the part of the detector perpendicular to the beam and hence regard the Poisson intensity as a function of p_T only.

Following the parametrization used in [51], we assume that the intensity function $f(p_\text{T})$ of the true Poisson process is given by

$$f(p_\text{T}) = LN_0 \left(\frac{p_\text{T}}{\text{GeV}}\right)^{-\alpha} \left(1 - \frac{2}{\sqrt{s}} p_\text{T} \cosh(y_\text{min})\right)^{\beta} e^{-\gamma/p_\text{T}}, \tag{7.4}$$

where L is the integrated luminosity, $\sqrt{s}$ is the center-of-mass energy, N_0, α, β and γ are free parameters and y_min is the minimum of $|y|$ on the rapidity range under consideration, in our case $y_\text{min} = 0$. For most values of p_T, this spectrum follows approximately the power law $p_\text{T}^{-\alpha}$ but for large values of p_T there is a kinematic cut-off at $\sqrt{s}/(2\cosh(y_\text{min}))$ which is the maximum possible jet transverse momentum for

[2]The integrated luminosity L is a measure of the amount of data collected in a physics experiment. It relates the cross section σ to the total number of events N via the relation $N = L\sigma$.

[3]The transverse momentum p_T is the component of the momentum vector which is perpendicular to the direction of the proton beam.

[4]The rapidity y is defined by $y = \tanh^{-1}\left(\frac{p_z}{E}\right)$, where p_z is the component of the momentum along the beam line and E is energy. Because of its invariance properties this is a handy variable in special relativity and can be roughly understood as a reparametrization of the polar angle θ between the jet and the proton beam.

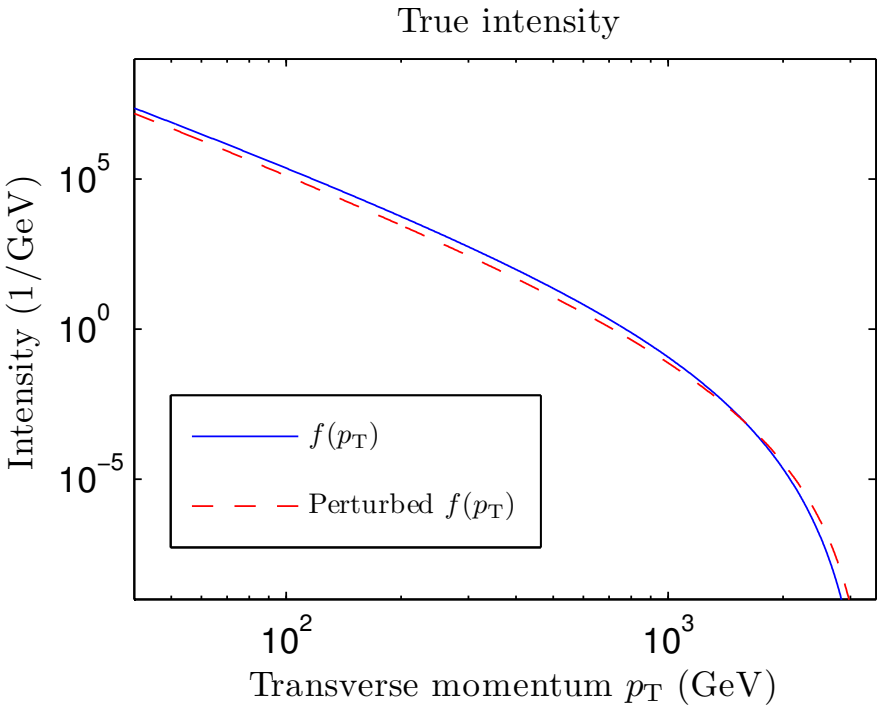


Figure 7.15: The intensity $f(p_T)$ used to generate the inclusive jet data and a slightly perturbed version of the intensity used for computing the smearing matrix $\boldsymbol{K}$.

a collision at the center-of-mass energy $\sqrt{s}$. The parameters of the model are set to $N_0 = 10^{14}$ pb/GeV, $\alpha = 5$, $\beta = 10$ and $\gamma = 10$ GeV. To enable comparison with the results in [51], the data are generated for an integrated luminosity of $L = 34$ pb^{-1} and a center-of-mass energy of $\sqrt{s} = 7$ TeV. We also consider a slightly perturbed version of this model with parameters $N_0 = 1.3 \cdot 10^{14}$ pb/GeV, $\alpha = 5.2$, $\beta = 8$ and $\gamma = 8$ GeV when computing the smearing matrix in order to avoid getting overly optimistic, unrealistic results by assuming the correct spectrum for this computation. The two intensities are shown in Figure 7.15.

We assume that the measurement of the transverse momentum p_T of the jets is smeared by the Gaussian $\mathcal{N}(0, \sigma(p_T)^2)$, where the standard deviation $\sigma(p_T)$ is given by the standard parametrization of calorimeter resolution

$$\sigma(p_T) = p_T \left(\frac{N}{p_T} \oplus \frac{S}{\sqrt{p_T}} \oplus C \right) = p_T \sqrt{\frac{N^2}{p_T^2} + \frac{S^2}{p_T} + C^2}.$$

The three terms of this parametrization are called the noise, stochastic and constant terms, respectively, and $\oplus$ is used to denote the summation of the squares of the terms. In this parametrization, the noise term is the dominant one for low p_T values, the stochastic term dominates for medium p_T values and for large p_T the constant term becomes the most important one. The parameters of this smearing were set to $N = 1$ GeV, $S = 1$ GeV$^{1/2}$ and $C = 0.05$.

In a measurement like this, it is crucial to choose the true space E and the

smeared space F appropriately. Since the integral of the intensity (7.4) diverges at the origin, it is not possible to take into account the whole spectrum down to vanishing p_{T} values. Because of this, we make the choice $E = [40\ \mathrm{GeV}, 3500\ \mathrm{GeV}]$ for the true space. Here the upper limit is given by the kinematic limit at $\sqrt{s} = 7$ TeV. The discretization of this space is first carried out using logarithmic binning with 20 bins. To avoid bins with no observations, the last 5 bins are then merged into a single large bin covering large p_{T} values which leaves us with $p = 16$ bins for the E-space. It would then feel natural to also choose the same interval as the smeared space F. However, this choice does not work in practice. The reason for this is that in reality, there is a significant contribution of events near the lower bound of the F-space that originate from true p_{T} values which are situated outside this lower bound in the E-space. If the lower bounds of the E- and F-spaces were made equal, this contribution would be neglected. The situation can be dealt with by setting the lower bound of the F-space to a higher value than the lower bound of the E-space. When this difference is made large enough, the smearing to the F-space from outside the lower bound of the true space becomes negligible and the model should form an adequate description of the real data of the experiment. It was found out that by setting the F-space to begin from the third bin of the discretized E-space, this smearing-from-the-outside effect is negligibly small leading to the choice[5] $F = [62.5\ \mathrm{GeV}, 3500\ \mathrm{GeV}]$. The discretization of the F-space is then carried out using the same binning as in the corresponding part of the E-space giving us $q = 14$ smeared bins.

When the E- and F-spaces are chosen like this, it often happens that a generated true event will never be observed. This is the case especially for the first couple of E-bins. To take this properly into account, let us for a moment assume that the observed space is the whole real line. In this case, the smeared observations would be described by

$$p^s_{\mathrm{T},i} = p_{\mathrm{T,i}} + E_i, \quad E_i \sim \mathcal{N}(0, \sigma(p_{\mathrm{T,i}})),$$

where the E_i are independent. Hence, we have

$$p(p^s_{\mathrm{T}}|p_{\mathrm{T}}) = \mathcal{N}(p^s_{\mathrm{T}}|p_{\mathrm{T}}, \sigma(p_{\mathrm{T}})^2).$$

However, only some of these smeared observations lie on the F-space. Following the notation of Section 2.1.4, we denote by p^*_{T} a true observation that ends up being observed. The smearing is then described by the truncated Gaussian

$$p(p^s_{\mathrm{T}}|p^*_{\mathrm{T}}) = \frac{1_F(p^s_{\mathrm{T}})\mathcal{N}(p^s_{\mathrm{T}}|p^*_{\mathrm{T}}, \sigma(p^*_{\mathrm{T}})^2)}{\int_F \mathcal{N}(p^s_{\mathrm{T}}|p^*_{\mathrm{T}}, \sigma(p^*_{\mathrm{T}})^2)\,\mathrm{d}p^s_{\mathrm{T}}}.$$

The losses caused by smearing of events to values outside of F, will then have to be taken into account in the efficiency $\varepsilon(p_{\mathrm{T}})$. Since the probability of losing an event at p_{T} is $1 - \int_F \mathcal{N}(p^s_{\mathrm{T}}|p_{\mathrm{T}}, \sigma(p_{\mathrm{T}})^2)\,\mathrm{d}p^s_{\mathrm{T}}$, the efficiency is given by

$$\varepsilon(p_{\mathrm{T}}) = \int_F \mathcal{N}(p^s_{\mathrm{T}}|p_{\mathrm{T}}, \sigma(p_{\mathrm{T}})^2)\,\mathrm{d}p^s_{\mathrm{T}},$$

[5]It might also be desirable to extend the F-space above 3500 GeV to take into account possible smearing to values higher than this but since it is extremely improbable to observe an event near the kinematic limit with this amount of data, we chose to ignore this effect.

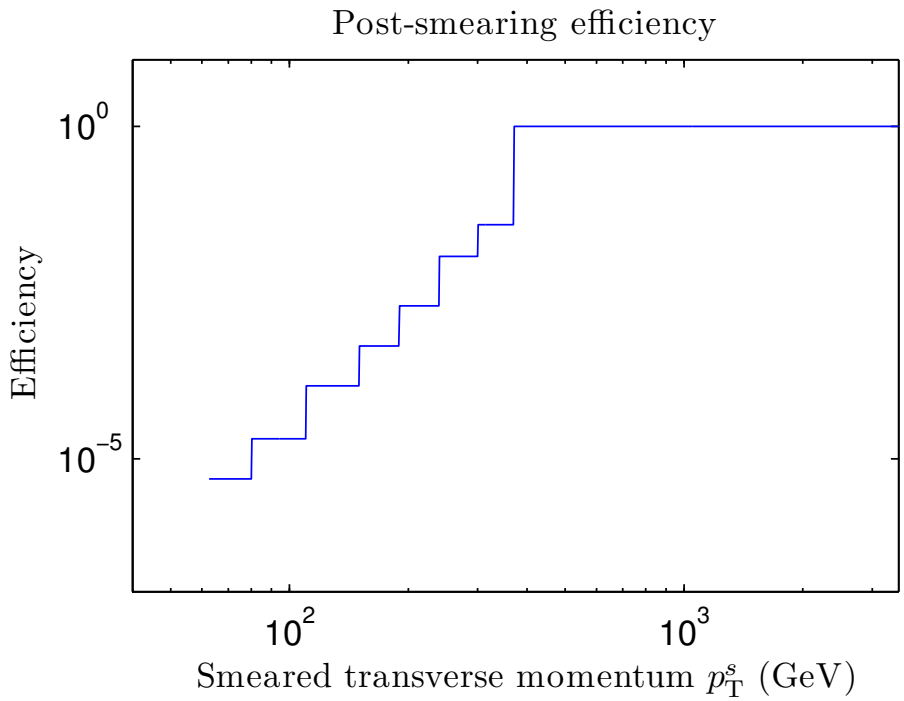


Figure 7.16: Post-smearing efficiency $\varepsilon_{\mathrm{PS}}(p_{\mathrm{T}}^s)$ for the inclusive jet data. The efficiency function emulates prescaling of the CMS jet triggers at low values of the transverse momentum p_{T}.

which takes into account losses due to smearing. What remains is to incorporate other detector inefficiencies into our model. Since it is very unlikely that an energetic jet on the rapidity region $|y| < 0.5$ escapes detection, we need not worry about non-detection of jets. What we do need to take into account, however, is the fact that for a power-law spectrum like this, it is not feasible to record every single jet on the low-p_{T} end of the distribution. Because of this, the detector is configured in such a way that only a fraction of the abundance of low-p_{T} jets are recorded. This is called *trigger prescaling* and is most naturally taken into account using a post-smearing efficiency $\varepsilon_{\mathrm{PS}}(p_{\mathrm{T}}^s)$ as discussed in Section 2.1.4. To emulate the jet triggers of the CMS experiment, we consider a piecewise-constant post-smearing efficiency with jumps at p_{T}^s values of 80, 110, 150, 190, 240, 300 and 370 GeV. The trigger prescales, or equivalently the detection efficiencies, we selected in such a way that the trigger is 100 % efficient above 370 GeV and then becomes increasingly prescaled for lower p_{T}^s values in such a way that the the lowest values of the observed smeared intensity are of the order of 10 % of the maximum reached at $p_{\mathrm{T}}^s = 370$ GeV. The corresponding post-smearing efficiency is shown in Figure 7.16.

Taking all these effects into account, the intensity function of the observed Pois-

son process is given by[6]

$$h(p_{\mathrm{T}}^s) = \int \varepsilon_{\mathrm{PS}}(p_{\mathrm{T}}^s) p(p_{\mathrm{T}}^s | p_{\mathrm{T}}^*) \varepsilon(p_{\mathrm{T}}^*) f(p_{\mathrm{T}}^*) \, \mathrm{d}p_{\mathrm{T}}^*$$
$$= \int 1_F(p_{\mathrm{T}}^s) \varepsilon_{\mathrm{PS}}(p_{\mathrm{T}}^s) \mathcal{N}(p_{\mathrm{T}}^s | p_{\mathrm{T}}^*, \sigma(p_{\mathrm{T}}^*)^2) f(p_{\mathrm{T}}^*) \, \mathrm{d}p_{\mathrm{T}}^*,$$

which can, if desired, be written in the form of Equation (2.6) as explained at the end of Section 2.1.4. From this, we see that the smearing kernel $k(p_{\mathrm{T}}, p_{\mathrm{T}}^s)$ is given by

$$k(p_{\mathrm{T}}, p_{\mathrm{T}}^s) = 1_F(p_{\mathrm{T}}^s) \varepsilon_{\mathrm{PS}}(p_{\mathrm{T}}^s) \mathcal{N}(p_{\mathrm{T}}^s | p_{\mathrm{T}}, \sigma(p_{\mathrm{T}})^2).$$

The resulting intensity function $h(p_{\mathrm{T}}^s)$ is shown in Figure 7.17. The "saw-like" structure of the function is a result of the trigger prescaling. This plot can also be used to verify that we have placed the lower bound of the F-space far enough from the lower bound of the E-space. If this was not the case, the intensity would start to dip at small values of p_{T}^s because there is a missing contribution of smeared events. The apparent linearity of the intensity on this log-log plot shows that the given binning adequately takes this effect into account.

Having access to both the true intensity $f(p_{\mathrm{T}})$ and the smeared intensity $h(p_{\mathrm{T}}^s)$, we can use Equations (2.7) and (2.8) to compute the means $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ of the true and smeared histograms, respectively. We then sample the observed histogram $\boldsymbol{y}$ from the Poisson distribution with mean $\boldsymbol{\mu}$ and, as before, the unfolding task is to use $\boldsymbol{y}$ to infer $\boldsymbol{\lambda}$.

The smearing matrix $\boldsymbol{K}$ is computed using the defining Equation (2.10) with the perturbed true intensity shown in Figure 7.15. This emulates the fact that before making the measurement, the best we can hope to do is to use an approximation of the true intensity for computing or simulating $\boldsymbol{K}$. In the case of a steeply falling spectrum, the piecewise-constant approximation (2.11), which we used previously in Section 7.1, is not appropriate since the true intensity changes significantly within each bin.

The condition number of the 14×16 smearing matrix $\boldsymbol{K}$ was estimated to be $\mathrm{cond}(\boldsymbol{K}) = 3.9 \cdot 10^5$. Hence, we expect unfolding of the inclusive jet spectrum to be an ill-posed task. The matrix has the largest off-diagonal contribution for the first few E-bins. Hence, we expect the ill-posedness to be the worst at the low-p_{T} part of the true histogram. The rank of $\boldsymbol{K}$ was computed to be 14 which means that the matrix has full row rank but is column-rank deficient. Hence, all the complications that appear in unfolding with column-rank deficient smearing matrices (see Chapter 4) apply to the unfolding task at hand.

7.2.2 Unfolding with Non-Uniform Binning

Up to now, we have presented the results of unfolding by showing the estimated values of the true means $\boldsymbol{\lambda}$ while often in reality we are actually interested in the true

[6]To be absolutely consistent with the notation, we should use the asterisk with p_{T}^s and h to denote that they refer to the values after the post-smearing thinning of the Poisson process, but, for brevity, we prefer to omit this from the notation in here.

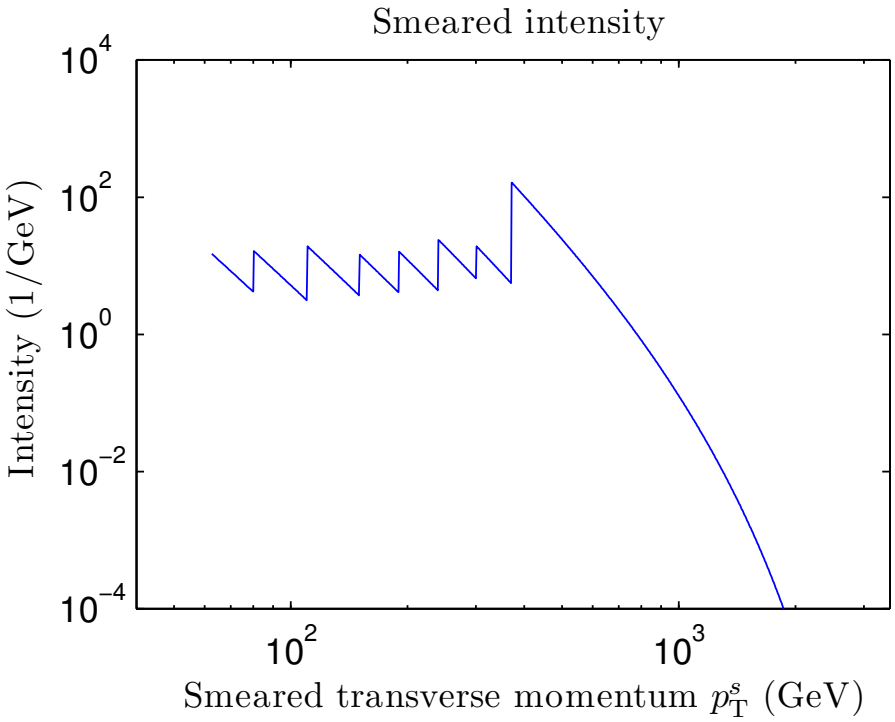


Figure 7.17: The smeared intensity function $h(p_{\mathrm{T}}^s)$ after taking into account both detector resolution and trigger prescaling. As a result of the triggering scheme, it is most probable to have an observation near $p_{\mathrm{T}}^s = 370$ GeV.

intensity function $f(x)$ or some normalized version of it as hinted by Equation (7.3). From Equation (2.9), we see that instead of $\boldsymbol{\lambda}$, we would rather be interested in inferring its scaled version

$$\boldsymbol{\lambda}_s = \left[\frac{\lambda_1}{C\nu(E_1)}, \frac{\lambda_2}{C\nu(E_2)}, \ldots, \frac{\lambda_p}{C\nu(E_p)}\right]^{\mathrm{T}}$$

which serves as a piecewise constant approximation of $f(x)$. Here C is some normalization constant. For example, in the case of cross section measurements, one often normalizes by the integrated luminosity by setting $C = L$.

When the E-space is discretized uniformly, i.e., $\nu(E_1) = \nu(E_2) = \cdots = \nu(E_p)$, it does not make a big difference if we estimate $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}_s$ since all the bins are scaled by the same coefficient and hence we are only talking about the global scaling of the spectrum. However, when the binning is non-uniform, it is important to consider $\boldsymbol{\lambda}_s$ instead of $\boldsymbol{\lambda}$ since looking at the plain Poisson means $\boldsymbol{\lambda}$ might give a completely wrong picture about the shape of the intensity function.

