TAMPERE UNIVERSITY OF TECHNOLOGY

Marko Järvenpää

# Bayesian Hierarchical Modelling for Image Processing Inverse Problems

Master of Science Thesis

Examiners: Prof. Robert Piché and
D.Tech. Simo Ali-Löytty
Examiners and topic approved in the
Science and Environmental Engineering
Faculty Council meeting
on 7 November 2012

# ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Science and Engineering
**JÄRVENPÄÄ, MARKO: Bayesian Hierarchical Modelling for Image Processing Inverse Problems**
Master of Science Thesis, 67 pages, 2 Appendix pages
March 2013
Major: Mathematics
Examiners: Prof. Robert Piché and D.Tech. Simo Ali-Löytty
Keywords: inverse problem, Bayesian statistics, hierarchical model, total variation, Gaussian scale mixture, deblurring

The main motivation of this work is to review and extend some recent ideas in Bayesian inverse problems especially in the context of practical image processing problems. Often these problems are solved using a deterministic setting and different optimisation algorithms. A Bayesian hierarchical model for total variation is presented in this thesis. This approach allows all the parameters of an inverse problem, including the "regularisation parameter", to be estimated simultaneously from the data.

The model is based on the characterisation of the Laplace density prior as a scale mixture of Gaussians. With different priors on the mixture variable, other total variation like regularisations are also obtained. All these priors have heavy tails that tend to induce sparsity, which has become an important topic in many areas of signal processing.

An approximation of the resulting posterior mean is found using a variational Bayes method. In addition, algorithms for computing just the maximum a posteriori estimate, although not a fully Bayesian approach, are presented. The methods are illustrated with examples of image deblurring, image denoising and inpainting, the first of which being the main application of this thesis.

Examples show that the methods generally work well for deblurring problems. Maximum a posteriori estimates preserve edges of "blocky" images well. The results given by variational Bayes method are more smooth than corresponding maximum a posteriori estimates which make it more suitable for problems where preserving the edges is not the top priority like deblurring smooth or partially blocky images. Variational Bayes method also makes Gibbs sampler redundant with this model as it is faster and gives slightly better results. As future work faster algorithms could be implemented as well as considering more complex and specialised models and more comprehensive simulations based on the ideas of this work.

# TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Teknis-luonnontieteellinen koulutusohjelma
**JÄRVENPÄÄ, MARKO: Hierarkkinen bayesiläinen malli inversio-ongelmiin kuvankäsittelyssä**
Diplomityö, 67 sivua, 2 liitesivua
Maaliskuu 2013
Pääaine: Matematiikka
Tarkastajat: Professori Robert Piché ja tohtori Simo Ali-Löytty
Avainsanat: inversio-ongelma, Bayesin menetelmät, hierarkkinen malli, total variation, Gaussin mikstuuri, kuvan terävöittäminen

Tämän työn tavoitteena on luoda katsaus ja viedä eteenpäin viimeaikaisia bayesiläiseen inversio-ongelmiin liittyviä ideoita. Käytännön sovelluksista erityisesti työssä keskitytään kuvankäsittelyongelmiin. Usein tällaisia ongelmia ratkaistaan käyttäen erilaisia deterministisiä optimointialgoritmeja. Tässä työssä sen sijaan esitellään hierarkkinen Bayes malli total variaatiolle. Tässä lähestymistavassa voidaan inversio-ongelmaan liittyvät parametrit kuten ”regularisaatioparametri” samanaikaisesti estimoida datasta.

Esitettävän mallin kulmakivi on tulos, jonka perusteella Laplace-priorijakauma voidaan esittää Gaussin mikstuurina. Asettamalla eri priorijakaumia mikstuurijakaumaksi, saadaan myös muita total variaation kaltaisia prioreja. Nämä priorit ovat paksuhäntäisiä ja ne tuottavat ”sparsity”-tyyppisiä tuloksia. Tähän aiheeseen liittyvät tutkimusaiheet ovat tulleet kiinnostaviksi monella signaalinkäsittelyn sovellusalueella.

Mallin tuloksena saatavan posteriorijakauman odotusarvon laskemiseen käytetään variaatioapproksimaatiota. Tämän lisäksi tässä työssä esitetään algoritmi maximum a posteriori –estimaatin laskemiseen, vaikka tämä lähestymistapa ei ole täysin bayesiläinen. Menetelmiä havainnollistetaan erityisesti kuvan terävöittämiseen liittyvillä esimerkeillä, mutta myös kohinan poistoa sekä inpainting-ongelmaa tutkitaan.

Tuloksista nähdään, että menetelmät toimivat hyvin etenkin kuvan terävöittämisongelmissa. Maximum a posteriori –estimaatti säilyttää kuvan väritasojen jyrkät rajat hyvin. Variaatioapproksimaatiolla saaduissa tuloksissa tällaiset väritasojen rajat jäävät pyöristyneimmiksi kuin edellä mainitussa menetelmässä. Variaatioapproksimaation antama ratkaisu sopii paremmin ongelmiin, joissa kuvissa olevien väritasojen rajojen säilyttäminen ei ole tärkeintä, kuten ”pehmeille” kuville. Tämä menetelmä on myös laskennallisesti nopeampi kuin Gibbs-näytteistykseen perustuva menetelmä ja tuottaa parempia tuloksia. Jatkossa voitaisiin keskittyä tämän työn pohjalta saataviin vielä kehittyneempiin malleihin ja suorittaa kattavampia simulaatioita.

# Preface

This work was carried out while working in Personal Positioning Algorithms Reaserch Group in Tampere University of Technology, Finland. The group was in Department of Mathematics and has now moved to Department of Automation and Engineering. The project is funded by Nokia Corporation.

I want to thank, first of all, to my supervisor Professor Robert Piché for introducing me to the fields of Bayesian methods and inverse problems, for the opportunity to work in this group and for the guidance. I also want to thank my supervisor D.Tech. Simo Ali-Löytty for useful comments that helped me to make this thesis better.

I want to thank my co-workers for interesting conversations and for answers to my questions. Especially I want to thank Juha Ala-Luhtala, who has studied robust estimation and some related topics to this work in this research group, for worthwhile comments and discussions that helped me to better understand many things. My thanks also goes to Professor Samuli Siltanen and D.Tech. Sampsa Pursiainen for interest and providing some useful material related to the topic of this thesis.

My sincere thanks goes also to friends, my parents and my brother for their support during my studies.

Tampere, 13th February 2013

Marko Järvenpää
marko.jarvenpaa@tut.fi

# Contents


**1 Introduction** **1**

**2 Preliminaries** **4**
2.1 Probability distributions and their properties . . . . . 4
2.2 Bayesian inference . . . . . 11
2.3 Bayesian hierarchical models . . . . . 13

**3 Bayesian inference algorithms** **15**
3.1 Gibbs sampler . . . . . 15
3.2 Variational Bayes . . . . . 16

**4 Regularisation methods** **19**
4.1 Tikhonov regularisation and Gaussian priors . . . . . 20
4.2 The Lasso and some generalisations . . . . . 23
4.3 Total Variation . . . . . 24

**5 Bayesian hierarchical TV regularisation** **28**
5.1 The linear model . . . . . 28
5.2 Bayesian total variation regularisation . . . . . 30
5.2.1 Hierarchical Model for TV prior . . . . . 30
5.2.2 Gibbs Sampler . . . . . 33
5.2.3 Coordinate Descent Method . . . . . 34
5.2.4 Variational Bayes . . . . . 35
5.3 Implementation details and discussion . . . . . 37

**6 Bayesian hierarchical TV regularisation in 2d** **40**
6.1 Anisotropic TV prior . . . . . 42
6.2 Isotropic TV prior . . . . . 43
6.3 Two-dimensional Laplace TV prior . . . . . 44
6.4 t-distribution TV prior . . . . . 46

**7 Image processing problems** **49**
7.1 Image deblurring . . . . . 50
7.2 Image denoising . . . . . 55
7.3 Image inpainting . . . . . 58

**8 Conclusions** **62**

**Bibliography** **64**

**A Derivations** **68**

# Symbols

| | |
|---|---|
| $\|\cdot\|$ | Vector norm |
| $\|\cdot\|_1$ | $L^1$ norm |
| $\|\cdot\|_2$ | Euclidean norm ($L^2$ norm) |
| $\|\cdot\|_P$ | Weighted norm so that $\|x\|_P^2 = x^T P x$ |
| $\approx$ | Approximation |
| $\in$ | Belongs to |
| $\nabla$ | Gradient |
| $\neq$ | Inequality |
| $\infty$ | Infinity |
| $\int$ | Integral |
| $\to$ | Limit |
| $\propto$ | Proportional to |
| $\subseteq$ | Subset |
| | |
| $A^{-1}$ | Inverse of matrix $A$ |
| $A^{1/2}$ | Square root of spd matrix $A$ so that $(A^{1/2})(A^{1/2})^T = A$ |
| $A$, $B$ | Matrices |
| $\arg\max_x$ | Argument that solves maximisation problem |
| $\arg\min_x$ | Argument that solves minimisation problem |
| $A^T$ | Transpose of matrix $A$ |
| $C_0^1(\Omega; \mathbb{R}^k)$ | Space of compactly supported continuously differentiable vector-valued functions on $\Omega$ |
| cosh | Hyperbolic cosine |
| $\det(A)$, $\|A\|$ | Determinant of matrix $A$ |
| $\frac{\mathrm{d}f}{\mathrm{d}x}$ | Derivative of $f$ respect to $x$ |
| diag | Diagonal matrix |
| $\mathbb{E}(\cdot)$ | Expectation |
| $\mathbb{E}_{\mathbf{x}}(\cdot)$ | Expectation with respect to random vector $\mathbf{x}$ |
| $\exp(\cdot)$ | Exponential function, $\exp(x) = \mathrm{e}^x$ |
| $\nabla \cdot f$ | Divergence of function $f$ |
| $\frac{\partial f}{\partial x}$ | Partial derivative of $f$ respect to $x$ |
| $\Gamma(\cdot)$ | Gamma function |
| $\Gamma_p(\cdot)$ | Multivariate Gamma function |
| I | Identity matrix |
| $\nabla_{ij}^h x$ | Difference for elements $(i, j+1)$ and $(i, j)$, $\nabla_{ij}^h x = x_{i,j+1} - x_{i,j}$ |
| $\nabla_{ij}^v x$ | Difference for elements $(i+1, j)$ and $(i, j)$, $\nabla_{ij}^v x = x_{i+1,j} - x_{i,j}$ |
| $\mathrm{KL}(q\|p)$ | Kullback-Leibler divergence of $q$ and $p$, respectively |
| $\mathrm{K}_p(\cdot)$ | Modified Bessel function of the second kind with parameter $p$ |

| | |
|---|---|
| $L^1(\Omega)$ | Space of integrable functions on $\Omega$ |
| ln | Natural logarithm |
| $\log_{10}$ | Logarithm to base 10 |
| min | Minimum value |
| $\mathsf{mode}(\cdot)$ | Mode |
| $\mathcal{N}(A)$ | Nullspace of $A$ |
| $p_{\mathbf{x}}(x)$, $p(x)$ | Probability density function of random vector $\mathbf{x}$ |
| $p_{\mathbf{x},\mathbf{y}}(x, y)$, $p(x, y)$ | Joint probability density function of random vectors $\mathbf{x}$ and $\mathbf{y}$ |
| $p_{\mathbf{x}\,\vert\,\mathbf{y}}(x \,\vert\, y)$, $p(x \,\vert\, y)$ | Conditional probability density function of random vector $\mathbf{x}$ given $\mathbf{y}$ |
| $q^*_{\mathbf{x}}(x)$, $q^*(x)$ | Optimal density for $\mathbf{x}$ in variational Bayes method |
| $\mathbb{R}$ | Real numbers |
| $\mathbb{R}_+$ | Positive real numbers |
| $\mathbb{R}^n$ | n-dimensional real valued vectors |
| $\mathbb{R}^{n\times k}$ | $n \times k$ matrix with real valued elements |
| $\mathrm{rank}(A)$ | Rank of matrix A |
| $\partial S$ | Boundary of set $S$ |
| sin | Sine function |
| sup | Supremum, least upper bound |
| $\mathrm{tr}(A)$ | Trace of matrix $A$ |
| $\mathbb{V}(\cdot)$ | Variance |
| $x_i$ | $i$th element of vector $x$ |
| $x_{-[i]}$ | All the other elements of vector $x$ except $i$th |
| $x_{i,j}$ | Element $(i, j)$ of resulting matrix when vector $x$ is unstacked |
| $X_{i,j}$ | Element $(i, j)$ of matrix $X$ |
| $x^{(t)}$ | $t$th sample drawn from some distribution |
| $\mathbf{x} \sim X$ | Random vector $\mathbf{x}$ follows the distribution $X$ |
| $x$, $y$ | Scalars or vectors |
| $\mathbf{x}$, $\mathbf{y}$ | Random variables or random vectors |
| $\mathbf{X}$, $\mathbf{Y}$ | Random matrices |
| $\mathbf{x} \,\vert\, (\mathbf{y} = y) \sim X$, $\mathbf{x} \,\vert\, y \sim X$ | $\mathbf{x}$ given $\mathbf{y}$ follows the distribution $X$ |

| | |
|---|---|
| $\mathrm{Exp}(\theta)$ | Exponential distribution with parameter $\theta$ |
| $\mathrm{Gamma}(\alpha, \beta)$ | Gamma distribution with parameters $\alpha$ and $\beta$ |
| $\mathrm{GIG}(a, b, p)$ | Generalised Inverse Gaussian distribution with parameters $a$, $b$ and $p$ |
| $\mathrm{IG}(\mu, \lambda)$ | Inverse Gaussian distribution with parameters $\mu$ and $\lambda$ |
| $\mathrm{InvGamma}(\alpha, \beta)$ | Inverse gamma distribution with parameters $\alpha$ and $\beta$ |
| $\mathrm{Laplace}(\mu, b)$ | Laplace distribution with parameters $\mu$ and $b$ |
| $\mathrm{MVLaplace}(\mu, \Sigma)$ | Multivariate Laplace distribution with parameters $\mu$ and $\Sigma$ |

| | |
|---|---|
| Normal$(\mu, \Sigma)$ | Normal (or Gaussian) distribution with parameters $\mu$ and $\Sigma$ |
| RIG$(\alpha, \beta)$ | Reciprocal inverse Gaussian distribution with parameters $\alpha$ and $\beta$ |
| $\mathrm{t}_\nu(\mu, \Sigma)$ | Student's t-distribution with parameters $\nu$, $\mu$ and $\Sigma$ |
| Wishart$(\Psi, \nu)$ | Wishart distribution with parameters $\Psi$ and $\nu$ |

# Abbreviations

| | |
|---|---|
| BSNR | Blurred signal-to-noise ratio |
| CM | Conditional mean estimate |
| DAG | Directed acyclic graph |
| ECM | Expectation conditional maximisation (algorithm) |
| EM | Expectation maximisation (algorithm) |
| GSM | Gaussian scale mixture |
| IAS | Iterative alternating sequential |
| iid | Independent and identically distributed |
| ISNR | Improvement in signal-to-noise ratio |
| KL | Kullback-Leibler divergence |
| Lasso | Least absolute shrinkage and selection operator |
| MAP | Maximum a posteriori estimate |
| MATLAB | Matrix Laboratory, a program for numerical computations |
| MCMC | Markov chain Monte Carlo |
| pdf | Probability density function |
| spd | Symmetric positive definite matrix |
| svd | Singular value decomposition |
| TV | Total variation |
| VB | Variational Bayes approximation |

# Remarks on notation

- We do not distinguish between scalars and vectors as it should always be evident from context which one each variable is. Similarly no notational difference is applied between random variables and random vectors. Term "random vector" is often also used when the variable can be either one, while random variable must be $1 \times 1$.

- Square root, inverse and division by a vector are defined componentwise for notational convenience. That is, if $x$ is a column $n$-vector it will be denoted

$$\sqrt{x} = [\sqrt{x_1}, \sqrt{x_2}, \ldots, \sqrt{x_n}]^T,$$
$$1/x = x^{-1} = [{x_1}^{-1}, {x_2}^{-1}, \ldots, {x_n}^{-1}]^T.$$

- Images with size $k \times n$ pixels are presented as columnwise stacked vectors through this thesis. However, matrix like indexing (for example $x_{i,j}$ refers to element or "pixel" $(i, j)$) is used. Also, we denote $N = kn$ which is naturally the number of pixels.

- Instead of "probability distribution" we often use just "distribution" or "density" rather loosely.

- All the integrals appearing this work are taken over the definition domain of the integrand if no integral limits are provided. All the integrals are Riemann integrals.

# Chapter 1

# Introduction

In this thesis Bayesian hierarchical methods for total variation regularisation are studied. The main objective is to present models that can be used to solve image deblurring tasks. We also briefly study image denoising and inpainting problems. There is a blurred and noisy image in the left in Figure 1.1. The problem is to be able to "deblur" the image so that it looks as close as possible the original unknown picture that is presented in the right in Figure 1.1. These image processing problems can be modelled as linear equations. However, due to noise and numerically problematic form of the blurring, obtaining a stable solution requires special regularisation methods. In this work we focus especially on statistical approaches where no additional parameters have to set by the user to obtain a deblurred image.

The classical methods for solving linear discrete system

$$y = Ax + \text{noise}, \tag{1.1}$$

where $A$ is a given (blurring) matrix, $x$ a vector (the image) to be solved and $y$ a given vector (original image), is to formulate it as a optimisation problem, in particular to a least squares problem. However, if the matrix has nontrivial nullspace, for instance, it has more columns than rows, or if the matrix is square but (close to) singular, it has either no unique solution or numerical problems emerge. This issue is typically dealt with by introducing a penalisation term and approximating the ill-posed problem with a problem that is well-posed. The problem is then to solve

$$\arg\min_x \{\|Ax - y\|^2 + \delta J(x)\}, \tag{1.2}$$

where $\delta$ is a regularisation parameter and $J(x)$ is a regularisation penalty, often $L^2$ or $L^1$ norm on $x$. The $L^2$ norm $J(x) = \|x\|_2^2$ is known as Tikhonov regularisation or Ridge regression. The $L^1$ norm $J(x) = \|x\|_1$ is often called the Lasso and was originally presented in [50]. If $\delta = 0$ then the problem reduces to the least squares method.

The goodness of the result, however, depends on how suitable a regularisation parameter $\delta$ was chosen. There exist several methods (see for example [52, Ch. 7] for further


Figure 1.1: (a) Noisy, blurred gray scale image of "Shepp-Logan phantom". (b) The original image which one wants to obtain given only the blurred image and knowledge about the blurring method.

discussion) to aid in choosing the regularisation parameter but there is no universally accepted method.

These optimisation problems can be interpreted as Bayesian inference problems. This is done by setting the prior density so that it corresponds to the regularisation penalty. For example setting a Gaussian prior leads to Tikhonov regularisation. The result of Bayesian inference is the whole posterior density. However, just the maximum a posteriori estimate is often computed. Furthermore, instead of setting regularisation parameter heuristically, in fully Bayesian models also these parameters are inferred from the data. This has the advantage that the user does not have to set it.

Total variation regularisation is a new popular alternative for restoring "blocky" images and it was initially presented in [48]. It penalises non-smoothness in the solutions while allowing occasional "jumps" and suits image deblurring problems better than Tikhonov or Lasso. The total variation regularisation term is, however, more difficult to deal with than the $L^2$ norm since it is not even differentiable everywhere. In this work total variation regularisation is studied in Bayesian setting by using Laplace prior which corresponds to the total variation penalty in corresponding minimisation problems. Exploiting the fact that the Laplace distribution can be presented as a Gaussian scale mixture ([3, 23, 31]) leads to a hierarchical model yielding a posterior density more feasible to deal with. This idea encourages to try other mixing densities that to the best of authors knowledge are not considered in literature. For example, Student's t-distribution is a Gaussian scale mixture and in this paper an alternative model combining t-distribution and total variation is proposed. As an extra we also obtain a hierarchical model for Lasso although we will focus mainly on total variation.

Although the resulting posterior distribution is intractable, the maximum a posteriori estimate is solved using direct maximisation of the posterior density in this work. Gibbs sampler update probabilities are also easily derived to give a Monte Carlo Markov Chain algorithm for the problem. However, sampling based algorithms are not considered in very detailed way. Instead, variational Bayes method as described

in [8, Ch. 10] or [38, Ch. 33] is used to approximate the posterior to obtain "analytic approximation" for the posterior mean and assess the uncertainty of the result. All these methods are compared empirically.

In literature there have been several studies related to the topic. A hierarchical Bayesian model for Lasso which is quite similar to the model used in this work was studied in [19, 43, 33]. In these papers the scale mixture idea was also used. In [33] also statistical model in the case featuring two penalisation terms, one $L^1$ and one total variation was analysed. Laplace priors are also considered in compressive sensing [6] and classification problems [28]. Fully Bayesian model of Tikhonov regularisation was studied in [26], where variational Bayes was used. Hierarchical models are also used in several inverse problem research projects, see for instance [10, 53, 41, 35]. Fast $L^1$ sampling methods have been considered for example in [36]. However, in these papers either no total variation model was considered or not everything was simultaneously estimated from the data or only sampling methods were proposed.

Bayesian models of total variation for image problems has been studied in [4, 5, 13, 10]. The ideas presented in this work are partly from these sources, though slightly different approach is taken. Also some ideas are extended and different variants of total variation are studied that allow exploiting Gaussian scale mixture property. In this work state-of-art or problem specific methods are not studied, some further related research is however found in papers [13, 9, 44]. There also exists several fast and popular minimisation algorithms (see for example [55, 24]) for solving total variation problems. There is a good summary of total variation and recent algorithms in [11].

The results obtained in this study are demonstrated by solving image processing problems for which total variation is suitable, namely deblurring. Also denoising and inpainting problems are briefly discussed. In deblurring problems considered in this work one wants to undo the blurring of an image with additive noise when the blurring kernel is known. All the needed parameters, regularisation parameter and variance of the Gaussian noise are estimated from the data in the model. In real life the blurring kernel may not be known but has to be estimated as well. This blind deconvolution problem was studied in [5] where similar concepts as in this work were used. In denoising problem only noise is removed and no blurring is removed. In inpainting problems in addition to denoising some pixels are unknown and must be restored using the information of the neighbouring pixels.

The contents of this thesis are the following. In next section some probability densities that are used in statistical models are presented. Also some basic principles of Bayesian inference and hierarchical models are briefly introduced. In Section 3 some inference algorithms are presented. After that, in Section 4 basic ideas of regularisation methods both from deterministic and Bayesian point of view are presented and briefly compared. Total variation is introduced in a deterministic framework. The main ideas of this work, the hierarchical model and inference, are presented in Section 5 and extended for two-dimensional case in Section 6. The results are illustrated in Section 7. Finally, after the examples the summary of this thesis is given and possible ideas for future work are discussed.

# Chapter 2

# Preliminaries

In this chapter some probability distributions and related results that are needed later in this work are presented. These probability densities are mainly needed in this work to model error or prior distributions and to be recognised when they emerge from posterior distributions. Thus only some basic relations are presented. Also basic ideas of Bayesian inference, which is the main tool to solve inverse problems in this thesis, are introduced. Remark that from now on both terms density and distribution are used to mean probability distribution and we will use these terms somewhat loosely.

## 2.1 Probability distributions and their properties

We start by presenting the multivariate normal density which is used commonly for observational errors. Normal density is perhaps the most used density in statistics and it is very tractable analytically. Due to the central limit theorem, the Gaussian densities are often good approximations to inherently non-Gaussian distributions when the observation is physically based on a large number of mutually independent random events [29]. In this thesis we use terms "normal" and "Gaussian" both to refer to this density.

**Definition 2.1.** A random $n$-vector $\mathbf{x}$ is said to have a multivariate normal or Gaussian distribution, denoted as $\mathbf{x} \sim \text{Normal}(\mu, \Sigma)$, with parameters $\mu$ and a $n \times n$ symmetric positive definite (spd) matrix $\Sigma$, if it has the probability density function (pdf)

$$p_{\mathbf{x}}(x) = \frac{1}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}. \tag{2.1}$$

The multivariate normal distribution has the following statistics

$$\mathbb{E}(\mathbf{x}) = \mathsf{mode}(\mathbf{x}) = \mu, \qquad \mathbb{V}(\mathbf{x}) = \Sigma. \tag{2.2}$$

A linear transformation of normal density is also normal.

**Theorem 2.2.** *If $m \times n$ matrix $A$ has full rank, $b$ is an $m$-vector and $\mathbf{x}$ is an $n$-dimensional random vector such that $\mathbf{x} \sim \text{Normal}(\mu, \Sigma)$, then the random vector $A\mathbf{x} + b \sim \text{Normal}(A\mu + b, A\Sigma A^T)$.*

*Proof.* The proof can be found in many textbooks. For example, see [47, Ch. 18]. □

**Remark 2.3.** Multivariate normal density could also be defined more generally so that $\Sigma$ does not need to be spd matrix. In this work these degenerate normal distributions are not used and that is why normal density is defined through the pdf.

Next we need to define some Bessel functions that will be encountered later when dealing with some more general densities. The second order modified Bessel function appearing later in the definitions of multidimensional Laplace and in generalised inverse Gaussian densities has the following integral presentation

$$\mathrm{K}_p(z) = \int_0^\infty \mathrm{e}^{-z\cosh(t)} \cosh(pt)\mathrm{d}t. \tag{2.3}$$

For positive $z$ the second order modified Bessel function has also the following integral formula

$$\mathrm{K}_p(z) = \frac{1}{2}\left(\frac{z}{2}\right)^p \int_0^\infty t^{-p-1} \exp\left(-\left(t + \frac{z^2}{4t}\right)\right) \mathrm{d}t. \tag{2.4}$$

The equivalence of (2.3) and (2.4) can be shown using a change of variables. In addition, the function clearly satisfies $\mathrm{K}_{-p}(z) = \mathrm{K}_p(z)$ for any $p > 0$. As a special case $p = 1/2$ we also have a simpler formula

$$\mathrm{K}_{1/2}(z) = \sqrt{\frac{\pi}{2}}\,\mathrm{e}^{-z} z^{-1/2}, \quad z > 0. \tag{2.5}$$

See, for instance [1, 31] and references therein for these and some additional properties and definitions. Some other properties and consequences of these functions will be discussed later.

The Generalised inverse Gaussian (GIG) distribution is a very general distribution family. The distribution was studied by Jorgensen in [27]. The following definition and properties are from this source. The GIG density can be defined with the following parametrisation.

**Definition 2.4.** A random variable $\mathbf{x} > 0$ has GIG distribution, denoted $\mathbf{x} \sim \text{GIG}(a, b, p)$, with parameters $a, b$ and $p$ if it has the pdf

$$p_{\mathbf{x}}(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \mathrm{e}^{-\frac{1}{2}(ax + \frac{b}{x})}, \quad x > 0, \tag{2.6}$$

where $K_p$ is the second order modified Bessel function. The normalisation constant follows from (2.4). The range of the parameters is

$$a > 0,\ b \geq 0,\ p > 0; \quad a > 0,\ b > 0,\ p = 0; \quad a \geq 0,\ b > 0,\ p < 0. \tag{2.7}$$

The GIG distribution is unimodal and skewed. The moments, mode and variance for GIG can be computed as given by the formulas in the following Proposition [27, pp. 7, 13–14]. The mean and variance have simplified formulas in the special cases $a = 0$ and $b = 0$ that follow by an asymptotic property of the modified Bessel function. These formulas are given in [27, pp. 13–14].

**Proposition 2.5.** *The mean, mode and variance of GIG distribution with parameters as in (2.7) are given by the formulas*

$$\mathbb{E}(\mathbf{x}^q) = \left(\frac{b}{a}\right)^{q/2} \frac{\mathrm{K}_{p+q}(\sqrt{ab})}{\mathrm{K}_p(\sqrt{ab})}, \quad q \in \mathbb{R}, \tag{2.8}$$

$$\mathsf{mode}(\mathbf{x}) = \begin{cases} \frac{(p-1)+\sqrt{(p-1)^2+ab}}{a}, & \text{if } a > 0, \\ \frac{b}{2(1-p)}, & \text{if } a = 0, \end{cases} \tag{2.9}$$

$$\mathbb{V}(\mathbf{x}) = \frac{b}{a}\left(\frac{\mathrm{K}_{p+2}(\sqrt{ab})}{\mathrm{K}_p(\sqrt{ab})} - \left(\frac{\mathrm{K}_{p+1}(\sqrt{ab})}{\mathrm{K}_p(\sqrt{ab})}\right)^2\right). \tag{2.10}$$

*Proof.* The formula for the central moments follows by direct integration and using the fact that the GIG pdf integrates to 1. The variance is then computed using $\mathbb{V}(\mathbf{x}) = \mathbb{E}(\mathbf{x}^2) - \mathbb{E}(\mathbf{x})^2$. The mode is computed by finding the unique zero of the derivative of the logarithm of the GIG pdf. □

Reciprocal Inverse Gaussian (RIG) and Inverse Gaussian (IG) densities are special cases of GIG. Setting $a = \alpha^2/\beta, b = \beta$ and $p = 1/2$ gives RIG and IG is obtained by setting $a = \lambda/\mu^2, b = \lambda$ and $p = -1/2$. Also gamma and inverse gamma densities are special cases of GIG which follow by setting $a = 2\beta, b = 0$ and $p = \alpha$ or $a = 0, b = 2\beta$ and $p = -\alpha$, respectively. Exponential distribution $\mathrm{Exp}(\theta)$ is the same as $\mathrm{Gamma}(1, \theta)$. These densities and some of their statistics are gathered in Tables 2.1 and 2.2. Note that $\Gamma(\cdot)$ is the gamma function and is defined as $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\mathrm{d}t$ for $x > 0$. IG and RIG densities have been studied in [51]. Different parametrisations are sometimes used for gamma and RIG densities and also many of the properties of these densities slightly differ then.

Let us next state some miscellaneous results related to these special cases that prove to be useful later in this work. If random variable $\mathbf{x}$ has distribution $\mathrm{GIG}(a, b, p)$ then $\mathbf{x}^{-1}$ has also GIG density, namely $\mathrm{GIG}(b, a, -p)$ [27, p. 7]. From this fact it follows that if $\mathbf{x} \sim \mathrm{IG}(\alpha/\beta, \alpha^2/\beta)$ and $\mathbf{y} = \mathbf{x}^{-1}$ then $\mathbf{y} \sim \mathrm{RIG}(\alpha, \beta)$. This result relates IG and RIG density. Similar connection exists between gamma and inverse gamma: if random variable $\mathbf{x} \sim \mathrm{Gamma}(\alpha, \beta)$ and $\mathbf{y} = \mathbf{x}^{-1}$ then $\mathbf{y} \sim \mathrm{InvGamma}(\alpha, \beta)$.

Also, if $\mathbf{x} \sim \mathrm{RIG}(\alpha, \beta)$ then

$$\mathbb{E}(\mathbf{x}^{-1}) = \frac{\alpha}{\beta}, \tag{2.11}$$

which is seen from Proposition 2.5 by setting $p = \frac{1}{2}$, $q = -1$ and using the symmetry of $\mathrm{K}_p$ on $p$. Some plots of RIG density are shown in Figure 2.1.

Table 2.1: Special cases of GIG distribution. All the parameters appearing in the formulas must be positive.

| $\mathbf{x} \sim$ | $p_{\mathbf{x}}(x)$ |
|---|---|
| $\mathrm{IG}(\mu, \lambda)$ | $\left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$ |
| $\mathrm{RIG}(\alpha, \beta)$ | $\frac{\alpha}{\sqrt{2\pi\beta}} \mathrm{e}^{2\alpha} x^{-\frac{1}{2}} \exp\left(-\frac{(\alpha x+\beta)^2}{2\beta x}\right)$ |
| $\mathrm{Exp}(\theta)$ | $\theta \mathrm{e}^{-\theta x}$ |
| $\mathrm{Gamma}(\alpha, \beta)$ | $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}$ |
| $\mathrm{InvGamma}(\alpha, \beta)$ | $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \mathrm{e}^{-\beta/x}$ |

Table 2.2: Some statistics of different distributions.

| $\mathbf{x} \sim$ | $\mathbb{E}(\mathbf{x})$ | $\mathsf{mode}(\mathbf{x})$ | $\mathbb{V}(\mathbf{x})$ |
|---|---|---|---|
| $\mathrm{IG}(\mu, \lambda)$ | $\mu$ | $\mu\left(\sqrt{1+\frac{9\mu^2}{4\lambda^2}} - \frac{3\mu}{2\lambda}\right)$ | $\frac{\mu^3}{\lambda}$ |
| $\mathrm{RIG}(\alpha, \beta)$ | $\frac{\beta(1+\alpha)}{\alpha^2}$ | $\frac{-\beta+\beta\sqrt{1+4\alpha^2}}{2\alpha^2}$ | use (2.10) |
| $\mathrm{Exp}(\theta)$ | $\theta^{-1}$ | 0 | $\theta^{-2}$ |
| $\mathrm{Gamma}(\alpha, \beta)$ | $\frac{\alpha}{\beta}$ | $\frac{\alpha-1}{\beta}$ for $\alpha > 1$ | $\frac{\alpha}{\beta^2}$ |
| $\mathrm{InvGamma}(\alpha, \beta)$ | $\frac{\beta}{\alpha-1}$ for $\alpha > 1$ | $\frac{\beta}{\alpha+1}$, | $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$ |

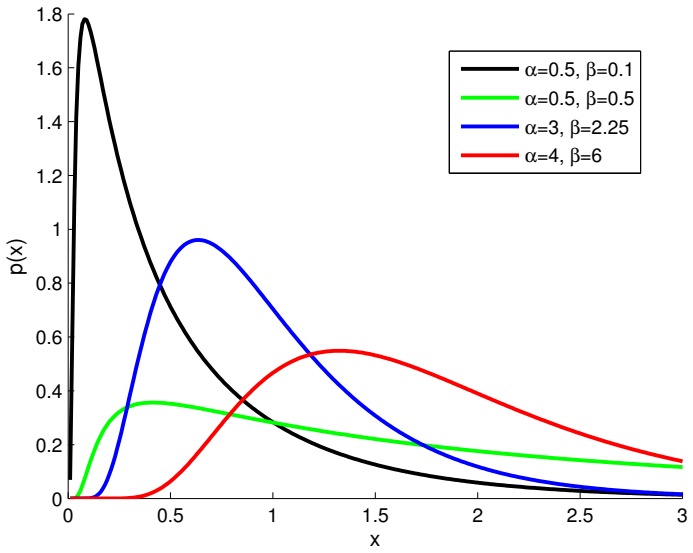


Figure 2.1: Some plots of RIG density.

**Definition 2.6.** A random $n$-vector $\mathbf{x}$ is said to have a multivariate t-distribution (or multivariate Student's t-distribution), denoted as $\mathbf{x} \sim \mathrm{t}_\nu(\mu, \Sigma)$, with parameters $\nu > 0$ (degree of freedom), $\mu$ and a $n \times n$ positive definite matrix $\Sigma$, if it has the pdf

$$p_{\mathbf{x}}(x) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})\nu^{n/2}\pi^{n/2}(\det(\Sigma))^{1/2}} \left(1 + \frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+n}{2}}. \tag{2.12}$$

The t-distribution is symmetric and behaves like normal density but it has heavier tails. For this reason it is often used in place of normal density when robustness to outliers is desired. The statistics of t-distribution are the following [8, Ch. 2.3.7].

$$\mathbb{E}(\mathbf{x}) = \mu, \text{ if } \nu > 1, \qquad \mathsf{mode}(\mathbf{x}) = \mu, \qquad \mathbb{V}(\mathbf{x}) = \frac{\nu}{\nu-2}\Sigma, \text{ if } \nu > 2. \tag{2.13}$$

A random vector $\mathbf{y}$ that follows t-distribution can be written as

$$\mathbf{y} = \mu + \frac{1}{\sqrt{\mathbf{r}}}\Sigma^{1/2}\mathbf{x}, \tag{2.14}$$

where $\mathbf{r} \sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ and $\mathbf{x} \sim \mathrm{Normal}(0, \mathrm{I})$ and $\mathbf{r}$, $\mathbf{x}$ are independent. The square root of matrix is defined so that $\Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma$. So t-distribution can be thought as a normal distribution with stochastic variance. Also, using Theorem 2.2 it can be seen that $\mathbf{y} \,|\, (\mathbf{r} = r) \sim \mathrm{Normal}(\mu, \frac{1}{r}\Sigma)$. The property is restated and proved in the next theorem. This Gaussian scale mixture (GSM) property can be used also to generate random variates.

**Theorem 2.7** (t-distribution as Gaussian scale mixture)**.** *If random vector* $\mathbf{y} \,|\, (\mathbf{r} = r) \sim \mathrm{Normal}(\mu, \frac{1}{r}\Sigma)$ *and* $\mathbf{r} \sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$*, then* $\mathbf{y} \sim \mathrm{t}_\nu(\mu, \Sigma)$*.*

*Proof.* Denoting $z = y - \mu$ it is easy to compute that

$$\begin{aligned}
p_{\mathbf{y}}(y) &= \int_0^\infty p_{\mathbf{y},\mathbf{r}}(y, r)\mathrm{d}r = \int_0^\infty p_{\mathbf{y}\,|\,\mathbf{r}}(y \,|\, r)p_{\mathbf{r}}(r)\mathrm{d}r \\
&= \int_0^\infty \frac{1}{(2\pi)^{n/2}(\det(\frac{1}{r}\Sigma))^{1/2}} \mathrm{e}^{-\frac{r}{2}z^T\Sigma^{-1}z} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} r^{\frac{\nu}{2}-1}\mathrm{e}^{-\frac{\nu r}{2}}\mathrm{d}r \\
&= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}\Gamma(\frac{\nu}{2})} \underbrace{\int_0^\infty r^{\frac{n+\nu}{2}-1}\mathrm{e}^{-\frac{1}{2}(z^T\Sigma^{-1}z+\nu)r}\mathrm{d}r}_{=\frac{\Gamma(\frac{n+\nu}{2})}{\left(\frac{1}{2}z^T\Sigma^{-1}z+\frac{1}{2}\nu\right)^{\frac{n+\nu}{2}}}} \\
&= \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})\nu^{n/2}\pi^{n/2}(\det(\Sigma))^{1/2}} \left(1 + \frac{1}{\nu}z^T\Sigma^{-1}z\right)^{-\frac{\nu+n}{2}},
\end{aligned}$$

which is the pdf of t-distribution. The integrand on the row 3 was observed to be gamma pdf without normalisation constant (see Table 2.1) and thus the integral on that line can be computed easily. □

**Remark 2.8.** Alternatively, the t-distribution can be characterised as a GSM using the connection $\mathbf{y} = \mu + \sqrt{\mathbf{r}}\Sigma^{1/2}\mathbf{x}$ with inverse gamma mixing density $\mathbf{r} \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu}{2})$. The proof of this follows directly from the proof of Theorem 2.7 by the connection of gamma and inverse gamma densities.

Let us define some other interesting and useful densities. In this work multivariate Laplace distribution is defined in the following way.

**Definition 2.9.** A random $n$-vector $\mathbf{x}$ is said to have a multivariate Laplace distribution, denoted as $\mathbf{x} \sim \text{MVLaplace}(\mu, \Sigma)$, with parameters $\mu$ and a $n \times n$ positive definite matrix $\Sigma$, if it has the pdf

$$p_{\mathbf{x}}(x) = \frac{2}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}} \frac{\mathrm{K}_{\frac{n}{2}-1}\left(\sqrt{2(x-\mu)^T\Sigma^{-1}(x-\mu)}\right)}{\left(\sqrt{\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right)^{\frac{n}{2}-1}}. \tag{2.15}$$

Function $\mathrm{K}_p$ is the modified Bessel function of the second kind with parameter $p \in \mathbb{R}$. This definition agrees with the one in [31, p. 235] but with the added location parameter $\mu$.

The multivariate Laplace distribution defined above is also a Gaussian scale mixture but with exponential mixing density. That is, it can be written as

$$\mathbf{y} = \mu + \sqrt{\mathbf{r}}\Sigma^{1/2}\mathbf{x}, \tag{2.16}$$

where $\mathbf{r} \sim \text{Exp}(1)$, $\mathbf{x} \sim \text{Normal}(0, \mathrm{I})$ and $\mathbf{r}$ and $\mathbf{x}$ are independent. This property in one-dimensional case was realised by Andrews and Mallows in 1974 in their paper [3], where also conditions for the existence of such relations was studied. The Laplace density could actually be defined using this GSM property. The next theorem is the generalisation to multidimensional case inspired by [18], where a slightly different version of the result below is given.

**Theorem 2.10** (Laplace as Gaussian scale mixture)**.** *If* $\mathbf{y} \,|\, (\mathbf{r} = r) \sim \text{Normal}(\mu, r\Sigma)$ *and* $\mathbf{r} \sim \text{Exp}(1)$*, then* $\mathbf{y} \sim \text{MVLaplace}(\mu, \Sigma)$.

*Proof.* The proof is straightforward calculation as in the t-distribution case.

$$\begin{aligned}
p_{\mathbf{y}}(y) &= \int_0^\infty p_{\mathbf{y}\,|\,\mathbf{r}}(y \,|\, r) p_{\mathbf{r}}(r) \mathrm{d}r \\
&= \int_0^\infty \frac{1}{(2\pi)^{n/2}(\det(r\Sigma))^{1/2}} e^{-\frac{1}{2r}z^T\Sigma^{-1}z} e^{-r} \mathrm{d}r \\
&= \frac{1}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}} \underbrace{\int_0^\infty r^{-n/2} e^{-\frac{1}{2}\left(2r + \frac{z^T\Sigma^{-1}z}{r}\right)} \mathrm{d}r}_{= 2\,\mathrm{K}_{1-\frac{n}{2}}\left(\sqrt{2z^T\Sigma^{-1}z}\right)\left(\frac{1}{2}z^T\Sigma^{-1}z\right)^{\frac{1}{2}\left(1-\frac{n}{2}\right)}} \\
&= \frac{2}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}} \frac{\mathrm{K}_{\frac{n}{2}-1}\left(\sqrt{2z^T\Sigma^{-1}z}\right)}{\left(\sqrt{\frac{1}{2}z^T\Sigma^{-1}z}\right)^{\frac{n}{2}-1}},
\end{aligned}$$

where we have denoted $z = y - \mu$. The integrand on line 3 was recognised as unnormalised GIG$(2, z^T\Sigma^{-1}z, 1 - \frac{n}{2})$ pdf (or alternatively a certain central moment of the inverse Gaussian distribution) and was computed using the fact that the density integrates to 1. □

The mean and variance of multivariate Laplace distribution can be computed from equation (2.16) using the fact that $\mathbf{x}$ and $\mathbf{r}$ are independent.

$$\begin{aligned}\mathbb{E}(\mathbf{y}) &= \mathbb{E}(\mu + \sqrt{\mathbf{r}}\Sigma^{1/2}\mathbf{x}) = \mathbb{E}(\mu) + \Sigma^{1/2}\mathbb{E}(\sqrt{\mathbf{r}})\mathbb{E}(\mathbf{x}) = \mu,\\ \mathbb{V}(\mathbf{y}) &= \mathbb{E}[(\mathbf{y} - \mathbb{E}(\mathbf{y}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))^T] = \mathbb{E}[\mathbf{r}(\Sigma^{1/2}\mathbf{x})(\Sigma^{1/2}\mathbf{x})^T]\\ &= \mathbb{E}(\mathbf{r})\mathbb{E}[(\Sigma^{1/2}\mathbf{x})(\Sigma^{1/2}\mathbf{x})^T] = 1 \cdot \Sigma = \Sigma.\end{aligned}$$

In one dimension the pdf with $\Sigma = 2/b^2 \in \mathbb{R}_+$ and $b > 0$ reduces to

$$p_{\mathbf{x}}(x) = \frac{b}{2}\mathrm{e}^{-b|x-\mu|}. \tag{2.17}$$

This one-dimensional Laplace density is denoted as Laplace$(\mu, b)$. This is easily seen by some simple calculations and using the fact (2.5). The second parameter was chosen in this specific way just for convenience. The mean is naturally $\mu$ and the variance is $2/b^2$. The density is sometimes also called the double-exponential density as it consists of two exponential curves. Laplace density has a sharp peak at its mean value, which will play an essential role in what follows later in this work. Some density functions with $\mu = 0$ are plotted in Figure 2.2.

**Definition 2.11.** A $p \times p$ positive-definite matrix $\mathbf{X}$ has Wishart distribution, denoted $\mathbf{X} \sim$ Wishart$(\Psi, \nu)$, with parameters $\mathbf{\Psi}$, which is $p \times p$ positive-definite matrix and $\nu > p - 1$, $\nu \in \mathbb{R}$ if it has the pdf

$$p_{\mathbf{X}}(X) = \frac{1}{2^{\frac{\nu p}{2}}(\det(\Psi))^{\nu/2}\Gamma_p(\frac{\nu}{2})}(\det(X))^{\frac{\nu-p-1}{2}}\mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\Psi^{-1}X)}, \tag{2.18}$$

where $\Gamma_p(\cdot)$ is multivariate gamma function and tr (trace) is the sum of diagonal elements. The mean and the mode are given by the formulas

$$\mathbb{E}(\mathbf{X}) = \nu\Psi, \quad \nu \geq p + 1, \qquad \mathsf{mode}(\mathbf{X}) = (\nu - p - 1)\Psi. \tag{2.19}$$

Wishart is a generalisation of Gamma distribution to positive definite matrices. In one dimension, the density Wishart$(v, n)$ clearly reduces to Gamma$(\frac{n}{2}, \frac{1}{2v})$. The properties of this matrix density is not studied here in more detail since in this work the Wishart distribution is used only as prior density for general covariance matrix $\Sigma$ of the multivariate normal density. In other words, we are only interested in its pdf and basic statistics since it is enough for this work. More properties and the definition of Wishart density through normal distributions can be found for instance in [2].

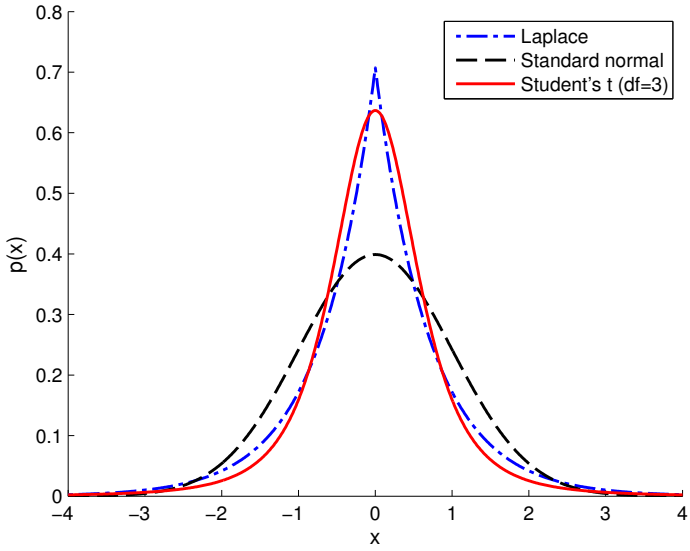


Figure 2.2: Comparison of Laplace, standard normal and t-distribution (with degree of freedom 3). All these densities are scaled to have mean zero and unit variance.

## 2.2 Bayesian inference

Bayesian inference can be seen as an "inverse problem". Namely, consider a statistical model describing the probability for obtaining some data $y$ and parameter $x$. The direct problem is to generate data given the model and parameters. The statistical inverse problem is, however, to describe parameters $x$ when the data $y$ is observed. In Bayesian statistics all the parameters are handled as random vectors. That is, a probability distribution describes the knowledge of $x$. The Bayesian inference is based on the Bayes' rule which is given by Theorem 2.12.

**Theorem 2.12** (Bayes' rule)**.** *Suppose that the $n$-dimensional random vector $\mathbf{x}$ has a known prior pdf $p_{\mathbf{x}}(x)$ and the data consists of observed value $y$ of an observable $k$-dimensional random vector $\mathbf{y}$ such that $p_{\mathbf{y}}(y) > 0$. Then the posterior pdf of $\mathbf{x}$ given the data $y$ is*

$$p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y) = \frac{p_{\mathbf{x}}(x)p_{\mathbf{y}\,|\,\mathbf{x}}(y\,|\,x)}{p_{\mathbf{y}}(y)}. \tag{2.20}$$

*Proof.* The proof can be found in many textbooks, see for example [29, Ch. 3.1]. □

The density $p_{\mathbf{x}}(x)$ is the prior density, or simply prior, that describes the initial knowledge of the value $\mathbf{x}$ before any data is observed. The conditional probability density $p_{\mathbf{y}\,|\,\mathbf{x}}(y\,|\,x)$, called likelihood, is the probabilistic model for obtaining the data $y$ if the unknown parameter $\mathbf{x}$ were known. The inference result, the posterior distribution $p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y)$ describes the knowledge about the parameter $x$ after taking the data into account and is the result of the inference problem.

The value

$$p_{\mathbf{y}}(y) = \int_{\mathbb{R}^n} p_{\mathbf{x}}(x)p_{\mathbf{y}\,|\,\mathbf{x}}(y\,|\,x)\,\mathrm{d}x, \tag{2.21}$$

that is sometimes called evidence, depends only on the data $y$ and it is only used for normalizing the pdf so that it integrates to 1. This constant value has no particular importance in inference but it has a role in Bayesian model comparison. Bayes' rule is thus often written simply as

$$p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y) \propto p_{\mathbf{x}}(x)p_{\mathbf{y}\,|\,\mathbf{x}}(y\,|\,x). \tag{2.22}$$

The prior is set to reflect the analyst's assumption or beforehand knowledge of the value to be inferred. Sometimes one sets improper prior to model initial knowledge of $x$. That is, a prior that is not a proper distribution in the sense that it does not integrate to 1. This is common especially if no clear prior knowledge exist. Conjugate priors that are densities from the same family of distributions as the likelihood are also commonly used for computational convenience. That is, they often produce simpler posterior densities while non-conjugate priors often lead to complicated posteriors and sampling methods are needed. Generally, with a lot of data the prior density plays no big role in determining the result. More information can be found in [45].

The whole posterior distribution is the solution of Bayesian inference. In one and two-dimensional cases one can plot the pdf to visualise the result. Possibly some marginals or credibility intervals can also be plotted to demonstrate the results. A posterior pdf with a narrow peak indicates that the mean or mode describes well the parameter while wide posterior pdf indicates that there is uncertainty about the result. However, it is often convenient to present some single numbers describing the posterior pdf. The following point estimates are often used the summarise the posterior.

MAP (maximum a-posteriori) estimate is defined as

$$x_{\mathrm{MAP}} = \mathsf{mode}(\mathbf{x}\,|\,y) = \arg\max_{x\in\mathbb{R}^n} p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y). \tag{2.23}$$

Basically, MAP estimate can be found by solving an optimisation problem often in high dimensional space using for example iterative, gradient based methods. However, there may be no solution to the problem or even when there exists a solution it may not be unique. That is to say, the posterior can be multimodal. There may be several local maximums in which case MAP may not describe the "location" of the pdf very well. Finding the highest mode can be difficult as global optimisation is generally a very difficult problem. MAP is closely related to maximum likelihood estimation, the difference is the prior density that is taken into account in MAP estimation.

Another commonly used estimator is the conditional mean (CM) which is also called posterior mean. CM is defined as

$$x_{\mathrm{CM}} = \mathbb{E}(\mathbf{x}\,|\,y) = \int_{\mathbb{R}^n} x\, p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y)\,\mathrm{d}x. \tag{2.24}$$

Conditional mean requires integration over often multidimensional domain. The conditional mean can also give better summary of a symmetric density with two distinct tops for example. CM is also known as the minimum mean square error estimator. Similarly as for the CM, one can define conditional covariance estimator. Also, the

conditional mean estimate described in this section may not exist, that is, the integral in (2.24) may not converge. Classical example is the Cauchy distribution which is the same as t-distribution with degree of freedom 1.

## 2.3 Bayesian hierarchical models

The main idea of hierarchical Bayesian models is to let the data determine the appropriate model used for the inversion of this data. That is, instead of determining the prior beforehand, some of the parameters in the prior are estimated from the data as well. This makes it possible to construct very flexible and "automated" models. Also parameters appearing in the likelihood can be estimated from the data instead of setting some fixed values for them.

Instead of considering prior $p_{\mathbf{x}}(x)$ (with some fixed parameters) for the random vector $\mathbf{x}$ to be inferred, one can consider prior $p_{\mathbf{x},\mathbf{r}}(x, r)$. Here $\mathbf{r}$ is hyperparameter with its own prior $p_{\mathbf{r}}(r)$ which is called hyperprior. Since $p_{\mathbf{x},\mathbf{r}}(x, r) = p_{\mathbf{x}\,|\,\mathbf{r}}(x\,|\,r)p_{\mathbf{r}}(r)$ integrating out the hyperparameter gives the prior for $\mathbf{x}$.

$$p_{\mathbf{x}}(x) = \int p_{\mathbf{x}\,|\,\mathbf{r}}(x\,|\,r)p_{\mathbf{r}}(r)\,\mathrm{d}r. \tag{2.25}$$

The posterior $p_{\mathbf{x},\mathbf{r}\,|\,\mathbf{y}}(x, r\,|\,y) \propto p_{\mathbf{x}\,|\,\mathbf{r}}(x\,|\,r)p_{\mathbf{r}}(r)p_{\mathbf{y}\,|\,\mathbf{x}}(y\,|\,x)$ contains now two types of parameters, some of which are of main interest and hyperparameters. Now, there are several ways to deal with this type of posterior, for example

- Solve the CM for $(\mathbf{x}, \mathbf{r})$. (Full-CM)
- Solve the MAP for $(\mathbf{x}, \mathbf{r})$. (Full-MAP)
- Marginalize $\mathbf{r}$, then solve $\hat{x} = \arg\max_x p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y)$. (Type I approach)
- Marginalize $\mathbf{x}$, then solve $\hat{r} = \arg\max_r p_{\mathbf{r}\,|\,\mathbf{y}}(r\,|\,y)$ and finally use $p_{\mathbf{x},\mathbf{r}\,|\,\mathbf{y}}(x, \hat{r}\,|\,y)$ to infer $\mathbf{x}$. (Empirical Bayes, evidence procedure, type II approach)
- Assume factorisation, for instance $p_{\mathbf{x},\mathbf{r}\,|\,\mathbf{y}}(x, r\,|\,y) \approx q_{\mathbf{x}}(x)q_{\mathbf{r}}(r)$ and find in some sense optimal $q_{\mathbf{x}}$ and $q_{\mathbf{r}}$. (Variational Bayes)

The names in brackets are not very settled. In this work we mainly consider the MAP and variational Bayes method. Full-CM is considered briefly for comparison and solved via sampling.