Luckily, if we are able to estimate $\boldsymbol{\lambda}$, it is straightforward to produce an estimator of $\boldsymbol{\lambda}_s$. In the case of point estimators, the natural way to proceed is to use

$$\hat{\boldsymbol{\lambda}}_s = \left[\frac{\hat{\lambda}_1}{C\nu(E_1)}, \frac{\hat{\lambda}_2}{C\nu(E_2)}, \ldots, \frac{\hat{\lambda}_p}{C\nu(E_p)}\right]^{\mathrm{T}}$$

as an estimator of $\boldsymbol{\lambda}_s$. The variance of each component of this estimator is given by

$$\mathrm{Var}[\hat{\lambda}_{s,i}|\boldsymbol{\lambda}] = \frac{1}{(C\nu(E_i))^2}\,\mathrm{Var}[\hat{\lambda}_i|\boldsymbol{\lambda}]\,.$$

Hence, the variance can be estimated with

$$\widehat{\mathrm{Var}}[\hat{\lambda}_{s,i}|\boldsymbol{\lambda}] = \frac{1}{(C\nu(E_i))^2}\widehat{\mathrm{Var}}[\hat{\lambda}_i|\boldsymbol{\lambda}]$$

leading us to estimate the standard deviation by

$$\widehat{\mathrm{Std}}[\hat{\lambda}_{s,i}|\boldsymbol{\lambda}] = \sqrt{\widehat{\mathrm{Var}}[\hat{\lambda}_{s,i}|\boldsymbol{\lambda}]} = \frac{1}{C\nu(E_i)}\sqrt{\widehat{\mathrm{Var}}[\hat{\lambda}_i|\boldsymbol{\lambda}]} = \frac{1}{C\nu(E_i)}\widehat{\mathrm{Std}}[\hat{\lambda}_i|\boldsymbol{\lambda}].$$

This means that the scaling of the estimator by some coefficient requires us to scale the error bars by the same coefficient.

In the case of Bayesian inference, we need to transform the marginal posterior $p(\lambda_i|\boldsymbol{y},\boldsymbol{\alpha})$ of λ_i to the marginal posterior $p(\lambda_{s,i}|\boldsymbol{y},\boldsymbol{\alpha})$ of $\lambda_{s,i}$. Denoting $g(\lambda_i) = p(\lambda_i|\boldsymbol{y},\boldsymbol{\alpha})$, the transformed posterior is given by $p(\lambda_{s,i}|\boldsymbol{y},\boldsymbol{\alpha}) = C\nu(E_i)g(C\nu(E_i)\lambda_{s,i})$. The lower bound $a_{s,i}$ of the $100(1-\alpha)\%$ central credible interval for $\lambda_{s,i}$ is then given by

$$\begin{aligned}\frac{\alpha}{2} = \int_{-\infty}^{a_{s,i}} p(\lambda_{s,i}|\boldsymbol{y},\boldsymbol{\alpha})\,\mathrm{d}\lambda_{s,i} &= \int_{-\infty}^{a_{s,i}} C\nu(E_i)g(C\nu(E_i)\lambda_{s,i})\,\mathrm{d}\lambda_{s,i} \\ &= \int_{-\infty}^{C\nu(E_i)a_{s,i}} p(\lambda_i|\boldsymbol{y},\boldsymbol{\alpha})\,\mathrm{d}\lambda_i.\end{aligned}$$

Hence, the lower bound a_i for λ_i is transformed into the lower bound $a_{s,i}$ for $\lambda_{s,i}$ via the relation $a_{s,i} = a_i/(C\nu(E_i))$. Similarly, the upper bound b_i and the posterior median $\hat{\lambda}_{\mathrm{med},i}$ are transformed via $b_{s,i} = b_i/(C\nu(E_i))$ and $\hat{\lambda}_{\mathrm{med},s,i} = \hat{\lambda}_{\mathrm{med},i}/(C\nu(E_i))$. Alternatively, when there is a sample available from the marginal posterior $p(\lambda_i|\boldsymbol{y},\boldsymbol{\alpha})$, this can be transformed into a sample from $p(\lambda_{s,i}|\boldsymbol{y},\boldsymbol{\alpha})$ by scaling each observation $\lambda_i^{(k)}$ by $1/(C\nu(E_i))$, that is $\lambda_{s,i}^{(k)} = \lambda_i^{(k)}/(C\nu(E_i))$. The sample $\{\lambda_{s,i}^{(k)}\}_{k=1}^N$ can then be used to compute the scaled median and central credible interval.

The non-uniform binning will also have be taken into account in the prior in order to penalize for the correct shape of the intensity. For example, when we wish to penalize for the norm of the solution, instead of $\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_0\|^2$, we should penalize for $\|\boldsymbol{L}_{\mathcal{E}}^0(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)\|^2$ in (5.12), where[7]

$$\boldsymbol{L}_{\mathcal{E}}^0 = \mathrm{diag}(1/\nu(E_1), 1/\nu(E_2), \ldots, 1/\nu(E_p)), \tag{7.5}$$

and we have denoted

$$\mathcal{E} = \{E_1, E_2, \ldots, E_p\}.$$

[7]There is no need to include the normalization coefficient C in $\boldsymbol{L}_{\mathcal{E}}^0$, since this can be absorbed into the hyperparameter α.

Similarly, the first-order finite-difference matrix $\boldsymbol{L}_1^1$ should be considered in the form

$$\boldsymbol{L}_{1,\mathcal{E}}^1 = \begin{bmatrix} -\frac{1}{h_1\nu(E_1)} & \frac{1}{h_1\nu(E_2)} & & & \\ & -\frac{1}{h_2\nu(E_2)} & \frac{1}{h_2\nu(E_3)} & & \\ & & \ddots & \ddots & \\ & & & -\frac{1}{h_{p-1}\nu(E_{p-1})} & \frac{1}{h_{p-1}\nu(E_p)} \end{bmatrix} \in \mathbb{R}^{(p-1)\times p},$$

where h_i is the distance between the center points of bins E_i and E_{i+1}. With the Dirichlet boundary condition for the right boundary, this becomes

$$\boldsymbol{L}_{2,\mathcal{E}}^1 = \begin{bmatrix} -\frac{1}{h_1\nu(E_1)} & \frac{1}{h_1\nu(E_2)} & & & \\ & -\frac{1}{h_2\nu(E_2)} & \frac{1}{h_2\nu(E_3)} & & \\ & & \ddots & \ddots & \\ & & & -\frac{1}{h_{p-1}\nu(E_{p-1})} & \frac{1}{h_{p-1}\nu(E_p)} \\ & & & & \frac{1}{h_p\nu(E_p)} \end{bmatrix} \in \mathbb{R}^{p\times p}. \tag{7.6}$$

This corresponds to assuming that the intensity is zero outside the right boundary of the true histogram, which we know to be the case for the inclusive jet spectrum because of the kinematic cut-off at $\sqrt{s}/(2\cosh(y_{\min}))$. Note that the appearance of h_p in $\boldsymbol{L}_{2,\mathcal{E}}^1$ requires us to define an imaginary pseudobin E_{p+1} outside of the right boundary of the true histogram.

Similar non-uniform finite-difference schemes are available for the second derivative as well. However, the second derivative requires a boundary condition for both ends of the true histogram which is problematic in the case of a power-law spectrum like the one shown in Figure 7.15. Hence, this topic will not be pursued further in here.

7.2.3 Unfolding Results

We now proceed to unfolding studies of the inclusive jet differential cross section data. In preliminary regularization studies of the spectrum, it was found out that we need to take into account the special nature of the first couple of bins of the true histogram. Namely, the two first true bins E_1 and E_2 serving as underflow bins to model the smearing from low p_{T} values, are only slightly constrained by the data via their contribution to the first few smeared bins. The same is true for the third true bin E_3 because there is significant migration of events from this bin to unobserved low p_{T} values. As a result, the standard regularization procedures are too strong for these bins. To solve the problem, we replace the standard 2-norm $\|\cdot\|$ in (5.12) with

a weighted 2-norm $\|\cdot\|_{\boldsymbol{w}}$ defined by

$$\|\boldsymbol{x}\|_{\boldsymbol{w}} = \|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{d}(w_i x_i)^2}, \quad \boldsymbol{x} \in \mathbb{R}^d$$

for a weight vector $\boldsymbol{w}$ with components $w_i > 0$, $i = 1, \ldots, d$. As earlier, we set $\boldsymbol{\lambda}_0 = \boldsymbol{0}$ and hence regularize the problem by considering the prior

$$p(\boldsymbol{\lambda}|\alpha) \propto 1_{\mathbb{R}^p_+}(\boldsymbol{\lambda}) \exp\left(-\alpha\|\boldsymbol{L}\boldsymbol{\lambda}\|^2_{\boldsymbol{w}}\right). \tag{7.7}$$

Similarly, Equation (6.4) on the M-step of the MCEM algorithm is replaced with

$$\alpha^{(k+1)} = \frac{1}{\frac{2}{pN}\sum_{i=1}^{N}\|\boldsymbol{L}\boldsymbol{\lambda}^{(i)}\|^2_{\boldsymbol{w}}}.$$

The choices we consider for $\boldsymbol{L}$ are the $p \times p$ matrices $\boldsymbol{L}^0_{\mathcal{E}}$ and $\boldsymbol{L}^1_{2,\mathcal{E}}$ defined by Equations (7.5) and (7.6). For the weights $\boldsymbol{w}$, we use

$$w_i = \frac{\int_F \int_{E_i} \mathcal{N}(p^s_{\mathrm{T}}|p_{\mathrm{T}}, \sigma(p_{\mathrm{T}})^2) f(p_{\mathrm{T}}) \,\mathrm{d}p_{\mathrm{T}} \,\mathrm{d}p^s_{\mathrm{T}}}{\int_{E_i} f(p_{\mathrm{T}}) \,\mathrm{d}p_{\mathrm{T}}}, \quad i = 1, \ldots, p. \tag{7.8}$$

This is the fraction of events of the true bin E_i that on average migrate to the observable smeared space F before taking into account the post-smearing efficiency $\varepsilon_{\mathrm{PS}}(p^s_{\mathrm{T}})$ and represents the amount of information about each bin E_i available in the observations. For realism, the weights are computed using the perturbed version of $f(p_{\mathrm{T}})$ shown in Figure 7.15 instead of the unknown correct intensity. For the given true binning, we find $w_1 \approx 0.01$, $w_2 \approx 0.20$ and $w_3 = 0.73$. For the rest of the bins, the weights are higher than 98 %.

For sampling from the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y}, \alpha)$, we employ the same sampling scheme as in the case of the Gaussian mixture model experiments with the important change that instead of Equations (7.1) and (7.2), we define the variances of the proposal density by

$$\sigma_i^2 = \left(\gamma \frac{y_{i-2}}{w_i \bar{\varepsilon}_{i-2}}\right)^2, \quad i = 1, \ldots, p, \tag{7.9}$$

where w_i is the weight (7.8) of the bin E_i and $\bar{\varepsilon}_i$ is the mean of the post-smearing efficiency function $\varepsilon_{\mathrm{PS}}$ over the smeared bin F_i

$$\bar{\varepsilon}_i = \frac{1}{\nu(F_i)} \int_{F_i} \varepsilon_{\mathrm{PS}}(p^s_{\mathrm{T}}) \,\mathrm{d}p^s_{\mathrm{T}}.$$

The values of $y_{-1}/\bar{\varepsilon}_{-1}$ and $y_0/\bar{\varepsilon}_0$ corresponding to bins E_1 and E_2 cannot be determined from the observations $\boldsymbol{y}$ and were instead extrapolated using splines from the values of $y_i/\bar{\varepsilon}_i$ for bins E_3 to E_7. The rationale for using (7.9) is the same as before with $\sigma^2_{2,i}$ but here we need to at least approximately take into account the post-smearing efficiency in order to scale the observations $\boldsymbol{y}$ to the same order of magnitude as $\boldsymbol{\lambda}$. In addition, we chose to scale with $1/w_i$ in order to increase the size

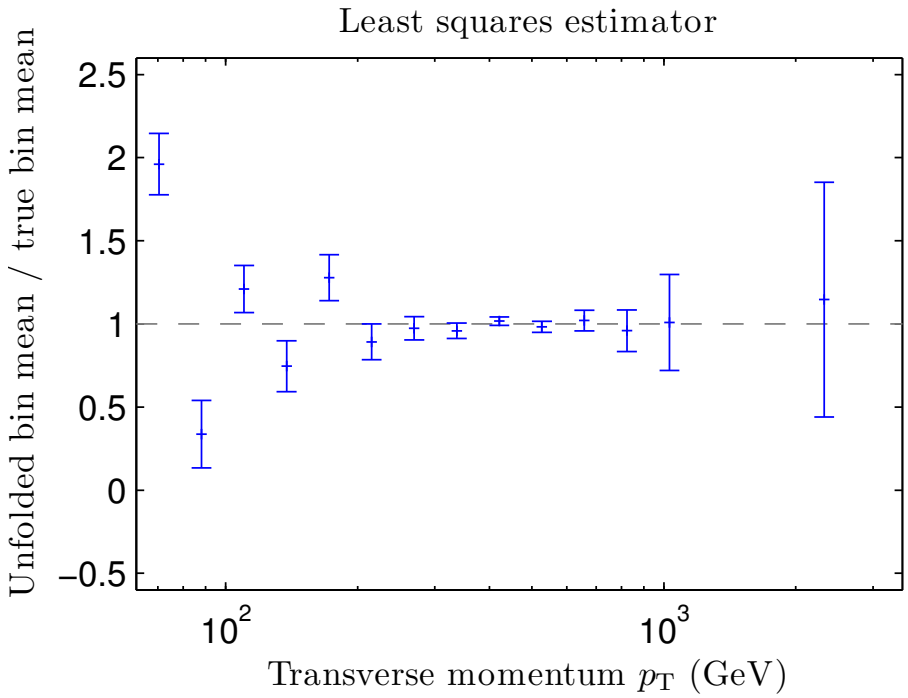


Figure 7.18: Unfolding of the inclusive jet differential cross section using the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y}$. The figure shows the component-wise ratio of the estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ to the correct mean $\boldsymbol{\lambda}$. The error bars are given by the estimated standard deviations $\widehat{\mathrm{Std}}[\hat{\lambda}_{\mathrm{LS},i}|\,\boldsymbol{\lambda}]$ followed by scaling with $1/\lambda_i$. The solution requires regularization for the small p_{T} values where the smearing is the strongest.

of the jumps for the uncertain first three true bins. Using a similar line of reasoning, the Metropolis–Hastings chain was started from $\lambda_i^{(1)} = y_{i-2}/\bar{\varepsilon}_{i-2},\ i = 1, \ldots, p$, which serves as our preliminary first guess for $\boldsymbol{\lambda}$.

The inclusive jet data will first be unfolded using Bayesian unfolding with the uniform non-negativity prior (5.11) and then using empirical Bayes unfolding for the truncated weighted-norm Gaussian smoothness prior defined in (7.7). For comparison, we also show the results for the least squares estimator.

The best way to demonstrate the ill-posedness of the least squares estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{y}$ is to show the ratio of $\hat{\boldsymbol{\lambda}}_{\mathrm{LS}}$ to the correct mean $\boldsymbol{\lambda}$. This is shown in Figure 7.18 along with error bars computed using (4.40) and scaled appropriately. We see that the estimator is clearly off the ideal value of unity especially for low p_{T} values where there is the strongest smearing and we expect to be able to tame these apparent oscillations by regularizing the problem. We can also see that clearly the error bars shown do not cover the correct value at unity. The most likely reason for this is that due to the column-rank deficiency of $\boldsymbol{K}$, the least squares estimator is biased (see Equation (4.17)) and hence the precautions outlined in Section 4.2.3 for using the estimated standard deviations to represent the uncertainty of a biased estimator apply in here.

To avoid any confusion, we have not included the indirectly observed bins E_1

and E_2 in Figure 7.18. The same is true for the rest of the figures that follow. These bins, which were included in the model to account for the smearing from the low p_{T} values, are poorly constrained by the observations which in general results in suboptimal inference results. These "helper bins" should be part of the unfolding procedure in order to produce correct results for the bins $E_3, \ldots, E_{16}$ but can be safely discarded from the final plots as they are not of interest to us in their own right.

Let us then unfold the inclusive jet spectrum using Bayesian unfolding with the uniform non-negativity prior. To this end, we sampled $N = 400\,000$ observations from the posterior with the step size $\gamma = 0.022$ giving us an acceptance rate of 32 %. The time series of the components of this Metropolis–Hastings chain are shown in Figure 7.19 which shows good convergence and mixing for all bins of the true histogram. The unfolded differential cross section shown in Figure 7.20 is obtained after scaling by $1/(L\nu(E_i))$ as explained in Section 7.2.2 and we see that the spectrum matches nicely with the desired result. To examine this result more closely, Figure 7.21 shows the ratios of the sampled λ_i's to their correct values. Paradoxically, the outcome where we have the smallest errors for medium values of p_{T} matches our expectations since the trigger prescaling increases the errors at low p_{T} values. Nevertheless, we see that clearly the posterior variance of the first few bins could be smaller.

We then proceed to perform empirical Bayes unfolding with the truncated weighted-norm Gaussian smoothness prior with $\boldsymbol{L} = \boldsymbol{L}^0_{\mathcal{E}}$. The parameters of the MCEM algorithm were set up as in the case of the Gaussian mixture model. That is, starting from $\alpha^{(0)} = 1 \cdot 10^{-4}$, we perform 30 EM iterations with sample size $N = 100\,000$ and then perform the final sampling with $M = 400\,000$ observations. Figure 7.22 shows that with the step size $\gamma = 0.022$, the iteration converges quickly to an MMLE of $\hat{\alpha}_{\mathrm{MMLE}} = 3.8 \cdot 10^{-12}$. The acceptance rate for the final sampling was 32 % and the time series of the corresponding Metropolis–Hastings chain shown in Figure 7.23 indicate no problems with the MCMC sampler. Due to the scales involved, it is difficult to see the differences of the various unfolding procedures by looking at the log-log plots of the plain cross section spectra. Hence, we show again in Figure 7.24 the ratio of the unfolded means to the true means which enables straightforward comparison with the earlier Figure 7.21. From this, we see that there is a major reduction in the posterior variance of the first few ill-posed bins of the true histogram. This is due to the additional regularization provided by the Gaussian smoothness prior.

We complete the inclusive jet experiments by penalizing for the first derivative in empirical Bayes unfolding by setting $\boldsymbol{L} = \boldsymbol{L}^1_{2,\mathcal{E}}$ in the Gaussian smoothness prior. Again, as shown in Figure 7.25, the MCEM algorithm converges with step size $\gamma = 0.02$ in approximately 10 iterations to an MMLE of $\hat{\alpha}_{\mathrm{MMLE}} = 2.4 \cdot 10^{-9}$. The acceptance rate of the final sampling was 31 % with the time series shown in Figure 7.26 indicating good mixing of the chain. The ratio plots of Figure 7.27 reveal that the additional regularization provided by the use of the first derivative has further reduced the size of the posterior uncertainty while still maintaining the median ratios close to unity. All in all, it seems that empirical Bayes unfolding with the

first derivative penalty $\boldsymbol{L} = \boldsymbol{L}^1_{2,\mathcal{E}}$ appears to provide an unfolded solution which has no major oscillations or artifacts and has a reasonable, well-understood uncertainty associated with it.

We conclude this section by noting that we have shown for comparison purposes in Figures 7.21(a), 7.24(a) and 7.27(a) the frequentist $\sqrt{n}$ errors of a perfect detector without any smearing. Because of the trigger prescaling, computation of these errors is not as straightforward for the inclusive jet data as earlier for the Gaussian mixture model data. In order to take the trigger prescaling into account, the reference errors shown in the ratio plots are computed using

$$\frac{\hat{\lambda}_{\mathrm{med},i}}{\lambda_i} \pm \frac{\sqrt{\bar{\varepsilon}_{f,i}\hat{\lambda}_{\mathrm{med},i}}}{\bar{\varepsilon}_{f,i}\lambda_i}, \tag{7.10}$$

where $\hat{\lambda}_{\mathrm{med},i}$ is the median of the ith marginal posterior, λ_i is the correct bin mean of the ith true bin and $\bar{\varepsilon}_{f,i}$ is the average efficiency of the ith bin given by

$$\bar{\varepsilon}_{f,i} = \frac{\int_{E_i} \varepsilon_{\mathrm{PS}}(p_{\mathrm{T}}) f(p_{\mathrm{T}})\,\mathrm{d}p_{\mathrm{T}}}{\int_{E_i} f(p_{\mathrm{T}})\,\mathrm{d}p_{\mathrm{T}}}, \quad i = 1, \ldots, 16.$$

The uncertainty in Equation (7.10) is formed by first scaling $\hat{\lambda}_{\mathrm{med},i}$ by the efficiency $\bar{\varepsilon}_{f,i}$ which gives us the corresponding number of Poisson counts after trigger prescaling. When these counts are used as an MLE for the prescaled Poisson mean, the error is given by $\sqrt{\bar{\varepsilon}_{f,i}\hat{\lambda}_{\mathrm{med},i}}$. This is then scaled by $1/\bar{\varepsilon}_{f,i}$ to give us an error for the unprescaled mean. Finally, the estimates are scaled by the correct means λ_i in order to produce the ratio plot.