The Gaussian scale mixtures that were already introduced in the previous section can be used in hierarchical models. The idea is that instead of specifying a constant variance according to preliminary information, the variance is stochastic with heavy-tailed hyperprior and is estimated from the data. We showed that taking exponential mixing density gives the Laplace prior and taking inverse gamma with certain parameters yields t-distribution. More general priors are obtained using GIG as mixing

density, however we did not present this result as the resulting prior is a generalised hyperbolic distribution which has very complicated pdf and thus it will not tell much about the prior. It can be shown (see [54]) that any heavy-tailed mixing density also yields a heavy-tailed prior.

Alternatively instead of focusing on the marginal of the prior that is computed using (2.25), which is (in multidimensional case) multivariate generalised hyperbolic distribution if the mixing density is GIG, we can focus on the effect of the hyperprior. By that I mean that for instance in the Laplace case we assume that the variances of a Gaussian prior are distributed as Exp(1). So they are generally rather small but sometimes can be quite high as the tails are heavy. So the prior tends to set several components to very close to the mean while allowing some larger components. In usual non-hierarchical models one can choose Gaussian prior but then it is beforehand specified by setting constant variance how the components will behave.

It is also possible to construct models with three or even more layers and "hyper-hyperpriors". For example a specific three layer model is presented in [44]. We refer to [40, 53, 35] for more general discussion about hierarchical models and methodology of inferring parameters. Next let us take a look at how to actually solve the point estimates in more complicated cases.

# Chapter 3

# Bayesian inference algorithms

Computing the CM estimate can be hard. Solving the mean for a density defined in multidimensional space requires integrating over a multidimensional domain. Often there is no analytical formula to use. In these multidimensional cases neither common quadrature methods are applicable since the computational cost increases rapidly as a function of the dimension. On the other hand, computing the MAP requires solving an optimisation problem which can be hard if there are latent variables like missing data. However, there exits several iterative techniques to solve MAP. We will focus more on CM.

In this chapter some methods to compute the CM estimate are briefly discussed. It is assumed that posterior density is known up to the normalisation constant.

## 3.1 Gibbs sampler

In Monte Carlo Markov Chain (MCMC) methods one generates a large number of samples from a distribution. It is useful when this pdf is multidimensional and the normalisation constant is unknown and intractable. Given independent samples $\{x^{(1)}, \ldots, x^{(N)}\}$ expectations such as the mean can be approximated as

$$\mathbb{E}(g(\mathbf{x}) \,|\, y) = \int g(x) p_{\mathbf{x}\,|\,\mathbf{y}}(x \,|\, y) \mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} g(x^{(i)}). \tag{3.1}$$

The Gibbs sampler is a MCMC algorithm to generate samples from multidimensional distribution and it is a specific case of Metropolis-Hastings method. The basic idea is that at each step of the algorithm only one component of the multidimensional random vector is changed by sampling from the corresponding one-dimensional conditional distribution that is obtained by keeping all the other parameters fixed. Sometimes these one-dimensional conditional distributions are easily obtained from the posterior and this is sometimes the case when dealing with hierarchical models. It is easier to

generate random samples from one-dimensional distributions. However, these conditional distributions need not necessarily be one-dimensional.

The Gibbs sampler algorithm for sampling from an $n$-dimensional posterior distribution $p_{\mathbf{x}\,|\,\mathbf{y}}(x\,|\,y)$ is presented as Algorithm 1. In this version of the algorithm one cycles through the indices in specific order. Another possibility is to choose the order of the updates at random.

**Algorithm 1:** Gibbs sampler (cyclic update)

Select some initial values $[x_1^{(0)}, \ldots, x_n^{(0)}]$ from the domain of $x$
**for** *i from 1 to N* **do**
  **for** *j from 1 to n* **do**
    $x_j^{(i)} \leftarrow$ sample from $p(x_j \,|\, x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_n^{(i-1)}, y)$
  **end**
**end**
Remove the first $N_1$ samples and return samples $\{x^{(N_1)}, \ldots, x^{(N)}\}$.

It usually takes some iterations before the sequence of samples $\{x^{(1)}, \ldots, x^{(N)}\}$ can be considered to be a set of random samples from the given distribution. The algorithm generates samples from the Markov chain that has the given distributions as its equilibrium distribution and it takes some time before this stationary chain is found. These "burn-in" samples are usually ignored. In addition, the sequential samples are correlated but usually all samples after the burn-in period are used in estimation nevertheless.

The law of large numbers implies that taking infinitely many samples the average converges to the conditional mean with probability one. The speed of converge does not, in principle, depend on the dimension of the problem. Even though the samples are not uncorrelated the convergence is guaranteed if the sequence of samples comes from ergodic Markov chain. However, in this work these convergence results are not studied in more detail. Further discussion of MCMC methods and their convergence as well as the proof for the Gibbs algorithm can be found for example in [46] or [29]. We next turn to methods that are not based on sampling.

## 3.2 Variational Bayes

Variational Bayes (VB) approximation can be used to approximate the mean and mode for different marginals of a posterior density. The idea of variational inference is to approximate a difficult posterior density to yield useful computational simplifications. Given the possibility to sample infinitely many samples the correct marginal density or some expectations could be achieved with MCMC methods. However, for large models it might be more sensible to compute "analytic approximation" since sampling based solution tend to be slow and thus only suitable for small scale problems. Also it can be hard to know if the sampler is producing samples from the correct density.

Expectation Maximisation (EM) which is a closely related method to VB, produces only the MAP estimate. In the EM algorithm one needs to evaluate the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables which can be infeasible. Thus approximations are needed. Compared to EM, VB yields information also of the uncertainty of the result and not just point estimates. For the rest of this section lower indices of densities will be left out for simplicity, for example we write $p(z \mid y)$ instead of $p_{\mathbf{z} \mid \mathbf{y}}(z \mid y)$.

The Kullback-Leibler divergence can be used to measure the closeness of two probability densities. It can be defined in the following way.

**Definition 3.1.** Let continuous distributions $q$ and $p$ be defined on the set $S$. Furthermore, assume that they are strictly positive on the set $S$. Then the Kullback-Leibler divergence or relative entropy (KL) between $q$ and $p$ is defined as the integral

$$\mathrm{KL}(q\|p) = -\int_S q(x) \ln\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{3.2}$$

Note that sometimes KL is defined without the minus sign and then the nominator and denominator are flipped. Kullback-Leibler divergence is positive for all probability densities $q$ and $p$, that is $\mathrm{KL}(q\|p) \geq 0$. The equality holds if and only if $q = p$ almost everywhere, see [32, Lemma 3.1]. Also, KL is not symmetric about its parameters, that is, generally $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$.

One can approximate the posterior distribution $p(z \mid y)$ with some distribution $q(z)$ which will be in some way restricted. Setting $q(z) = p(z \mid y)$ would obviously minimize the unrestricted problem. So one wants to find pdf $q(z)$ so that the Kullback-Leibler divergence

$$\mathrm{KL}(q\|p) = -\int q(z) \ln\left(\frac{p(z \mid y)}{q(z)}\right) \mathrm{d}z, \tag{3.3}$$

is minimised, where $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ are parameters (and latent variables) and $y$ is the data. One can show that the following decomposition holds

$$\ln(p(y)) = \mathrm{L}(q) + \mathrm{KL}(q\|p), \tag{3.4}$$

where

$$\mathrm{L}(q) = \int q(z) \ln\left(\frac{p(z, y)}{q(z)}\right) \mathrm{d}z \tag{3.5}$$

and $\mathrm{KL}(q\|p)$ is as in (3.3). Since $\ln(p(\mathbf{y}))$ is constant, minimizing Kullback-Leibler can be achieved by maximizing the lower bound $\mathrm{L}(q)$.

We can assume that a probability distribution $q(z)$ above with parameters (and latent variables) that can be also grouped such that $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_g)$, is restricted to be factorized as

$$q_{\mathbf{z}}(z) = \prod_{i=1}^{g} q_{\mathbf{z}_i}(z_i). \tag{3.6}$$

Here the parameters $\mathbf{z}_1, \ldots, \mathbf{z}_g$ need not be of the same size or one-dimensional. This form of variational inference is often called a mean field approximation. Note that the idea is to approximate the joint density and that the independent densities $q_{\mathbf{z}_i}(z_i)$, for which a shorter notation $q_i(z_i)$ will be used for the rest of this section, can be poor approximations for the real marginal densities [21].

The objective is to find such densities $q_i, i = 1, \ldots, g$ that minimize $\mathrm{KL}(q||p)$ and maximize the corresponding lower bound. The optimal pdfs $q_i^*$ can be found by computing the following expectations (see for instance [21] or [8, pp. 464 – 466] for derivation)

$$\ln q_j^*(z_j) = \mathbb{E}_{\mathbf{z}_i : i \neq j}[\ln p(z, y)] + c = \int \ln p(z, y) \prod_{i \neq j} q_i(z_i) \, \mathrm{d}z_i + c. \tag{3.7}$$

In the above formula $p(z, y)$ is the joint distribution of $\mathbf{z}$ and the data $\mathbf{y}$ and $c$ is constant with respect to the current random vector to be solved. The expectation is taken over all variables except the $j$th. The constant is related to the normalisation term and it is not needed to be computed since we know that $q_i$'s are normalised pdfs.

The parameters of the distributions $q_j^*$ will usually depend on expectations with respect to other distributions $q_i^*$ for $i \neq j$. So in practise the parameters are solved iteratively. That is, once the unknown distributions are obtained, one starts with some initial values for the unknown parameters of these pdfs and updates them in cyclic way using the current estimates for the other densities until some stopping criteria is satisfied. The algorithm is guaranteed to converge. [8]

# Chapter 4

# Regularisation methods

In inverse problems one usually wants to interpret some indirect physical measurements of an unknown object of interest. For example in X-ray tomography typical inverse problem is to reconstruct three-dimensional structure of patients insides given some X-ray images. In this case there may not be enough data to be able to make the reconstruction without special methods. In this work we focus on image blurring which can be said to be quite classical example of inverse problem. The inverse problems are usually much harder to solve than the corresponding direct problems.

To gain better understanding what makes some problem an inverse problem let us revise the notion of well-posed problem by Hadamard:

- The problem must have solution (existence).
- The problem must have at most one solution (uniqueness).
- The solution must depend continuously on data (stability).

Inverse problem fails to satisfy one or more of these conditions. In linear inverse problems (1.1) that are dealt with here, the problems arise with the second and third conditions. The matrix may not have full rank and thus there exists no unique inverse or even if it has the inversion of this matrix can be numerically unstable. That is, small measurement errors cause the direct solution attempt fail.

Next some popular regularisation methods, Tikhonov regularisation, Lasso and total variation regularisation are briefly introduced. For Tikhonov regularisation both minimisation and statistical approach, that is, Gaussian priors, are presented. For the Lasso and total variation mainly non-statistical approaches are introduced and hierarchical Bayesian statistical models for these cases are considered in later chapters.

## 4.1 Tikhonov regularisation and Gaussian priors

Let $y$ be an $n$-vector of observations, $x$ is a $k$-vector of unknown parameters, and $A$ is a known constant $n \times k$ matrix. The Tikhonov regularised solution for the equation $y = Ax + \epsilon$, where $\epsilon$ models noise, is the vector that is the solution to

$$x_{\text{Tikh}} = \underset{x \in \mathbb{R}^k}{\arg\min} \{\|Ax - y\|_2^2 + \delta \|x\|_2^2\} \tag{4.1}$$

for some regularisation parameter $\delta > 0$. If $\delta = 0$ then (4.1) simplifies to the standard least squares minimisation problem. The parameter $\delta$ can be used to tune the balance between small residual and small $L^2$ norm for the solution vector. In inverse problems there may be infinitely many solutions for the corresponding least squares problem and one of the roles of the penalty term is to make the solution unique.

**Theorem 4.1.** *The Tikhonov regularised solution for (4.1) is given by*

$$x_{Tikh} = V D_\delta^+ U^T y, \tag{4.2}$$

*where $A = UDV^T$ is the singular value decomposition (svd) of $A$ with orthogonal matrices $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{k \times k}$ and diagonal matrix $D \in \mathbb{R}^{n \times k}$ with singular values $d_1, \ldots, d_r \geq 0$, where $r = \min(k, n)$, on its diagonal, and*

$$D_\delta^+ = \operatorname{diag}\left(\frac{d_1}{d_1^2 + \delta}, \ldots, \frac{d_r}{d_r^2 + \delta}\right) \in \mathbb{R}^{k \times n}. \tag{4.3}$$

*Proof.* See [49, p. 33]. □

From theorem 4.1 we can see that the regularisation parameter makes the inverse of $D$ better conditioned as singular values tend to vary a lot in the case of almost nonsingular matrix. This fact makes numerical computation very unstable. The solution has also another formula which is given by

**Theorem 4.2.** *The solution to the minimisation problem (4.1) satisfies*

$$x_{Tikh} = (A^T A + \delta I)^{-1} A^T y. \tag{4.4}$$

*Proof.* The proof can be found in [49, p. 39]. □

This solution is usually faster to compute than the one that requires computing the svd. If $\delta = 0$ then the solution to the least squares problem is also given by Theorem 4.2. Next we will look at the statistical approach to this problem.

**Theorem 4.3.** *Let $\mathbf{x}$ and $\boldsymbol{\epsilon}$ be mutually independent $k$ and $n$-dimensional random vectors, respectively, with Gaussian densities*

$$\mathbf{x} \sim \text{Normal}(x_0, \Sigma_0), \quad \boldsymbol{\epsilon} \sim \text{Normal}(0, P) \tag{4.5}$$

*with positive definite covariance matrices* $\Sigma_0 \in \mathbb{R}^{k\times k}$ *and* $P \in \mathbb{R}^{n\times n}$*. Furthermore, assume that there exists a linear model* $\mathbf{y} = A\mathbf{x} + \boldsymbol{\epsilon}$ *with known matrix* $A \in \mathbb{R}^{n\times k}$ *for noisy measurement* $\mathbf{y}$*. Then the posterior is*

$$\mathbf{x} \,|\, (\mathbf{y} = y) \sim \text{Normal}(x_{\text{post}}, \Sigma_{\text{post}}), \tag{4.6}$$

*where*

$$\Sigma_{\text{post}} = (\Sigma_0^{-1} + A^T P^{-1} A)^{-1}, \tag{4.7}$$
$$x_{\text{post}} = \Sigma_{\text{post}}(A^T P^{-1} y + \Sigma_0^{-1} x_0). \tag{4.8}$$

*Proof.* These formulas are obtainable by computing the product of the two Gaussian densities. By Bayes' rule we get

$$\begin{aligned} p_{\mathbf{x}\,|\,\mathbf{y}}(x \,|\, y) &\propto p_{\mathbf{y}\,|\,\mathbf{x}}(y \,|\, x) p_{\mathbf{x}}(x) \\ &\propto \mathrm{e}^{-\frac{1}{2}(y-Ax)^T P^{-1}(y-Ax) - \frac{1}{2}(x-x_0)^T \Sigma_0^{-1}(x-x_0)} \\ &= \mathrm{e}^{-\frac{1}{2}(x^T\Sigma_0^{-1}x + x^T A^T P^{-1} A x - 2x^T \Sigma_0^{-1} x_0 - 2x^T A^T P^{-1} y) + c_1} \\ &= \mathrm{e}^{-\frac{1}{2}(x^T Q x - 2x^T Q Q^{-1}(\Sigma_0^{-1} x_0 + A^T P^{-1} y)) + c_1} \\ &= \mathrm{e}^{-\frac{1}{2}(x-\hat{x})^T Q (x-\hat{x}) + c_2}, \end{aligned}$$

where we defined $Q = \Sigma_0^{-1} + A^T P^{-1} A$ which is clearly positive definite and thus invertible and $\hat{x} = Q^{-1}(A^T P^{-1} y + \Sigma_0 x_0)$. The constants $c_1$ and $c_2$ do not depend on $x$ so the formula on the last line can be recognised as $\text{Normal}(\hat{x}, Q^{-1})$ from which the formulas for $\Sigma_{\text{post}}$ and $x_{\text{post}}$ follow. □

Using the matrix inversion lemma (see appendix A) or by computations as in [29, Ch. 3.4] one can write the equations for the posterior mean and covariance matrix in the form

$$x_{\text{post}} = x_0 + \Sigma_0 A^T (A\Sigma_0 A^T + P)^{-1}(y - Ax_0), \tag{4.9}$$
$$\Sigma_{\text{post}} = \Sigma_0 - \Sigma_0 A^T (A\Sigma_0 A^T + P)^{-1} A\Sigma_0, \tag{4.10}$$

which is sometimes more feasible to use. This is especially so if matrix $A$ has more columns than rows. Now, assuming the prior $\mathbf{x} \sim \text{Normal}(0, (\lambda \mathrm{I})^{-1})$ and white noise model $\boldsymbol{\epsilon} \sim \text{Normal}(0, (\nu \mathrm{I})^{-1})$ result (4.8) gives the familiar equation for the posterior mean

$$x_{\text{post}} = (A^T A + \delta \mathrm{I})^{-1} A^T y, \tag{4.11}$$

where $\delta = \lambda/\nu$. This re-interpretation gives new insight into the choice of the regularisation parameter $\delta$. It is the ratio of the noise and prior variances. The MAP and CM estimates are the same in the case of pure Gaussian densities. It is also worth emphasizing that the statistical approach offers also information about the credibility of the result, for example the variance as given by (4.7).

The Tikhonov regularisation can also be generalised by using penalty function $J(x) = \delta\|Lx\|_2^2$ in the place of $J(x) = \delta\|x\|_2^2$ in (4.1). The matrix $L$ is typically a discretised

differential operator. It can be shown that the solution to this generalised Tikhonov regularisation satisfies

$$x_{\text{Tikh}} = (A^T A + \delta L^T L)^{-1} A^T y. \tag{4.12}$$

Corresponding prior in the statistical model would be

$$p_{\mathbf{x}}(x) \propto \mathrm{e}^{-\frac{\lambda}{2}\|Lx\|^2} = \mathrm{e}^{-\frac{\lambda}{2}x^T L^T L x}, \tag{4.13}$$

and it can be seen that Theorem 4.3 also gives the result (4.12) with $x_0 = 0$ and $\Sigma_0^{-1} = \lambda L^T L$. These priors are called Gaussian smoothness priors and in particular those prior models that have structural information encoded in them are useful. However, in the Tikhonov regularisation case the matrix $L$ need not necessarily have full column rank in which case matrix $L^T L$ is not invertible. We have the following result.

**Theorem 4.4.** *Consider linear model* $\mathbf{y} = A\mathbf{x} + \boldsymbol{\epsilon}$ *as before, where* $\mathbf{x}$ *and* $\boldsymbol{\epsilon}$ *be mutually independent* $k$ *and* $n$*-dimensional random vectors and* $\boldsymbol{\epsilon} \sim \text{Normal}(0, P)$ *with positive definite covariance matrix* $P \in \mathbb{R}^{n\times n}$*. Let* $L \in \mathbb{R}^{k\times n}$ *so that* $\mathcal{N}(L) \cap \mathcal{N}(A) = \{0\}$*. Then*

$$p_{\mathbf{x}\,|\,\mathbf{y}}(x \,|\, y) \propto \mathrm{e}^{-\frac{1}{2}((y-Ax)^T P^{-1}(y-Ax) + \|Lx\|_2^2)} \tag{4.14}$$

*defines a Gaussian density with positive definite covariance matrix* $\Sigma_{\text{post}}$ *and mean* $x_{\text{post}}$ *as given by (4.7) and (4.8) but with setting* $\Sigma_0^{-1} = L^T L$*.*

*Proof.* Denote $Q = A^T P^{-1} A + L^T L$. If $x \in \mathcal{N}(Q)$ then

$$x^T Q x = \|Lx\|_2^2 + \|P^{-1/2} A x\|_2^2 = 0.$$

Since $\mathcal{N}(P^{-1/2}) = \{0\}$ we can see that $x \in \mathcal{N}(L) \cap \mathcal{N}(A) = \{0\}$ and thus $x = 0$ implying that $Q$ has trivial nullspace. So $Q$ is invertible and we can also see that it is symmetric and positive definite. The rest follows immediately from Theorem 4.3 by setting $\Sigma_0^{-1} = L^T L$, $x_0 = 0$ and since $Q$ is positive definite. □

If $A$ and $L$ have common nontrivial kernel then the resulting density is degenerate and the inverse problem becomes underdetermined. That is, there exists no unique solution. [29].

In the optimisation problem version of Tikhonov regularisation the regularisation parameter $\delta$ was considered to be known and also in the previous statistical approach the variance of the Gaussian error term was assumed to be known. There are several (deterministic) methods, for instance, Morozov discrepancy principle, L-curve method and generalised cross validation [52, 25] for choosing "correct" regularisation parameter. However, it is possible to estimate it as well as the error variance from the data simultaneously as was the idea of the hierarchical models.

## 4.2 The Lasso and some generalisations

Another popular regularised least squares method is the Lasso (least absolute shrinkage and selection operator) which uses $L^1$ norm as the regularisation penalty. It was first presented by Tibshirani [50] in 1996. The problem is to solve

$$\begin{aligned} &\arg\min_{x\in\mathbb{R}^k} \|Ax-y\|_2^2 \\ &\text{s.t.} \quad \|x\|_1 \leq t, \end{aligned} \tag{4.15}$$

where $\|x\|_1 = \sum_{i=1}^k |x_i|$ and $t \geq 0$ is a tuning parameter. This method promotes sparsity, that is, it tends to produce results in which several components of $x$ are driven to zero. On the other hand Tikhonov regularisation tends to produce solutions with small but generally nonzero elements. This is why Lasso suits well for compressive sensing problems. The Lasso problem can also be written as

$$\arg\min_{x\in\mathbb{R}^k} \{\|Ax-y\|_2^2 + \delta\|x\|_1\}, \tag{4.16}$$

where $\delta$ is regularisation parameter. These two approaches can be related using Lagrange multipliers. What is interesting in the context of this work is that the $L^1$ penalty can be interpreted as a zero mean Laplace prior density just like the Tikhonov regularisation term was related to Gaussian prior. The sparsity property of the Lasso is due to the fact that the Laplace density has more probability mass near the mean and has heavier tails than the Gaussian density. The prior corresponding to the Lasso penalty term is

$$p_{\mathbf{x}}(x) = \frac{\delta}{2}\exp\left(-\delta\sum_{i=1}^k |x_i|\right). \tag{4.17}$$

There exists, however, no standard distribution for the resulting posterior like in the corresponding case of Gaussian prior. Luckily, a hierarchical model can be constructed by setting Gaussian prior and setting the exponential hyperprior for the variance. This hierarchical Bayesian Lasso model is not considered here since it appears as a specific case of total variation regularisation model in one-dimensional case later in this work.

The Lasso method has also several extensions that are introduced here very briefly. (See for example [40] for some general algorithms.) They also have hierarchical counterparts in addition to the minimisation problems. In *elastic net* one wants to find the solution to

$$\arg\min_{x\in\mathbb{R}^k} \{\|Ax-y\|_2^2 + \delta_1\|x\|_1 + \delta_2\|x\|_2^2\}, \tag{4.18}$$

that is both $L^1$ and $L^2$ penalties are simultaneously placed. Another modern variant is *fused Lasso*, which is a minimisation problem

$$\arg\min_{x\in\mathbb{R}^k} \{\|Ax-y\|_2^2 + \delta_1\|x\|_1 + \delta_2\sum_i |x_{i+1}-x_i|\}, \tag{4.19}$$

In fused Lasso both $L^1$ and total variation penalties are applied. The parameters $\delta_1$ and $\delta_2$ are positive tuning parameters. Hierarchical Bayesian models for both of these methods as for the standard Lasso have been considered in literature, see for example [33], where the need for these kind of variants is justified.

## 4.3 Total Variation

The total variation (TV) of a function $f : [0, 1] \to \mathbb{R}$ can be defined as

$$\mathrm{TV}(f) = \sup \sum_i |f(x_{i+1}) - f(x_i)|, \tag{4.20}$$

where the supremum is taken over all partitions $0 = x_0 < x_1 < \ldots < x_n = 1$. If the function $f$ is smooth one can multiply and divide the right-hand side of (4.20) with $\Delta x_i = x_{i+1} - x_i$ and after taking the limit $\Delta x_i \to 0$ obtain the following formulation for TV [52].

$$\mathrm{TV}(f) = \int_0^1 \left| \frac{\mathrm{d}f}{\mathrm{d}x} \right| \mathrm{d}x. \tag{4.21}$$

More generally, let $f \in L^1(\Omega)$ which is the space of integrable functions on $\Omega \subseteq \mathbb{R}^k$. TV functional can be defined as

$$\mathrm{TV}(f) = \sup \left\{ \int_\Omega f \, \nabla \cdot g \, \mathrm{d}x \,\middle|\, g = (g_1, \ldots, g_k) \in C_0^1(\Omega; \mathbb{R}^k), \|g(x)\|_2 \leq 1 \right\}. \tag{4.22}$$

The test function space $C_0^1(\Omega; \mathbb{R}^k)$ consists of continuously differentiable vector-valued functions on $\Omega$ that vanish at the boundary $\partial\Omega$ and $\nabla \cdot g$ is the divergence of function $g$. If function $f$ satisfies $\mathrm{TV}(f) < \infty$ then it is said to have bounded variation.

The natural way to extend total variation to two-dimensional case from (4.21) is to consider

$$\mathrm{TV}(f) = \int_0^1 \int_0^1 \|\nabla f\|_2 \, \mathrm{d}x \, \mathrm{d}y \tag{4.23}$$

for smooth function $f$ defined on $[0, 1] \times [0, 1]$, where $\nabla f = [\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}]^T$ is the gradient of $f$. It can also be shown that (4.22) simplifies to (4.23) in the case of smooth function $f$. For more detailed treatment we refer to [52, Ch. 8].