When we compare the reference errors computed this way to the 68.27 % credible intervals produced by the Bayesian techniques, we see that in all cases the reference errors are smaller than the Bayesian errors. This matches our intuitive expectation of smearing increasing the uncertainty of the measurement with respect to the ideal detector. We also see that the stronger the regularization, the closer the Bayesian errors are to the reference errors.

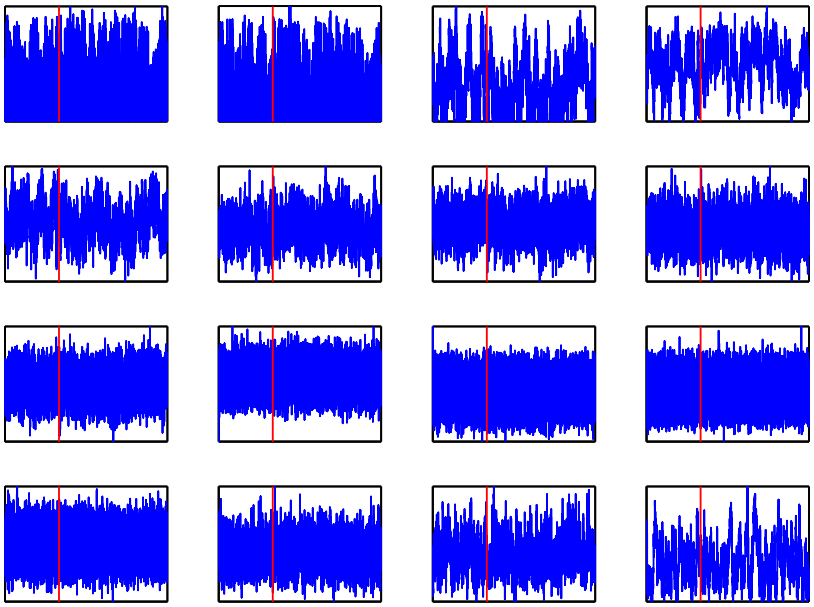


Figure 7.19: Time series of the components λ_i of the Metropolis–Hastings chain with the uniform non-negativity prior. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.

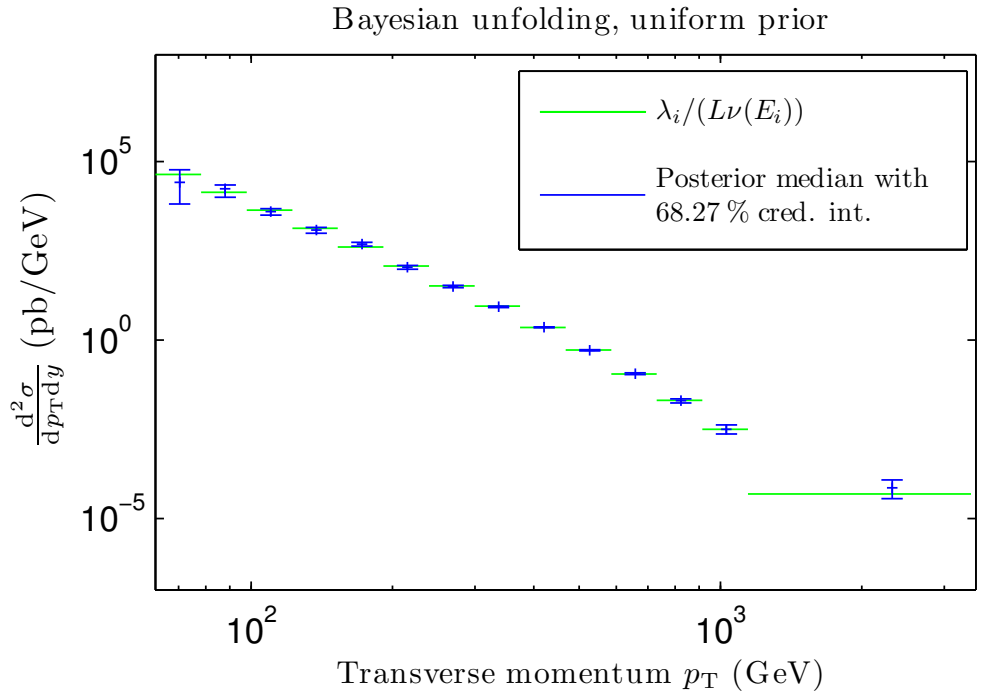


Figure 7.20: Comparison of the results of Bayesian unfolding of the inclusive jet differential cross section to the correct value of $\boldsymbol{\lambda}$ when using the uniform non-negativity prior. The figure shows the bin means λ_i scaled with the bin width $\nu(E_i)$ and normalized by the integrated luminosity L. The blue error bars indicate the central 68.27 % credible intervals which are to be compared to the correct values shown by the green histogram.

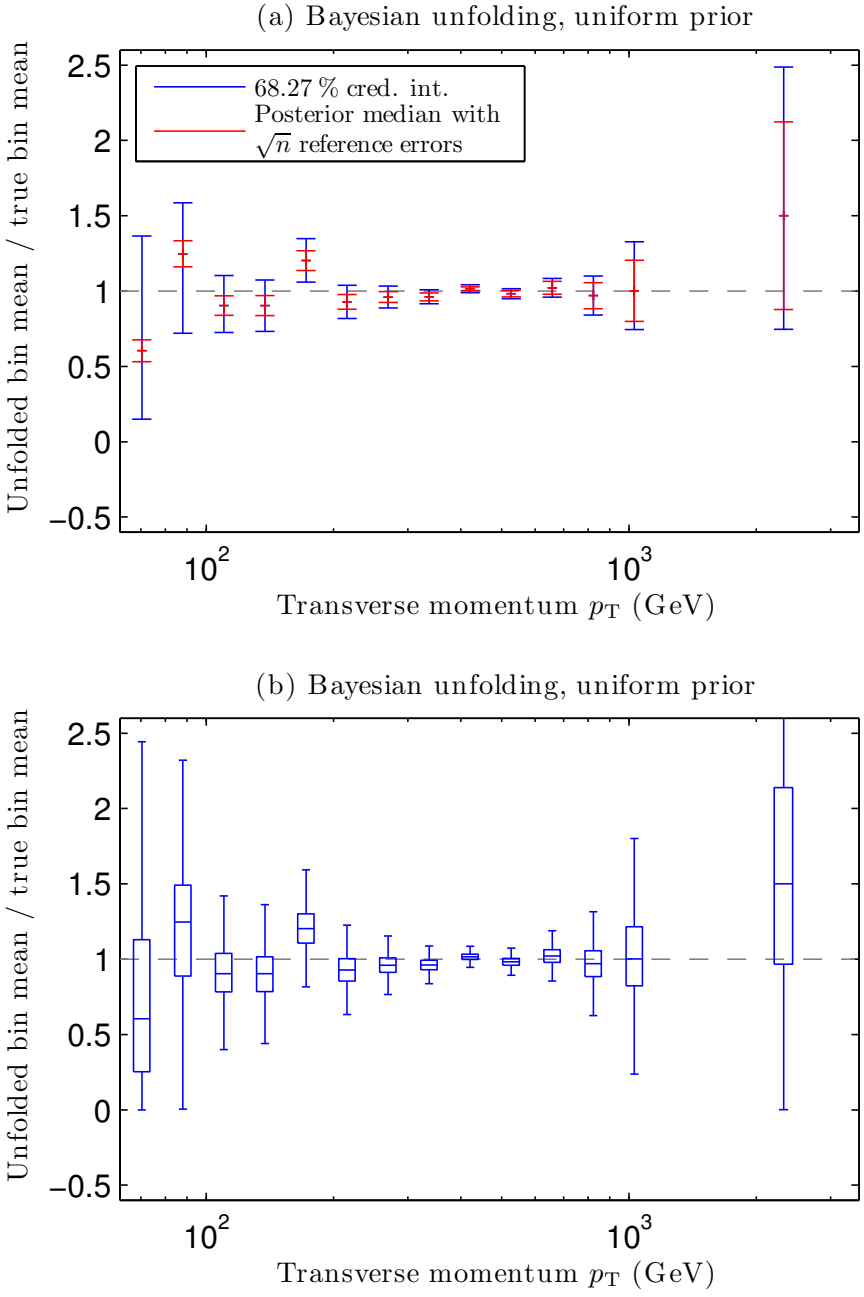


Figure 7.21: Comparison of the results of Bayesian unfolding of the inclusive jet differential cross section to the correct value of $\boldsymbol{\lambda}$ when using the uniform non-negativity prior. The results are shown for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y})$ after dividing by the correct bin means. In Figure (a), the blue error bars indicate the central 68.27 % credible intervals of these ratios. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of an ideal detector without any smearing. Figure (b) shows a box plot of the ratios where the horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

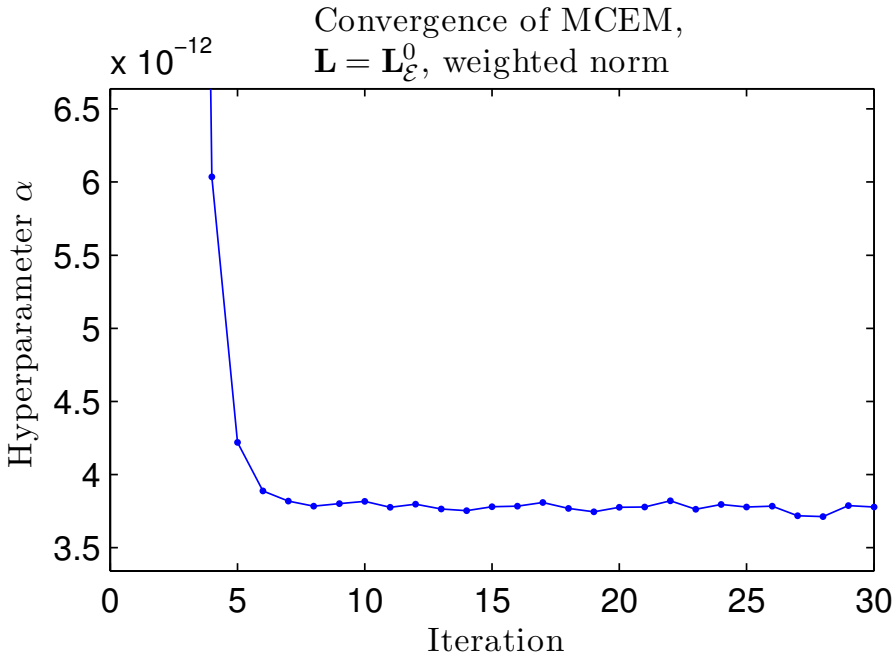


Figure 7.22: Convergence of the MCEM algorithm in empirical Bayes unfolding for the truncated weighted-norm Gaussian smoothness prior penalizing for the norm of the solution with $\boldsymbol{L} = \boldsymbol{L}^0_{\mathcal{E}}$.

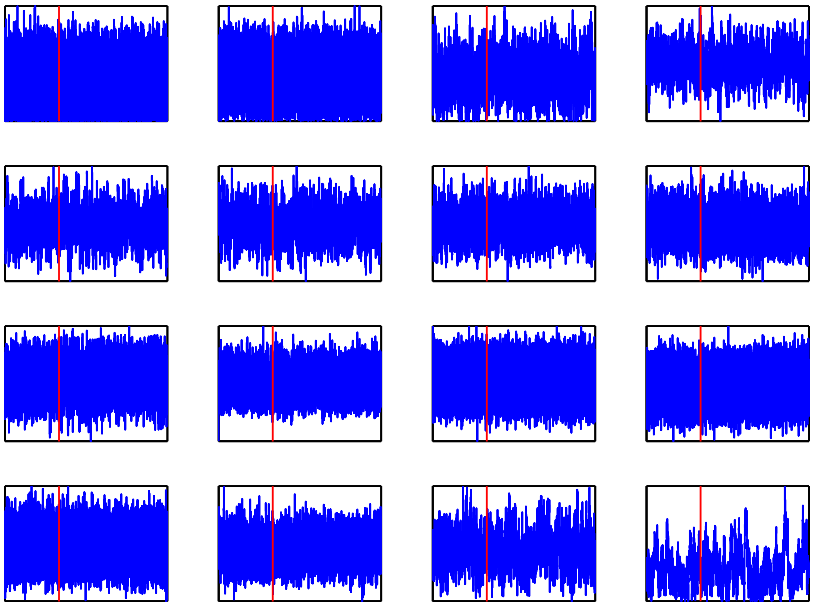
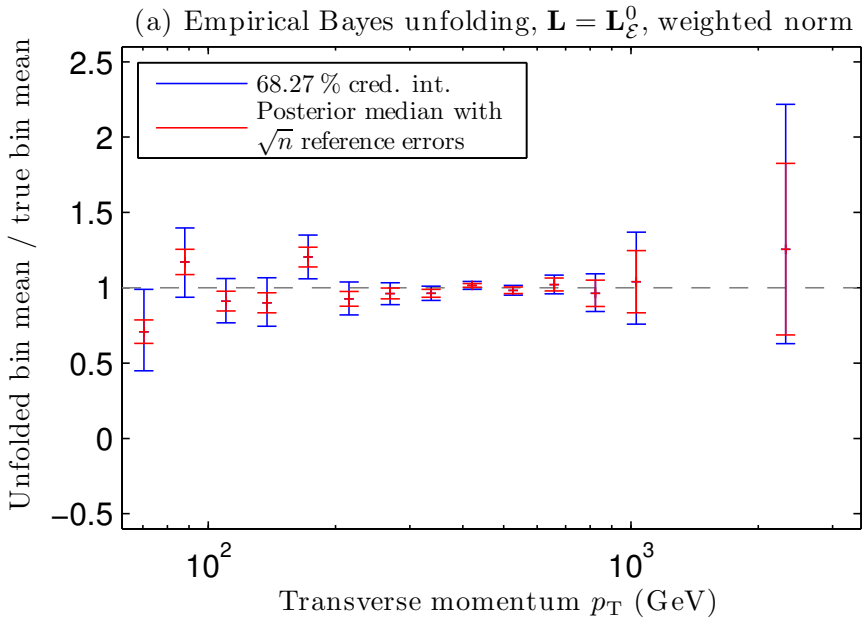


Figure 7.23: Time series of the components λ_i of the Metropolis–Hastings chain for the MMLE obtained from the MCEM algorithm in empirical Bayes unfolding for the truncated weighted-norm Gaussian smoothness prior penalizing for the norm of the solution with $\boldsymbol{L} = \boldsymbol{L}^0_{\mathcal{E}}$. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.



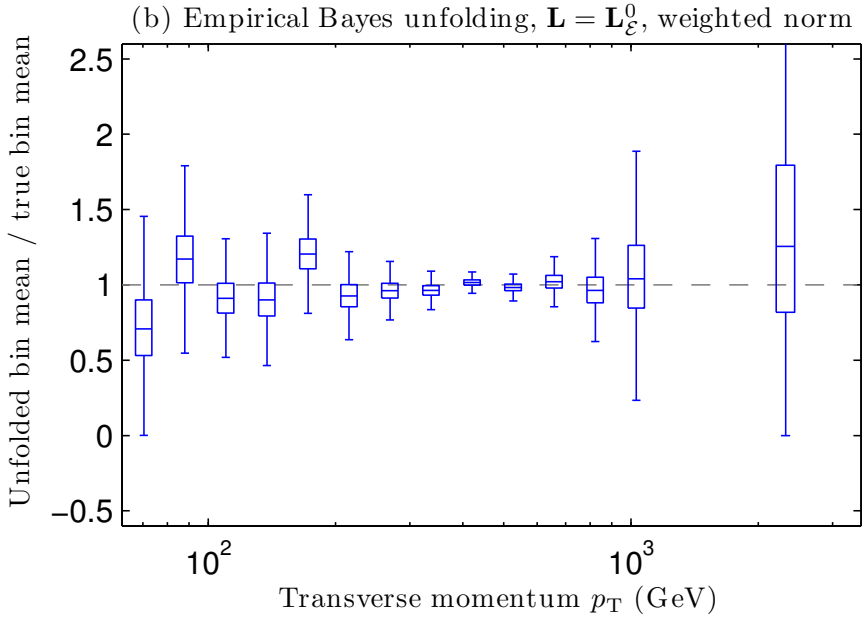


Figure 7.24: Comparison of the results of empirical Bayes unfolding of the inclusive jet differential cross section to the correct value of $\boldsymbol{\lambda}$ when using the truncated weighted-norm Gaussian smoothness prior penalizing for the norm of the solution with $\boldsymbol{L} = \boldsymbol{L}^0_{\mathcal{E}}$. The results are shown for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_{\mathrm{MMLE}})$ after dividing by the correct bin means. In Figure (a), the blue error bars indicate the central 68.27 % credible intervals of these ratios. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of an ideal detector without any smearing. Figure (b) shows a box plot of the ratios where the horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

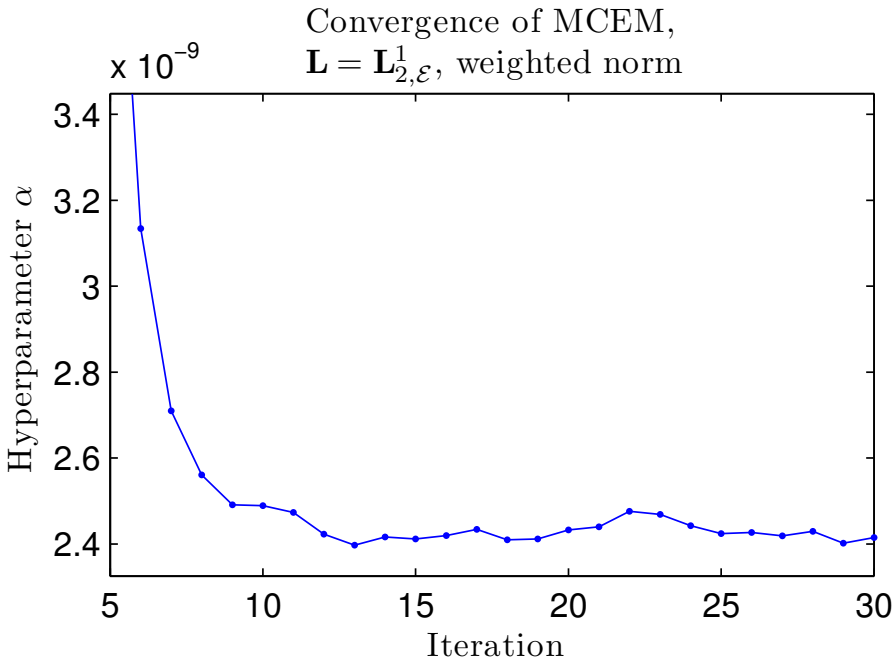


Figure 7.25: Convergence of the MCEM algorithm in empirical Bayes unfolding for the truncated weighted-norm Gaussian smoothness prior penalizing for the first derivative of the solution with $\boldsymbol{L} = \boldsymbol{L}^1_{2,\mathcal{E}}$.