The total variation regularisation tends to produce very good reconstructions of "blocky images". These blocky images are nearly piecewise constant with jump discontinuities and the length of curves on which the discontinuities occur is relatively small. The total variation has also a geometric interpretation. TV as in (4.21) and (4.23) can be interpreted as the lateral surface of the graph of $f$. For example, let $S$ be a region with a smooth boundary $\partial S$ contained in unit square. Let $f(x, y) = h > 0$ in the interior of $S$ and $f(x, y) = 0$ in the exterior. Then $\mathrm{TV}(f)$ is the length of $\partial S$ multiplied by the height $h$ of the discontinuity of $f$. [52, p. 148] This is demonstrated in Figure 4.1.

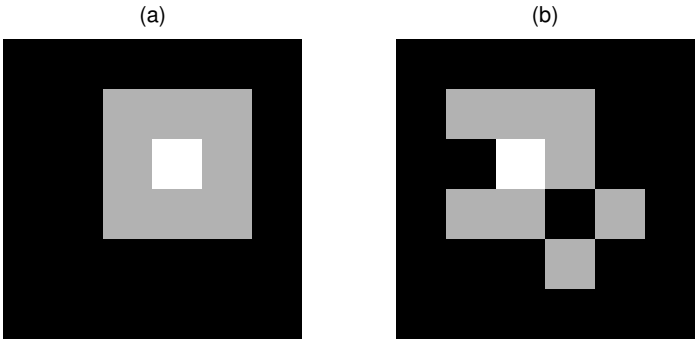


Figure 4.1: The piecewise defined function in (a) has smaller total variation than the one in (b) since the total length of the jump borders is clearly smaller in (a).

However, we want to consider TV in discrete domain since it allows to develop TV in statistical framework. For numerical computations continuous variables must be discretised anyway but theory is often carried through in continuous domain when considering deterministic approach. Anyway, suppose that the function $f = f_{ij}$ is defined on equispaced grid characterised by points $x_{ij}, i = 1, ..., k,\ j = 1, ..., n$ with spacings $\Delta x$ and $\Delta y$ in two dimensions. The following notation is used from now on to denote horizontal and vertical differences between "pixels": $\nabla^v_{ij} f = f_{i+1,j} - f_{i,j}$ and $\nabla^h_{ij} f = f_{i,j+1} - f_{i,j}$. Now (4.23) can be discretised so that ones obtains

$$\mathrm{TV}(f) = \sum_{i=1}^{k} \sum_{j=1}^{n} \sqrt{\left(\frac{\nabla^v_{ij} f}{\Delta x}\right)^2 + \left(\frac{\nabla^h_{ij} f}{\Delta y}\right)^2} \tag{4.24}$$

with some boundary conditions applied for the "overindexing" terms. If we consider equispaced two-dimensional grid so that $\Delta x = \Delta y$ then these constant delta terms can be neglected since they are absorbed to the regularisation parameter and in Bayesian model to the normalisation constant. Similarly approximating the derivative in one-dimensional case leads to discretisation for (4.21) of the form

$$\mathrm{TV}(f) = \sum_i |f_{i+1} - f_i|. \tag{4.25}$$

We now focus on TV in two-dimensional case. The discretised two-dimensional version of total variation functional is called isotropic TV and it is here defined by

$$\mathrm{TV}_{iso}(f) = \sum_{i=1}^{k} \sum_{j=1}^{n} \sqrt{\left(\nabla^v_{ij} f\right)^2 + \left(\nabla^h_{ij} f\right)^2}. \tag{4.26}$$

Another discretisation option is the TV variant

$$\mathrm{TV}(f) = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\{ \left| \frac{\nabla^v_{ij} f}{\Delta x} \right| + \left| \frac{\nabla^h_{ij} f}{\Delta y} \right| \right\}. \tag{4.27}$$

Again, for simplicity one can assume $\Delta x = \Delta y$ and neglect these two terms in (4.27). The terms that "overindex" are defined by some boundary conditions. This version of TV is called anisotropic TV and after neglecting those delta terms it can be written as

$$\mathrm{TV}_{aniso}(f) = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\{ \left| \nabla_{ij}^{v} f \right| + \left| \nabla_{ij}^{h} f \right| \right\}. \tag{4.28}$$

This $L^1$ norm version can be seen as an approximation to the "real" TV functional as in (4.23). Sometimes it is even wrongly presented as the actual TV penalty. This anisotropic version might be somewhat easier to deal with, though neither are differentiable everywhere. These discrete TV versions are presented, for example, in [55]. This $L^1$ version of TV is obviously related to the Lasso. Each absolute value term in the summation can be seen as zero mean Laplace distribution when considering the differences of neighbouring elements of $f$ as random variables. While the isotropic TV is seen as the "real" generalisation for TV, the anisotropic form may also be more suitable for some more complex structures than two-dimensional grid of an image. In some applications it may be desirable that some arbitrary differences of components are penalised.

The discrete total variation task from the optimisation point of view is the following problem. Again, the regularisation parameter is $\delta$.

$$\arg\min_{x} \{\|Ax - y\|_2^2 + \delta \mathrm{TV}(x)\}, \tag{4.29}$$

In (4.29) $\mathrm{TV}(x)$ can be chosen to be one of the TV penalties considered previously. Note that $x$ is used whenever we consider discretised model and $f$ refers to function. Due to the fact that the TV penalty is not differentiable at origin, one of the basic methods for solving (4.29) is to approximate the TV functional. For example in two-dimensional isotropic TV case this is often done by using the following penalty

$$J_\beta(f) = \int_0^1 \int_0^1 \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \beta^2}\, \mathrm{d}x\, \mathrm{d}y, \tag{4.30}$$

for some small constant $\beta > 0$. The penalty in (4.30) can be discretised. This leads to a minimisation problem that is differentiable everywhere and thus easier to solve using numerical methods. Some minimisation algorithms like steepest descent or Newton's method can be directly applied to solve (4.29) with this penalty. These algorithms require computing the gradient and some require the Hessian matrix. For the details see for example [52]. Naturally similar approximation can be done in one-dimensional case. This approximation is quite good for small $\beta$, larger values of $\beta$ have the effect of rounding of sharp edges.

There exists several more sophisticated and faster algorithms that need smaller number of iterations for convergence than those basic iterative methods. To mention some recent and fast methods, consider generalised proximal gradient method [55] and Split Bregman method [24]. It is also possible to transform the one-dimensional and also the two-dimensional anisotropic TV problem into a linearly constrained quadratic

programming optimisation task [49, pp. 44–45]. There exists several algorithms to tackle this kind of optimisation problems.

Here TV regularisation was considered as a deterministic optimisation problem. In the next section a Bayesian hierarchical model for total variation regularisation is derived.

# Chapter 5

# Bayesian hierarchical TV regularisation

In this section it is shown how to model TV in Bayesian setting. The model presented here is slightly different than the one for the Lasso in papers [43, 33]. We will also consider more general case with GIG mixing density. The case with total variation prior, that is, Laplace prior is then obtained as a special case. In addition different methodology to infer the conditional mean and MAP estimates are considered. This chapter is only about one-dimensional TV regularisation and extensions to two-dimensional cases are considered in the next chapter.

## 5.1 The linear model

We will start by considering linear Gaussian observation model as before but with more general error covariance matrix. Setting $\Sigma = \nu \mathrm{I}$ will give the white noise error.

$$\mathbf{y} \,|\, (\mathbf{x} = x, \mathbf{\Sigma} = \Sigma) \sim \mathrm{Normal}(Hx, \Sigma^{-1}). \tag{5.1}$$

Here $y$ is an $n$-vector of observations (data), $\mathbf{x}$ is a $k$-vector of unknown model coefficients, $k \times k$ matrix $\mathbf{\Sigma}$ is a precision parameter, and $H$ is a given $n \times k$ matrix with $n \geq k = \mathrm{rank}(H)$. This model can be as well written as

$$\mathbf{y} = H\mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \,|\, (\mathbf{\Sigma} = \Sigma) \sim \mathrm{Normal}(0, \Sigma^{-1}). \tag{5.2}$$

The one-dimensional discrete TV prior on the coefficients $\mathbf{x}$ with no boundary conditions applied is

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x \,|\, \lambda) \propto \lambda^{\frac{k-1}{2}} \mathrm{e}^{-\sqrt{\lambda} \sum_{i=1}^{k-1} |x_{i+1} - x_i|}. \tag{5.3}$$

As argued in Section 4.3 this prior penalises oscillations while allowing occasional jumps. The hyperparameter $\boldsymbol{\lambda}$ controls the overall "strength" of the penalisation.

The priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$ are

$$\boldsymbol{\lambda} \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda), \qquad \boldsymbol{\Sigma} \sim \text{Wishart}(m_\nu, V_\nu), \tag{5.4}$$

where the parameters $\alpha_\lambda, \beta_\lambda, m_\nu$ are positive constants and $V_\nu$ is spd matrix. We will derive the results using these but later improper priors that do not require setting additional tuning parameters will be used. In the white noise case setting $m_\nu = 2\alpha_\nu, V_\nu = 1/(2\beta_\nu)$ gives a gamma prior, that is

$$\boldsymbol{\nu} \sim \text{Gamma}(\alpha_\nu, \beta_\nu), \tag{5.5}$$

so that white noise case can be considered. Now with the likelihood (5.1) and the priors (5.4) the posterior density is, by Bayes' law (when actually considering the white noise error case (5.5) for ease of demonstration),

$$\begin{aligned}
&p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda}\,|\,\mathbf{y}}(x,\nu,\lambda\,|\,y) \\
&\propto\; p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda}}(x,\nu,\lambda)\, p_{\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda}}(y\,|\,x,\nu,\lambda) \\
&=\; p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x\,|\,\lambda)\, p_{\boldsymbol{\lambda}}(\lambda)\, p_{\boldsymbol{\nu}}(\nu)\, p_{\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\nu}}(y\,|\,x,\nu) \\
&\propto\; \lambda^{\frac{k-3}{2}+\alpha_\lambda} \nu^{\frac{n}{2}+\alpha_\nu-1} \mathrm{e}^{-\left(\frac{\nu}{2}\|y-Hx\|_2^2+\sqrt{\lambda}\sum_{i=1}^{k-1}|x_{i+1}-x_i|+\beta_\lambda\lambda+\beta_\nu\nu\right)}.
\end{aligned} \tag{5.6}$$

The priors were chosen to be gamma distributions due to conjugacy. However, it is possible to set $\alpha_\lambda = 0$ and $\beta_\lambda = 0$. This "vague" prior

$$p_{\boldsymbol{\lambda}}(\lambda) \propto \lambda^{-1} \tag{5.7}$$

can be used if parameters $\alpha_\lambda$ and $\beta_\lambda$ are not to be tuned. Similar vague prior can be chosen for $\boldsymbol{\nu}$ as well. Note that these priors are improper as they do not integrate to 1. This fact might imply that the posterior is also improper which makes the estimation somewhat open for criticism. Anyway, we will not care about this at this stage as further analysis would require very technical computations. In practise one can also use some small but nonzero values for the parameters of these gamma densities.

If the noise and penalisation parameters $\nu, \lambda$ are known, the computation of the MAP estimate can be obtained by computing the minimum of

$$\frac{1}{2}\|y - Hx\|_2^2 + \frac{\sqrt{\lambda}}{\nu}\sum_{i=1}^{k-1}|x_{i+1} - x_i|. \tag{5.8}$$

This computation is not trivial due to absolute values, see for example [52]. The selection of the "regularisation parameter" $\frac{\sqrt{\lambda}}{\nu}$ is also a challenge. However, as shall be shown in the next section, the introduction of more hyperparameters makes the whole problem much easier to solve!

## 5.2 Bayesian total variation regularisation

### 5.2.1 Hierarchical Model for TV prior

In the discrete TV prior (5.3), the model coefficients' differences are conditionally independent Laplace random variables, that is,

$$\mathbf{x}_{i+1} - \mathbf{x}_i \,|\, (\boldsymbol{\lambda} = \lambda) \overset{\text{iid}}{\sim} \text{Laplace}(0, \sqrt{\lambda}). \tag{5.9}$$

As mentioned, the Laplace distribution is a scale mixture of Gaussians (Theorem 2.10). So, in place of (5.3), one can use the prior

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda},\mathbf{r}}(x \,|\, \lambda, r) \propto \left(\frac{\lambda^{k-1}}{r_1 r_2 \cdots r_{k-1}}\right)^{1/2} \mathrm{e}^{-\frac{\lambda}{2}\sum_{i=1}^{k-1}\frac{(x_{i+1}-x_i)^2}{2r_i}}. \tag{5.10}$$

The difference with the previous model (5.3) is the addition of hyperparameters $\mathbf{r}_1, \ldots, \mathbf{r}_{k-1}$. These additional hyperparameters are a-priori independent of $\boldsymbol{\lambda}$ and have as prior distribution

$$\mathbf{r}_i \overset{\text{iid}}{\sim} \text{Exp}(1). \tag{5.11}$$

By marginalising the new hyperparameters out of (5.10) gives the TV prior (5.3). This hierarchical model is presented in Figure 5.1.

For notational convenience in the following, (5.10) is rewritten in the form

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda},\mathbf{r}}(x \,|\, \lambda, r) \propto \left(\frac{\lambda^{k-1}}{r_1 r_2 \cdots r_{k-1}}\right)^{1/2} \mathrm{e}^{-\frac{\lambda}{2}\|R^{-1}Dx\|_2^2}, \tag{5.12}$$

where

$$D = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \tag{5.13}$$

is the $(k-1) \times k$ discrete difference operator and $R = \text{diag}(\sqrt{2r}) \in \mathbb{R}^{(k-1)\times(k-1)}$ with the convention used also later in this paper that the square root is taken component wise. One can see that these satisfy

$$\|R^{-1}Dx\|_2^2 = \sum_{i=1}^{k-1} \frac{(x_{i+1} - x_i)^2}{2r_i}. \tag{5.14}$$

The augmented hierarchical model can be summarised by the directed acyclic graph (DAG) shown in Figure 5.1.

Note that in (5.3) and (5.9) replacing the differences $x_{i+1} - x_i$ with $x_i$ and including all the $k$ components of $x$ (and consequently adding extra hyperparameter $\mathbf{r}_k$) gives similar hierarchical model for the Lasso. It is thus noted already at this stage that all

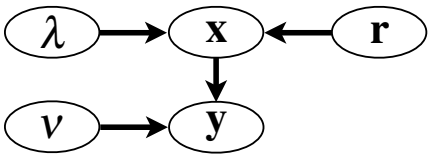


Figure 5.1: Graphical model of TV regularisation.

the computations and results to follow in one-dimensional case hold for the Lasso, one just have to replace all the $k-1$ differences $x_{i+1}-x_i$ with the $k$ components $x_i$ and set $D = \mathrm{I} \in \mathbb{R}^{k\times k}$.

We will use zero boundary conditions in this TV model. One can also consider periodic boundary conditions by including additional penalty term $x_k - x_1$ into (5.3) and adding a corresponding extra row to matrix $D$. Another option would be to use zero boundary condition by adding additional term – say – component $x_1$ to matrix $D$ which would be then of the form

$$D = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}. \tag{5.15}$$

In addition, then a new hyperparameter $\mathbf{r}_k$ must be introduced for the hierarchical model. Again, all the computations and results to follow will remain essentially the same and are not considered in detail in this work. We will though use periodic boundary conditions in 2d case.

**Remark 5.1.** Although we are interested in TV and related Laplace prior we will use the general $\mathrm{GIG}(a,b,p)$ mixing density in the derivations to follow. One can always revert back to the Laplace GSM prior by simply using $\mathrm{GIG}(2,0,1) = \mathrm{Exp}(1)$. This way we will obtain many more interesting priors than just this one!

The full posterior for the hierarchical TV model is, by using Bayes' law

$$\begin{aligned}
&p_{\mathbf{x},\boldsymbol{\Sigma},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(x,\Sigma,\lambda,r\,|\,y) \\
&\propto p_{\mathbf{x},\boldsymbol{\Sigma},\boldsymbol{\lambda},\mathbf{r}}(x,\Sigma,\lambda,r)\, p_{\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\Sigma},\boldsymbol{\lambda},\mathbf{r}}(y\,|\,x,\Sigma,\lambda,r) \\
&= p_{\mathbf{x}\,|\,\boldsymbol{\lambda},\mathbf{r}}(x\,|\,\lambda,r)\, p_{\boldsymbol{\lambda}}(\lambda)\, p_{\mathbf{r}}(r)\, p_{\boldsymbol{\Sigma}}(\Sigma)\, p_{\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\Sigma}}(y\,|\,x,\Sigma) \\
&\propto \lambda^{\frac{k-1}{2}+\alpha_\lambda-1}|\Sigma|^{\frac{m_\nu-n}{2}} \prod_{i=1}^{k-1} r_i^{p-\frac{3}{2}}\, \mathrm{e}^{-\frac{1}{2}\|y-Hx\|_\Sigma^2-\frac{\lambda}{2}\|R^{-1}Dx\|_2^2} \\
&\quad \cdot \mathrm{e}^{\frac{1}{2}a\sum_{i=1}^{k-1} r_i-\frac{1}{2}b\sum_{i=1}^{k-1} r_i^{-1}-\beta_\lambda\lambda-\frac{1}{2}\mathrm{tr}(V_\nu^{-1}\Sigma)} \qquad (5.16) \\
&= \lambda^{\frac{k-1}{2}+\alpha_\lambda-1}|\Sigma|^{\frac{m_\nu-n}{2}} \prod_{i=1}^{k-1} r_i^{p-\frac{3}{2}}\, \mathrm{e}^{-\frac{1}{2}(x-\hat{x})^T Q(x-\hat{x})-\frac{1}{2}(y^T\Sigma y-\hat{x}^T Q\hat{x})} \\
&\quad \cdot \mathrm{e}^{-\frac{1}{2}a\sum_{i=1}^{k-1} r_i-\frac{1}{2}b\sum_{i=1}^{k-1} r_i^{-1}-\beta_\lambda\lambda-\frac{1}{2}\mathrm{tr}(V_\nu^{-1}\Sigma)}, \qquad (5.17)
\end{aligned}$$

where $\hat{x} = Q^{-1}H^T\Sigma y$ and $Q = H^T\Sigma H + \lambda D^T R^{-2} D$. The equations (5.16) and (5.17) are equivalent since $Q = Q^T$ and $Q$ is positive definite (and thus invertible) and that is why

$$\begin{aligned}
&\|y - Hx\|_\Sigma^2 + \lambda \|R^{-1}Dx\|_2^2 \\
&= (Hx)^T\Sigma Hx - (Hx)^T\Sigma y - y^T\Sigma Hx + y^T\Sigma y + \lambda x^T D^T R^{-2} Dx \\
&= x^T(H^T\Sigma H + \lambda D^T R^{-2} D)x - x^T H^T\Sigma y - y^T\Sigma Hx + y^T\Sigma y \\
&= x^T Qx - x^T Q Q^{-1} H^T\Sigma y - y^T\Sigma H Q^{-1} Qx + y^T\Sigma y \\
&= x^T Qx - x^T Q\hat{x} - \hat{x}^T Qx + \hat{x}^T Q\hat{x} + y^T\Sigma y - \hat{x}^T Q\hat{x} \\
&= (x - \hat{x})^T Q(x - \hat{x}) + y^T\Sigma y - \hat{x}^T Q\hat{x}. \qquad (5.18)
\end{aligned}$$

This connection could have been seen by using Theorem 4.3 as well.

In general, there are several ways to deal with a posterior as above. For example one could try to marginalise hyperparameters $\Sigma$ and $\lambda$ or use the empirical Bayes approach. However, because the hierarchical model has conjugate priors, statistical inference is easily accomplished using a Gibbs sampler or variational Bayes algorithm. Also EM based solution or direct minimisation algorithms can be employed to solve just the MAP estimate.

We will next limit to the white noise case, that is, we set $\Sigma = \nu\mathrm{I}$ and $m_\nu = 2\alpha_\nu, V_\nu = 1/(2\beta_\nu)$. Now that the posterior (5.16) is derived the generalisation is easy to make. One basically just needs to replace the corresponding gamma densities with Wishart densities and carefully check the other differences to occur. Notice that now we can write

$$\hat{x} = Q^{-1}H^T y, \qquad (5.19)$$
$$Q = H^T H + \tfrac{\lambda}{\nu} D^T R^{-2} D \qquad (5.20)$$

and the posterior reverts to

$$\begin{aligned}
&p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(x,\nu,\lambda,r\,|\,y) \\
&\propto\; p_{\mathbf{x}\,|\,\boldsymbol{\lambda},\mathbf{r}}(x\,|\,\lambda,r)\, p_{\boldsymbol{\lambda}}(\lambda)\, p_{\mathbf{r}}(r)\, p_{\boldsymbol{\nu}}(\nu)\, p_{\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\nu}}(y\,|\,x,\nu) \\
&\propto\; \lambda^{\frac{k-1}{2}+\alpha_\lambda-1} \nu^{\frac{n}{2}+\alpha_\nu-1} \prod_{i=1}^{k-1} r_i^{p-\frac{3}{2}}\, \mathrm{e}^{-\frac{\nu}{2}\|y-Hx\|_2^2 - \frac{\lambda}{2}\|R^{-1}Dx\|_2^2} \\
&\quad \cdot\, \mathrm{e}^{-\frac{1}{2}a\sum_{i=1}^{k-1} r_i - \frac{1}{2}b\sum_{i=1}^{k-1} r_i^{-1} - \beta_\lambda\lambda - \beta_\nu\nu} \qquad (5.21) \\
&=\; \lambda^{\frac{k-1}{2}+\alpha_\lambda-1} \nu^{\frac{n}{2}+\alpha_\nu-1} \prod_{i=1}^{k-1} r_i^{p-\frac{3}{2}}\, \mathrm{e}^{-\frac{\nu}{2}(x-\hat{x})^T Q(x-\hat{x}) - \frac{1}{2}(y^T\Sigma y - \hat{x}^T Q\hat{x})} \\
&\quad \cdot\, \mathrm{e}^{-\frac{1}{2}a\sum_{i=1}^{k-1} r_i - \frac{1}{2}b\sum_{i=1}^{k-1} r_i^{-1} - \beta_\lambda\lambda - \beta_\nu\nu}, \qquad (5.22)
\end{aligned}$$

which has slightly simpler form.

### 5.2.2 Gibbs Sampler

The Gibbs Sampler update distributions are evident by inspection of the full posterior (5.21) and (5.22). Simply looking at the form of the conditionals the following densities (all of which were discussed in Section 2) will emerge.

$$\mathbf{x} \,|\, \nu, \lambda, r, y \;\sim\; \text{Normal}(Q^{-1}H^T y, (\nu Q)^{-1}), \tag{5.23a}$$

$$\boldsymbol{\nu} \,|\, x, \lambda, r, y \;\sim\; \text{Gamma}(\tfrac{n}{2} + \alpha_\nu,\, \tfrac{1}{2}\|y - Hx\|_2^2 + \beta_\nu), \tag{5.23b}$$

$$\boldsymbol{\lambda} \,|\, x, \nu, r, y \;\sim\; \text{Gamma}(\tfrac{k-1}{2} + \alpha_\lambda,\, \tfrac{1}{2}\|R^{-1}Dx\|_2^2 + \beta_\lambda), \tag{5.23c}$$

$$\mathbf{r}_i \,|\, x, \nu, \lambda, r_{-[i]}, y \;\sim\; \text{GIG}\left(a, \tfrac{1}{2}\lambda(x_{i+1} - x_i)^2 + b, p - \tfrac{1}{2}\right), \quad i = 1, \ldots, k-1. \tag{5.23d}$$

The notation $r_{-[i]}$ means all the other components $r_j$ except the $i$th. In the case of Laplace density (that is $a = 2, b = 0, p = 1$) we see that

$$\mathbf{r}_i \,|\, x, \nu, \lambda, r_{-[i]}, y \;\sim\; \text{RIG}\left(\sqrt{\lambda}|x_{i+1} - x_i|, \tfrac{1}{2}\lambda(x_{i+1} - x_i)^2\right), \quad i = 1, \ldots, k-1. \tag{5.24}$$

This leads to the following simple sampling algorithm which is given below as Algorithm 2.

**Algorithm 2:** Gibbs sampler for TV regularisation in 1d.

Given $H$ and $y$:
assign starting values for $\lambda^{(0)}$, $\nu^{(0)}$, $r^{(0)}$
**for** *t from 1 to* $n_t$ **do**
  $R \leftarrow \text{diag}(\sqrt{2r^{(t-1)}})$
  $Q \leftarrow H^T H + \frac{\lambda^{(t-1)}}{\nu^{(t-1)}} D^T R^{-2} D$
  $x^{(t)} \leftarrow$ sample from $\text{Normal}(Q^{-1}H^T y, (\nu^{(t-1)}Q)^{-1})$
  $\nu^{(t)} \leftarrow$ sample from $\text{Gamma}(\frac{n}{2} + \alpha_\nu, \frac{1}{2}\|y - Hx^{(t)}\|_2^2 + \beta_\nu)$
  $\lambda^{(t)} \leftarrow$ sample from $\text{Gamma}(\frac{k-1}{2} + \alpha_\lambda, \frac{1}{2}\|R^{-1}Dx^{(t)}\|_2^2 + \beta_\lambda)$
  **for** *i from 1 to* $k-1$ **do**
    $r_i^{(t)} \leftarrow$ sample from $\text{RIG}(\sqrt{\lambda^{(t)}}|x_{i+1}^{(t)} - x_i^{(t)}|, \frac{1}{2}\lambda^{(t)}(x_{i+1}^{(t)} - x_i^{(t)})^2)$
  **end**
**end**

Sampling from multivariate normal or gamma distribution is common and easy task, see for example [17]. However, RIG is somewhat exotic density that does not arise in many applications. In order to sample from $\text{RIG}(\alpha, \beta)$ one can first draw a sample from $\text{IG}(\alpha/\beta, \alpha^2/\beta)$ and then compute the reciprocal of this value. There is a sampling method for inverse Gaussian distribution $\text{IG}(\mu, \lambda)$ that is based on the many-to-one transformation. This algorithm is presented in Algorithm 3 and is justified in [17, pp. 145–149].

Generating samples from GIG density can be done by using an algorithm as described in [16]. This algorithm is based on the reparametrisation of GIG into a density with two parameters and using the ratio sampling method that is presented in [15, p. 60]. However, the presented Gibbs sampler for a large image is slow since one needs to

**Algorithm 3:** Sampling from IG($\mu, \lambda$) density.

```
n ← sample from standard normal distribution
y ← n^2
x ← μ + μ^2 y/(2λ) − (μ/(2λ)) sqrt(4μλy + μ^2 y^2)
z ← sample from uniform distribution on [0, 1]
if z ≤ μ/(μ + x) then
    return x
else
    return μ^2/x
end
```

sample from GIG or RIG density numerous times and also at each step invert the matrix $Q$. Next we will focus on non-sampling techniques.

### 5.2.3 Coordinate Descent Method

The MAP estimate for the posterior derived earlier can be found for example by using coordinate descent method. The procedure is to maximize the posterior respect to each variable at a time keeping the other variables fixed. One can loop through the variables this way using the newest values of parameters at each step. The algorithm converges to a local optimum with some assumptions and the method is derivative-free. [37, Ch. 8.9]. A coordinate descent based algorithm for the Lasso can be found in [40, p. 441]. The same idea is also called IAS (iterative alternating sequential) in some inverse problems literature and we will mainly use this abbreviation to refer this technique for finding the MAP estimate from now on. In this TV model cycling through variables (or "coordinates") is done via the following equations.

$$x = (H^T H + \tfrac{\lambda}{\nu} D^T R^{-2} D)^{-1} H^T y, \tag{5.25a}$$

$$\nu = \frac{n - 2 + 2\alpha_\nu}{\|y - Hx\|_2^2 + 2\beta_\nu}, \tag{5.25b}$$

$$\lambda = \frac{k - 3 + 2\alpha_\lambda}{\|R^{-1} D x\|_2^2 + 2\beta_\lambda}, \tag{5.25c}$$

$$r_i = \frac{p - \frac{3}{2} + \sqrt{\left(p - \frac{3}{2}\right)^2 + a\left(\frac{1}{2}\lambda(x_{i+1} - x_i)^2 + b\right)}}{a}, \quad i = 1, \ldots, k-1, \tag{5.25d}$$

where $R = \operatorname{diag}(\sqrt{2r})$ with the convention that the square root is taken component wise. These formulas actually follow from the unique modes of the conditional densities. In the Laplace case the last equation simplifies to

$$r_i = -\tfrac{1}{4} + \tfrac{1}{4}\sqrt{1 + 4\lambda(x_{i+1} - x_i)^2}, \quad i = 1, \ldots, k-1. \tag{5.26}$$

The result is a kind of reweighted least-squares algorithm. The first formula can be seen as Tikhonov regularisation formula with $L = R^{-1}D$ and the regularisation parameter $\frac{\lambda}{\nu}$. Considering the formula as the corresponding minimisation problem it

is seen that each difference $x_{i+1}-x_i$ is penalised with different weight that is computed through other three equations.

The problem of this formula is, however, if $x_i$ and $x_{i-1}$ become almost the same for some index $i$, then the corresponding latent variable $r_i$ will be small. This makes solving the linear system for $x$ computationally problematic since some weights become very large. This issue and how to avoid it are discussed later in more detail.

Since the gradient and the Hessian are feasible to be computed also other kind of optimisation algorithms could be employed. Anyway, coordinate descent method is simple to code and it also converged reasonable fast in experiments. It can be also compared to other actually quite similar and related methods like variational Bayes and EM which are presented and discussed next.

### 5.2.4 Variational Bayes

Variational Bayes method can be used to approximate the full posterior derived earlier in the following way

$$p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,y}(x,\nu,\lambda,r\,|\,y) \approx q_{\mathbf{x}}(x)q_{\boldsymbol{\nu}}(\nu)q_{\boldsymbol{\lambda}}(\lambda)q_{\mathbf{r}}(r). \tag{5.27}$$

We will denote the expectation, for example, with respect to random vectors $\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}$ as $\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}}$ when all the other variables are kept fixed. A bar over random vector denotes its mean and $c_i$'s denote values that are constants with respect to current variables. These constants are not necessary to be computed. Notice also that matrix $R$ depends on $\mathbf{r}$. With these conventions it can be calculated that

$$\begin{aligned}
\ln q^*_{\mathbf{x}}(x) &= \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}}\left[\ln p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(x,\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,y)\right] + c_1 \\
&= \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}}\Big[\left(\tfrac{k-1}{2}+\alpha_\lambda-1\right)\ln\boldsymbol{\lambda} + \left(\tfrac{n}{2}+\alpha_\nu-1\right)\ln\boldsymbol{\nu} + \left(p-\tfrac{3}{2}\right)\sum_{i=1}^{k-1}\ln\mathbf{r}_i \\
&\quad -\tfrac{\boldsymbol{\nu}}{2}\|y-Hx\|_2^2 - \tfrac{\boldsymbol{\lambda}}{2}\|R^{-1}Dx\|_2^2 - \tfrac{1}{2}a\sum_{i=1}^{k-1}\mathbf{r}_i - \tfrac{1}{2}b\sum_{i=1}^{k-1}\mathbf{r}_i^{-1} - \beta_\lambda\boldsymbol{\lambda} - \beta_\nu\boldsymbol{\nu}\Big] + c_1 \\
&= -\tfrac{1}{2}\bar{\nu}\|y-Hx\|_2^2 - \tfrac{1}{2}\bar{\lambda}\mathbb{E}_{\mathbf{r}}(\|R^{-1}Dx\|_2^2) + c_2 \\
&= -\tfrac{1}{2}(\bar{\nu}\|y-Hx\|_2^2 + \bar{\lambda}\|\bar{R}^{-1}Dx\|_2^2) + c_2 \\
&\overset{(5.18)}{=} -\tfrac{1}{2}(\bar{\nu}(x-\hat{x})^T\bar{Q}(x-\hat{x}) + \bar{\nu}(y^Ty - \hat{x}^T\bar{Q}\hat{x})) + c_2 \\
&= -\tfrac{\bar{\nu}}{2}(x-\hat{x})^T\bar{Q}(x-\hat{x}) + c_3,
\end{aligned}$$

where we have denoted $\hat{x} = \bar{Q}^{-1}H^Ty$, $\bar{Q} = H^TH + \frac{\bar{\lambda}}{\bar{\nu}}D^T\bar{R}^{-2}D$ and $\bar{R} = \operatorname{diag}\left(\sqrt{2/\mathbb{E}(\mathbf{r}^{-1})}\right)$ and thus $\bar{R}^{-2} = \frac{1}{2}\operatorname{diag}\left(\mathbb{E}(\mathbf{r}^{-1})\right)$ with the convention that the operations on $\mathbf{r}$ are to be done component wise. The fourth equality can be seen to be true by using linearity of expectation and result (5.14). So we obtain

$$\mathbf{x} \sim \text{Normal}(\hat{x}, (\bar{\nu}\bar{Q})^{-1}). \tag{5.28}$$

Next the distribution for $\boldsymbol{\nu}$ is derived. We can compute that

$$\begin{aligned}\ln q^*_{\boldsymbol{\nu}}(\nu) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\lambda},\mathbf{r}}\left[\ln p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(\mathbf{x},\nu,\boldsymbol{\lambda},\mathbf{r}\,|\,y)\right] + c_1 \\ &= \left(\frac{n}{2}+\alpha_\nu - 1\right)\ln\nu - \frac{\nu}{2}\mathbb{E}_{\mathbf{x}}(\|y-H\mathbf{x}\|_2^2) - \beta_\nu\nu + c_2.\end{aligned}$$

Using some properties of expectation and trace one can see that

$$\begin{aligned}\mathbb{E}(\|y-H\mathbf{x}\|_2^2) &= \mathbb{E}(y^Ty - y^TH\mathbf{x} - \mathbf{x}^TH^Ty + \mathbf{x}^TH^TH\mathbf{x}) \\ &= \mathbb{E}(y^Ty) - \mathbb{E}(y^TH\mathbf{x}) - \mathbb{E}(\mathbf{x}^TH^Ty) + \mathbb{E}(\mathrm{tr}(\mathbf{x}^TH^TH\mathbf{x})) \\ &= y^Ty - y^TH\bar{x} - \bar{x}^TH^Ty + \mathbb{E}(\mathrm{tr}(H\mathbf{x}\mathbf{x}^TH^T)) \\ &= y^Ty - y^TH\bar{x} - \bar{x}^TH^Ty + \mathrm{tr}(H\mathbb{E}(\mathbf{x}\mathbf{x}^T)H^T) \\ &= y^Ty - y^TH\bar{x} - \bar{x}^TH^Ty + \mathrm{tr}(H\bar{x}\bar{x}^TH^T) + \mathrm{tr}(H\mathbb{V}(\mathbf{x})H^T) \\ &= y^Ty - y^TH\bar{x} - \bar{x}^TH^Ty + \bar{x}^TH^TH\bar{x} + \mathrm{tr}(\mathbb{V}(\mathbf{x})H^TH) \\ &= \|y-H\bar{x}\|_2^2 + \mathrm{tr}(\mathbb{V}(\mathbf{x})H^TH).\end{aligned}$$

Thus the distribution for $\boldsymbol{\nu}$ is

$$\boldsymbol{\nu} \sim \mathrm{Gamma}\left(\tfrac{n}{2}+\alpha_\nu, \tfrac{1}{2}\|y-H\bar{x}\|_2^2 + \tfrac{1}{2}\mathrm{tr}(\mathbb{V}(\mathbf{x})H^TH) + \beta_\nu\right). \tag{5.29}$$

Derivation of the distribution for $\boldsymbol{\lambda}$ is quite similar as it was for the precision $\boldsymbol{\nu}$. Since it holds that

$$\mathbb{E}_{\mathbf{x},\mathbf{r}}(\|R^{-1}D\mathbf{x}\|_2^2) = \frac{1}{2}\sum_{i=1}^{k-1}\mathbb{E}(\mathbf{r}_i^{-1})\mathbb{E}((\mathbf{x}_{i+1}-\mathbf{x}_i)^2), \tag{5.30}$$

which is easily verified using (5.14), the result is

$$\boldsymbol{\lambda} \sim \mathrm{Gamma}\left(\frac{k-1}{2}+\alpha_\lambda, \frac{1}{4}\sum_{i=1}^{k-1}\mathbb{E}(\mathbf{r}_i^{-1})\mathbb{E}((\mathbf{x}_{i+1}-\mathbf{x}_i)^2) + \beta_\lambda\right). \tag{5.31}$$

Writing out the square term and using the linearity of expectation shows that

$$\mathbb{E}((\mathbf{x}_{i+1}-\mathbf{x}_i)^2) = \mathbb{E}(\mathbf{x}_{i+1}^2) + \mathbb{E}(\mathbf{x}_i^2) - 2\mathbb{E}(\mathbf{x}_{i+1}\mathbf{x}_i). \tag{5.32}$$

Given the mean and covariance matrix of $\mathbf{x}$, this statistic can be computed easily summing corresponding components of $\mathbb{E}(\mathbf{x}\mathbf{x}^T) = \mathbb{V}(\mathbf{x}) + \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^T$. The means for $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ can now be computed easily given the mean and variance of $\mathbf{x}$ since the densities are gamma. For $\boldsymbol{\lambda}$ and for $\mathbf{x}$ one also needs the moment $\mathbb{E}(\mathbf{r}_i^{-1})$.

The components of $\mathbf{r}$ do not depend on each other and it is seen that $q_{\mathbf{r}}(r) = \prod_{i=1}^{k-1} q_{\mathbf{r}_i}(r_i)$. For each of the components $\mathbf{r}_i$ we obtain GIG density. The derivation goes as follows.

$$\begin{aligned}\ln q^*_{\mathbf{r}_i}(r_i) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}_{-[i]}}[\ln p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,y)] + c_1 \\ &= \left(p-\frac{3}{2}\right)\ln r_i - \frac{\bar{\lambda}}{2}\mathbb{E}_{\mathbf{x},\mathbf{r}_{-[i]}}\left(\sum_{i=1}^{k-1}\frac{(\mathbf{x}_{i+1}-\mathbf{x}_i)^2}{2r_i}\right) - \frac{1}{2}ar_i - \frac{1}{2}br_i^{-1} + c_2\end{aligned}$$

$$= \left(p - \frac{3}{2}\right) \ln r_i - \frac{1}{2r_i}\left(\frac{\bar{\lambda}}{2}\mathbb{E}_{\mathbf{x}}((\mathbf{x}_{i+1} - \mathbf{x}_i)^2) + b\right) - \frac{1}{2}ar_i + c_3$$

From above it is seen that

$$\mathbf{r}_i \sim \text{GIG}\left(a, \frac{\bar{\lambda}}{2}\mathbb{E}((\mathbf{x}_{i+1} - \mathbf{x}_i)^2) + b, p - \frac{1}{2}\right), \quad i = 1, ..., k-1. \tag{5.33}$$

In the special case where the prior is initially Laplace, GIG reverts to RIG as in the case of Gibbs sampler conditional densities and we find

$$\mathbf{r}_i \sim \text{RIG}\left(\sqrt{\bar{\lambda}\mathbb{E}((\mathbf{x}_{i+1} - \mathbf{x}_i)^2)}, \frac{1}{2}\bar{\lambda}\mathbb{E}((\mathbf{x}_{i+1} - \mathbf{x}_i)^2)\right), \quad i = 1, ..., k-1. \tag{5.34}$$

Given the parameters, the moment $\mathbb{E}(\mathbf{r}_i^{-1})$ in RIG case can be computed from the following formula

$$\mathbb{E}(\mathbf{r}_i^{-1}) = \frac{2}{\sqrt{\bar{\lambda}\mathbb{E}((\mathbf{x}_{i+1} - \mathbf{x}_i)^2)}}. \tag{5.35}$$

Similar formula for more general case (5.33) can be obtained from Proposition 2.5.

Starting with some initial values for unknown parameters in the above pdfs and updating them one at a time using estimates from previous updates, one will end up with optimal distributions for $\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\lambda}$ and $\mathbf{r}$. As a result one can extract the mean (which is the same as the mode in this normal distribution case) and evaluate the variance of $\mathbf{x}$. Furthermore, one can for example plot the densities of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ to analyse the inference result. The resulting algorithm is presented as Algorithm 4 below.

**Algorithm 4:** VB algorithm for TV regularisation in 1d.

set initial values of parameters of densities $q^0_{\mathbf{r}_j}(r_j), j = 1, \ldots, k-1$, $q^0_{\boldsymbol{\nu}}(\nu)$ and $q^0_{\boldsymbol{\lambda}}(\lambda)$
**for** $i = 1, 2, \ldots$ *until stopping criteria is met* **do**
  use the newest values of the parameters at each step:
  solve parameters for $q^i_{\mathbf{x}}(x)$ using (5.35) and (5.28)
  solve parameters for $q^i_{\boldsymbol{\nu}}(\nu)$ using (5.29)
  solve parameters for $q^i_{\boldsymbol{\lambda}}(\lambda)$ using (5.31)
  **for** $j$ *from 1 to* $k-1$ **do**
    solve parameters for $q^i_{\mathbf{r}_j}(r_j)$ using (5.34)
  **end**
**end**
return the parameters of $q^{\text{end}}_{\mathbf{x}}(x)$, $q^{\text{end}}_{\boldsymbol{\nu}}(\nu)$, $q^{\text{end}}_{\boldsymbol{\lambda}}(\lambda)$ and $q^{\text{end}}_{\mathbf{r}_j}(r_j), j = 1, \ldots, k-1$

## 5.3 Implementation details and discussion

As mentioned in Section 5.2.3 in Laplace case it is expected that some differences of neighbouring pixels tend to go zero causing infinite weights for corresponding components of $r$ in (5.26). This feature makes computing the inverse of $Q$ numerically

difficult as infinite values must be handled. Let $T = D^T R^{-2} D$. This singularity issue can be solved (to some extent) by using connection

$$Q^{-1} = (H^T H + \tfrac{\lambda}{\nu} T)^{-1} = (T^{-1} H^T H + \tfrac{\lambda}{\nu} \mathrm{I})^{-1} T^{-1}. \tag{5.36}$$

Alternatively one can use matrix inversion lemma as in Theorem 4.3 to get rid off the issue. For these to work it is required that $T = D^T R^{-2} D$ is invertible which is clearly true in the Lasso case since $D = \mathrm{I}$ and thus $T$ is diagonal. It also works in TV case but with certain boundary conditions. This in turn may lead to a problem called "zero-locking" since if some components become zero they will stay at zero. In practise those weights can be also artificially limited using some small positive threshold. What is more, the issue can be avoided using hyperprior that is almost as Exp(1) but will not make $r$ to go zero in equation (5.25d). For example one can use $\mathrm{GIG}(2, 0.001, 1)$. Of course then the prior is no more Laplace but some other heavy-tailed approximation to it that has no sharp peak at origin. This is also related to approximation $|x| \approx \sqrt{x^2 + \beta^2}$ which can be considered in one-dimensional optimisation problem case. In some applications related issue is handled by removing these zero-locked coefficients from the computations when they go close to zero. Notice that in VB or Gibbs sampler case this effect is not an issue and it seems to only arise in the case of MAP estimate.

Inverting the covariance matrix of $\mathbf{x}$ is required in VB iterative formulas. It can be done faster by assuming bccb (block circulant with circulant blocks) structure for covariance matrix and inverting it in Fourier domain (see [4]). Then it is not needed to save the whole matrix in memory. For the MAP estimate only a linear system has to be computed which is evidently faster and either can be done in Fourier domain or if the matrix does not have any nice structure to exploit, iterative techniques such as the conjugate gradient method with preconditioning can be used [52, Ch. 5]. Compared to Tikhonov regularisation formula, one needs to solve this linear system as many times as iteration steps are needed for convergence. These techniques become very important in the case of 2d images. For example in the case of $256 \times 256$ image one would need to construct and invert $256^2 \times 256^2$ matrix which is practically not possible without special methods.

The similarity between the Gibbs sampler conditional densities, IAS equations and the independent densities of VB method can not be ignored. For instance, in VB case the means and certain moments with respect to other independent densities appear where in corresponding conditional densities one uses samples drawn from the other corresponding conditional densities. The Expectation Maximisation (EM) method has a generalisation called Expectation Conditional Maximisation (ECM) [39] which could also have been used and would have produced quite similar formulas as the IAS or VB method. In EM based approaches usually the hyperparameters are considered as latent variables. In ECM the maximisation step is done in the style of IAS by maximising the log-likelihood with respect to each variable. Unlike in standard EM it is enough that the value of the objective function increases but the exact maximum is not needed to be solved. The method still converges. EM algorithm has been applied for the Lasso case in  cite Figueiredo2003. EM based methods, however, are not considered in this

work in more detail since we have derived an algorithm for the MAP using the IAS method.

Setting $\mathrm{GIG}(0, w, -\frac{w}{2}) = \mathrm{InvGamma}(\frac{w}{2}, \frac{w}{2})$ as the mixing density leads to hierarchical Student's t-distribution model. In this model one needs to set the degree of freedom $w$ for the prior. Since $a = 0$ we see that the resulting densities for hyperparameters are also inverse gamma densities with a mode that does not feature singularity issues. For small degree of freedom this prior is expected to produce sparser solutions than with larger values.

Next we extend these ideas to the two-dimensional case, where things turn out to be a little more complicated.

# Chapter 6

# Bayesian hierarchical TV regularisation in 2d

In the previous chapter the hierarchical Bayesian model was introduced and several ways to compute point estimates from the resulting posterior density were presented. In this chapter the same procedure is extended to two-dimensional case since one of the main applications of total variation is image problems that are naturally in two-dimensional setting.

There are several ways to generalise the concept of TV to two dimensional space, mainly the isotropic and anisotropic TV as we saw in Section 4.3. In this work the anisotropic TV is considered in more detail. Isotropic TV is used in papers [4] and [5] in Bayesian setting. In these papers an inequality trick is used to tackle the difficult formula of isotropic TV. However, in this work two more priors that can be seen either as approximations or alternatives (we could call them "TV like" priors) to TV are considered and presented. The derivations of the results goes mostly in the same way as in one-dimensional case and thus all the steps are not represented in detail but we just describe the main steps. Also due to brevity and since the generalisation should be rather evident, the general case with GIG mixing density is not presented. We also only consider Gaussian white noise error which usually has been done in literature too.

For the rest of this section a linear model as in Section 5.1 is considered. However, an image with size $k \times n$ pixels is assumed here and the matrix $H$ will be of size $kn \times kn$. Images $x$ and $y$ are both $k \times n$ matrices as well but notice that for the rest of this section we will consider them as column wise stacked vectors having the length $N = kn$ without any special notation. For notational simplicity indexing based on presentation of these stacked vectors as matrices is used. That is, the pixel $(i, j)$ of the image $x$ is denoted as $x_{i,j}$.

Periodic boundary conditions will be used in this chapter. That is,

$$x_{i,1} = x_{i,n+1}, \quad i = 1, \ldots, k \tag{6.1a}$$

$$x_{1,j} = x_{k+1,j}, \quad j = 1, \ldots, n. \tag{6.1b}$$

In addition, there are $2N$ differences of neighbouring pixels that are penalised when using these boundary conditions. Other kind of boundary conditions could be also considered but in this work in all two-dimensional cases periodic boundary conditions are used. This choice makes the formulas slightly less cluttered as boundaries do not need to be handled differently. Matrix $D \in \mathbb{R}^{2N \times N}$ consists of two $N \times N$ blocks corresponding the differences of components of $x$ in row wise and column wise directions. That is, $D = [D_v^T, D_h^T]^T$, where $D_v$ is a $N \times N$ matrix consisting of vertical differences and similarly $N \times N$ matrix $D_h$ contains the horizontal differences. This difference matrix is assumed throughout this section.

Contours for different priors that are studied closer in the following pages are plotted in Figure 6.1. We can see that all the densities have more "probability mass" near the origin than Gaussian density in (a). Thus they endeavour more sparse solutions than Gaussian. The anisotropic TV (which is a product of two one-dimensional Laplace densities) and well as the product of two t-distributions also tend to produce solution for which also each component is often close to zero while the rest are rotationally invariant. They tend to yield sparse solutions for the sum of squared components while anisotropic TV tends to produce solutions that are close to coordinate axes. Heavy tails imply that single large values are more common than in Gaussian case.

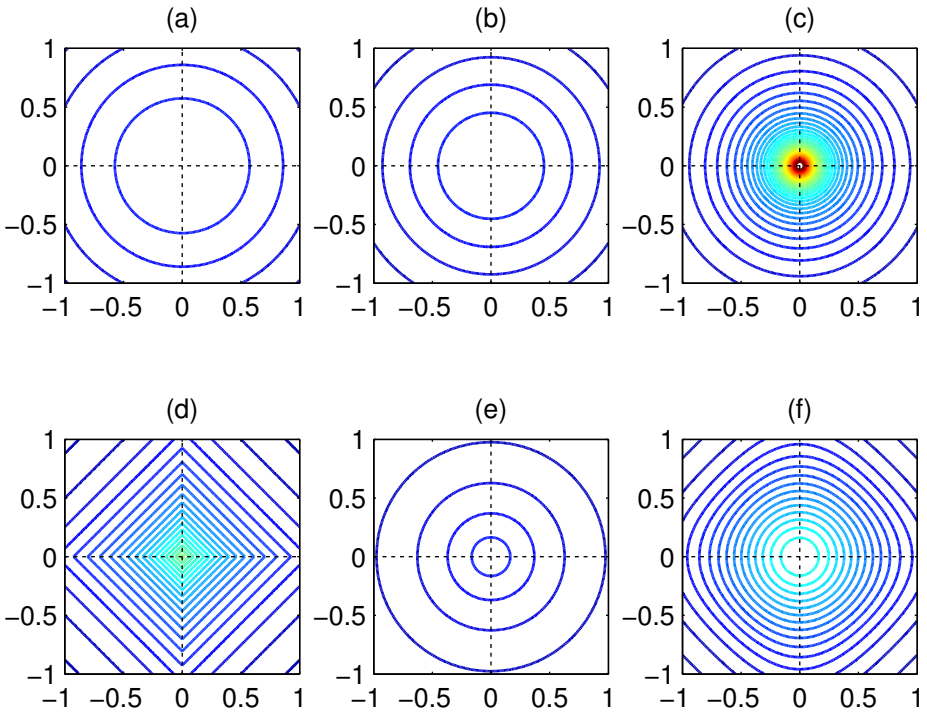


Figure 6.1: Priors with zero mean and covariance I $\in \mathbb{R}^2$ (except that isotropic TV is not scaled properly). (a) Standard normal, (b) two-dimensional t-distribution with degree of freedom 3, (c) Laplace density, (d) Anisotropic TV, (e) Isotropic TV, (f) product of two t-distributions both having degree of freedom 3.

## 6.1 Anisotropic TV prior

The anisotropic TV prior is studied first. Clearly each term in the penalty

$$\mathrm{TV}_{aniso}(x) = \sum_{i=1}^{k}\sum_{j=1}^{n}\left\{\left|\nabla_{ij}^{v}x\right| + \left|\nabla_{ij}^{h}x\right|\right\} \tag{6.2}$$

is related to a one-dimensional Laplace density used in the one-dimensional case. So this approach leads to a similar Bayesian hierarchical model as the one-dimensional TV case, mostly only the dimensions are changed. To be more precise, one needs to set hyperparameters $\mathbf{r}_{i,j,l}$ for $i = 1, \ldots, k$, $j = 1, \ldots, n$, $l = 1, 2$, where $i$ and $j$ refer to a pixels and $l$ either to vertical or horizontal difference of adjacent pixels. Consequently, diagonal matrix $R$ consisting of these weights will be of size $2N \times 2N$.

The prior that corresponds (6.2) is

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x\,|\,\lambda) \propto \lambda^{N}\mathrm{e}^{-\sqrt{\lambda}\sum_{i=1}^{k}\sum_{j=1}^{n}\left\{\left|\nabla_{ij}^{v}x\right|+\left|\nabla_{ij}^{h}x\right|\right\}}. \tag{6.3}$$

Similarly as in one-dimensional case, we will employ $\mathrm{Gamma}(\alpha_\lambda, \beta_\lambda)$ and $\mathrm{Gamma}(\alpha_\nu, \beta_\nu)$ priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$, respectively, and the likelihood (5.1) but with white noise. The GSM property for each term in the double sum is exploited (see equations (5.9) and (5.10)). The resulting posterior distribution is consequently

$$\begin{aligned}
&p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(x,\nu,\lambda,r\,|\,y)\\
&\propto \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i,j,l} r_{i,j,l}^{-1/2}\,\mathrm{e}^{-\frac{\nu}{2}\|y-Hx\|_2^2-\frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=1}^{n}\left\{\frac{(\nabla_{ij}^{v}x)^2}{2r_{i,j,1}}+\frac{(\nabla_{ij}^{h}x)^2}{2r_{i,j,2}}\right\}-\sum_{i,j,l} r_{i,j,l}-\beta_\lambda\lambda-\beta_\nu\nu}\\
&= \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i,j,l} r_{i,j,l}^{-1/2}\,\mathrm{e}^{-\frac{\nu}{2}\|y-Hx\|_2^2-\frac{\lambda}{2}\|R^{-1}Dx\|_2^2-\sum_{i,j,l} r_{i,j,l}-\beta_\lambda\lambda-\beta_\nu\nu}\\
&= \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i,j,l} r_{i,j,l}^{-1/2}\,\mathrm{e}^{-\frac{\nu}{2}(x-\hat{x})^TQ(x-\hat{x})-\frac{\nu}{2}(y^Ty-\hat{x}^TQ\hat{x})-\sum_{i,j,l} r_{i,j,l}-\beta_\lambda\lambda-\beta_\nu\nu}.
\end{aligned} \tag{6.4}$$

Matrix $Q$ and vector $\hat{x}$ are as in one-dimensional case and are given by formulas (5.19) and (5.20), though now $R$ is $2N \times 2N$ diagonal matrix and the diagonal elements are $\sqrt{2r_{i,j,l}}$.