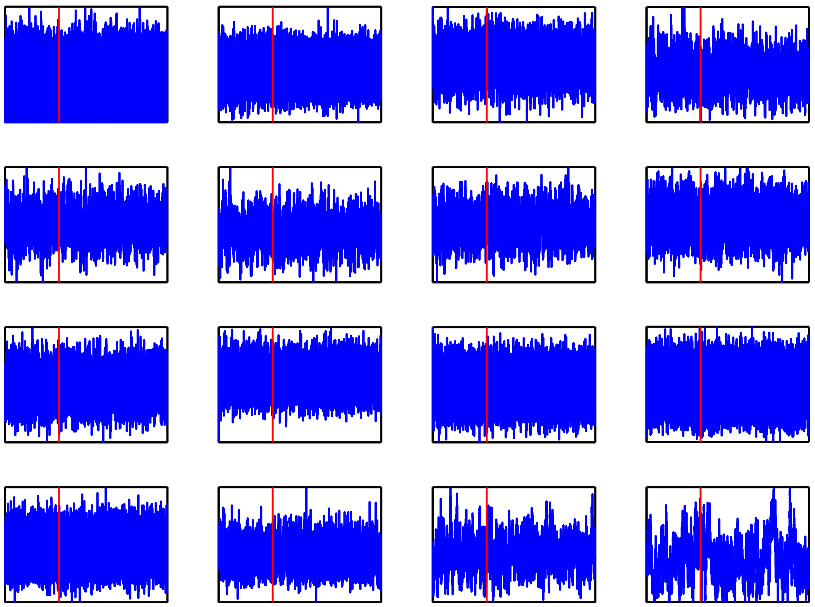
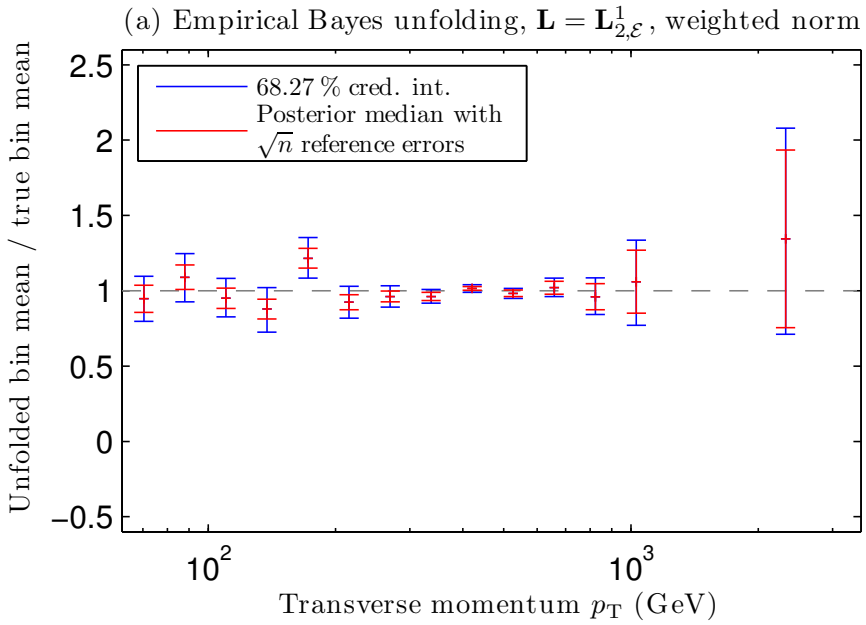


Figure 7.26: Time series of the components λ_i of the Metropolis–Hastings chain for the MMLE obtained from the MCEM algorithm in empirical Bayes unfolding for the truncated weighted-norm Gaussian smoothness prior penalizing for the first derivative of the solution with $\boldsymbol{L} = \boldsymbol{L}^1_{2,\mathcal{E}}$. The component indices i increase from left to right and top to bottom. The vertical red lines show the length of the burn-in.



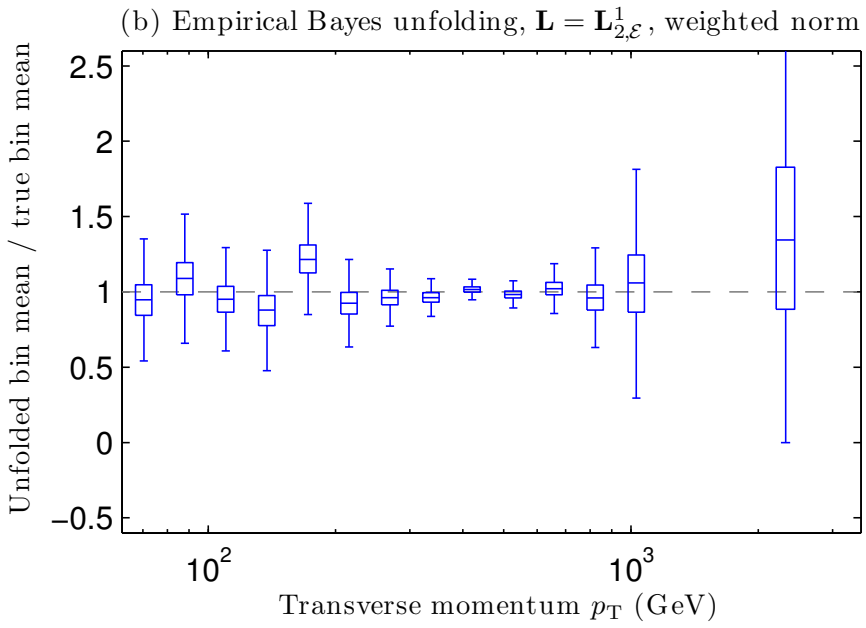


Figure 7.27: Comparison of the results of empirical Bayes unfolding of the inclusive jet differential cross section to the correct value of $\boldsymbol{\lambda}$ when using the truncated weighted-norm Gaussian smoothness prior penalizing for the first derivative of the solution with $\boldsymbol{L} = \boldsymbol{L}^1_{2,\mathcal{E}}$. The results are shown for samples from the marginal posteriors $p(\lambda_i|\boldsymbol{y}, \hat{\alpha}_\mathrm{MMLE})$ after dividing by the correct bin means. In Figure (a), the blue error bars indicate the central 68.27 % credible intervals of these ratios. For comparison purposes, the red error bars show the $\sqrt{n}$ errors of an ideal detector without any smearing. Figure (b) shows a box plot of the ratios where the horizontal lines are the sample medians and the boxes show the interquartile ranges while the whiskers extend to the smallest (largest) datum still within 1.5 times the interquartile range from the lower (upper) quartile.

Chapter 8

Discussion and Conclusions

To conclude this thesis, we first discuss possible directions for future unfolding studies in Section 8.1. We then give in Section 8.2 a concise list of observations that were made in the course of this work regarding the high energy physics unfolding problem. In a number of cases, we also give recommendations on good unfolding practices and refer the reader to the relevant literature or the appropriate parts of this thesis. We then summarize this work in Section 8.3.

8.1 Directions for Future Work

We have focused in this work on the discrete version of the unfolding problem with a particular emphasis on the error estimation of the solution using empirical Bayes techniques. Nevertheless, there are several aspects of the problem that were not discussed in detail. In this section, we outline general directions for future unfolding studies and discuss a number of potential improvements to the empirical Bayes unfolding technique presented in Chapter 6.

The main emphasis of this thesis has been on the discrete version of the unfolding problem. However, the Poisson process formulation of the problem, as explained in Section 2.1, enables us to also study the continuous problem

$$h(\boldsymbol{y}) = (Kf)(\boldsymbol{y}) = \int k(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{8.1}$$

where f is the unknown intensity function of the true Poisson process and h is the intensity function of the smeared Poisson process. In such a case, unfolding could proceed by first estimating h using the smeared observations $\boldsymbol{Y}_i$. The resulting estimator $\hat{h}$ is then substituted in (8.1) and we solve the equation $\hat{h} = Kf$ to find an estimator of f.

There are at least two reasons why the continuous version of the problem is more appealing than the discrete one. Firstly, we saw in Section 2.1.5 that the discrete smearing matrix $\boldsymbol{K}$ depends on the unknown intensity f. This is a problem since knowledge of $\boldsymbol{K}$ is needed to estimate f in the first place. We solved this issue by using various approximations of $\boldsymbol{K}$ to remove the dependency on f, but this is clearly not a completely satisfactory solution. Note, however, that the dependence on f only

appears because of the discretization of the problem. That is, in the continuous case, the smearing kernel k does not depend on f. Hence, by considering the continuous case, determination of k can arguably be disentangled from the inference of f.

The second benefit of continuous methods is related to problems in choosing the binning of the discrete histograms. An attentive reader might have noticed that we have not discussed in detail how one should choose the binnings of the true histogram $\boldsymbol{x}$ and the smeared histogram $\boldsymbol{y}$. This is because optimal choice of these binnings appears to be a very challenging problem. The reason for this is that the discretization already contributes towards regularization of the problem. Hence, the parameters p and q, i.e., the number of bins in the true and observed histograms, can be thought of as additional regularization parameters and, presumably, the optimal choice of these parameters is intertwined with the optimal choice of the main regularization parameter of the unfolding algorithm itself. The situation clearly gets even more complicated if one allows for non-uniform binning of the histograms. In HEP applications, the choice of the histogram binnings has, thus far, been rather arbitrary and the use of continuous methods would remove such arbitrariness from the problem and also allow using only one layer of regularization with a single adjustable regularization parameter.

An aspect of the real-life unfolding problem that we have also ignored in the earlier parts of this thesis is related to the fact that we have assumed the smearing kernel k to have a fixed known value. However, in a real-world particle physics experiment, the smearing kernel k has to be estimated either from Monte Carlo simulations or from calibration measurements. Hence, we would first have to come up with a way of estimating k and then using the estimator $\hat{k}$ in the unfolding procedure. In this case, the uncertainty in $\hat{k}$ should also contribute to the uncertainty of the unfolded intensity $\hat{f}$ (or the uncertainty of the unfolded histogram $\hat{\boldsymbol{\lambda}}$ in the discrete case). It is likely that a global Bayesian approach to the problem could again allow for well-founded error estimation of the solution. The details on how this should be done are left as a subject of future unfolding studies.

These long-term considerations aside, there are also several ways of improving the empirical Bayes unfolding technique presented in Chapter 6. Clearly, one of the drawbacks of the method is that the choice of the proposal density in the Metropolis–Hastings sampler requires a lot of manual fine-tuning and cross-checking and, even so, the results are not guaranteed to be optimal. The use of Gibbs sampling would largely solve this problem but, as noted in Section 5.2, the form of the posterior of the unfolding problem is such that efficient Gibbs sampling is not feasible. However, it might still be possible to reduce, or avoid altogether, the expensive numerical computations required by the Gibbs sampler by considering some suitable approximation of the full Gibbs sapling scheme. Probably the most promising idea for improving the sampling scheme would be to use adaptive MCMC techniques (see, e.g., [10, Section 3.4.4]), where the sampler is adapted to the shape of the posterior based on the output of the earlier iterations of the algorithm. Several technical improvements of the standard Metropolis–Hastings algorithm are also possible. For example, the sampler would probably behave better, especially for the bins with low bin contents, if the proposals were forced to always be non-negative. This could be done either by

imposing a reflecting barrier at zero for the proposals or by considering a proposal density with a non-negative support, such as, for example, the log-normal density.

In empirical Bayes unfolding, we have also ignored the uncertainty associated with the marginal maximum likelihood estimator $\hat{\boldsymbol{\alpha}}_{\mathrm{MMLE}}$ of the hyperparameters $\boldsymbol{\alpha}$. Since the uncertainty on $\boldsymbol{\alpha}$ contributes to the uncertainty on $\boldsymbol{\lambda}$, taking this effect properly into account would increase the errors associated with the empirical Bayes solution. Several techniques for incorporating this uncertainty into the empirical Bayes approach have been proposed in the literature, see, e.g., [10, Section 5.4]. One attractive possibility would be the bootstrap technique of Laird and Louis [36], but this could turn out be computationally unfeasible unless significant improvements to the MCMC sampling scheme can be made. An alternative option would be to compute an estimated covariance matrix for $\hat{\boldsymbol{\alpha}}_{\mathrm{MMLE}}$, which can be readily obtained from the EM iteration [43, Chapter 4], and then use this information to quantify the uncertainty on $\boldsymbol{\alpha}$. A detailed analysis of these ideas will again be left as a subject of future studies.

We also focused in this work on one particular class of prior distributions, namely the Gaussian smoothness priors with the non-negativity constraint given by Equation (5.12). In particular, in order to be able to normalize the truncated prior density for the MCEM algorithm, we were forced to set $\boldsymbol{\lambda}_0 = \mathbf{0}$ which, in some sense, means that our reference spectrum is a histogram with no observations. Especially in the case of a steeply falling power-law spectrum such as the inclusive jet spectrum analyzed in Section 7.2, this might not be the most appropriate choice. This is because the true solution is known to have significantly non-zero bin contents as well as first- and second-order derivatives. As a result, it would be interesting to see if, by dropping the non-negativity requirement of the prior, it was possible to include also $\boldsymbol{\lambda}_0$ in the MCEM algorithm and hence find the marginal maximum likelihood estimate of both α and $\boldsymbol{\lambda}_0$.

Clearly, in addition to the Gaussian smoothness priors, a number of other possibilities can also be considered. In fact, in the Bayesian analysis, the choice of the prior density should be regarded as a modeling question where as much problem-specific information about the plausible solutions as possible should be incorporated into the unfolding procedure. Especially with such challenging spectra as the steeply falling inclusive jet spectrum considered in Section 7.2, the performance of Bayesian unfolding could potentially by greatly improved by considering problem-specific priors. For example, in the inclusive jet analysis, one could consider a prior which is concentrated around the theoretical spectrum as predicted by the theory of quantum chromodynamics. If the problem-specific prior involves any free hyperparameters, and in many cases it does, then empirical Bayes unfolding with the MCEM algorithm can, at least in principle, be used to find their data-driven estimates.

Let us furthermore note that despite being the standard general-purpose prior for regularizing ill-posed problems in the literature on statistical inverse problem, the Gaussian smoothness prior (5.12) might not be the optimal choice when a non-negativity constraint of the solution is involved. Note, for example, that without any constraints the Gaussian smoothness prior is symmetric with respect to the mean $\boldsymbol{\lambda}_0$ but this is no longer true when the non-negativity constraint is included.

Hence, densities that are inherently non-negative such as the multivariate log-normal density might be more natural and appropriate choices when such constraints are involved.

Finally, we note that there is a number of technical issues related to the implementation of the MCEM algorithm in empirical Bayes unfolding that were not studied in detail. Firstly, in the computational experiments of Chapter 7, we ran the algorithm for a predetermined number of iterations and verified its convergence by plotting the iterates as a function of the iteration number. Clearly, if a fully automated unfolding algorithm is desired, one should implement a proper stopping rule for the iteration. In addition, it is often recommended in the MCEM literature that the posterior sample size N is kept small in the beginning of the iteration and then increased as the iteration proceeds and more accuracy is needed. These issues are discussed, e.g., in [7].

8.2 Observations and Recommendations

In this section, we list a number of observations and related recommendations about current unfolding practices in high energy physics. The observations are accompanied with references to the literature and the relevant parts of this thesis.

- It is often claimed (see Section 11.2 in [13]) that the estimator $\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$ is the maximum likelihood estimator of the true histogram $\boldsymbol{\lambda}$ when the smearing matrix $\boldsymbol{K}$ is invertible. However, this is in general not true because $\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$ is not guaranteed to satisfy the non-negativity constraint $\boldsymbol{\lambda} \geq \boldsymbol{0}$. In fact, as explained in Section 4.1, there is no closed-form solution to the maximum likelihood problem (4.5). Nevertheless, the MLE always exists and can be found using the EM iteration.

- As shown in Section 4.1.2, the "Bayesian" D'Agostini iteration of [16] is equivalent to the EM iteration for the maximum likelihood estimator of $\boldsymbol{\lambda}$. That is, there is nothing "Bayesian" about the method. Furthermore, in astronomy and optics, the same iteration is known as the Lucy–Richardson deconvolution [40, 50].

- As a result of the two observations above, contrary to common belief, the D'Agostini iteration will not, in general, converge to the solution given by $\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$.

- The SVD unfolding method described in [28] corresponds to a certain generalization of Tikhonov regularization (see Equation (4.38)). Furthermore, SVD only provides an approximate solution to this problem. If an exact solution is desired, one should either use the Moore–Penrose pseudoinverse or the generalized singular value decomposition [2, Section 5.4].

- As we see, the terminology used in high energy physics for the different unfolding techniques is not consistent with the terminology used in statistics for

the same algorithms. Table 8.1 summarizes the naming conventions of the two fields.

- Out of the commonly used unfolding techniques, only the D'Agostini iteration and Bayesian unfolding enforce non-negativity of the solution. In particular, Tikhonov regularization and its generalizations can give solutions with negative bin contents.

- When one imposes the smoothness of the unfolded histograms in Tikhonov regularization and Bayesian techniques, it is important to take boundary conditions properly into account in the discretized differential operator $\boldsymbol{L}$. If this is not done properly, it is easy to implicitly require a solution that vanishes on the boundaries even though it is known that in reality this should not be the case. See Section 4.2.2 for a discussion about the relationship between the boundary conditions and the choice of $\boldsymbol{L}$.

- When non-uniform binning is used, it is also crucial to change $\boldsymbol{L}$ to reflect this. For example, the SVD method, as described in [28], does not seem to properly take this into account. See Section 7.2.2 for a discussion on how non-uniform binning can be incorporated in $\boldsymbol{L}$.

- Quite interestingly, the D'Agostini iteration with early stopping does not explicitly enforce any boundary conditions. However, it is not immediately clear what should be the physical interpretation of the regularization provided by this approach.

- As described in Section 4.2.3, it is possible to have errors smaller than $\sqrt{n}$ in frequentist unfolding if only the estimated standard deviation of the estimator is used to construct the error bars. This is because regularization makes the estimators biased in which case the estimated standard deviations cannot be regarded as approximate confidence intervals for the true solution $\boldsymbol{\lambda}$. In fact, the errors constructed this way can be made arbitrarily small by increasing the strength of the regularization.

- The techniques presented in Section 4.2.3 for error estimation of TSVD and Tikhonov regularization are not applicable to the D'Agostini iteration because of its nonlinearity. D'Agostini [16] provides a way of estimating the uncertainty of the solution but his calculations have been criticized by Adye [1]. Clearly, error estimation of the D'Agostini iteration has not been fully settled yet.

- Whenever likelihood-based unfolding techniques such as the D'Agostini iteration or Bayesian unfolding are used, one should use the plain Poisson-distributed event counts as the input to the algorithm. In particular, when these counts are scaled to correct for, e.g., trigger efficiency, the scaled observations are no longer Poisson distributed and cannot be used in the Poisson likelihood.

- Unfolding by bin-by-bin correction factors does not seem to have a counterpart in the statistical inverse problems literature. This is perhaps because the method corrects for the "efficiency" of each bin instead of bin-to-bin migrations. As such, the method introduces a significant, undesired bias for the Monte Carlo model used for deriving the correction factors and hence should be avoided altogether.

- It often happens that events can get smeared both into and out of the observed histogram near its boundaries. If this is the case, then it is important to take this into account in the unfolding procedure. For a demonstration of how this can be done, see the computational experiment of Section 7.2.

- Bayesian unfolding as described in [11] requires an essentially arbitrary choice of the prior density. Empirical Bayes unfolding described in Chapter 6 provides a partial solution to this problem by allowing one to choose the "strength" of the prior $p(\boldsymbol{\lambda}|\boldsymbol{\alpha})$ using the observed data $\boldsymbol{y}$. The choice of the parametric family of priors $\{p(\boldsymbol{\lambda}|\boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$ is still left to the discretion of the analyst, but the amount of subjectivity that remains is comparable to choosing a certain frequentist regularization technique or penalty term.

- When Metropolis–Hastings sampling is used in Bayesian unfolding, it is essential to verify the convergence and mixing of the chain as described in Section 5.2 and demonstrated in practice in Chapter 7. If the convergence and mixing is not satisfactory, the proposal density will have to be modified accordingly. As a word of caution, the original paper [11] on Bayesian unfolding does not discuss these important issues in detail.

- In all the computational experiments presented in Chapter 7, the uncertainty of the solution given by empirical Bayes unfolding was, as expected, larger than the $\sqrt{n}$ errors of an ideal detector. There are, however, no guarantees for this to be true in the general case. This is because regularization uses information from adjacent bins to constrain the possible values of each unfolded bin and it is possible that the use of such additional information enables reduction of the uncertainty to a value smaller than $\sqrt{n}$. However, as shown by the computational experiments, this seems to rarely happen in practice.

- Unfolding algorithms should never be used as black boxes. This is because analysis-specific information has to be always incorporated in the procedure to obtain correct results. When choosing and customizing the unfolding procedure, the factors that one should take into account include, for example, information about the expected shape of the distribution, handling of the histogram boundaries and the desired way of treating detector inefficiencies.

Table 8.1: The naming conventions of common unfolding techniques in high energy physics and the names of the corresponding algorithms in the statistics literature. The matrix inversion technique refers to the estimator $\hat{\boldsymbol{\lambda}} = \boldsymbol{K}^{-1}\boldsymbol{y}$, while the D'Agostini iteration is described in [16] and the SVD technique in [28].