Conditional densities, IAS and variational Bayes iteration formulas should be evident by comparing the above posterior to the one-dimensional case. The posterior is of the same form, only sums and products in which hyperparameters $r_{i,j,l}$ appear have more terms and the first parameter of gamma distribution for both $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ has changed. The conditional densities are presented below.

$$\mathbf{x}\,|\,\nu,\lambda,r,y \;\sim\; \mathrm{Normal}(Q^{-1}H^Ty, (\nu Q)^{-1}), \tag{6.5a}$$

$$\boldsymbol{\nu}\,|\,x,\lambda,r,y \;\sim\; \mathrm{Gamma}(\tfrac{N}{2}+\alpha_\nu,\ \tfrac{1}{2}\|y-Hx\|_2^2+\beta_\nu), \tag{6.5b}$$

$$\boldsymbol{\lambda}\,|\,x,\nu,r,y \;\sim\; \mathrm{Gamma}(N+\alpha_\lambda,\ \tfrac{1}{2}\|R^{-1}Dx\|_2^2+\beta_\lambda), \tag{6.5c}$$

$$\mathbf{r}_{i,j,l} \,|\, x, \nu, \lambda, r_{-[i,j,l]}, y \;\sim\; \mathrm{RIG}\left(\sqrt{\lambda}|d_{i,j,l}|, \tfrac{1}{2}\lambda d_{i,j,l}^2\right), \quad i = 1, \ldots, k,\ j = 1, \ldots, n,\ l = 1, 2, \tag{6.5d}$$

where $d_{i,j,1} = \nabla_{ij}^v x$ and $d_{i,j,2} = \nabla_{ij}^h x$. The coordinate descent iterative formulas follow immediately by taking the modes of the conditional densities. The GIG mixing density generalisation is also evident and is obtained by replacing the RIG with corresponding GIG density. From this generalisation one can obtain easily a prior in which all differences of $x$ are t-distributed with some degree of freedom $w$. VB formulas also follow immediately by replacing certain values in the conditional densities with expectations so we will not present them here.

## 6.2 Isotropic TV prior

A fully Bayesian model featuring the isotropic TV prior (6.6) has been analysed and used in [4] and [5], where VB was used to solve an approximation to the conditional mean of the image. The prior used in both of these papers is of the form

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x \,|\, \lambda) \propto \lambda^{\frac{N}{2}} \mathrm{e}^{-\lambda \mathrm{TV}_{iso}(x)}. \tag{6.6}$$

There exists no obvious method like the GSM property to get rid of the square root in isotropic TV penalty (4.26). The trick is to use inequality

$$\sqrt{w} \leq \frac{w + z}{2\sqrt{z}}, \tag{6.7}$$

which holds for $w \geq 0, z > 0$, to find a lower bound for the posterior. Then instead of minimising the KL between the product of the independent densities and the actual posterior, the idea is to minimise an upper bound for it, that is, KL between the product of independent densities and the lower bound of the posterior. This leads to VB formulas with an additional step in which one solves an optimising problem for $z$ that follows from using (6.7). The equality in (6.7) holds if $z = w$ which makes this optimisation subproblem simple. The idea of this additional step is to make the upper bound "tight" at each iteration step.

This idea is also closely related to majorization-minimization (MM) methods in which one constructs, for example, a quadratic upper bound for a complicated function and minimises it instead of the more difficult function. These MM methods have been proposed in [7, 42, 20] for TV regularisation, though, Bayesian approach was not considered in these papers. It is not fully clear how the regularisation parameter $\lambda$ should be chosen in this case since the prior is improper (see discussion in [42]). In GSM case it is chosen so that each difference of neighbouring pixels has proper Laplace density but this can not be used in the case of isotropic TV as there is no GSM property to be exploited.

Instead of presenting this isotropic TV model as briefly described above in the framework of this paper, a prior which is based on two-dimensional Laplace density which is GSM, is studied in the next section.

## 6.3 Two-dimensional Laplace TV prior

Multivariate Laplace is scale mixture of multivariate normal distribution as stated in Theorem 2.10. In the case of independent components, that is $\Sigma = \frac{2r}{\lambda}\mathrm{I}$, the result can be written so that if $\mathbf{x} \,|\, (\mathbf{r} = r, \boldsymbol{\lambda} = \lambda) \sim \text{Normal}(0, \frac{2r}{\lambda}\mathrm{I})$ and $\mathbf{r} \sim \text{Exp}(1)$ then $\mathbf{x} \,|\, (\boldsymbol{\lambda} = \lambda) \sim \text{MVLaplace}(0, \frac{2}{\lambda}\mathrm{I})$. In this formulation, $m$-dimensional random vector $\mathbf{x} \,|\, (\boldsymbol{\lambda} = \lambda)$ has the pdf

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x \,|\, \lambda) = \frac{\lambda}{(2\pi)^{m/2}} \lambda^{m/2-1} \frac{\mathrm{K}_{m/2-1}(\sqrt{\lambda}\|x\|_2)}{(\sqrt{\lambda}\|x\|_2)^{m/2-1}}. \tag{6.8}$$

Recall that function $\mathrm{K}_p$ is the modified Bessel function of the second kind with parameter $p \in \mathbb{R}$. In two-dimensional case ($m = 2$) this pdf reduces to

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x \,|\, \lambda) = \frac{\lambda}{2\pi} \mathrm{K}_0\left(\sqrt{\lambda}\sqrt{x_1^2 + x_2^2}\right), \tag{6.9}$$

where $x = [x_1, x_2]^T \in \mathbb{R}^2$. Notice that contrary to one-dimensional Laplace density, this bivariate Laplace density has a singularity at origin. This is also seen in Figure 6.1 (c).

This result encourages to try and study the following two-dimensional TV prior

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x \,|\, \lambda) \propto \lambda^N \prod_{i=1}^{k} \prod_{j=1}^{n} \mathrm{K}_0\left(\sqrt{\lambda}\sqrt{(x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2}\right), \tag{6.10}$$

that is based on bivariate Laplace density. The boundary conditions $x_{1,j} = x_{k+1,j}$ and $x_{i,1} = x_{i,n+1}$ are applied as previously. We will again use notation $\nabla_{ij}^v x = x_{i+1,j} - x_{i,j}$ and $\nabla_{ij}^h x = x_{i,j+1} - x_{i,j}$ to obtain shorter and hopefully clearer formulas.

Thanks to GSM property, in the place of (6.10) one can use prior

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda},\mathbf{r}}(x \,|\, \lambda, r) \propto \lambda^N \prod_{i=1}^{k} \prod_{j=1}^{n} r_{i,j}^{-1} \, \mathrm{e}^{-\frac{\lambda}{2} \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{\left(\nabla_{ij}^h x\right)^2 + \left(\nabla_{ij}^v x\right)^2}{2r_{i,j}} - \sum_{i=1}^{k} \sum_{j=1}^{n} r_{i,j}}, \tag{6.11}$$

where $\mathbf{r}_{i,j}$ are added hyperparameters and $\mathbf{r}_{i,j} \sim \text{Exp}(1)$.

As before, Gamma priors for $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ and (white noise version of) likelihood (5.1) are used. By Bayes' law, this approach leads to the posterior

$$\begin{aligned} &p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(x,\nu,\lambda,r\,|\,y) \\ &\propto \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i=1}^{k}\prod_{j=1}^{n} r_{i,j}^{-1}\,\mathrm{e}^{-\frac{\nu}{2}\|y-Hx\|_2^2-\frac{\lambda}{2}\|P^{-1}Dx\|_2^2-\sum_{i=1}^{k}\sum_{j=1}^{n} r_{i,j}-\beta_\lambda\lambda-\beta_\nu\nu} \\ &= \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i=1}^{k}\prod_{j=1}^{n} r_{i,j}^{-1}\,\mathrm{e}^{-\frac{\nu}{2}(x-\hat{x})^TQ(x-\hat{x})-\frac{\nu}{2}(y^Ty-\hat{x}^TQ\hat{x})-\sum_{i=1}^{k}\sum_{j=1}^{n} r_{i,j}-\beta_\lambda\lambda-\beta_\nu\nu}, \end{aligned} \tag{6.12}$$

where $\hat{x} = Q^{-1}H^Ty$, $Q = H^TH + \frac{\lambda}{\nu}D^TP^{-2}D$, $P = \mathrm{diag}([\sqrt{2r},\sqrt{2r}]) \in \mathbb{R}^{2N\times 2N}$. Here only half of number of hyperparameters in anisotropic TV model are needed. Anyway, also this model suffers from the "singularity issue" in MAP equations and some tricks are to be used to avoid division by arbitrarily small numbers.

The posterior has the same form as before and the conditional densities are almost as in one-dimensional case. The only interesting change is the emergence of GIG also in this bivariate Laplace prior case, which can be seen from the posterior formula. The hyperparameters $r_{i,j}$ now have exponent $-1$ instead of $-1/2$. The conditional distributions are

$$\begin{aligned} \mathbf{x}\,|\,\nu,\lambda,r,y &\sim \mathrm{Normal}(Q^{-1}H^Ty,(\nu Q)^{-1}), && \text{(6.13a)}\\ \boldsymbol{\nu}\,|\,x,\lambda,r,y &\sim \mathrm{Gamma}(\tfrac{N}{2}+\alpha_\nu,\ \tfrac{1}{2}\|y-Hx\|_2^2+\beta_\nu), && \text{(6.13b)}\\ \boldsymbol{\lambda}\,|\,x,\nu,r,y &\sim \mathrm{Gamma}(N+\alpha_\lambda,\ \tfrac{1}{2}\|P^{-1}Dx\|_2^2+\beta_\lambda), && \text{(6.13c)}\\ \mathbf{r}_{i,j}\,|\,x,\nu,\lambda,r_{-[i,j]},y &\sim \mathrm{GIG}\left(2,\tfrac{1}{2}\lambda\Big(\left(\nabla_{ij}^h x\right)^2+\left(\nabla_{ij}^v x\right)^2\Big),0\right) \\ &\qquad i=1,\dots,k,\ j=1,\dots,n. && \text{(6.13d)} \end{aligned}$$

The formulas for the MAP using IAS method are seen easily from the conditional densities. One just needs to compute the modes of the conditional densities.

Derivation of variational Bayes formulas goes in similar fashion as in one-dimensional case. The results, that is, the optimal distributions $q_{\mathbf{x}}, q_{\boldsymbol{\nu}}, q_{\boldsymbol{\lambda}}$ and $q_{\mathbf{r}}$ are

$$\begin{aligned} \mathbf{x} &\sim \mathrm{Normal}(\hat{x},(\bar{\nu}\bar{Q})^{-1}), && \text{(6.14a)}\\ \boldsymbol{\nu} &\sim \mathrm{Gamma}\left(\tfrac{N}{2}+\alpha_\nu,\tfrac{1}{2}\|y-H\bar{x}\|_2^2+\tfrac{1}{2}\mathrm{tr}(\mathbb{V}(\mathbf{x})H^TH)+\beta_\nu\right), && \text{(6.14b)}\\ \boldsymbol{\lambda} &\sim \mathrm{Gamma}\left(N+\alpha_\lambda,\tfrac{1}{4}\sum_{i=1}^{k}\sum_{j=1}^{n}\mathbb{E}\left(\mathbf{r}_{i,j}^{-1}\right)\mathbb{E}\Big(\left(\nabla_{ij}^h\mathbf{x}\right)^2+\left(\nabla_{ij}^v\mathbf{x}\right)^2\Big)+\alpha_\lambda\right), && \text{(6.14c)}\\ \mathbf{r}_{i,j} &\sim \mathrm{GIG}\left(2,\tfrac{1}{2}\bar{\lambda}\mathbb{E}\Big(\left(\nabla_{ij}^h\mathbf{x}\right)^2+\left(\nabla_{ij}^v\mathbf{x}\right)^2\Big),0\right),\quad i=1,\dots,k,\ j=1,\dots,n, && \text{(6.14d)} \end{aligned}$$

where $\hat{x} = \bar{Q}^{-1}H^Ty$, $\bar{Q} = H^TH + \frac{\bar{\lambda}}{\bar{\nu}}D^T\bar{P}^{-2}D$ and $\bar{P}^{-2} = \frac{1}{2}\,\mathrm{diag}\left([\mathbb{E}(\mathbf{r}^{-1}),\mathbb{E}(\mathbf{r}^{-1})]\right)$. The expectations $\mathbb{E}(\mathbf{r}_{i,j}^{-1})$ needed for $\bar{Q}$ can be computed using the formula

$$\mathbb{E}(\mathbf{r}_{i,j}^{-1}) = \sqrt{\frac{2}{b}}\,\frac{\mathrm{K}_1(\sqrt{2b})}{\mathrm{K}_0(\sqrt{2b})}, \tag{6.15}$$

where $b = \frac{1}{2}\bar{\lambda}\mathbb{E}\Big( \Big(\nabla_{ij}^h \mathbf{x}\Big)^2 + \Big(\nabla_{ij}^v \mathbf{x}\Big)^2 \Big)$. The formula (6.15) follows from Proposition 2.5. The expectation formula in $b$ and also appearing in the density of $\boldsymbol{\lambda}$ can be computed given the mean and variance of $\mathbf{x}$.

There is no well-defined corresponding minimisation problem for this two-dimensional Laplace TV prior. This fact follows from the singularity of the two-dimensional Laplace density at origin. It is easily seen that the corresponding penalty would tend to minus infinity for a result that has all the pixels the same. The one-dimensional Laplace density is, however, defined and finite for all values in $\mathbb{R}$.

As a more general case, the GIG mixing density could be used but we limited to Laplace prior. As a special case quite similar formulas for two-dimensional Student's t-distribution prior could have been obtained. In next section we will derive this using another characterisation of GSM and see some other properties of this t-distribution TV.

## 6.4 t-distribution TV prior

The t-distribution in two-dimensional case with zero mean, covariance matrix $\Sigma = \frac{1}{\lambda}\mathrm{I}$ and degrees of freedom $w$ has the pdf proportional to

$$\lambda \left[1 + \frac{\lambda}{w}\left(x_1^2 + x_2^2\right)\right]^{-\frac{w+2}{2}}.$$

Parameter $w$ is kept fixed here. Changing the value of $w$ corresponds to changing the shape of the prior density. The GSM property gives reason to consider the prior

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda}}(x\,|\,\lambda) \propto \prod_{i=1}^{k}\prod_{j=1}^{n} \lambda \left[1 + \frac{\lambda}{w}\left((\nabla_{ij}^v x)^2 + (\nabla_{ij}^h x)^2\right)\right]^{-\frac{w+2}{2}}, \tag{6.16}$$

where $w > 0$ is the degree of freedom corresponding the t-distribution. In this section this prior is studied closer.

Given $\lambda, \nu$ and degree of freedom $w$ we see that

$$\begin{aligned}
&\arg\max_{x\in\mathbb{R}^N} \left\{ \mathrm{e}^{-\frac{\nu}{2}\|y-Hx\|_2^2} \prod_{i=1}^{k}\prod_{j=1}^{n} \lambda \left[1 + \frac{\lambda}{w}\left((\nabla_{ij}^v x)^2 + (\nabla_{ij}^h x)^2\right)\right]^{-\frac{w+2}{2}} \right\} \\
&= \arg\min_{x\in\mathbb{R}^N} \left\{ \frac{\nu}{2}\|y-Hx\|_2^2 - \ln \prod_{i=1}^{k}\prod_{j=1}^{n} \left[1 + \frac{\lambda}{w}\left((\nabla_{ij}^v x)^2 + (\nabla_{ij}^h x)^2\right)\right]^{-\frac{w+2}{2}} \right\} \\
&= \arg\min_{x\in\mathbb{R}^N} \left\{ \frac{1}{2}\|y-Hx\|_2^2 + \frac{w+2}{2\nu} \sum_{i=1}^{k}\sum_{j=1}^{n} \ln \left[1 + \frac{\lambda}{w}\left((\nabla_{ij}^v x)^2 + (\nabla_{ij}^h x)^2\right)\right] \right\} \\
&= \arg\min_{x\in\mathbb{R}^N} \left\{ \frac{1}{2}\|y-Hx\|_2^2 + \delta_1 \sum_{i=1}^{k}\sum_{j=1}^{n} \ln \left[1 + \delta_2\left((\nabla_{ij}^v x)^2 + (\nabla_{ij}^h x)^2\right)\right] \right\},
\end{aligned} \tag{6.17}$$

which is the corresponding minimisation problem. The parameters $\delta_1$ and $\delta_2$ control the regularisation strength and the form of the penalty term, respectively.

Since t-distribution is GSM with Gamma($\frac{w}{2}, \frac{w}{2}$) mixing density (see Theorem 2.7), the hierarchical prior

$$p_{\mathbf{x}\,|\,\boldsymbol{\lambda},\mathbf{r}}(x\,|\,\lambda,r) \propto \lambda^N \prod_{i=1}^{k}\prod_{j=1}^{n} r_{i,j}\, \mathrm{e}^{-\frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} r_{i,j}\left((\nabla_{ij}^v x)^2+(\nabla_{ij}^h x)^2\right)}, \tag{6.18}$$

with hyperparameters $\mathbf{r}_{i,j}$ is used in place of (6.16) in the same manner as in previous derivations. We could have got quite similar results if we had applied GIG mixing density in previous section. Anyway, we will present this and the results here although they are quite similar as in other method considered in this work. The hyperparameters $\mathbf{r}_{i,j} > 0$ have gamma densities and so they have pdfs

$$p_{\mathbf{r}_{i,j}}(r_{i,j}) \propto r_{i,j}^{\frac{w}{2}-1}\mathrm{e}^{-\frac{w}{2}r_{i,j}}. \tag{6.19}$$

Using the same approach as previously we end up with the posterior

$$\begin{aligned}
&p_{\mathbf{x},\boldsymbol{\nu},\boldsymbol{\lambda},\mathbf{r}\,|\,\mathbf{y}}(x,\nu,\lambda,r\,|\,y)\\
&\propto \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i=1}^{k}\prod_{j=1}^{n} r_{i,j}^{\frac{w}{2}}\,\mathrm{e}^{-\frac{\nu}{2}\|y-Hx\|_2^2-\frac{\lambda}{2}\|PDx\|_2^2-\frac{w}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} r_{i,j}-\beta_\lambda\lambda-\beta_\nu\nu}\\
&= \lambda^{N+\alpha_\lambda-1}\nu^{\frac{N}{2}+\alpha_\nu-1}\prod_{i=1}^{k}\prod_{j=1}^{n} r_{i,j}^{\frac{w}{2}}\,\mathrm{e}^{-\frac{\nu}{2}(x-\hat{x})^TQ(x-\hat{x})-\frac{\nu}{2}(y^Ty-\hat{x}^TQ\hat{x})-\frac{w}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} r_{i,j}-\beta_\lambda\lambda-\beta_\nu\nu},
\end{aligned} \tag{6.20}$$

where $\hat{x} = Q^{-1}H^Ty$, $Q = H^TH + \frac{\lambda}{\nu}D^TP^2D$ and $P^2 = \operatorname{diag}\left([r^T, r^T]\right)$. Once again, the posterior has essentially the same form as in previous cases and the derivations of the results to follow are mainly very similar. Although it might bore the reader to see similar formulas appear once more we will state the results here. So, the conditional distributions for Gibbs sampler are the following.

$$\begin{aligned}
\mathbf{x}\,|\,\nu,\lambda,r,y &\sim \text{Normal}(Q^{-1}H^Ty, (\nu Q)^{-1}), && \text{(6.21a)}\\
\boldsymbol{\nu}\,|\,x,\lambda,r,y &\sim \text{Gamma}(\tfrac{N}{2}+\alpha_\nu,\ \tfrac{1}{2}\|y-Hx\|_2^2+\beta_\nu), && \text{(6.21b)}\\
\boldsymbol{\lambda}\,|\,x,\nu,r,y &\sim \text{Gamma}(N+\alpha_\lambda,\ \tfrac{1}{2}\|PDx\|_2^2+\beta_\lambda), && \text{(6.21c)}\\
\mathbf{r}_{i,j}\,|\,x,\nu,\lambda,r_{-[i,j]},y &\sim \text{Gamma}\left(\frac{w+2}{2}, \frac{\lambda}{2}\left[(\nabla_{ij}^v x)^2+(\nabla_{ij}^h x)^2\right]+\frac{w}{2}\right), && \text{(6.21d)}\\
&\qquad i=1,\ldots,k,\ j=1,\ldots,n. && \text{(6.21e)}
\end{aligned}$$

From above we can see that instead of GIG distribution we now have gamma density (though as argued in Chapter 2, gamma is a special case of GIG). Anyway, this is good news since gamma density is more common and sampling from gamma is faster than from GIG. However, sampling methods are not the main focus of this work.

The IAS method leads to the following iteration formulas for the MAP estimate. These formulas again follow from the conditional densities.

$$x = (H^T H + \tfrac{\lambda}{\nu} D^T P^2 D)^{-1} H^T y, \tag{6.22a}$$

$$\nu = \frac{N - 2 + 2\alpha_\nu}{\|y - Hx\|_2^2 + 2\beta_\nu}, \tag{6.22b}$$

$$\lambda = \frac{2N - 2 + 2\alpha_\lambda}{\|PDx\|_2^2 + 2\beta_\lambda}, \tag{6.22c}$$

$$r_{i,j} = \frac{w}{\lambda \left[ (\nabla_{ij}^v x)^2 + (\nabla_{ij}^h x)^2 \right] + w}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n. \tag{6.22d}$$

The conditional mean estimate can be estimated using variational Bayes as in previous section. The only major difference compared to previous calculations is the emergency of the gamma density for the hyperparameters.

It is well known that letting the degree of freedom $w$ approach infinity, the t-distribution approaches normal distribution. Consequently, from the above computations we can get a hierarchical model for Tikhonov regularisation with penalty $J(x) = \|Dx\|_2^2$. The prior for $\mathbf{x} \,|\, (\boldsymbol{\lambda} = \lambda)$ transforms to Gaussian by neglecting the use of GSM property, setting all $r_{i,j} = 1$ and taking the limit $w = \infty$ in posterior formula.

For example, let us examine the IAS formulas. Now, letting $w \to \infty$ it can be seen that $r_{i,j} \to 1$ and consequently one obtains

$$x = (H^T H + \tfrac{\lambda}{\nu} D^T D)^{-1} H^T y, \qquad \lambda = \frac{2N - 2 + 2\alpha_\lambda}{\|Dx\|_2^2 + 2\beta_\lambda}. \tag{6.23}$$

The formula for $\nu$ will not change from (6.22b). So we obtain a Tikhonov regularisation formula with $L = D$ and regularisation parameter is $\delta = \frac{\lambda}{\nu}$. Thus we obtain a method to perform Tikhonov regularisation in which the regularisation parameter is estimated from the data as well.

Let us finally make a brief summary of the many TV variants of this chapter. First we considered anisotropic TV which followed quite clearly from the hierarchical TV model in one-dimension. We could have made similar derivations using t-distributions or more generally using GIG mixing density. We skipped isotropic TV and, instead, presented TV prior that is based on two-dimensional Laplace density. Finally we presented t-distribution TV prior, which was based on two-dimensional t-distribution. We noticed that as a special case we obtain a hierarchical model for Tikhonov regularisation. All these are presented in Figure (6.1) on page 41.

# Chapter 7

# Image processing problems

In this section the methodology and algorithms presented in earlier sections are applied in one and two-dimensional image problems. We focus mainly on image deblurring problem but denoising and inpainting problems are also briefly approached. We perform the simulations by applying blur and noise to selected test images and see if the algorithms can be used to restore the original images successfully. We will use additive white Gaussian noise in all the examples. The following quantities are used to criticise and compare the results. The blurred signal-to-noise ratio is defined here as

$$\mathrm{BSNR} = 10 \log_{10} \left( \frac{\|Hx\|_2^2}{N\sigma^2} \right), \tag{7.1}$$

where $H$ is the blurring matrix, $x$ is the original image, $N = kn$ size of the image and $\sigma^2$ is the variance of the additive white noise. Small values of BSNR indicate higher noise level and thus more challenging deconvolution problem.

The improvement in signal-to-noise ratio is defined as

$$\mathrm{ISNR} = 10 \log_{10} \left( \frac{\|x - y\|_2^2}{\|x - \hat{x}\|_2^2} \right), \tag{7.2}$$

where $x$ is the original image, $y$ is the observed image and $\hat{x}$ is the estimated image. This quantity is used to quantify the deblurring performance. The higher the value the better the result.

There are several priors (that follow by setting different GIG mixing densities) that can be used in the following examples. However, we mainly test priors that are hierarchical presentations of Laplace and t-distributions instead of testing a huge number of possibilities for best possible deblurring performance. Also, we mainly focus on "cartoony" images that have jump discontinuities. Some smooth images are also used. Whenever we talk about "Laplace prior" or "t-distribution prior" in this section we actually mean some TV or TV "like" prior presented in preceding sections. With t-distribution we used degree of freedom 3 even though some other value could have worked better in some situations. All the results were implemented and results computed with

MATLAB. Also improper priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ were used except when stated otherwise.

## 7.1 Image deblurring

The one-dimensional deblurring problem is specified by Fredholm first kind integral equation of convolution type which can be presented as

$$g(x) = \int_0^1 k(x - x')f(x')\,\mathrm{d}x', \quad 0 < x < 1. \tag{7.3}$$

Here $g$ represents the blurred image, $f$ is the original image and $k$ presents the blurring effect and is called kernel. In all examples a Gaussian kernel is used:

$$k(x) = c\,\exp(-x^2/(2\gamma^2)), \tag{7.4}$$

with positive parameters $c$ and $\gamma$ that control the blurring effect. The direct problem would be to form a blurred image $g$ given the kernel $k$ and original image $f$. The inverse problem is to restore the image $f$ when the kernel $k$ and blurred image $g$ are known.

This problem can be discretised by applying midpoint quadrature. Then one obtains a familiar linear equation

$$y = Hf + \epsilon, \tag{7.5}$$

where the blurring matrix $H$ has entries

$$H_{ij} = \frac{c}{n}\,\exp\left(-\frac{((i-j)/n)^2}{2\gamma^2}\right), \quad 1 \le i, j \le n. \tag{7.6}$$

In practise the error term $\epsilon$ is also considered and it is on the right side of (7.5). This error is caused by discretisation approximations and measurement errors of the image. This problem can not be solved for large systems simply by inverting $H$. This naive reconstruction attempt will likely fail since the blurring matrix is increasingly ill-conditioned in large dimensional discretisations and the errors in measurements get amplified causing very unreliable and bad results.

Here we consider the example from Vogel [52] with the one-dimensional "image" defined on the interval $[0, 1]$:

$$x_{\text{Image }1}(t) = \begin{cases} 0.75, & \text{if } 0.1 < t < 0.25, \\ 0.25, & \text{if } 0.3 < t < 0.32, \\ \sin^4(2\pi t), & \text{if } 0.5 < t < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{7.7}$$

and $\epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$. The interval $[0, 1]$ is divided into $n = k = 100$ equidistant intervals and the parameters for the blurring are $\gamma = 0.05$ and $c = 1/(\gamma\sqrt{2\pi})$. (In

the case of Image 2 #2 we use $\gamma = 0.1$.) The discretisation is, however, performed in larger grid to avoid inverse crime. Linear interpolation is used to revert back to the original grid. The total variation priors are tested with this example. The results are also compared to solutions computed with a deterministic minimisation algorithm. Constrained quadratic programming was used to solve the minimum of TV as in [49, p. 44–45] and the regularisation parameter was manually chosen to be such that gave best improvement in signal-to-noise ratio. Tikhonov regularisation with $L = D$ and regularisation parameter chosen to give practically the best possible result was also used for comparison.

Some results with two different noise levels and different total variation methods are summed up in Table 7.1. Some reconstructions are presented in Figures 7.1, 7.2 and 7.3. Image 1 is (7.7) and image 2 as (7.7) but with the smooth curve on $(0.5, 1)$ replaced by a blocky piece of signal.

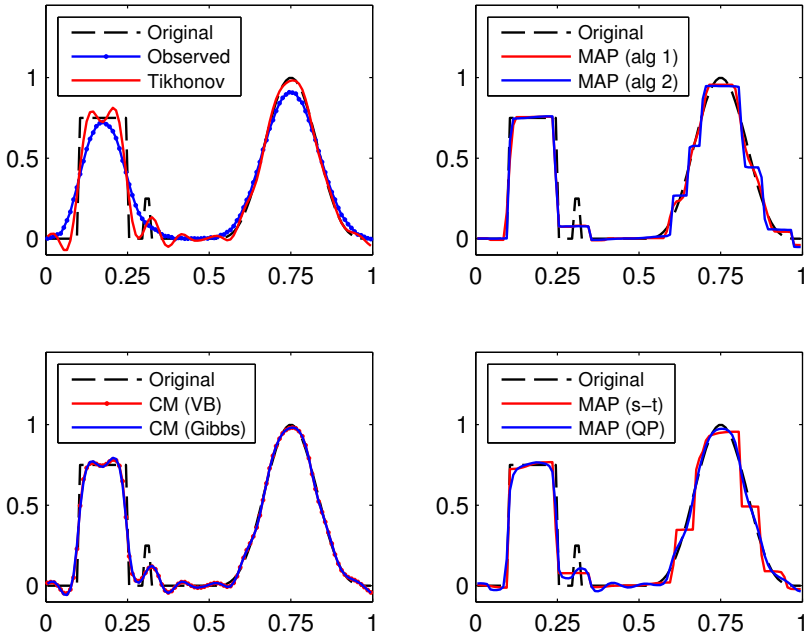


Figure 7.1: TV regularisation of one-dimensional blurred image (Image 1) with 40 dB noise level. MAP estimates worked well for blocky signal but also "staircasing" effect is seen with the smooth curve on interval $(0.5, 1)$. CM estimates did not preserve the sharp edges so well but have no trouble with the smooth curve.

Results show that the estimation of parameters using the introduced hierarchical model works and the results are mostly only slightly worse than the results obtained by deterministic minimisation algorithm. Note that in the deterministic optimisation case the regularisation parameter was manually chosen so this method has a clear advantage over the model in which the parameters are estimated and no additional heuristics is applied. We also see several other interesting things: There is very little difference between CM estimates. Practically it does not matter in this case whether the results are computed using VB or the Gibbs sampler algorithm and in fact VB

Table 7.1: Results of deblurring in 1d with different TV models. Here alg 1 refers to the Laplace prior in which zero-going components were limited to a small values, alg 2 is the approximate TV solution computed using the algorithm with GIG$(1, 0.001, 2)$ mixing density. Abbreviation s-t refers to t-distribution prior and QP refers to the deterministic minimisation algorithm.

| | | Image 1 | Image 2 #1 | Image 2 #2 |
|---|---|---|---|---|
| BSNR | Method | ISNR (dB) | ISNR (dB) | ISNR (dB) |
| 40 dB | Tikhonov | 3.60 | 3.01 | 4.47 |
| | MAP (alg 1) | 4.76 | 5.91 | 8.86 |
| | MAP (alg 2) | 3.40 | 6.84 | 8.99 |
| | CM (VB) | 4.83 | 3.98 | 4.58 |
| | CM (Gibbs) | 4.30 | 3.52 | 4.43 |
| | MAP (s-t) | 2.82 | 7.25 | 7.62 |
| | MAP (QP) | 6.01 | 8.40 | 9.05 |
| 30 dB | Tikhonov | 2.88 | 2.81 | 4.20 |
| | MAP (alg 1) | 3.23 | 6.06 | 4.63 |
| | MAP (alg 2) | 2.77 | 7.26 | 4.32 |
| | CM (VB) | 3.82 | 3.60 | 4.46 |
| | CM (Gibbs) | 3.45 | 3.22 | 4.27 |
| | MAP (s-t) | 1.53 | 5.80 | 5.56 |
| | MAP (QP) | 3.88 | 9.02 | 6.51 |

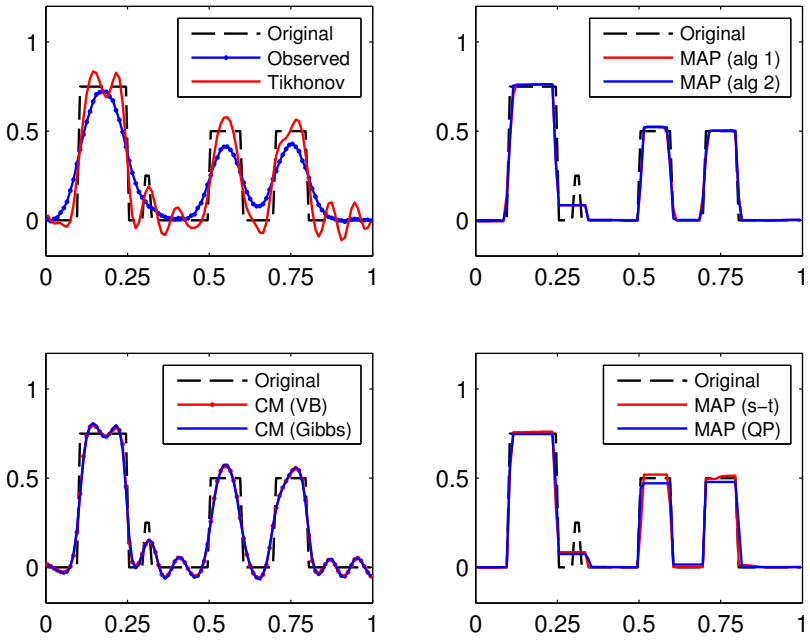


Figure 7.2: TV regularisation of one-dimensional blurred image (Image 2 #1). Here the noise level was 40 dB. The MAP estimate can clearly preserve the edges well.

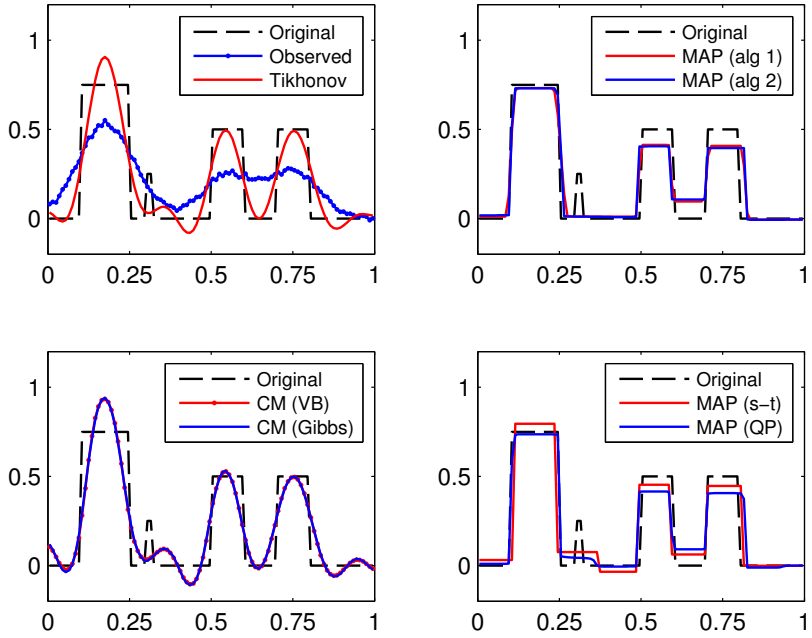


Figure 7.3: TV regularisation of one-dimensional blurred image (Image 2 #2). Here we have more blurring and more noise (30 dB) than in Figure 7.2.

seems to give slightly better results. In Gibbs sampler 10000 samples were generated and the samples looked uncorrelated so the result should be reliable. Note also that while the means seem to agree quite well, the densities itself can be different. The CM estimates were not much better than Tikhonov regularisation though. Especially with larger blurring levels the results did not differ much.

Also, we can see that the MAP estimates work better for blocky signals than CM which tend to produce smoother solutions and did not preserve the edges well. With the smooth piece of the signal in Image 1 the MAP estimates, on the other hand, produced "staircasing" effect, which is one of the typical disadvantages of TV. The regularisation parameter seems to be estimated larger than what it should be in this case. Because of this the results of Image 1 do not favour the hierarchical TV model. Using for example GIG$(1, 1/2, 2)$ mixing density this effect would be less severe but then the sharp corners would be more round. The regularisation model in which t-distribution was used produced also surprisingly good results although the density is not sharp peaked unlike Laplace density. It gave very "sparse" results but produced strong staircasing effect with the smooth curve in Figure 7.1.

The IAS and VB algorithms converged reasonable fast in these tests. Usually fewer than 15 iterations were needed, though for the VB a smaller stopping tolerance was applied and approximately 70 or less iterations were needed. Note also that ISNR will not always tell the whole truth about the preservation of blocky edges. If this is the top priority then definitely some TV method should be preferred over for example Tikhonov regularisation. In the hierarchical model MAP estimate is then the one to use. The error in Tikhonov and CM case is mostly due to smoothness while with MAP estimates for TV especially in noisy reconstructions it is due to errors in the

places of jumps or the level of a flat area gets slightly wrong value. Also the MAP estimates tend to differ from time to time and they depend on the noise quite a lot while other methods gave more consistent results. This also made the comparison somewhat problematic.

Let us now consider two-dimensional image examples. The image blurring model in two-dimensional space is described by the formula

$$g(x,y) = \int_0^1 \int_0^1 k(x-x', y-y') f(x', y') \, \mathrm{d}x' \, \mathrm{d}y'. \tag{7.8}$$

Discretisation of this integral equation leads to standard linear problem as in the one-dimensional example (7.5). We will skip the details and take a look at the examples.

We test the algorithms with three images. The first is from Hansen [25] and is here called "Image 3", the second is "Shepp-Logan phantom" and third is the famous "Lena" image. All the images are gray scale images but we applied some nice color palette for the Image 3. This image is of size $26 \times 26$, the second and third are $200 \times 200$ pixel images (Phantom200 and Lena200). We test also with $50 \times 50$ Lena and Phantom images (Phantom50 and Lena50). We use $7 \times 7$ Gaussian blur mask and two noise levels.

Some results are summarised in Table 7.2. Some reconstructions are shown in Figures 7.4, 7.5, 7.6 and 7.7. We only computed the reconstructions for the two larger images using MAP estimates which only require solving a linear system. No fast code to invert the matrices for the VB (or Gibbs sampler) were implemented as discussed in Section 5.3. We can see that the MAP estimates give nice results with the "blocky" Image 3 and Shepp-Logan phantom images while with the Lena image some staircasing effects are evident. For the smooth Lena image the regularisation parameter was not estimated as one might want. With some tuning better results could have been obtained. With the smaller Lena image the CM estimate computed using the VB algorithm, on the other hand, gave good results. Gibbs sampler was not tested since it performed slightly worse as VB and it is clearly very slow. Again, we also note the very good performance of t-distribution prior model in the case of blocky images. This t-distribution prior was the version that is product of two one-dimensional t-distributions. With larger degree of freedom also the results with the smooth Lena image would have been better. Note that even though numbers show that Tikhonov regularisation did quite well in some cases, it actually produced artifacts and left some smoothness to the images.

The difference between the performance of anisoTV and (two-dimensional) Laplace is rather small. However, there is some small but interesting difference. Taking a very close look at Figure 7.6 shows that the anisoTV produces blocky but somewhat jagged borders while with Laplace the borders are blocky but less jagged. This does not show much in ISNR values, though. On the other hand Laplace produced some artefacts near the center point of red plus-sign with Image 3 and that is why it got inferior results compared to anisoTV. The difference is subtle and not much of important meaning though.

Some final remarks are also worth mentioning. The Barzilai-Borwein optimisation algorithm [49, pp. 47–48] for isotropic TV that was used for comparison was not very optimised and it is also based on the approximation (4.30). We chose $\beta^2 = 1/1000$. The method is related to steepest descent method. As future work one might want to implement some other algorithm for more fair comparison as well as find a fast way to invert the matrices in the VB case so that larger images could be tested. Let us next take a look at related problem, image denoising.

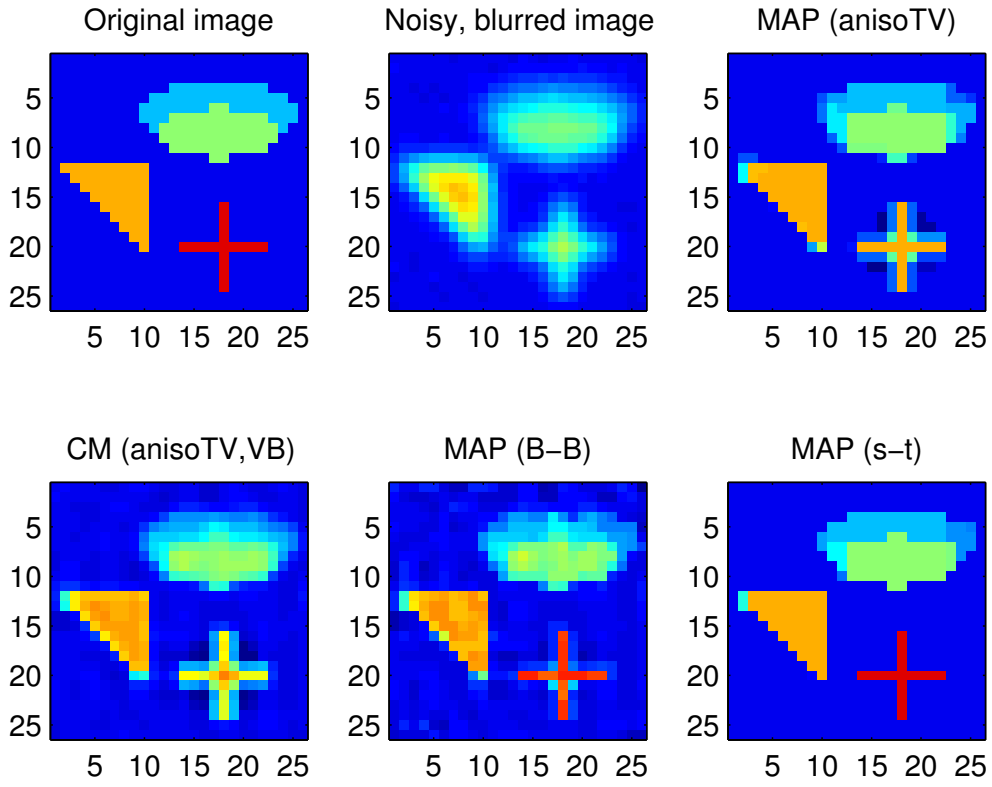


Figure 7.4: An image deblurring example with an artificial "Image 3". The noise level was 40 dB.

## 7.2 Image denoising

The image denoising is a special case of deblurring, the matrix $H$ is just set to identity matrix I. That is, one simply wants to remove the excessive noise from the image. So, the problem from the deterministic point of view is to find a solution to the optimisation problem

$$\arg\min_{x \in \mathbb{R}^N} \left\{ \|y - x\|_2^2 + \delta \mathrm{TV}(x) \right\}. \tag{7.9}$$

Total variation approach for this type of problem was first proposed in [48]. Other models for this type of problems naturally exist. The corresponding statistical model is

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{Normal}(0, (\boldsymbol{\nu}\mathrm{I})^{-1}), \tag{7.10}$$

Table 7.2: Results of 2d deblurring. For the MAP of anisoTV and (two-dimensional) Laplace we actually used GIG$(1, 0.001, 2)$ mixing density. B-B is the Barzilai-Borwein minimisation algorithm for isotropic TV.

| | | Phantom50 | Phantom200 | Lena50 | Lena200 |
|---|---|---|---|---|---|
| BSNR | Method | ISNR (dB) | ISNR (dB) | ISNR (dB) | ISNR (dB) |
| 40 dB | Tikhonov | 5.97 | 4.25 | 6.37 | 4.64 |
| | MAP (anisoTV) | 11.78 | 10.11 | 6.70 | 4.89 |
| | CM (anisoTV,VB) | 10.14 | - | 7.28 | - |
| | MAP (s-t) | 16.53 | 25.80 | 6.67 | 3.23 |
| | CM (s-t,VB) | 15.33 | - | 6.48 | - |
| | MAP (Laplace) | 14.68 | 9.20 | 6.56 | 4.64 |
| | CM (Laplace,VB) | 11.08 | - | 7.27 | - |
| | MAP (B-B) | 12.27 | - | 6.97 | - |
| 30 dB | Tikhonov | 3.18 | 3.44 | 4.05 | 3.52 |
| | MAP (anisoTV) | 7.13 | 5.57 | 2.87 | 2.82 |
| | CM (anisoTV,VB) | 4.21 | - | 4.21 | - |
| | MAP (s-t) | 9.67 | 11.02 | 2.67 | 1.32 |
| | CM (s-t,VB) | 8.05 | - | 3.54 | - |
| | MAP (Laplace) | 6.56 | 5.62 | 3.15 | 2.68 |
| | CM (Laplace,VB) | 5.45 | - | 4.28 | - |
| | MAP (B-B) | 7.12 | - | 3.09 | - |

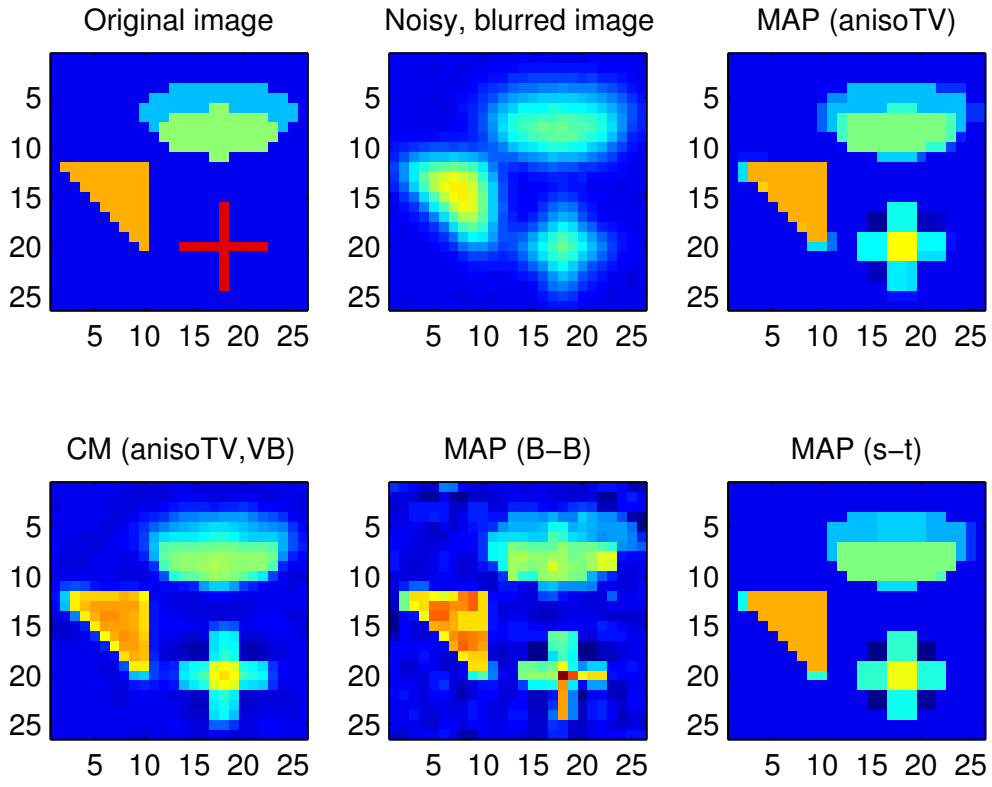


Figure 7.5: An image deblurring example with an artificial "Image 3". The noise level was 30 dB which caused restoring the red plus-sign to become unsuccessful.

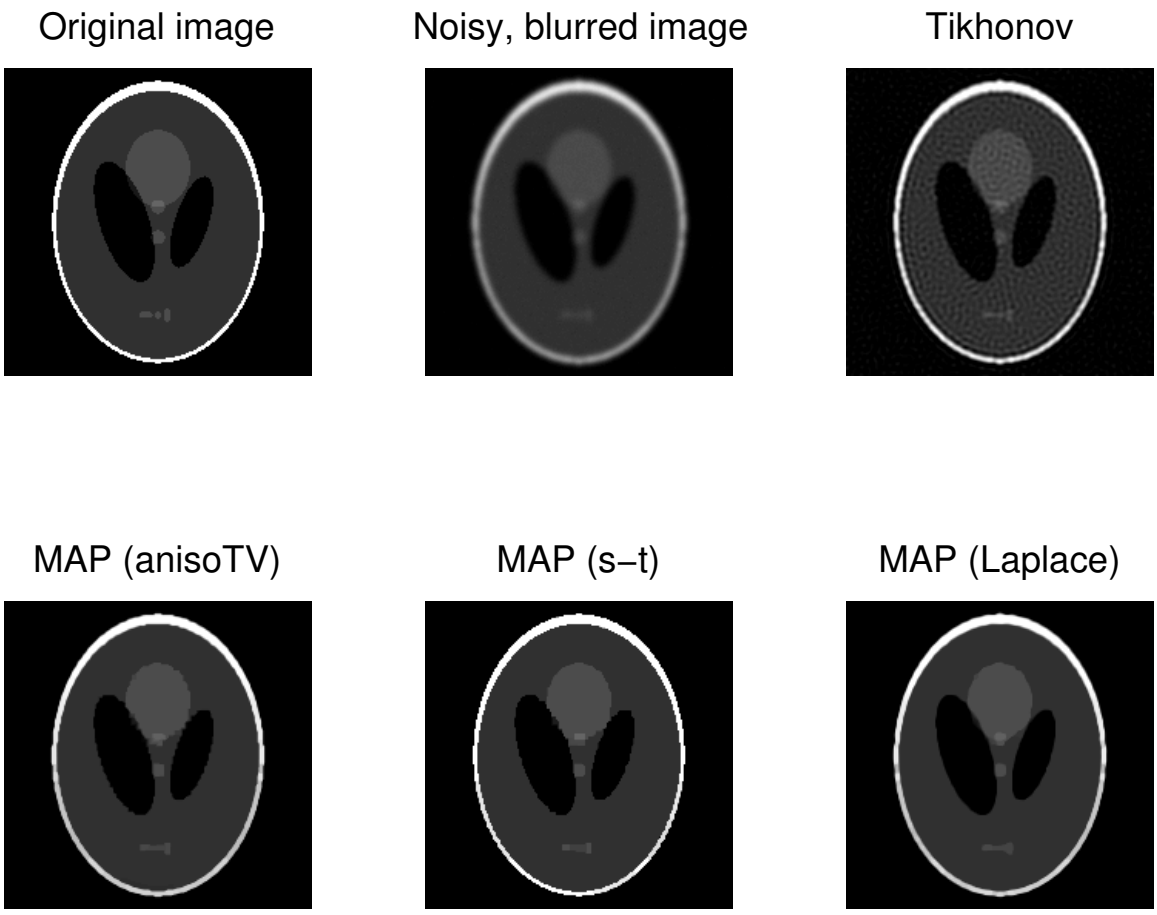


Figure 7.6: Deblurring the Shepp-Logan phantom image.

with total variation related prior for $\mathbf{x} \,|\, \boldsymbol{\lambda}$ and gamma priors for $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$. The algorithms presented in this work tackle this problem directly, one just needs to set $H = \mathrm{I}$.

As a simple example we consider $36 \times 36$ artificial image (the same Image 3 as in previous section) with 10% white Gaussian noise added. Some demonstrative results are shown in Figure 7.8.