Name in HEP	Name in statistics
Matrix inversion	Method of moments estimation
Iterative Bayesian unfolding / D'Agostini iteration	Expectation-maximization algorithm
SVD unfolding	Generalized Tikhonov regularization

8.3 Concluding Remarks

We have analyzed in this thesis the high energy physics unfolding problem from a statistical point of view. We first formulated a mathematical model for the problem using the theory of Poisson point processes. We then investigated in detail the tools provided by both the frequentist and Bayesian paradigms of statistics for solving the problem. It was shown that the bias of regularized point estimators makes error estimation of frequentist unfolding challenging, while the main issue with Bayesian methods is the choice of the regularization strength imposed by the prior density. To solve these issues, we proposed using an empirical Bayes unfolding technique which combines elements of the frequentist and Bayesian approaches. We derived a Monte Carlo EM algorithm for finding the marginal maximum likelihood estimate of the hyperparameters of the regularizing prior density and then used the resulting posterior to construct Bayesian credible intervals for the solution. The desired performance of the method was verified with computational experiments using simulated data. In addition to proposing a novel, well-performing unfolding technique derived starting from the first principles of the problem, a number of insights were gained about good unfolding practices which will benefit future physics analyses using unfolding as part of the analysis chain.

References

[1] T. Adye. Corrected error calculation for iterative Bayesian unfolding. `http://hepunx.rl.ac.uk/~adye/software/unfold/bayes_errors.pdf`, 2011. Retrieved on June 27, 2012.

[2] R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter Estimation and Inverse Problems*. Elsevier, 2005.

[3] M. Bazaraa, H. Sherali, and C. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley-Interscience, 3rd edition, 2006.

[4] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer-Verlag, 2nd edition, 2003.

[5] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 2nd edition, 1986.

[6] G. Bohm and G. Zech. *Introduction to Statistics and Data Analysis for Physicists*. Deutsches Elektronen-Synchrotron, 2010.

[7] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(1):265–285, 1999.

[8] S. P. Brooks and G. O. Roberts. Convergence assesment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8:319–335, 1998.

[9] S. L. Campbell and C. D. Meyer, Jr. *Generalized Inverses of Linear Transformations*. Dover Publications, 1979.

[10] B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, 3rd edition, 2009.

[11] G. Choudalakis. Fully Bayesian unfolding. arXiv:1201.4612v4 [physics.data-an], 2012.

[12] R. D. Cousins. Why isn't every physicist a Bayesian? *American Journal of Physics*, 63(5):398–410, 1995.

[13] G. Cowan. *Statistical Data Analysis*. Oxford University Press, 1998.

[14] D. Cox and V. Isham. *Point Processes*. Chapman & Hall/CRC, 1980.

[15] P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[16] G. D'Agostini. A multidimensional unfolding method based on Bayes' theorem. *Nuclear Instruments and Methods A*, 362:487–498, 1995.

[17] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes, Volume 1: Elementary Theory and Methods*. Springer-Verlag, 2nd edition, 2003.

[18] B. N. Datta. *Numerical Linear Algebra and Applications*. SIAM, 2nd edition, 2010.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[20] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 2000.

[21] S. N. Evans and P. B. Stark. Inverse problems as statistics. *Inverse Problems*, 18(4):R55, 2002.

[22] F. Garwood. Fiducial limits for the Poisson distribution. *Biometrika*, 28(3/4):437–442, 1936.

[23] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004.

[24] G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, 3rd edition, 1996.

[25] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992.

[26] D. A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, 1997.

[27] J. G. Heinrich. Coverage of error bars for Poisson data. CDF Public Note CDF/MEMO/STATISTICS/PUBLIC/6438, 2003.

[28] A. Höcker and V. Kartvelishvili. SVD approach to data unfolding. *Nuclear Instruments and Methods in Physics Research A*, 372:469–481, 1996.

[29] J. Jacod and P. Protter. *Probability Essentials*. Springer-Verlag, 2nd edition, 2004.

[30] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[31] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, 2005.

[32] O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, 2nd edition, 2002.

[33] A. F. Karr. *Point Processes and Their Statistical Inference*. Marcel Dekker, 2nd edition, 1991.

[34] J. Kingman. *Poisson Processes*. Clarendon Press, 1993.

[35] K. Knight. *Mathematical Statistics*. Chapman & Hall/CRC, 2000.

[36] N. M. Laird and T. A. Louis. Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82(399):739–757, 1987.

[37] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, 8(2):306–316, 1984.

[38] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. SIAM, 1995.

[39] R. A. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

[40] L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79(6):745–754, 1974.

[41] L. Lyons. Unfolding: Introduction. In H. B. Prosper and L. Lyons, editors, *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, CERN-2011-006, pages 225–228, CERN, Geneva, Switzerland, 17–20 January 2011.

[42] A. Malinverno and V. A. Briggs. Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics*, 69(4):1005–1016, 2004.

[43] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 2nd edition, 2008.

[44] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[45] L. Mirsky. *An Introduction to Linear Algebra*. Dover Publications, 1990.

[46] Y. Mitsuhata. Adjustment of regularization in ill-posed linear inverse problems by the empirical Bayes approach. *Geophysical Prospecting*, 52(3):213–239, 2004.

[47] V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Mathematics Doklady*, (7):414–417, 1966.

[48] H. N. Mülthei and B. Schorr. On an iterative method for the unfolding of spectra. *Nuclear Instruments and Methods in Physics Research A*, 257:371–377, 1987.

[49] R.-D. Reiss. *A Course on Point Processes*. Springer-Verlag, 1993.

[50] W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

[51] S. Chatrchyan et al., CMS Collaboration. Measurement of the inclusive jet cross section in pp collisions at $\sqrt{s} = 7$ TeV. *Physical Review Letters*, 107:132001, 2011.

[52] J. Shao. *Mathematical Statistics*. Springer, 2nd edition, 2003.

[53] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2):113–122, 1982.

[54] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):3–23, 1993.

[55] Z. Szkutnik. Unfolding intensity function of a Poisson process in models with approximately specified folding operator. *Metrika*, 52:1–26, 2000.

[56] Y. Vardi, L. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.

[57] E. Veklerov and J. Llacer. Stopping rule for the MLE algorithm based on statistical hypothesis testing. *IEEE Transactions on Medical Imaging*, 6(4):313–319, 1987.

[58] C. R. Vogel. *Computational Methods for Inverse Problems*. SIAM, 2002.

[59] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[60] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[61] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

[62] G. A. Young and R. L. Smith. *Essentials of Statistical Inference*. Cambridge University Press, 2005.

Appendix A

Mathematical Background

This appendix gives an overview of the mathematical tools used in this thesis. The topics covered are measure-theoretic probability theory in Section A.1, the mathematical theory of statistical inference in Section A.2, linear algebra with an emphasis on the properties of the Moore–Penrose pseudoinverse in Section A.3 and an introduction to inverse problems in Section A.4. Our treatment is based on well-known, established literature on each particular subject with the relevant references given at the beginning of each section.

A.1 Introduction to Probability Theory

This section is devoted to providing a review of the main concepts and results of probability theory that are used throughout this thesis. These results can be found on any standard text book on probability theory, such as Jacod and Protter [29] or Kallenberg [32]. Unless otherwise indicated, our treatment here is based on [52, Chapter 1].

The concept of a measure is central to mathematical probability theory. In order to proceed with the definition, we first need to introduce the concept of a σ-algebra.

Definition A.1. Let $\mathcal{F}$ be a collection of subsets of some set Ω. Then $\mathcal{F}$ is called a *σ-algebra* on Ω if it has the following properties:

(i) $\emptyset \in \mathcal{F}$

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c = \Omega \backslash A$ is the complement of A.

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\cup A_i \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is called a *measurable space*. In measure theory, the set Ω is simply some space of interest and the σ-algebra $\mathcal{F}$ contains the measurable subsets of this space. In probability theory, however, these abstract objects have an intuitive interpretation. Namely, the set Ω is called the *sample space* and its elements represent all the different possible outcomes of the random experiment we are studying. For example, in the case of tossing a 6-face die, we could set $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Similarly, the elements of the σ-algebra $\mathcal{F}$ represent different *events* for which we can assign a probability. For instance, in the case of the die, the event corresponding to getting an even outcome, would be the set $\{2, 4, 6\} \subset \Omega$.

When the space of interest Ω is the d-dimensional real space $\mathbb{R}^d$, there is a certain σ-algebra which often turns out to be handy. This is the *Borel σ-algebra*, which is defined to be the smallest σ-algebra containing all the open sets of $\mathbb{R}^d$. We denote this σ-algebra on $\mathbb{R}^d$ by $\mathcal{B}^d$ and the elements of $\mathcal{B}^d$ are called the *Borel sets* of $\mathbb{R}^d$. When $E \subset \mathbb{R}^d$ is a Borel set and $\mathcal{B}_E = \{E \cap B : B \in \mathcal{B}^d\}$, one can show that $\mathcal{B}_E$ is a σ-algebra on E and hence $(E, \mathcal{B}_E)$ forms a measurable space. Furthermore, it is easy to see that when E is countable, $\mathcal{B}_E = \mathcal{P}(E)$, where $\mathcal{P}(E)$ is the power set of E.

A measure is then a way of assigning a "size" to an element of $\mathcal{F}$.

Definition A.2. Let $(\Omega, \mathcal{F})$ be a measurable space. A *measure* μ on $\mathcal{F}$ is a mapping from $\mathcal{F}$ to the extended non-negative real line

$$\mu : \mathcal{F} \to \mathbb{R}_+ \cup \{+\infty\}$$

with the following properties:

(i) $\mu(\emptyset) = 0$

(ii) If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

The triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*. The measure μ is *finite* if $\mu(A) < \infty$ for all $A \in \mathcal{F}$. Moreover, a finite measure μ is called a *probability measure* when $\mu(\Omega) = 1$. In this case, we denote the measure by P and call the triple $(\Omega, \mathcal{F}, P)$ a *probability space*. The interpretation of the probability measure P is that it assigns a probability $P(A) \in [0, 1]$ to each event A in the σ-algebra $\mathcal{F}$.

On the measurable space $(\mathbb{R}^d, \mathcal{B}^d)$, it is natural to work with a measure μ which satisfies the conventional notion of volume for Cartesian products of intervals

$$\mu([a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]) = (b_1 - a_1) \cdot (b_2 - a_2) \cdots (b_d - a_d).$$

It can be shown that there is a unique measure ν, called the *Lebesgue measure*, which satisfies this condition[1]. The Lebesgue measure ν is used exclusively throughout this thesis when the space of interest is the d-dimensional real space $\mathbb{R}^d$ or some Borel subset of this space.

Another important example of a measure is the *counting measure* ϱ. It is defined on a measurable space $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is the power set $\mathcal{P}(\Omega)$ of Ω, by counting the

[1]The Lebesgue measure ν can actually be defined for a larger family of sets than the Borel σ-algebra $\mathcal{B}^d$ and one can show that all Lebesgue measurable sets form a σ-algebra which has $\mathcal{B}^d$ as a sub-σ-algebra. However, for our needs, the measure space $(\mathbb{R}^d, \mathcal{B}^d, \nu)$ is general enough.

number of elements in the set $A \in \mathcal{F}$. That is, $\varrho(A)$ is the number of elements in $A \in \mathcal{F}$ when A is finite and for an infinite A, we set $\varrho(A) = \infty$.

A third example of a measure that we often encounter is the *Dirac measure* δ_x at $x \in \Omega$. It is defined for any σ-algebra $\mathcal{F}$ of subsets of Ω by

$$\delta_x(A) = 1_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}, \quad A \in \mathcal{F},$$

where 1_A is the *indicator function* of the set A.

It is often convenient to think of Ω as some abstract underlying space of outcomes ω and regard the actual observations as a measurable function of these outcomes. This gives rise to the concept of a random variable. Let us first define what we mean by a measurable function.

Definition A.3. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces. A function $f : \Omega \to \Lambda$ is said to be *measurable* if the preimages $f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\} = \{f \in B\}$ belong to $\mathcal{F}$ for all $B \in \mathcal{G}$.

When in the definition above $(\Lambda, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$, where $\mathcal{B} = \mathcal{B}^1$ is the Borel σ-algebra of $\mathbb{R}$, we call f a *Borel function*. Furthermore, when $(\Omega, \mathcal{F}, P)$ is a probability space, we call a measurable function $X : \Omega \to \Lambda$ a $(\Lambda, \mathcal{G})$*-valued random element*. An important special case arises when Λ is the d-dimensional real space $\mathbb{R}^d$ or some Borel subset E of this space and $\mathcal{G}$ is chosen to be the Borel σ-algebra $\mathcal{B}_E$ associated with this space. In this case, we call the measurable function $X := \boldsymbol{X}$ an E*-valued random variable* or just a *random variable*[2].

It is conventional to denote random variables by capital letters, such as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \ldots$, in which case we refer to the mapping from Ω to $E \subset \mathbb{R}^d$. The value $\boldsymbol{X}(\omega)$ that this function obtains is then often denoted by the corresponding lower case letter, e.g., $\boldsymbol{X}(\omega) = \boldsymbol{x}$ or $\boldsymbol{X} = \boldsymbol{x}$ for short. When it is clear from the context whether we mean the random variable or just its random realization, we denote the both using the lower case letters $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots$ in order to avoid awkward and clumsy notation.

We next introduce the important concept of the distribution of a random element.

Definition A.4. Let X be a $(\Lambda, \mathcal{G})$-valued random element. Then the *distribution* of X, denoted by P_X, is

$$P_X(B) = (P \circ X^{-1})(B) = P(X^{-1}(B)) = P(X \in B), \quad B \in \mathcal{G}.$$

Proposition A.5. *P_X is a probability measure on $\mathcal{G}$.*

Proof. Let $B_1, B_2, \ldots \in \mathcal{G}$ be disjoint. Then

$$P_X(\cup_i B_i) = P(X^{-1}(\cup_i B_i)) = P(\cup_i X^{-1}(B_i)) = \sum_i P(X^{-1}(B_i)) = \sum_i P_X(B_i).$$

Since, in addition, $P_X(\emptyset) = P(\emptyset) = 0$, we see that P_X is a measure. Furthermore, $P_X(\Lambda) = P(\Omega) = 1$ and hence P_X is a probability measure. □

[2]Boldface notation is used whenever $d > 1$.

Hence, the distribution P_X is the probability measure which gives the probability of having the value of X in the set B, $P_X(B) = P(X \in B)$. Here the measurability of X guarantees that the set $X^{-1}(B) = \{X \in B\}$ belongs to the σ-algebra $\mathcal{F}$ where the probability measure P is defined. When the distribution of X is given by P_X, we write $X \sim P_X$.

When $\boldsymbol{X}$ is a random variable, it is often convenient to characterize its distribution $P_{\boldsymbol{X}}$ using the cumulative distribution function.

Definition A.6. Let $\boldsymbol{X}$ be an $\mathbb{R}^d$-valued random variable. The *cumulative distribution function* (cdf) of $\boldsymbol{X}$, denoted by $F_{\boldsymbol{X}}$, is then defined by

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{X}}(x_1, \ldots, x_d) = P_{\boldsymbol{X}}((-\infty, x_1] \times \cdots \times (-\infty, x_d]), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{A.1}$$

One can in fact show that there is a one-to-one correspondence between the cdf $F_{\boldsymbol{X}}$ and the distribution $P_{\boldsymbol{X}}$ and hence we can use either one to describe the random variable $\boldsymbol{X}$.

There are two important classes of random variables that we often encounter. An E-valued random variable $\boldsymbol{X}$, where E is a Borel set of $\mathbb{R}^d$, is said to be *continuous* if its distribution $P_{\boldsymbol{X}}$ can be written in the form

$$P_{\boldsymbol{X}}(A) = \int_A p_{\boldsymbol{X}} \, \mathrm{d}\nu = \int_A p_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \quad A \in \mathcal{B}_E, \tag{A.2}$$

where ν is the Lebesgue measure and $p_{\boldsymbol{X}} : E \to \mathbb{R}_+$ is a non-negative Borel function called the *probability density function* (pdf) of $\boldsymbol{X}$. Since $P_{\boldsymbol{X}}(E) = P(\Omega) = 1$, we see that the pdf $p_{\boldsymbol{X}}$ of a continuous random variable $\boldsymbol{X}$ integrates to unity,

$$\int_E p_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 1.$$

Another common class of random variables is that of *discrete* random variables which are E-valued random variables $\boldsymbol{X}$ where $E = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$ is countable subset of $\mathbb{R}^d$. It follows that the distribution $P_{\boldsymbol{X}}$ of $\boldsymbol{X}$ can be written as

$$P_{\boldsymbol{X}}(A) = \int_A p_{\boldsymbol{X}} \, \mathrm{d}\varrho = \sum_{\boldsymbol{x}_i \in A} p_{\boldsymbol{X}}(\boldsymbol{x}_i), \quad A \in \mathcal{P}(E), \tag{A.3}$$

where ϱ is the counting measure and the function $p_{\boldsymbol{X}} : E \to [0, 1]$ is called the *probability mass function* (pmf) of $\boldsymbol{X}$. As above, from $P_{\boldsymbol{X}}(E) = P(\Omega) = 1$, we find $\sum_{i=1}^{\infty} p_{\boldsymbol{X}}(\boldsymbol{x}_i) = 1$. By noting that $P(\boldsymbol{X} = \boldsymbol{x}_i) = P_{\boldsymbol{X}}(\{\boldsymbol{x}_i\}) = p_{\boldsymbol{X}}(\boldsymbol{x}_i)$, we see that the value $p_{\boldsymbol{X}}(\boldsymbol{x}_i)$ of the pmf has a straightforward interpretation as the probability of the realization $\boldsymbol{X} = \boldsymbol{x}_i$.

When there is no risk of confusion, we often use the shorthand notations $p(\boldsymbol{x}), p(\boldsymbol{y}), p(\boldsymbol{z}), \ldots$ to denote the pdfs or pmfs $p_{\boldsymbol{X}}(\boldsymbol{x}), p_{\boldsymbol{Y}}(\boldsymbol{y}), p_{\boldsymbol{Z}}(\boldsymbol{z}), \ldots$ of the random variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \ldots$ Note also that although most random variables encountered in practical applications are either discrete or continuous, it is easy to define

random variables that fall on neither of these two categories. For example, the rectified Gaussian distribution, which has the form

$$P_{\boldsymbol{X}}(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x + \frac{1}{2}\delta_0(A), \quad A \in \mathcal{B}_{[0,\infty)},$$

is neither continuous nor discrete.

An important theorem by Radon and Nikodym can often be used to infer the existence of the pdf of a random variable $\boldsymbol{X}$. To provide the statement of the theorem, we first need the technical definition of a σ-finite measure.

Definition A.7. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then μ is said to be *σ-finite*, if there exists a sequence $A_1, A_2, \ldots \in \mathcal{F}$ such that $\cup A_i = \Omega$ and $\mu(A_i) < \infty$ for all i.

Clearly, any finite measure μ is also σ-finite. The Lebesgue measure ν is also σ-finite since $\mathbb{R}^d = \cup A_n$, where $A_n = (-n, n) \times \cdots \times (-n, n)$ with $\nu(A_n) = (2n)^d < \infty$. On the other hand, the counting measure ϱ is σ-finite if and only if Ω is countable.

We also need the definition of absolute continuity.

Definition A.8. Let λ and μ be two measures on the measurable space $(\Omega, \mathcal{F})$. The measure λ is said to be *absolutely continuous* with respect to μ, denoted by $\lambda \ll \mu$, if $\lambda(A) = 0$ for every $A \in \mathcal{F}$ for which $\mu(A) = 0$. In other words, $\lambda \ll \mu$ if

$$\mu(A) = 0 \Rightarrow \lambda(A) = 0, \quad \forall A \in \mathcal{F}.$$

When λ is absolutely continuous with respect to the Lebesgue measure ν, we often simply say that λ is absolutely continuous without explicitly writing down the Lebesgue measure ν.

With these definitions, we are now fully equipped to state the Radon–Nikodym theorem.

Theorem A.9. (Radon–Nikodym theorem) *Let λ and μ be two measures on the measurable space $(\Omega, \mathcal{F})$. If $\lambda \ll \mu$ and μ is σ-finite, then there exists a μ-almost everywhere unique non-negative Borel function $f : \Omega \to \mathbb{R}_+$ such that*

$$\lambda(A) = \int_A f\,\mathrm{d}\mu, \quad \forall A \in \mathcal{F}. \tag{A.4}$$

Proof. See Theorem 32.2 in [5]. □

The function f in Equation (A.4) is called the *density* (or the *Radon–Nikodym derivative*) of λ with respect to μ. Note also that (A.4) implies $\lambda \ll \mu$. That is, $\lambda \ll \mu$ is a necessary condition for (A.4) but, for sufficiency, the additional assumption on σ-finiteness of μ is required.

Comparing the statement of the Radon–Nikodym theorem with Equation (A.2) and noting that the Lebesgue measure is σ-finite, we see that an E-valued random variable $\boldsymbol{X}$ is continuous if and only if the distribution $P_{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure ν, that is, $P_{\boldsymbol{X}} \ll \nu$. Furthermore, it is easily seen that, whenever E is countable, $P_{\boldsymbol{X}}$ cannot be absolutely continuous with respect

to the Lebesgue measure ν. Hence, discrete random variables cannot be continuous. However, it always holds that $P_{\boldsymbol{X}} \ll \varrho$, where ϱ is the counting measure on $\mathcal{B}_E$. This can be seen from

$$\varrho(A) = 0 \Leftrightarrow A = \emptyset \Rightarrow P_{\boldsymbol{X}}(A) = 0, \quad \forall A \in \mathcal{B}_E.$$

Since, for countable E, the counting measure ϱ is σ-finite and $\mathcal{B}_E = \mathcal{P}(E)$, we see that the Radon–Nikodym theorem gives a formal justification for the existence of the pmf of a discrete random variable in Equation (A.3).

There are various ways of summarizing the information contained in the distribution $P_{\boldsymbol{X}}$ of an E-valued random variable $\boldsymbol{X}$, where $E \subset \mathbb{R}^d$ is a Borel set. The location of the values of the random variable can be described using the *expectation* $\mathrm{E}[\boldsymbol{X}]$ defined by the integral

$$\mathrm{E}[\boldsymbol{X}] = \int_\Omega \boldsymbol{X}(\omega)\,\mathrm{d}P(\omega) = \int_E \boldsymbol{x}\,\mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x}) \tag{A.5}$$

assuming that these integrals exist and are finite for each component of $\boldsymbol{X}$. If this is the case, $\boldsymbol{X}$ is said to be *integrable*. The expectation of $\boldsymbol{X}$ is also called the *mean* or *expected value* of $\boldsymbol{X}$.

Similarly, the spread of the values of $\boldsymbol{X}$ around its expectation $\mathrm{E}[\boldsymbol{X}]$ can be summarized using the *covariance matrix*

$$\mathrm{Cov}[\boldsymbol{X}] = \mathrm{E}\big[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^{\mathrm{T}}\big]$$

given that all the elements of this expectation and $\mathrm{E}[\boldsymbol{X}]$ are finite. By writing out the product, we see that the covariance is equivalently also given by

$$\mathrm{Cov}[\boldsymbol{X}] = \mathrm{E}\big[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big] - \mathrm{E}[\boldsymbol{X}]\mathrm{E}[\boldsymbol{X}]^{\mathrm{T}}.$$

The diagonal elements of $\mathrm{Cov}[\boldsymbol{X}]$ are the *variances* of the components X_i. That is, the variance of X_i, denoted by $\mathrm{Var}[X_i]$, is given by

$$\mathrm{Var}[X_i] = \mathrm{E}\big[(X_i - \mathrm{E}[X_i])^2\big] = \mathrm{E}\big[X_i^2\big] - \mathrm{E}[X_i]^2.$$

The square root of the variance is called the *standard deviation* of X_i and denoted by $\mathrm{Std}[X_i]$. That is

$$\mathrm{Std}[X_i] = \sqrt{\mathrm{Var}[X_i]} = \sqrt{\mathrm{E}[(X_i - \mathrm{E}[X_i])^2]}.$$

The covariance has the following useful transformation property:

$$\mathrm{Cov}[\boldsymbol{A}\boldsymbol{X}] = \boldsymbol{A}\,\mathrm{Cov}[\boldsymbol{X}]\,\boldsymbol{A}^{\mathrm{T}}, \tag{A.6}$$

where $\boldsymbol{X}$ is a d-dimensional random variable and $\boldsymbol{A}$ is a matrix with d columns. This result can be easily shown as follows:

$$\begin{aligned}
\mathrm{Cov}[\boldsymbol{A}\boldsymbol{X}] &= \mathrm{E}\big[(\boldsymbol{A}\boldsymbol{X} - \mathrm{E}[\boldsymbol{A}\boldsymbol{X}])(\boldsymbol{A}\boldsymbol{X} - \mathrm{E}[\boldsymbol{A}\boldsymbol{X}])^{\mathrm{T}}\big] \\
&= \mathrm{E}\big[\boldsymbol{A}(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\big] \\
&= \boldsymbol{A}\mathrm{E}\big[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^{\mathrm{T}}\big]\boldsymbol{A}^{\mathrm{T}} \\
&= \boldsymbol{A}\,\mathrm{Cov}[\boldsymbol{X}]\,\boldsymbol{A}^{\mathrm{T}},
\end{aligned}$$

Table A.1: Properties of the discrete distributions used in this thesis. Certain values of the parameters of these distributions will give rise to the expression 0^0 which should be interpreted as 1.

Distribution	Properties	
Binomial	Notation	$\mathrm{Bin}(p, n)$
	Parameters	success probability $p \in [0, 1]$; number of trials $n \in \mathbb{N}_0$
	pmf	$\binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$
	Expectation	np
	Variance	$np(1-p)$
Multinomial	Notation	$\mathrm{Mult}(\boldsymbol{p}, n)$
	Parameters	probabilities $\boldsymbol{p} = \left[p_1, \ldots, p_d\right]^{\mathrm{T}}, p_i \in [0, 1]$, $\sum_i p_i = 1$; number of trials $n \in \mathbb{N}_0$
	pmf	$\frac{n!}{x_1! \cdots x_d!} p_1^{x_1} \cdots p_d^{x_d}$, $x_i \in \{0, 1, \ldots, n\}$, $\sum_i x_i = n$
	Expectation	$n\boldsymbol{p}$
	Covariance	$n\left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^{\mathrm{T}}\right)$
Poisson	Notation	$\mathrm{Poisson}(\lambda)$
	Parameters	mean $\lambda \geq 0$
	pmf	$\frac{\lambda^x}{x!} e^{-\lambda}$, $x = 0, 1, 2, \ldots$
	Expectation	λ
	Variance	λ

where we have used the linearity of the expectation operator defined by Equation (A.5).

There is a number of common probability distributions that are often encountered in practice. Table A.1 summarizes the properties of the standard discrete distributions used in this thesis, while Table A.2 provides the same information for continuous distributions.

It is often desirable to consider the *conditional distribution* of a random variable $\boldsymbol{X}$ given the value $\boldsymbol{y}$ of another random variable $\boldsymbol{Y}$. We note that a rigorous definition of a conditional distribution is a relatively involved topic requiring substantial use of σ-algebras and conditional expectations. Here we simply content ourselves with stating that given an E-valued random variable $\boldsymbol{X}$ and an F-valued random variable $\boldsymbol{Y}$, there is way of defining the conditional distribution $P_{\boldsymbol{X}}(\cdot|\boldsymbol{Y} = \boldsymbol{y})$ which, for any fixed $\boldsymbol{y} \in F$, is a probability measure on $\mathcal{B}_E$ and has the interpretation as the distribution of the random variable $\boldsymbol{X}$ given the realization $\boldsymbol{y}$ of the random variable $\boldsymbol{Y}$. For more details, we refer the reader to [52, Section 1.4].

An important theorem known as Bayes' rule relates the densities of the conditional distributions $P_{\boldsymbol{X}}(\cdot|\boldsymbol{Y} = \boldsymbol{y})$ and $P_{\boldsymbol{Y}}(\cdot|\boldsymbol{X} = \boldsymbol{x})$.

Table A.2: Properties of the continuous distributions used in this thesis.

Distribution	Properties	
Chi-square	Notation	$\chi^2(k)$
	Parameters	degrees of freedom $k \in \mathbb{N}$
	pdf	$\frac{1}{\Gamma(k/2)2^{k/2}} x^{k/2-1} e^{-x/2}$, $x \geq 0$
	Expectation	k
	Variance	$2k$
Multivariate Gaussian	Notation	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
	Parameters	mean $\boldsymbol{\mu} \in \mathbb{R}^d$; covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ (positive definite)
	pdf	$\frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$, $\boldsymbol{x} \in \mathbb{R}^d$
	Expectation	$\boldsymbol{\mu}$
	Covariance	$\boldsymbol{\Sigma}$
Uniform	Notation	$U(a, b)$
	Parameters	bounds $a, b \in \mathbb{R}$, $a < b$
	pdf	$\frac{1}{b-a}$, $x \in [a, b]$
	Expectation	$\frac{a+b}{2}$
	Variance	$\frac{(b-a)^2}{12}$

Theorem A.10. (Bayes' rule) *Let $\boldsymbol{X}$ be an E-valued random variable and $\boldsymbol{Y}$ an F-valued random variable, where E and F are Borel sets of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and assume that we know the distributions $P_{\boldsymbol{Y}}(\cdot|\boldsymbol{X} = \boldsymbol{x})$ and $P_{\boldsymbol{X}}$. Let μ be a σ-finite measure and assume that $P_{\boldsymbol{Y}}(\cdot|\boldsymbol{X} = \boldsymbol{x}) \ll \mu$. Denote the resulting density by $p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$ and assume that this is jointly measurable with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$. Furthermore, assume that $P_{\boldsymbol{X}} \ll \lambda$ for a σ-finite measure λ and denote the resulting density by $p_{\boldsymbol{X}}(\boldsymbol{x})$. Then $P_{\boldsymbol{X}}(\cdot|\boldsymbol{Y} = \boldsymbol{y}) \ll \lambda$ and the density of $P_{\boldsymbol{X}}(\cdot|\boldsymbol{Y} = \boldsymbol{y})$ with respect to λ is given by*

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{y}) = \frac{p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x})}{m(\boldsymbol{y})}, \quad \boldsymbol{x} \in E, \ \boldsymbol{y} \in F, \tag{A.7}$$

where the denominator is given by

$$m(\boldsymbol{y}) = \int_E p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{x})$$

and assumed to be strictly positive for all $\boldsymbol{y} \in F$.

Proof. See Theorem 4.1 in [52]. □

Bayes' rule is also known as *Bayes' theorem* or *Bayes' formula*. Since the denominator $m(\boldsymbol{y})$ in (A.7) can be interpreted as the marginal density of $\boldsymbol{Y}$ with respect

to μ, we may also write

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y}) = \frac{p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})}{p_{\boldsymbol{Y}}(\boldsymbol{y})}$$

and when there is no risk of confusion, we use the shorthand notation

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}. \tag{A.8}$$

When $\boldsymbol{x}$ is a continuous random variable, we may set $\lambda = \nu$, where ν is the Lebesgue measure, in which case the marginal density of $\boldsymbol{y}$ is given by

$$p(\boldsymbol{y}) = \int_E p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

In Bayes' rule (A.8), $\boldsymbol{x}$ usually represents the unknowns that we are interested in inferring and $\boldsymbol{y}$ the observed data. Because of this, $p(\boldsymbol{x}|\boldsymbol{y})$ is known as the *posterior density* of the unknowns $\boldsymbol{x}$ given the observations $\boldsymbol{y}$ and $p(\boldsymbol{x})$ as the *prior density* of the unknowns. Furthermore, $p(\boldsymbol{y}|\boldsymbol{x})$ is referred to as the *likelihood* of $\boldsymbol{x}$. For more information on the interpretation and use of Bayes' rule in statistical inference, see Sections A.2 and 5.1.

We conclude this section by introducing the concept of independence.

Definition A.11. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a finite collection of random variables with a joint cdf $F_{(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}$ defined by Equation (A.1). The random variables $\boldsymbol{X}_i$ are said to be *independent*, denoted by $\perp\!\!\!\perp \boldsymbol{X}_i$, if the joint cdf factorizes with respect to the variables

$$F_{(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} F_{\boldsymbol{X}_i}(\boldsymbol{x}_i), \tag{A.9}$$

where $F_{\boldsymbol{X}_i}$ is the marginal cdf of $\boldsymbol{X}_i$.

This definition extends to countable collections of random variables as follows:

Definition A.12. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be a countable collection of random variables. If the factorization (A.9) holds for any finite collection of these random variables, then the $\boldsymbol{X}_i$ are said to be *independent*, denoted by $\perp\!\!\!\perp \boldsymbol{X}_i$.

Let us note that when densities exist, the random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent if and only if the joint density factorizes

$$p_{(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} p_{\boldsymbol{X}_i}(\boldsymbol{x}_i).$$

The intuitive interpretation of independence is that having information about one random variable, does not help us deducing the value of the other. Namely, if two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y}) = p_{\boldsymbol{X}}(\boldsymbol{x})$.

If the factorization holds only when conditioned on another random variable, the random variables are said to be *conditionally independent*. That is, when densities

exist, the random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are said to be conditionally independent given $\boldsymbol{Y}$ if

$$p_{(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{Y} = \boldsymbol{y}) = \prod_{i=1}^{n} p_{\boldsymbol{X}_i}(\boldsymbol{x}_i|\boldsymbol{Y} = \boldsymbol{y}).$$

This is denoted by $\perp\!\!\!\perp \boldsymbol{X}_i|\boldsymbol{Y}$.

We often encounter a countable collection of random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ which are independent and all follow the same distribution $P_{\boldsymbol{X}}$, that is,

$$\perp\!\!\!\perp \boldsymbol{X}_i \quad \text{and} \quad \boldsymbol{X}_i \sim P_{\boldsymbol{X}},\ i = 1, 2, \ldots$$

Such a collection of random variables is said to be *independent and identically distributed* (i.i.d.) and denoted by $\boldsymbol{X}_i \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{X}}$. A collection of i.i.d. random variables has the important property that, by the law of large numbers, their arithmetic mean tends towards their common expectation.

Theorem A.13. (Strong law of large numbers) *Let $X_1, X_2, \ldots$ be i.i.d. random variables. Assume that the X_i are integrable and denote $\mu = \mathrm{E}[X_i]$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu,$$

where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence (i.e. convergence with probability 1).

Proof. See Theorem 1.13(ii) in [52]. □

A.2 Statistical Inference

This section is devoted to introducing the central principles of *statistical inference* where the goal is to use the observed data to extract information about some quantities of interest. Our treatment here is based on the references [21, 35, 52].

Let $(\Lambda, \mathcal{G})$ be a measurable space, where Λ is a separable metric space, and assume that we observe in our experiment a $(\Lambda, \mathcal{G})$-valued random element X with an unknown distribution P_X. Assume, however, that we know that P_X belongs to some family of probability measures $P_X \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where the index set Θ is assumed to be a non-empty subset of a separable Banach space U. The family $\mathcal{P}$ is called a *statistical model* and the elements $\theta \in \Theta$ represent possible "theories" about the state of nature. By assumption, for some unknown true $\theta \in \Theta$, we have $P_X = P_\theta$ and hence $X \sim P_\theta$. The quadruple $(\Theta, \mathcal{P}, \Lambda, \mathcal{G})$ is called a *statistical experiment* indexed by Θ and forms the basis of any statistical inference task. Furthermore, by convention, the statistical model $\mathcal{P}$ is called *parametric* if $U = \mathbb{R}^d$, otherwise the model is *nonparametric*.

The goal of statistical inference is to learn something about θ given a realization of X. The features of θ that we are interested in are called parameters.

Definition A.14. Let $(\Theta, \mathcal{P}, \Lambda, \mathcal{G})$ be a statistical experiment. Then a continuous mapping $g : \Theta \to V$, where V is a separable metric space, is called a *parameter* of the experiment.

If we are interested in θ itself, we could take $V = U$ and let g to be the identity mapping. In a more general setting, $g(\theta)$ could be, for example, the norm or some projection of θ.

A parameter $g(\theta)$ is said to be identifiable if distinct values of the parameter produce distinct probability measures P_θ.

Definition A.15. Let $g(\theta)$ be a parameter of the statistical experiment $(\Theta, \mathcal{P}, \Lambda, \mathcal{G})$. The parameter $g(\theta)$ is then said to be identifiable if

$$\Big(g(\theta_1) \neq g(\theta_2)\Big) \Rightarrow \Big(P_{\theta_1} \neq P_{\theta_2}\Big), \quad \forall \theta_1, \theta_2 \in \Theta.$$

Since all statistical inference is based on the data X, we often consider functions of the data. Such functions are called statistics.

Definition A.16. Let X be a $(\Lambda, \mathcal{G})$-valued random element in the statistical experiment $(\Theta, \mathcal{P}, \Lambda, \mathcal{G})$. Then any measurable, known function T of X is called a *statistic*.

As a transformation of the random variable X, any statistic $T(X)$ is also a random variable with a distribution $P_{T(X)}$. A statistic which is constructed in order to learn something about an unknown parameter $g(\theta)$ is called a point estimator.

Definition A.17. Let $g : \Theta \to V$ be a parameter in the statistical experiment $(\Theta, \mathcal{P}, \Lambda, \mathcal{G})$ and assume that the observations X have the distribution P_θ for some index $\theta \in \Theta$, that is $X \sim P_\theta$. Then a statistic $T : \Lambda \to V$ whose primary purpose is to estimate $g(\theta)$ is called a *point estimator* of $g(\theta)$ and denoted by $\hat{g}$.

From this definition, it is clear that in principle any statistic T could be a point estimator $\hat{g}$ of $g(\theta)$, but clearly, some statistics are better estimators of $g(\theta)$ than others. Hence, we need tools for both constructing estimators $\hat{g}$ and for evaluating their performance. For simplicity, we assume for the rest of this section that $V = \mathbb{R}^d$ and thus we write $\hat{\boldsymbol{g}}$ for $\hat{g}$ and $\boldsymbol{g}(\theta)$ for $g(\theta)$.

One of the simplest ways of measuring the performance of $\hat{\boldsymbol{g}}$ in estimating $\boldsymbol{g}(\theta)$ is to see how well the estimator performs on average. This can be measured using the bias of the estimator.

Definition A.18. Let $\hat{\boldsymbol{g}}$ be a point estimator of $\boldsymbol{g}(\theta)$. The *bias* of $\hat{\boldsymbol{g}}$ is then defined by

$$\text{bias}(\hat{\boldsymbol{g}}) = \text{E}[\hat{\boldsymbol{g}} - \boldsymbol{g}(\theta)|\theta] = \text{E}[\hat{\boldsymbol{g}}|\theta] - \boldsymbol{g}(\theta)$$

provided that the expectations are well defined. Here $\text{E}[\cdot|\theta]$ denotes expectation with respect to the measure P_θ.

When $\text{bias}(\hat{\boldsymbol{g}}) = \mathbf{0}$, the estimator $\hat{\boldsymbol{g}}$ is said to be *unbiased* which means that on average, the value of $\hat{\boldsymbol{g}}$ coincides with the true value of $\boldsymbol{g}(\theta)$.

Although small or vanishing bias is clearly a desirable feature of an estimator $\hat{\boldsymbol{g}}$, it does not give a complete picture about its distribution. Namely, it only characterizes the location of $\hat{\boldsymbol{g}}$ but does not tell us anything about the spread of its distribution. To see how the values of $\hat{\boldsymbol{g}}$ vary around its expectation, we need to look at the covariance

$$\text{Cov}[\hat{\boldsymbol{g}}|\theta] = \text{E}\big[(\hat{\boldsymbol{g}} - \text{E}[\hat{\boldsymbol{g}}|\theta])(\hat{\boldsymbol{g}} - \text{E}[\hat{\boldsymbol{g}}|\theta])^{\text{T}} \big| \theta\big].$$

Clearly, in the best-case scenario, we would have an estimator $\hat{\boldsymbol{g}}$ for which both $\text{bias}(\hat{\boldsymbol{g}})$ and the diagonal elements of $\text{Cov}[\hat{\boldsymbol{g}}|\theta]$ are small in which case the distribution of $\hat{\boldsymbol{g}}$ is concentrated around the desired value $\boldsymbol{g}(\theta)$. The mean squared error is a performance measure which provides exactly this information.

Definition A.19. Let $\hat{\boldsymbol{g}}$ be a point estimator of $\boldsymbol{g}(\theta)$. Then the *mean squared error* (MSE) of $\hat{\boldsymbol{g}}$ is defined to be

$$\text{MSE}[\hat{\boldsymbol{g}}] = \text{E}\big[\|\hat{\boldsymbol{g}} - \boldsymbol{g}(\theta)\|^2|\theta\big].$$

In some sense, the MSE measures how far on average the estimator $\hat{\boldsymbol{g}}$ is from the true value $\boldsymbol{g}(\theta)$. The following proposition relates the MSE to the bias and covariance of $\hat{\boldsymbol{g}}$.