It can be seen that the reconstruction is fairly good in Figure 7.8. The minimisation algorithm did as well as our methods that estimate everything from the image. However, in the case of the fully hierarchical model, difficulties with estimating correct values for the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ were encountered. Often the regularisation parameter was estimated to be far too small and the denoised image had very little noise removed. The results were also sensitive to initial guess. Thus for proper results one may need to set these parameters manually even though it is against the motivation of this work and also the principles of fully Bayesian inference. In the MAP and VB equations the values (means in the VB case) of $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ can be simply replaced by fixed values. However, with some good initial values and/or well chosen gamma priors for $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$, the algorithms often did seem to converge to good results but generally some tuning and trying was needed so one could have set these parameters manually

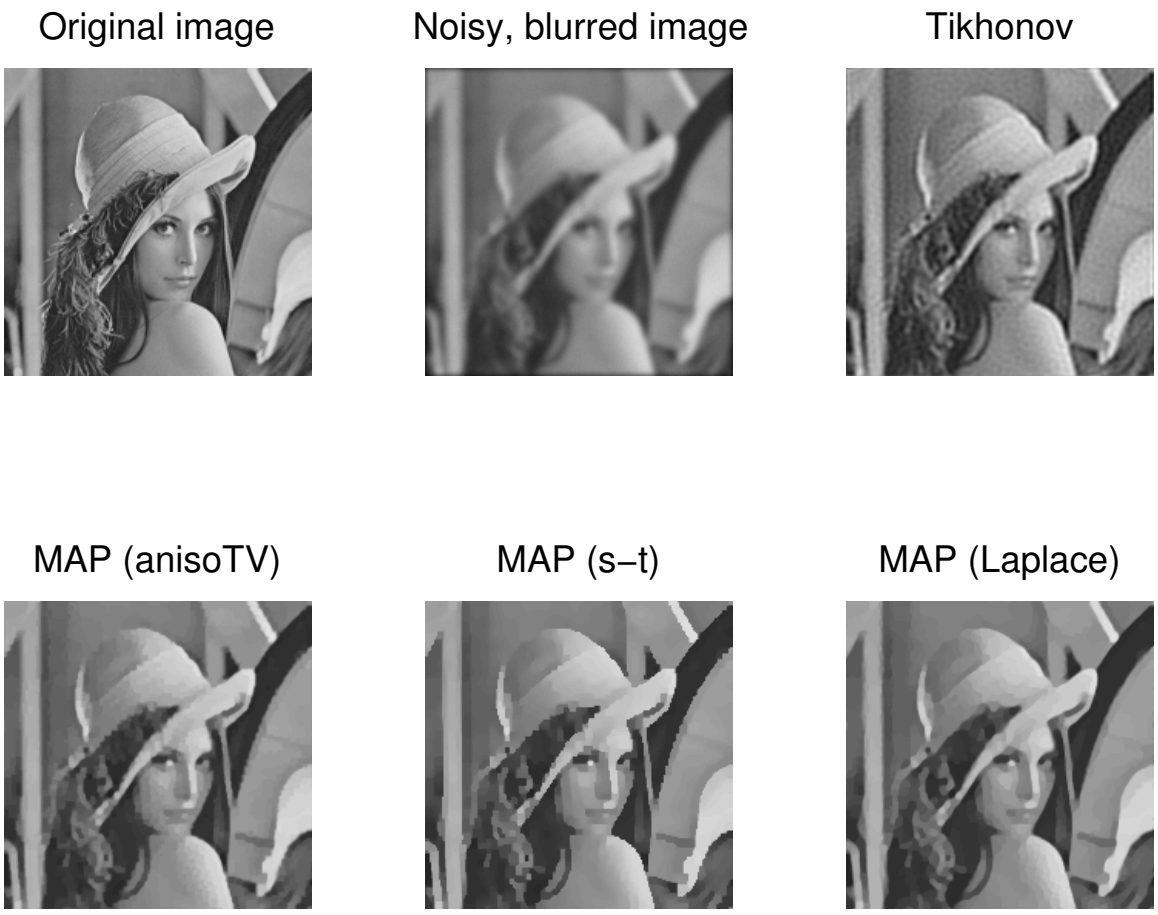


Figure 7.7: Deblurring the Lena image.

as well. Next let us take a look at the inpainting problem which also shows similar observations as the denoising case.

## 7.3 Image inpainting

As the last example, inpainting problem is considered. Here the objective is the same as in denoising but in addition some parts of the image are corrupted or missing. So one must reconstruct the image in the missing area using the data outside of the missing domain. One may wish to estimate scenery behind some small object that one wants to remove from the image, which leads to this problem as well. For instance, removing thin objects or text is quite common inpainting problem in practise. TV suits well for inpainting problems since it tends to recover sharp edges instead of smoothing the missing area as classical methods based on Laplace's equation tend to do [22]. Here we only consider one simple example. The inpainting problem can be written as a optimisation problem in our discrete case as

$$\underset{x\in\mathbb{R}^N}{\arg\min}\ \left\{\|x_{\text{meas}} - y_{\text{meas}}\|_2^2 + \delta \text{TV}(x)\right\}, \tag{7.11}$$

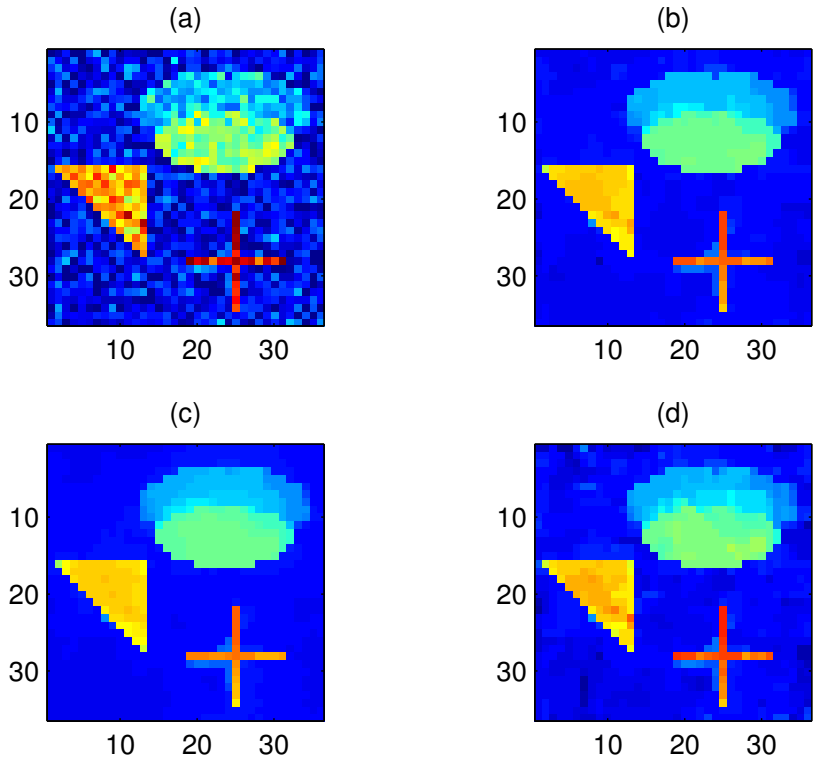


Figure 7.8: Image denoising example. (a) Noisy image, (b) MAP estimate (anisoTV), (c) restoration with VB (anisoTV), (d) isotropic TV (Barzilai-Borwein optimisation algorithm).

with $y = [y_{\text{meas}}^T, y_{\text{miss}}^T]^T$, where $y_{\text{meas}}$ refers to noisy, measured pixels and $y_{\text{miss}}$ are the missing (or obviously corrupted) pixels that are not used at all. Similarly we split the image to be estimated $x = [x_{\text{meas}}^T, x_{\text{miss}}^T]^T$. This problem is a discrete version of one in [12]. So the total variation term is applied to the whole image, while the quadratic fidelity term is applied only to areas for which the pixels in measured image are available. The hierarchical methods presented in this work can be used to solve this problem with some slight changes. To be more precise, applying the white noise version of the likelihood as in the equation (5.1) only to pixels that are observed and letting the missing pixels be "handled" by the TV prior.

Image inpainting in considered here only as an extra example and the methods presented in this work are not compared to other methods like it was done in deblurring problem. The artificial image has size $26 \times 26$ pixels and the white noise was added having the standard deviation of 5% and 10% of the maximum value of the image. The missing areas are small rectangles in this simple example. Some of the results are shown in Figures 7.9 and 7.10.

We can see that the reconstruction works fairly well in Figure 7.9 and 7.10. Here CM estimate seems to be more permissive against some noise and the filling the missing gaps happens perhaps smoother way when compared to MAP estimate. This behaviour is seen in Figure 7.10. The MAP estimate is, as already noticed in previous simulations, very strict against smoothness and eliminates all noise and preserves edges very well.

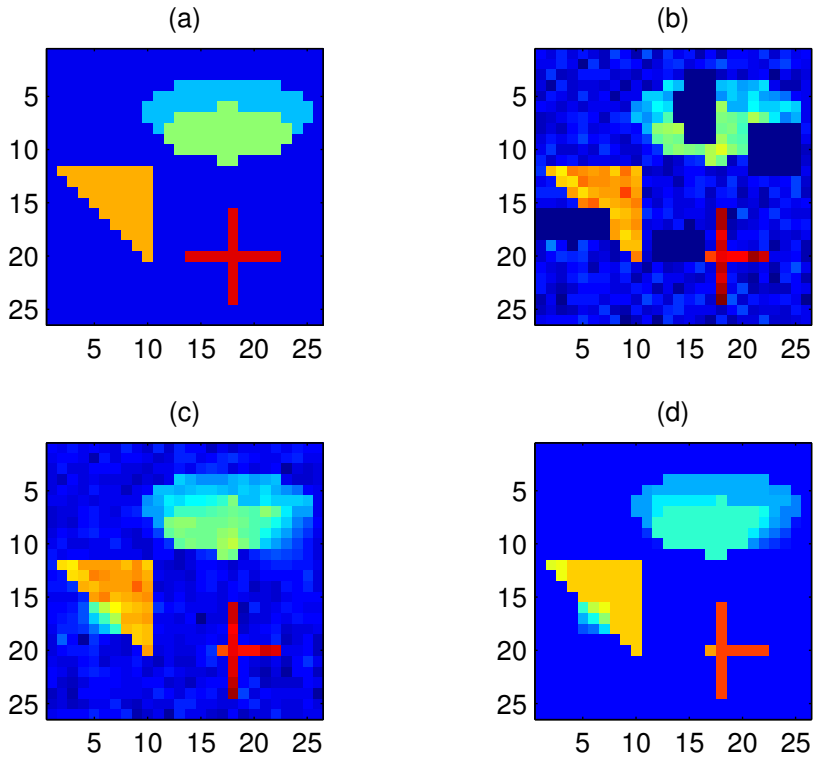


Figure 7.9: An example run on an image inpainting problem with 5% of noise. Dark blue sections present the missing domain. (a) Original image, (b) noisy and partially corrupted image, (c) reconstructed image using anisotropic TV and VB approximation for CM, (d) as in c-part but MAP using IAS method.

On the other hand we see a another typical problem associated with TV: For example the red plus-sign, while preserved quite well, has lost some of its color. This "loss of contrast" issue is another typical problem of TV regularisation.

The algorithms suffer from the same problems as in the denoising case. While it is often possible to make the algorithms converge to a desired solution, some tuning of initial values or setting parameters manually is needed. In Figure 7.10 the parameters for noise and strength of TV were set manually. In Figure 7.9 there was no problem with convergence. Convergence behaviour also seems to be sensitive to the size of noise and also the size of missing areas. Some further study and simulations are thus needed to see if it would be possible to get rid of this problematic behaviour. However, sometimes it might be desired to be able to manually control the strength of the denoising effect as TV easily destroys detailed textures and causes loss of contrast. The need for setting the parameters manually is then a good possibility. Also, how can one make an automatic method for choosing proper noise removal strength if one does not know how much of noise should be removed in the first place.

We also see that all the methods work very well in the case of very little noise. This is since it does not matter how the regularisation parameter is estimated in this case. The missing area will get filled according to the observed boundary points and the fact that no or very little denoising is done is not important since there is no need

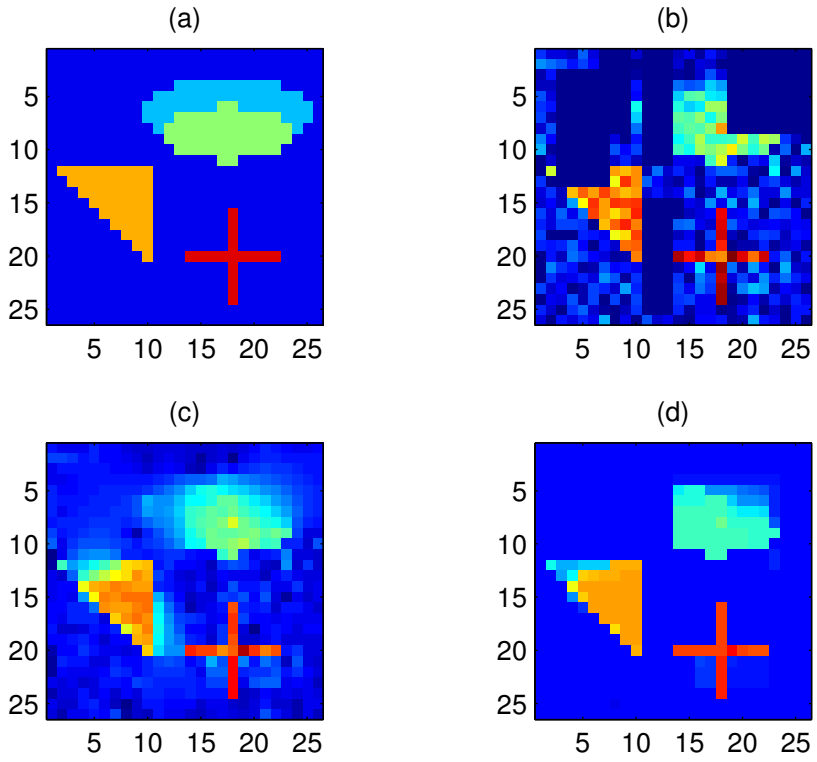


Figure 7.10: Another example of an inpainting problem. In this example the variance and TV penalty strength were set manually and not estimated. Noise level is 10%. (a) Original image, (b) noisy and partially corrupted image, (c) reconstructed image using anisotropic TV and VB approximation for CM, (d) as in c-part but MAP using IAS method.

for denoising anyway. However, in this case there is no reason to use this kind of statistical methods but apply the TV penalty only for the missing domain.

We tested the algorithms rather briefly with denoising and inpainting problems and only with a simple and small image and no proper comparison to other methods was presented. Some other models or more careful implementation might be a topic for future work.

# Chapter 8

# Conclusions

In this work total variation regularisation was studied in Bayesian context. The results derived in this work were applied in several image processing problems. Often these problems are solved via different optimisation algorithms which usually require some additional methods for determining values for tuning parameters. In this work, however, the total variation penalty function was considered as a Laplace prior distribution and the posterior for the model was derived exploiting the Gaussian scale mixture property. Also other TV like priors were considered. Algorithms for the conditional mean and maximum a posteriori estimates were then derived. Usually in literature only the MAP estimate is computed or some MCMC sampling method is used for CM. Here we used variational Bayes method for deriving formulas for CM. What was also tried was to see if all the parameters could be simultaneously and successfully inferred from the data.

The model in which all the parameters are estimated worked well in deblurring case and practically yielded comparable results to a deterministic algorithm for which the regularisation parameter was hand tuned for best possible result. In denoising and inpainting problems algorithms sometimes converged to obviously unwanted results. So sometimes the parameters had to be chosen manually as in deterministic approach or the initial values had to be chosen with care to obtain properly reconstructed images. While this breaks the motivation of this work and principles of Bayesian inference the methods produced then good results. Also in literature there has not been studies of for example fully "Bayesian inpainting" (as far as we know) and these problems are indeed difficult to solve.

It was noticed that VB and Gibbs sampling produced practically the same reconstructions in our model. Comparison of the MAP and (approximative) CM estimates as given by variational Bayes or sampling method showed that the CM tends to yield more smooth and less of "sparsity promoting" image reconstructions. This kind of observations has also been made for example in [28]. This is beneficial is some cases like with smooth images as it is less likely to produce unwanted "staircasing". With images that were fully blocky VB mostly produced inferior results than MAP but CM

was more stable to noise. The VB is also computationally much more heavy, on the other hand in the Laplace MAP case "singularity issues" had to be avoided.

All in all, it can be said that MAP estimates of the model worked successfully in deblurring case, especially in small noise conditions and for blocky images. VB can also be used for more smooth images and it is better and faster choice than using the Gibbs sampler. Initial results with denoising and inpainting problems were not fully successful but these additional problems were not the main topic of this thesis. It might also be possible to get them to work well. In general the problem of either choosing regularisation parameter or in the Bayesian approach estimating all parameters successfully and simultaneously is still more or less difficult question and in that sense while the results of this work were not perfect the results were generally quite promising.

While not an issue related to this work, TV is not invariant on discretisation in certain sense as studied in [34] and [14]. Roughly speaking, making the computational grid infinitely dense while keeping measurements fixed causes the conditional mean as well as the MAP estimates to either converge to useless zero or smooth solution breaking the edge-preserving property or to diverge. This issue has led to studies of sparsity promoting priors that have this discretisation invariance property. For example Besov space priors that are constructed on wavelets and are closely related to TV as studied in [30] are such. It might be interesting, although clearly much more difficult since those methods have much more complicated structure, to study if similar work as in this thesis could be done in that case.

In this work we only considered Gaussian noise. However, also Laplace, Student's t or some other heavy tailed distribution noise which is GSM could be applied quite easily. This would lead to hierarchical model for the noise as well and should not make derivations or computer implementation much more complicated compared to methods in this work or related literature. Finally that model could be tested using some heavy-tailed density noise, or for example Poisson noise.

The algorithms that were derived and implemented in this work were not optimized for best possible speed or performance for real image processing problems since the main objective was in presenting background, theoretical work and simple simulations. So these algorithms could be coded to perform faster by studying and implementing Fourier-based methods that should make computations related to convolution operations faster and are commonly used in "real" image processing applications. In addition, we used GIG mixing density and tried to present derivations at a general level. Still, we mainly focused on TV hierarchical models that were related to Laplace and t-distribution. As such, the other priors could be studied and more comprehensive simulations could be done.

It would be also possible to consider hierarchical models with more "layers" leading to more complicated but possible even more automated methods. That is, one would not need to choose between different TV or TV like priors but the best prior for the problem would also be estimated. Finding a new application or research field to which apply either these or related methods would also be interesting.

# Bibliography


[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions*. Dover, New York, 1965.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, USA, third edition, 2003.

[3] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974.

[4] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Total Variation Image Restoration and Parameter Estimation Using Variational Posterior Distribution Approximation. In *International Conference on Image Processing (1)*, pages 97–100. IEEE, 2007.

[5] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Variational Bayesian Blind Deconvolution Using a Total Variation Prior. *IEEE Transactions on Image Processing*, 18(1):12–26, 2009.

[6] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Image Processing*, 19(1):53–63, January 2010.

[7] J.M. Bioucas-Dias, M.A.T. Figueiredo, and J.P. Oliveira. Total variation-based image deconvolution: a majorization-minimization approach. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 2. IEEE, May 2006.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[9] L. Bornn, R. Gottardo, and A. Doucet. Grouping Priors and the Bayesian Elastic Net, January 2010. [accessed on 20.1.2013]. Available at: `http://arxiv.org/abs/1001.4083v1`.

[10] D. Calvetti and E. Somersalo. Recovery of Shapes: Hypermodels and Bayesian Learning. In *Journal of Physics: Conference Series*, volume 124, 2008.

[11] A. Chambolle, V. Caselles, M. Novaga, D. Cremers, and T. Pock. An introduction to Total Variation for Image Analysis. Lecture Notes of Summer School of Sparsity in Linz in September 2009. [accessed on 20.1.2013]. Available at: `http://hal.archives-ouvertes.fr/hal-00437581/en/`.

[12] T. F. Chan and J. Shen. Mathematical Models for Local Nontexture Inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043, December 2001.

[13] G. K. Chantas, N. P. Galatsanos, R. Molina, and A. K. Katsaggelos. Variational Bayesian Image Restoration With a Product of Spatially Weighted Total Variation Image Priors. *IEEE Transactions on Image Processing*, 19(2):351–362, 2010.

[14] S. Comelli. A novel class of priors for edge-preserving methods in Bayesian inversion. Italian diploma thesis (mathematics), Universita degli Studi di Milano, 2011.

[15] J. S. Dagpunar. *Principles of random variate generation.* Clarendon Press, Oxford, 1988.

[16] J. S. Dagpunar. An easily implemented generalized inverse Gaussian generator. *Communications in Statistics - Simulation and Computation*, 18(2):703–710, 1989.

[17] L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1986.

[18] T. Eltoft, T. Kim, and T. Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5), June 2006.

[19] M. A. T. Figueiredo. Adaptive Sparseness for Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159, September 2003.

[20] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization-Minimization Algorithms for Wavelet-Based Image Restoration. *IEEE Transactions on Image Processing*, 16(12):2980–2991, 2007.

[21] C. Fox and S. Roberts. A Tutorial on Variational Bayesian Inference. *Artificial Intelligence Review*, 38(2):85–95, 2012.

[22] P. Getreuer. Total Variation Inpainting using Split Bregman. *Image Processing On Line*, 2012.

[23] T. Gneiting. Normal scale mixtures and dual probability densities. *The Journal of Statistical Computation and Simulation*, 59(1):375–384, 1997.

[24] T. Goldstein and S. Osher. The Split Bregman Method for L1-Regularized Problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.

[25] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion.* SIAM Monographs on Mathematical Modeling and Computation. SIAM, Philadelphia, 1997.

[26] B. Jin and J. Zou. Hierarchical Bayesian inference for ill-posed problems via variational method. *Journal of Computational Physics*, 229(19):7317–7343, September 2010.

[27] B. Jorgensen. *Statistical Properties of the Generalized Inverse Gaussian Distribution*, volume 9 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1982.

[28] A. Kabán. On Bayesian classification with Laplace priors. *Pattern Recognition Letters*, 28(10):1271–1282, 2007.

[29] J. P. Kaipio and E. Somersalo. *Computational and Statistical Methods for Inverse Problems*. Springer, 2004.

[30] V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen. Sparsity-promoting Bayesian inversion. *Inverse Problems*, 28, 2012.

[31] S. Kotz, T. J. Kozubowski, and K. Podgórski. *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Progress in Mathematics Series. Birkhäuser, 2001.

[32] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[33] M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

[34] M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537–1563, 2004.

[35] F. Lucka. Hierarchical Bayesian approaches to the inverse problem of EEG/MEG current density reconstruction. German diploma thesis (mathematics), Institute for Computational and Applied Mathematics, University of Muenster, March 2011.

[36] F. Lucka. Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors. *Inverse Problems*, 28 (12):125012, September 2012.

[37] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. Springer, New York, third edition, 2008.

[38] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[39] X. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, June 1993.

[40] K. P. Murphy. *Machine Learning: a Probabilistic Perspective*. The MIT Press, USA, 2012.

[41] A. Nummenmaa, T. Auranen, M. S. Hämäläinen, I. P. Jääskeläinen, J. Lampinen, M. Sams, and A. Vehtari. Hierarchical Bayesian estimates of distributed MEG sources: Theoretical aspects and comparison of variational and MCMC methods. *NeuroImage*, (3):947–966, 2007.

[42] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. T. Figueiredo. Review: Adaptive total variation image deblurring: A majorization-minimization approach. *Signal Processing*, 89(9):1683–1693, September 2009.

[43] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), June 2008.

[44] N. L. Pedersen, D. Shutin, C. N. Manchón, and B. N. Fleury. Sparse Estimation using Bayesian Hierarchical Prior Modeling for Real and Complex Models, April 2012. [accessed on 10.1.2013]. Available at: `http://arxiv.org/abs/1108.4324v2`.

[45] A. Penttinen and R. Piché. Bayesian methods, 2010. Tampere University of Technology. Lecture Notes. [accessed on 5.12.2012]. Available at: `http://URN.fi/URN:NBN:fi:tty-201012161393`.

[46] C. P. Robert. *Monte Carlo Statistical Methods*. Springer, New York, third edition, 2002.

[47] G. Roussas. *A Course in Mathematical Statistics*. Academic Press, USA, second edition, 1997.

[48] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[49] S. Siltanen. Computational inverse problems, February 2010. University of Helsinki. Lecture Notes. [accessed on 5.12.2012]. Available at: `http://wiki.helsinki.fi/pages/viewpage.action?pageId=63749375`.

[50] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[51] M. C. K. Tweedie. Statistical properties of inverse Gaussian distributions. I. *Annals of Mathematical Statistics*, 28(2):362–377, 1956.

[52] C. R. Vogel. *Computational Methods for Inverse Problems*. Number 10 in Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2002.

[53] D. P. Wipf and S. S. Nagarajan. A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(35):947–966, 2009.

[54] Y. Yu. On Normal Variance-Mean Mixtures, June 2011. [accessed on 20.1.2013]. Available at: `http://arxiv.org/pdf/1106.2333v1`.

[55] W. Zuo and Z. Lin. A Generalized Accelerated Proximal Gradient Approach for Total-Variation-Based Image Restoration. *IEEE Transactions on Image Processing*, 20(10):2748–2759, 2011.

# Appendix A

# Derivations

We show that the result called matrix inversion lemma in Section 4.1 holds. This identity has several names, for example Sherman–Morrison–Woodbury formula and there are several ways to show that it indeed holds, for example using Schur complements. After that we see that

$$x_{\text{post}} = x_0 + \Sigma_0 A^T (A\Sigma_0 A^T + P)^{-1}(y - Ax_0) = \Sigma_{\text{post}}(A^T P^{-1} y + \Sigma_0^{-1} x_0), \tag{A.1}$$

$$\Sigma_{\text{post}} = \Sigma_0 - \Sigma_0 A^T (A\Sigma_0 A^T + P)^{-1} A\Sigma_0 = (\Sigma_0^{-1} + A^T P^{-1} A)^{-1}. \tag{A.2}$$

**Lemma A.1.** *If $B, U, C$ and $V$ are such matrices that all the products appearing below are defined and the required inverses exist then*

$$(B + UCV)^{-1} = B^{-1} - B^{-1}U(C^{-1} + VB^{-1}U)^{-1}VB^{-1}. \tag{A.3}$$

*Proof.* A direct computation shows that

$$\begin{aligned}
&(B + UCV)(B^{-1} - B^{-1}U(C^{-1} + VB^{-1}U)^{-1}VB^{-1}) \\
&= BB^{-1} - BB^{-1}U(VB^{-1}U + C^{-1})^{-1}VB^{-1} + UCVB^{-1} \\
&\quad - UCVB^{-1}U(VB^{-1}U + C^{-1})^{-1}VB^{-1} \\
&= \mathrm{I} + UCVB^{-1} - (U + UCVB^{-1}U)(VB^{-1}U + C^{-1})^{-1}VB^{-1} \\
&= \mathrm{I} + UCVB^{-1} - UC(VB^{-1}U + C^{-1})(VB^{-1}U + C^{-1})^{-1}VB^{-1} \\
&= \mathrm{I} + UCVB^{-1} - UCVB^{-1} \\
&= \mathrm{I}.
\end{aligned}$$

□

Substituting $B = \Sigma_0^{-1}$, $U = A^T$, $C = P^{-1}$ and $V = A$ into (A.3) gives the second identity (A.2).

The first identity follows from somewhat tedious computations. We denote $Q = A\Sigma_0 A^T + P$. The computation goes as follows.

$$
\begin{aligned}
&(\Sigma_0^{-1} + A^T P^{-1} A)^{-1}(A^T P^{-1} y + \Sigma_0^{-1} x_0) \\
&\overset{(A.2)}{=} (\Sigma_0 - \Sigma_0 A^T (A\Sigma_0 A^T + P)^{-1} A\Sigma_0)(A^T P^{-1} y + \Sigma_0^{-1} x_0) \\
&= \quad x_0 - \Sigma_0 A^T Q^{-1} A\Sigma_0 A^T P^{-1} y - \Sigma_0 A^T Q^{-1} A x_0 + \Sigma_0 A^T P^{-1} y \\
&= \quad x_0 + \Sigma_0 A^T Q^{-1}((QP^{-1} - A\Sigma_0 A^T P^{-1}) y - A x_0) \\
&= \quad x_0 + \Sigma_0 A^T Q^{-1}((A\Sigma_0 A^T P^{-1} + PP^{-1} - A\Sigma_0 A^T P^{-1}) y - A x_0) \\
&= \quad x_0 + \Sigma_0 A^T Q^{-1}(y - A x_0).
\end{aligned}
$$

Now we can see that (A.1) indeed holds.