Proposition A.20. *Let $\hat{\boldsymbol{g}}$ be a point estimator of $\boldsymbol{g}(\theta)$. Then the MSE can be decomposed as follows:*

$$\text{MSE}[\hat{\boldsymbol{g}}] = \text{tr}(\text{Cov}[\hat{\boldsymbol{g}}|\theta]) + \|\text{bias}(\hat{\boldsymbol{g}})\|^2,$$

where $\text{tr}(\text{Cov}[\hat{\boldsymbol{g}}|\theta]) = \sum_{i=1}^{d} \text{Cov}[\hat{\boldsymbol{g}}|\theta]_{ii} = \sum_{i=1}^{d} \text{Var}[\hat{g}_i|\theta]$ *is the trace of the covariance matrix.*

Proof. A simple computation gives

$$\begin{aligned}
\text{MSE}[\hat{\boldsymbol{g}}] &= \text{E}\big[\|\hat{\boldsymbol{g}} - \boldsymbol{g}(\theta)\|^2|\theta\big] \\
&= \text{E}\left[\sum_{i=1}^{d} \left(\hat{g}_i - g_i(\theta)\right)^2 \,\middle|\, \theta\right] \\
&= \sum_{i=1}^{d} \big(\text{E}\big[\hat{g}_i^2|\theta\big] - 2g_i(\theta)\text{E}[\hat{g}_i|\theta] + g_i(\theta)^2\big) \\
&= \sum_{i=1}^{d} \big(\text{E}\big[\hat{g}_i^2|\theta\big] - \text{E}[\hat{g}_i|\theta]^2 + \text{E}[\hat{g}_i|\theta]^2 - 2g_i(\theta)\text{E}[\hat{g}_i|\theta] + g_i(\theta)^2\big) \\
&= \sum_{i=1}^{d} \big(\text{Var}[\hat{g}_i|\theta] + \left(\text{E}[\hat{g}_i|\theta] - g_i(\theta)\right)^2\big) \\
&= \sum_{i=1}^{d} \text{Var}[\hat{g}_i|\theta] + \sum_{i=1}^{d} \text{bias}(\hat{g}_i)^2 \\
&= \text{tr}(\text{Cov}[\hat{\boldsymbol{g}}|\theta]) + \|\text{bias}(\hat{\boldsymbol{g}})\|^2.
\end{aligned}$$

□

We see that if we agree to use the MSE to measure the performance of the estimator, then we would like to minimize both the bias and the variance of our estimator $\hat{\boldsymbol{g}}$. It turns out that often these two properties are intertwined in such a way that increasing the bias decreases the variance and vice versa. This phenomenon is called the *bias-variance trade-off*. As a result, the global optimum of the MSE, if it exists, is often attained using a biased estimator with the benefit of a decrease in the variance. As seen in Chapter 4, this observation plays an essential role in solving inverse problems.

Although in the frequentist paradigm of statistics any statistic $\boldsymbol{T}(X)$ could serve as an estimator of $\boldsymbol{g}(\theta)$, there are a number of standard recipes that are often used to construct estimators. One of the most commonly encountered recipes is *maximum likelihood estimation*. For simplicity, assume that U and V are real spaces with possibly different dimensions and that $(\Lambda, \mathcal{G}) = (E, \mathcal{B}_E)$, where E is a Borel set of yet another real space. Hence, the observations $\boldsymbol{X}$, the indices $\boldsymbol{\theta}$ and the parameters $\boldsymbol{g}(\boldsymbol{\theta})$ are all real-valued vectors. Let us first define the likelihood function.

Definition A.21. Assume that for all $\boldsymbol{\theta} \in \Theta$ the probability measure $P_{\boldsymbol{\theta}}$ is absolutely continuous and denote its density by $p(\boldsymbol{x}|\boldsymbol{\theta})$. For a fixed $\boldsymbol{x} \in E$, we then call this density, when evaluated as a function of $\boldsymbol{\theta}$, the *likelihood function* $L(\boldsymbol{\theta})$, that is,

$$L : \Theta \to \mathbb{R}_+, \boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}).$$

The maximum likelihood estimator is then the estimator that maximizes the likelihood.

Definition A.22. The *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$ is a statistic $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood $L(\boldsymbol{\theta})$, that is,

$$L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta.$$

More generally, the MLE of the parameter $\boldsymbol{g}(\boldsymbol{\theta})$ is the estimator $\hat{\boldsymbol{g}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

Hence, the MLE corresponds, in some sense, to the value of $\boldsymbol{g}(\boldsymbol{\theta})$ which is the most likely to have produced the observations $\boldsymbol{X} = \boldsymbol{x}$.

In addition to maximum likelihood estimation, another standard procedure for constructing point estimators is a technique called the *method of moments*. Assume that we have n i.i.d. real-valued observations $X_1, \ldots, X_n \in \mathbb{R}$ from the distribution $P_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, and that we are interested in inferring $\boldsymbol{\theta}$ itself, that is $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. Assume further that $\mathrm{E}\big[|X_i|^d|\boldsymbol{\theta}\big] < \infty$ and denote

$$\mu_j = \mu_j(\boldsymbol{\theta}) = \mathrm{E}\big[X_i^j|\boldsymbol{\theta}\big], \quad j = 1, \ldots, d.$$

Then, by the law of large numbers (Theorem A.13),

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad j = 1, \ldots, d$$

can be used as an estimator of μ_j. In method of moments estimation, we equate these *sample moments* $\hat{\mu}_j$ with the theoretical moments μ_j in order to find an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, that is,

$$\hat{\mu}_j = \mu_j(\boldsymbol{\theta}), \quad j = 1, \ldots, d.$$

Any solution $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ of these d equations is then called a *method of moments estimator* of $\boldsymbol{\theta}$. This framework generalizes to multivariate observations $\boldsymbol{X}_i$ in a natural way by equating the theoretical and estimated moments of the components X_i.

The point estimator $\hat{\boldsymbol{g}}$ summarizes our understanding about the unknown $\boldsymbol{g}(\theta)$ using a single element of $\mathbb{R}^d$. Instead of this, it would often be more informative to find a subset of $\mathbb{R}^d$ which is likely to contain the true value $\boldsymbol{g}(\theta)$ because the spread of such a set then enables us to quantify our uncertainty regarding the true value of the parameter. In the frequentist paradigm, such information is provided by a confidence set for $\boldsymbol{g}(\theta)$.

Definition A.23. Let $\boldsymbol{g} : \Theta \to \mathbb{R}^d$ be a parameter in the statistical experiment $(\Theta, \mathcal{P}, \Lambda, \mathcal{G})$ with observations $X \sim P_\theta$, $\theta \in \Theta$. Then the set-valued function $C(X) \in \mathcal{B}^d$ is called a *confidence set* for $\boldsymbol{g}(\theta)$ at $100(1-\alpha)\%$ *confidence level*, where $\alpha \in (0, 1)$ is a fixed constant, if

$$\inf_{\theta \in \Theta} P_\theta(\boldsymbol{g}(\theta) \in C(X)) \geq 1 - \alpha.$$

In other words, the confidence set $C(X)$ is a set-valued random element which, for any $\theta \in \Theta$, contains the value $\boldsymbol{g}(\theta)$ with a probability of at least $1 - \alpha$. It is important to note that this does not mean that for a given observed $X = x$ and confidence set $C(x)$, there would be a $100(1-\alpha)\%$ probability that $\boldsymbol{g}(\theta)$ is included in $C(x)$. Instead, this means that if the experiment was repeated infinitely many times, then at least $100(1-\alpha)\%$ of the observed confidence sets would include the true parameter $\boldsymbol{g}(\theta)$.

When $g(\theta) \in \mathbb{R}$, it is natural to consider interval-valued confidence set $C(X) = [a(X), b(X)]$, where $a(X)$ and $b(X)$, $a(X) < b(X)$, are real-valued statistics. Such a confidence set is called a *confidence interval* for $g(\theta)$.

Apart from the suggestive notation, we have up to now made no reference to θ as a random element. In fact, in the frequentist paradigm of statistics, θ has always a fixed non-stochastic value. That is, even though θ is unknown, we do not consider probability measures for θ. In the Bayesian paradigm, this restriction is dropped and θ is regarded as a Θ-valued random element. The probability measures over θ are then understood to represent our "degree of belief" about the fixed true value of θ.

The Bayesian paradigm provides an alternative framework for constructing both point and interval estimators for the unknown parameters. For simplicity, let us assume that $(\Lambda, \mathcal{G}) = (E, \mathcal{B}_E)$, where E is a Borel set of $\mathbb{R}^m$, and that Θ is a Borel set of $U = \mathbb{R}^n$. Assume further that the parameter of interest is $\boldsymbol{\theta}$ itself and that, for every $\boldsymbol{\theta} \in \Theta$, the probability measure $P_{\boldsymbol{\theta}}$ has a density $p(\boldsymbol{x}|\boldsymbol{\theta})$ which is jointly measurable with respect to both $\boldsymbol{x}$ and $\boldsymbol{\theta}$. By assuming the existence of a known *prior density* $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, we can then use Bayes' rule (A.8) to compute the *posterior*

density $p(\boldsymbol{\theta}|\boldsymbol{x})$ of the parameter $\boldsymbol{\theta}$ given the observations $\boldsymbol{x}$,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_\Theta p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Theta,\ \boldsymbol{x} \in E \tag{A.10}$$

assuming that $p(\boldsymbol{x}) > 0$. In (A.10), the second equality holds when the parameter $\boldsymbol{\theta}$ is continuous. In the discrete case, the integration is with respect to the counting measure and can be replaced by a summation over $\boldsymbol{\theta}$.

The posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ represents our degree of belief about the unknown parameter $\boldsymbol{\theta}$ after observing the data $\boldsymbol{x}$. In the Bayesian paradigm, any point estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ should be somehow based on the posterior. Such estimators include, for example, the *posterior mean estimator*

$$\hat{\boldsymbol{\theta}} = \mathrm{E}[\boldsymbol{\theta}|\boldsymbol{x}] = \int_\Theta \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta}$$

and the *maximum a posteriori estimator*

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta}|\boldsymbol{x})$$

provided that these are well-defined quantities. In the one-dimensional case, also the median of the posterior is a viable point estimator of the unknown parameter.

The Bayesian approach generalizes to inference of $\boldsymbol{g}(\boldsymbol{\theta})$ is a straightforward manner. Namely, since the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ induces a posterior $p(\boldsymbol{g}(\boldsymbol{\theta})|\boldsymbol{x})$ for any transformation $\boldsymbol{g} : \Theta \to \mathbb{R}^d$, we can use this induced posterior to infer $\boldsymbol{g}(\boldsymbol{\theta})$.

The Bayesian analogue of a frequentist confidence set is called a credible set.

Definition A.24. Let $p(\boldsymbol{g}(\boldsymbol{\theta})|\boldsymbol{x})$ be the posterior density of the parameter $\boldsymbol{g}(\boldsymbol{\theta}) \in \mathbb{R}^d$ given the observations $\boldsymbol{x}$. Then a $100(1-\alpha)\%$ *credible set* for $\boldsymbol{g}(\boldsymbol{\theta})$ is any set $C \in \mathcal{B}^d$ which satisfies

$$P(\boldsymbol{g}(\boldsymbol{\theta}) \in C|\boldsymbol{x}) = \int_C p(\boldsymbol{g}(\boldsymbol{\theta})|\boldsymbol{x})\,\mathrm{d}\boldsymbol{g}(\boldsymbol{\theta}) = \int_{\boldsymbol{g}^{-1}(C)} p(\boldsymbol{\theta}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta} \geq 1-\alpha, \tag{A.11}$$

where $\boldsymbol{g}^{-1}(C)$ is the preimage of C and $\alpha \in (0,1)$ is a fixed constant.

In the one-dimensional case, where $g(\boldsymbol{\theta}) \in \mathbb{R}$, a credible set of the form $C = [a,b]$ is called a *credible interval* for $g(\boldsymbol{\theta})$.

As opposed to frequentist confidence sets, the defining property (A.11) of Bayesian credible sets does make a probability statement about the unknown parameter $\boldsymbol{g}(\boldsymbol{\theta})$. That is, the credible set C is a set which contains the true value of $\boldsymbol{g}(\boldsymbol{\theta})$ with a probability of $1-\alpha$, where the probability should be interpreted in the Bayesian sense as describing our subjective degree of belief about the value of $\boldsymbol{g}(\boldsymbol{\theta})$.

The question of how to choose the prior density $p(\boldsymbol{\theta})$ is of central importance in Bayesian inference. We focus here on presenting the terminology related to different choices of the prior and refer the reader to Section 5.3 for a more detailed discussion on prior densities in the case of the unfolding problem. In purely Bayesian thinking, the prior should reflect our subjective a priori knowledge about $\boldsymbol{\theta}$ before observing

the data $\boldsymbol{x}$. Such priors are called *informative*. On the other hand, in cases where such subjective knowledge is not available or one is aiming at drawing as objective inferences as possible, it is desirable to try to choose the prior in such a way that it only gives a very vague picture about the values of $\boldsymbol{\theta}$. Such priors are called *uninformative*. Finally, at times, the prior $p(\boldsymbol{\theta})$ is chosen in such a way that it cannot be normalized to be a probability density function, i.e., $\int_\Theta p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \infty$, but such *improper* priors are not considered a problem as long as the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ can be normalized to be a probability density function.

A.3 Elements of Linear Algebra

This section provides a number of definitions and results in linear algebra that are used throughout this thesis and especially in Chapter 4. The results presented in this section are compiled using references [26, 31, 9, 45, 24].

We often regard an $m \times n$ real matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ as a *linear mapping*[3] from $\mathbb{R}^n$ to $\mathbb{R}^m$. That is,

$$\boldsymbol{A} : \mathbb{R}^n \to \mathbb{R}^m, \boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}. \tag{A.12}$$

There are two fundamental subspaces associated with the matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Firstly, the *kernel* or the *null space* of $\boldsymbol{A}$, denoted by $\ker(\boldsymbol{A})$, is the set of vectors of the domain that are mapped to zero,

$$\ker(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

Secondly, the *range* or the *image* of $\boldsymbol{A}$, denoted by $\mathrm{ran}(\boldsymbol{A})$, is the set of all values that the mapping can attain,

$$\mathrm{ran}(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^m : \exists \boldsymbol{x} \in \mathbb{R}^n \text{ s.t. } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}\}.$$

By definition, the mapping (A.12) is surjective when $\mathrm{ran}(\boldsymbol{A}) = \mathbb{R}^m$. Furthermore, (A.12) is injective if and only if $\ker(\boldsymbol{A}) = \{\boldsymbol{0}\}$.

The set of vectors that are orthogonal to all vectors of a given subspace U is called the orthogonal complement of U.

Definition A.25. Let U be a subspace of $\mathbb{R}^d$. Then the *orthogonal complement* of U, denoted by $U^\perp$, is the set

$$U^\perp = \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^\mathrm{T}\boldsymbol{y} = 0,\ \forall \boldsymbol{y} \in U\}.$$

It is easy to check that $U^\perp$ is a subspace of $\mathbb{R}^d$ and that $(U^\perp)^\perp = U$.

The following proposition gives two useful relations between the range and kernel of $\boldsymbol{A}$ and the range and kernel of its transpose $\boldsymbol{A}^\mathrm{T}$.

[3] Linear mappings can also be defined in a more abstract setting without making reference to matrices. An abstract linear mapping $T : V \to W$ between two finite-dimensional vector spaces can then be represented by a matrix $\boldsymbol{A}$ after fixing the bases of the spaces V and W.

Proposition A.26. *Let* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. *Then*

(i) $\operatorname{ran}(\boldsymbol{A})^{\perp} = \ker(\boldsymbol{A}^{\mathrm{T}})$,

(ii) $\ker(\boldsymbol{A})^{\perp} = \operatorname{ran}(\boldsymbol{A}^{\mathrm{T}})$.

Proof. To prove (i), let $\boldsymbol{x} \in \mathbb{R}^m$. Then

$$\begin{aligned}\boldsymbol{x} \in \ker(\boldsymbol{A}^{\mathrm{T}}) &\Leftrightarrow \left\langle \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{y}\right\rangle = 0,\ \forall \boldsymbol{y} \in \mathbb{R}^n \\ &\Leftrightarrow \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\rangle = 0,\ \forall \boldsymbol{y} \in \mathbb{R}^n \\ &\Leftrightarrow \boldsymbol{x} \in \operatorname{ran}(\boldsymbol{A})^{\perp}.\end{aligned}$$

Hence, $\operatorname{ran}(\boldsymbol{A})^{\perp} = \ker(\boldsymbol{A}^{\mathrm{T}})$. Claim (ii) then follows by setting $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{T}}$ in (i) and taking the orthogonal complement of both side of the equation. □

Let $\boldsymbol{A}$ be an $m \times n$ real matrix. Then the *column rank* of $\boldsymbol{A}$ is the dimension of the space spanned by the n columns of $\boldsymbol{A}$. If the column rank of $\boldsymbol{A}$ is n, we say that $\boldsymbol{A}$ has *full column rank*. Similarly, the dimension of the space spanned by the m rows of $\boldsymbol{A}$ is called the *row rank* of $\boldsymbol{A}$, and when the row rank is m, we say that $\boldsymbol{A}$ has *full row rank*. The linear mapping (A.12) is injective if and only if $\boldsymbol{A}$ has full column rank. Similarly, (A.12) is surjective if and only if $\boldsymbol{A}$ has full row rank. It can be shown (see Theorem 4.4.1 of [26]) that for any matrix $\boldsymbol{A}$ the column rank and row rank are equal. This common value is called the *rank* of $\boldsymbol{A}$ and denoted by $\operatorname{rank}(\boldsymbol{A})$.

Analogously with the standard 2-norm of a vector $\boldsymbol{x}$ defined by $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$, it is possible to define a matrix norm for a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

Definition A.27. For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the norm $\|\boldsymbol{A}\|$ is defined by

$$\|\boldsymbol{A}\| = \sup_{\{\boldsymbol{x}\in\mathbb{R}^n : \|\boldsymbol{x}\|=1\}} \|\boldsymbol{A}\boldsymbol{x}\| = \sup_{\boldsymbol{x}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

It is easy to check that the matrix norm is indeed a norm. Furthermore, it is immediate from the definition that $\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|$.

Out of the various possible decomposition of a matrix $\boldsymbol{A}$, the singular value decomposition has a central role in Chapter 4 of this thesis.

Theorem A.28. (Singular value decomposition) *Every matrix* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *admits the decomposition*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}},$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ *and* $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ *are orthogonal matrices and* $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ *is a diagonal matrix with non-negative diagonal elements* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$.

Proof. See Theorem A.1 in [31]. □

The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are called the left and right *singular vectors* of $\boldsymbol{A}$, respectively, and the diagonal elements σ_i of $\boldsymbol{\Sigma}$ are called the *singular values* of $\boldsymbol{A}$.

Proposition A.29. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and assume that $\boldsymbol{A}$ has p strictly positive singular values. That is, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > \sigma_{p+1} = \cdots = \sigma_{\min(m,n)} = 0$. Then, $p = \operatorname{rank}(\boldsymbol{A})$.*

Proof. See Corollary 21.12.2 in [26]. □

Another important tool that often turns out to be useful is the Moore–Penrose pseudoinverse $\boldsymbol{A}^\dagger$ of a matrix $\boldsymbol{A}$, which generalizes the notion of matrix inversion to singular and non-square matrices.

Definition A.30. Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. The *Moore–Penrose pseudoinverse* of $\boldsymbol{A}$ (or *pseudoinverse* for short), denoted by $\boldsymbol{A}^\dagger$, is the $n \times m$ real matrix which satisfies the following four criteria:

(i) $\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}$,

(ii) $\boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{A}^\dagger$,

(iii) $(\boldsymbol{A}\boldsymbol{A}^\dagger)^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{A}^\dagger$,

(iv) $(\boldsymbol{A}^\dagger\boldsymbol{A})^{\mathrm{T}} = \boldsymbol{A}^\dagger\boldsymbol{A}$.

Theorem A.31. *For every matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the Moore–Penrose pseudoinverse $\boldsymbol{A}^\dagger \in \mathbb{R}^{n\times m}$ exists and is unique.*

Proof. See Theorem 20.1.1 in [26]. □

It is easy to verify using Definition A.30 that in the following special cases the pseudoinverse $\boldsymbol{A}^\dagger$ can be easily computed using the original matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$:

(i) When $\boldsymbol{A}$ is a non-singular square matrix, $\boldsymbol{A}^\dagger = \boldsymbol{A}^{-1}$.

(ii) When $\boldsymbol{A}$ has full column rank, i.e., $\operatorname{rank}(\boldsymbol{A}) = n$, $\boldsymbol{A}^\dagger = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$.

(iii) When $\boldsymbol{A}$ has full row rank, i.e., $\operatorname{rank}(\boldsymbol{A}) = m$, $\boldsymbol{A}^\dagger = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$.

In addition, when $\boldsymbol{A} = \operatorname{diag}\big(a_1, \ldots, a_{\min(m,n)}\big)_{m\times n}$, that is, an $m \times n$ diagonal matrix with diagonal elements $a_1, \ldots, a_{\min(m,n)}$, the pseudoinverse is given by $\boldsymbol{A}^\dagger = \operatorname{diag}\Big(a_1^\dagger, \ldots, a_{\min(m,n)}^\dagger\Big)_{n\times m}$, where

$$a_i^\dagger = \begin{cases} \frac{1}{a_i}, & \text{if } a_i \neq 0 \\ 0, & \text{if } a_i = 0 \end{cases}.$$

In the general case, the pseudoinverse can be easily computed using the singular value decomposition.

Proposition A.32. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a matrix with the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$. The Moore–Penrose pseudoinverse $\boldsymbol{A}^\dagger$ of $\boldsymbol{A}$ is then given by*

$$\boldsymbol{A}^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{U}^{\mathrm{T}}. \tag{A.13}$$

Proof. The claim is easily proved by verifying that (A.13) satisfies the four criteria of Definition A.30. □

The range and kernel of the pseudoinverse $\boldsymbol{A}^\dagger$ are equal to the range and kernel of $\boldsymbol{A}^\mathrm{T}$, respectively.

Proposition A.33. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then*

(i) $\mathrm{ran}(\boldsymbol{A}^\dagger) = \mathrm{ran}(\boldsymbol{A}^\mathrm{T})$,

(ii) $\ker(\boldsymbol{A}^\dagger) = \ker(\boldsymbol{A}^\mathrm{T})$.

Proof. See Theorem 20.5.1 in [26]. □

The pseudoinverse also has the following two useful properties:

Proposition A.34. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then the following hold:*

(i) If $\boldsymbol{A}$ has full column rank, $\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{I}$.

(ii) If $\boldsymbol{A}$ has full row rank, $\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{I}$.

Proof.

(i) When $\boldsymbol{A}$ has full column rank, $\boldsymbol{A}^\dagger = (\boldsymbol{A}^\mathrm{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathrm{T}$. Hence

$$\boldsymbol{A}^\dagger\boldsymbol{A} = (\boldsymbol{A}^\mathrm{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathrm{T}\boldsymbol{A} = \boldsymbol{I}.$$

(ii) When $\boldsymbol{A}$ has full row rank, $\boldsymbol{A}^\dagger = \boldsymbol{A}^\mathrm{T}(\boldsymbol{A}\boldsymbol{A}^\mathrm{T})^{-1}$. Hence

$$\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{A}\boldsymbol{A}^\mathrm{T}(\boldsymbol{A}\boldsymbol{A}^\mathrm{T})^{-1} = \boldsymbol{I}.$$ □

The pseudoinverse can also be used to construct orthogonal projections which are defined as follows:

Definition A.35. Let U be a subspace of $\mathbb{R}^d$. Then the matrix $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ is the *orthogonal projection* onto U if it has the following properties:

(i) $\mathrm{ran}(\boldsymbol{P}) = U$,

(ii) $\boldsymbol{P}^2 = \boldsymbol{P}$,

(iii) $\boldsymbol{P}^\mathrm{T} = \boldsymbol{P}$.

From this definition, it follows that $\boldsymbol{P}\boldsymbol{x} \in U$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{x} - \boldsymbol{P}\boldsymbol{x} \in U^\perp$. Furthermore, it can be shown that the orthogonal projection onto a given subspace U is unique. Hence, the interpretation of the orthogonal projection $\boldsymbol{P}$ is that it maps any vector $\boldsymbol{x}$ to its unique best approximation among the vectors of the subspace U.

The orthogonal projection $\boldsymbol{P}$ onto U can also be used to construct the orthogonal projection onto $U^\perp$.

Proposition A.36. *Let U be a subspace of $\mathbb{R}^d$ and $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ be the orthogonal projection onto U. Then $\boldsymbol{I} - \boldsymbol{P}$ is the orthogonal projection onto $U^\perp$.*

Proof. See Corollary 12.5.10 in [26]. □

We are now ready to state the following result for the pseudoinverse $\boldsymbol{A}^\dagger$:

Proposition A.37. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then the following hold:*

(i) $\boldsymbol{A}\boldsymbol{A}^\dagger$ *is the orthogonal projection onto* $\mathrm{ran}(\boldsymbol{A})$.

(ii) $\boldsymbol{A}^\dagger\boldsymbol{A}$ *is the orthogonal projection onto* $\mathrm{ran}(\boldsymbol{A}^\mathrm{T})$.

Proof. See Theorem 20.5.1 in [26]. □

In fact, Proposition A.37 can be used as a definition of the pseudoinverse $\boldsymbol{A}^\dagger$ and it can be shown to be equivalent to our four criteria for the pseudoinverse given in Definition A.30.

Corollary A.38. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then the following hold:*

(i) $\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\dagger$ *is the orthogonal projection onto* $\ker(\boldsymbol{A}^\mathrm{T})$.

(ii) $\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A}$ *is the orthogonal projection onto* $\ker(\boldsymbol{A})$.

Proof.

(i) By Propositions A.36 and A.37, $\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\dagger$ is the orthogonal projection onto $\mathrm{ran}(\boldsymbol{A})^\perp$, which, by Proposition A.26, is equal to $\ker(\boldsymbol{A}^\mathrm{T})$.

(ii) Using the same proposition as above, $\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A}$ is the orthogonal projection onto $\mathrm{ran}(\boldsymbol{A}^\mathrm{T})^\perp = \ker(\boldsymbol{A})$. □

We conclude this section by introducing Gram matrices which are used in the proof of Proposition 6.1.

Definition A.39. Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then the $n \times n$ square matrix $\boldsymbol{G} = \boldsymbol{A}^\mathrm{T}\boldsymbol{A}$ is called the *Gram matrix* of $\boldsymbol{A}$.

The Gram matrix has the following useful properties:

Proposition A.40. *Let $\boldsymbol{G}$ be the Gram matrix of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then* $\mathrm{rank}(\boldsymbol{G}) = \mathrm{rank}(\boldsymbol{A})$.

Proof. See Theorem 5.5.4 in [45]. □

Proposition A.41. *The Gram matrix $\boldsymbol{G}$ of the matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is positive semidefinite.*

Proof. Since $\boldsymbol{G} = \boldsymbol{A}^\mathrm{T}\boldsymbol{A}$, we have

$$\boldsymbol{x}^\mathrm{T}\boldsymbol{G}\boldsymbol{x} = \boldsymbol{x}^\mathrm{T}\boldsymbol{A}^\mathrm{T}\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^\mathrm{T}\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|^2 \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \qquad \square$$

A.4 Inverse Problems

This section provides a conceptual introduction to *inverse problems* based on references [20, 31] unless otherwise mentioned. An inverse problem is an umbrella term for a wide range of problems where the mapping from some quantity of primary interest to the observations is well-behaved but inversion of this mapping to find the quantity of interest given the observations is not straightforward. More concretely, let H_1 and H_2 be separable Hilbert spaces and let $K : H_1 \to H_2$ be a bounded linear operator, called the *forward operator*, which maps the quantity of interest f into the observations h. That is, we are interested in solving the linear operator equation

$$h = Kf. \tag{A.14}$$

Clearly, when h is known exactly and K is invertible, the solution is given by $f = K^{-1}h$. The problem is that, in virtually all practical applications, h is only known approximately. This can often be modeled by regarding h as the parameter of some appropriate statistical model and thinking that we observe a realization of this model instead of h itself. For example, in the unfolding problem, h is the intensity function of a Poisson point process and the data we observe is a single realization of this process. Whatever the mechanism linking h to the actual observations, the common outcome is that we need to construct an estimator $\hat{h}$ of h based on the observed data and then try to solve

$$\hat{h} = Kf. \tag{A.15}$$

Let us denote the (approximate) solution of this equation by $\hat{f}$. From (A.15), it is clear that three types of issues could occur in the inversion process:

(i) It could happen that $\hat{h} \notin \operatorname{ran}(K)$ in which case $\hat{f}$ cannot be an exact solution of (A.15).

(ii) It could happen that for a given $\hat{h}$, the solution $\hat{f}$ is not unique.

(iii) It could happen that the solution $\hat{f}$ is a discontinuous (or nearly discontinuous) function of $\hat{h}$.

When the problem at hand is affected by one or more of these issues, it is said to be *ill-posed*.

Issues (i) and (ii) are not, in general, too restrictive. Firstly, if K is bijective, we obviously need not worry about them. Secondly, when K is not surjective, issue (i) can often be circumvented by looking at some approximate generalized solution of (A.15) instead of the exact one. A classical example is the least squares estimation of an unsatisfiable system of linear equations $\hat{\boldsymbol{h}} = \boldsymbol{K}\boldsymbol{f}$. Similarly, issue (ii), which arises when K is not injective, can in most cases be solved by imposing some criterion to pick out a desired solution out of the multiple possibilities. For example, the least squares solution of $\hat{\boldsymbol{h}} = \boldsymbol{K}\boldsymbol{f}$ is, in general, not unique, but choosing the solution with the smallest norm results in a unique solution $\hat{\boldsymbol{f}}_{\mathrm{LS}}$ given by the pseudoinverse, $\hat{\boldsymbol{f}}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\hat{\boldsymbol{h}}$.

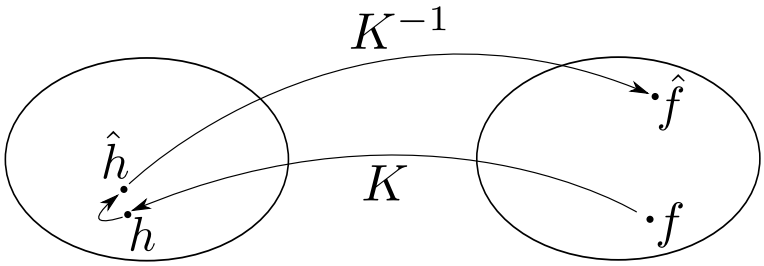


Figure A.1: Illustration of an ill-posed inverse problem. Assume that we would like to solve the equation $h = Kf$, but instead of exact h, we are only able to construct an estimator $\hat{h}$ of h. Because the (generalized) inverse mapping K^{-1} is discontinuous or nearly discontinuous, even small perturbations with respect to the correct h may cause the solution $\hat{f}$ to lie far away from the desired value f.

Issue (iii), however, is a lot more tricky. This is because it means that small perturbations in $\hat{h}$ can result in arbitrarily large changes in $\hat{f}$. This effect is illustrated in Figure A.1. Since perturbations are always present in any estimator $\hat{h}$, this is a major practical issue. As a result, inverse problems are problems that are mainly characterized by a forward operator which has a discontinuous or nearly discontinuous (generalized) inverse operator. In practice, this often means that the solution $\hat{f}$ corresponding to the naïve inversion of the forward operator K exhibits large oscillations or other distinct undesired artifacts. Put in the language of statistics, the estimator $\hat{f}$ has an excessively large variance.

When the Hilbert spaces H_1 and H_2 are infinite dimensional, the (generalized) inverse of K might be discontinuous in the sense of the ordinary definition of continuity. On the other hand, when H_1 and H_2 are finite dimensional, any linear mapping from H_2 back to H_1 is continuous. One could be tempted to believe that this solves the above-mentioned issue (iii), but unfortunately this is not the case. The reason for this is that even though a continuous mapping cannot make arbitrary jumps, it is still allowed to change rapidly as a function of the data, which could cause significant numerical problems.

Fortunately, there are ways of quantifying the degree of ill-posedness of such nearly discontinuous linear mappings. In the following treatment, which is motivated by [18, Theorem 8.10], we restrict ourselves to the finite-dimensional case. That is, consider the finite-dimensional version of Equation (A.14) given by

$$\boldsymbol{h} = \boldsymbol{K}\boldsymbol{f}, \tag{A.16}$$

where $\boldsymbol{K} \in \mathbb{R}^{m\times n}$, $\boldsymbol{h} \in \mathbb{R}^m$ and $\boldsymbol{f} \in \mathbb{R}^n$, and suppose that we decide to find the solution by solving the corresponding least squares problem. As explained in Section 4.2, the minimum norm least squares solution of (A.16) is given by the pseudoinverse

$$\boldsymbol{f}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\boldsymbol{h}.$$

To see how stable this solution is with respect to perturbations in $\boldsymbol{h}$, assume that we use the observations to construct an estimator $\hat{\boldsymbol{h}}$ of $\boldsymbol{h}$ and compute the corresponding

least squares solution

$$\hat{\boldsymbol{f}}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}\hat{\boldsymbol{h}}.$$

When $\boldsymbol{h}$ is replaced by $\hat{\boldsymbol{h}}$, the change in the solution is

$$\begin{aligned}\hat{\boldsymbol{f}}_{\mathrm{LS}} - \boldsymbol{f}_{\mathrm{LS}} &= \boldsymbol{K}^{\dagger}(\hat{\boldsymbol{h}} - \boldsymbol{h}) \\ &= (\boldsymbol{K}^{\dagger} - \boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{K}^{\dagger} + \boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{K}^{\dagger})(\hat{\boldsymbol{h}} - \boldsymbol{h}) \\ &= \boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{K}^{\dagger}(\hat{\boldsymbol{h}} - \boldsymbol{h}),\end{aligned}$$

where the last equality follows from Definition A.30. By Proposition A.37, $\boldsymbol{K}\boldsymbol{K}^{\dagger}$ is the orthogonal projection onto $\mathrm{ran}(\boldsymbol{K})$. Let us thus denote $\boldsymbol{P} = \boldsymbol{K}\boldsymbol{K}^{\dagger}$, whence

$$\hat{\boldsymbol{f}}_{\mathrm{LS}} - \boldsymbol{f}_{\mathrm{LS}} = \boldsymbol{K}^{\dagger}(\boldsymbol{P}\hat{\boldsymbol{h}} - \boldsymbol{P}\boldsymbol{h}).$$

It then follows that

$$\|\hat{\boldsymbol{f}}_{\mathrm{LS}} - \boldsymbol{f}_{\mathrm{LS}}\| \leq \|\boldsymbol{K}^{\dagger}\|\|\boldsymbol{P}\hat{\boldsymbol{h}} - \boldsymbol{P}\boldsymbol{h}\|.$$

Furthermore,

$$\|\boldsymbol{P}\boldsymbol{h}\| = \|\boldsymbol{K}\boldsymbol{K}^{\dagger}\boldsymbol{h}\| \leq \|\boldsymbol{K}\|\|\boldsymbol{K}^{\dagger}\boldsymbol{h}\| = \|\boldsymbol{K}\|\|\boldsymbol{f}_{\mathrm{LS}}\|.$$

Hence, we have the following upper bound for the relative change of the least squares solution

$$\frac{\|\hat{\boldsymbol{f}}_{\mathrm{LS}} - \boldsymbol{f}_{\mathrm{LS}}\|}{\|\boldsymbol{f}_{\mathrm{LS}}\|} \leq \|\boldsymbol{K}\|\|\boldsymbol{K}^{\dagger}\|\frac{\|\boldsymbol{P}\hat{\boldsymbol{h}} - \boldsymbol{P}\boldsymbol{h}\|}{\|\boldsymbol{P}\boldsymbol{h}\|}. \tag{A.17}$$

This means that we can guarantee that when $\boldsymbol{h}$ is replaced by the estimator $\hat{\boldsymbol{h}}$, the perturbation in the least squares solution is small if $\hat{\boldsymbol{h}}$ is close to $\boldsymbol{h}$ *and* $\|\boldsymbol{K}\|\|\boldsymbol{K}^{\dagger}\|$ is small.

This motivates the following definition.

Definition A.42. Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then, the *condition number* of $\boldsymbol{A}$, denoted by $\mathrm{cond}(\boldsymbol{A})$, is defined to be

$$\mathrm{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{\dagger}\|.$$

It can be shown [9, Section 10.4] that the condition number is equivalently given by

$$\mathrm{cond}(\boldsymbol{A}) = \frac{\sigma_1}{\sigma_p}, \tag{A.18}$$

where σ_1 is the largest and σ_p the smallest strictly positive singular value of $\boldsymbol{A}$. Note that by Proposition A.29, $p = \mathrm{rank}(\boldsymbol{A})$. It immediately follows from (A.18) that for all non-zero matrices $\boldsymbol{A}$, we have $\mathrm{cond}(\boldsymbol{A}) \geq 1$.

In light of the discussion above, the condition number can be used to characterize the ill-posedness of the finite-dimensional problem (A.16). That is, we expect the solution of the problem to be unstable when $\mathrm{cond}(\boldsymbol{K}) \gg 1$.

Naturally, when $\boldsymbol{K}$ is an invertible square matrix, we have $\boldsymbol{K}^{\dagger} = \boldsymbol{K}^{-1}$, $\boldsymbol{P} = \boldsymbol{I}$ and $\boldsymbol{f}_{\mathrm{LS}} = \boldsymbol{f}$. Hence, Equation (A.17) simplifies to

$$\frac{\|\hat{\boldsymbol{f}} - \boldsymbol{f}\|}{\|\boldsymbol{f}\|} \leq \|\boldsymbol{K}\|\|\boldsymbol{K}^{-1}\|\frac{\|\hat{\boldsymbol{h}} - \boldsymbol{h}\|}{\|\boldsymbol{h}\|} = \mathrm{cond}(\boldsymbol{K})\frac{\|\hat{\boldsymbol{h}} - \boldsymbol{h}\|}{\|\boldsymbol{h}\|}$$

and we see that the condition number cond($\boldsymbol{K}$) characterizes the sensitivity of the solution of the satisfiable linear system of equations $\boldsymbol{h} = \boldsymbol{K}\boldsymbol{f}$ to perturbations in $\boldsymbol{h}$.

Stabilization of an ill-posed inverse problem, where the solution $\hat{f}$ is not stable with respect to the estimator $\hat{h}$, is called *regularization* of the problem. This means that we try to find an approximate solution of (A.15) which is stable with respect to the observed data. This is usually done by enforcing some desired properties of the solution, such as smoothness or small norm. While various different approaches for regularizing an ill-posed problem have been proposed, most regularization techniques can be classified under the following three broad categories:

1. **Regularized frequentist point estimators**, where one modifies the standard frequentist solution $\hat{f}$ so that it has some desired properties that the correct solution is known to possess. The strength of the regularization is controlled using some *regularization parameter* δ. Tikhonov regularization and truncated singular value decomposition, both of which are discussed in detail in Section 4.2, are classical examples of this approach.

2. **Truncated iterative methods**, where one considers an iteration which, in the asymptotic limit, is known to return the unregularized solution $\hat{f}$ of (A.15). Regularization is then imposed by stopping the iteration prematurely before the solution starts to have unphysical properties, such as large oscillations. The earlier the iteration is stopped, the stronger the regularization. An example of this type of regularization is the EM algorithm with early stopping which is presented in Section 4.1.

3. **Bayesian methods**, where the inverse problem is formulated as a Bayesian inference problem and the solution is based on the Bayesian posterior distribution for the unknown. Regularization is imposed by selecting the prior in Bayes' rule in such a way that it places emphasis on physically plausible solutions. The strength of the regularization is then controlled via the spread of the prior density. Bayesian and related empirical Bayes regularization techniques are discussed in detail in Chapters 5 and 6.

The common theme in all these methods is that one has to strike a balance between satisfying Equation (A.15) and enforcing desired properties of the solution. When frequentist point estimators are used, this balance in controlled by the regularization parameter δ; in iterative techniques, by deciding when to stop the iteration; and in Bayesian techniques, by choosing the spread of the prior. Since it may have a significant effect on the final solution, choosing the appropriate regularization strength is of central importance when solving ill-posed inverse problems. This issue is discussed in the context of regularized point estimators and truncated iterative methods in Section 4.3, while in the Bayesian framework, empirical Bayes techniques discussed in Chapter 6 provide an elegant way of solving the problem.