Politecnico Di Torino

Master Degree in Physics of Complex Systems

Image Reconstruction via Expectation Propagation with auxiliary informative variables

Supervisor:

Prof. Alfredo Braunstein

Anna Paola Muntoni

Author:

Francesco Udine

Academic year 2021/2022

Abstract

Tomography is an imaging technique that allows one to reconstruct object sections by analyzing particular penetrating waves; this method finds applications in different areas of science. The mathematical procedure used to reconstruct images is generally called tomographic reconstruction; when X-rays are exploited, the overall procedure is called Computed Tomography scan (CT scan). Historically, this technique was first idealized by Johann Radon in 1917, when he introduced the so-called Radon transform: he showed that a cross-section of an image can be reconstructed in a single step through an infinite set of its projections. The first practical application in CT scan dates back to 1971, more than fifty years later. Nowadays, image reconstruction algorithms use, conversely, a finite number of projections, and iterative reconstruction methods. However, the original problem is mapped into an ill-posed linear problem, so one has to add more information, such as image regularizations, to find a feasible solution. In recent years, it has been possible to formulate the problem of image reconstruction in a Bayesian context, that is, in a probabilistic framework. Here, one has to find the pixel assignment that maximizes a posterior probability distribution. The huge advantage carried by this framework is that of encompassing the ability to introduce non-convex priors which reflect the typical characteristics of the images. However, this makes the problem intractable for general optimization methods, i.e. linear programming. Thanks to the Expectation Propagation method, it is possible to deal with these posteriors using approximations. Moreover, the ability to deal with non-convex functions allows us to introduce designed priors according to the empirical behavior of some pixel functions. This can be done for the so-called pixel-difference variables or for some other auxiliary variables obtained through a linear operator. Therefore, the idea behind the present work is to create a dataset of images somewhat similar to the tomographic ones and to apply a linear transform on the pixel variables: by studying the empirical behavior of the auxiliary variables, we try to deduce the functional form of the priors describing these auxiliary variables. In this thesis, we apply the Haar transform to the images of the dataset and we study the behavior of these auxiliary variables to define priors describing the Haar coefficients. Then, we compare the performances obtained by some implementations of the EP algorithm (differing in the priors used) in several measurement regimes. The results shown in the simulations confirm the validity of using auxiliary variables in the EP algorithm, in particular the pixel-difference variables.

Contents

1 Introduction

Tomography is a medical imaging procedure used to reconstruct the cross section of an object from a set of measurements (X-ray attenuations or other penetrating waves) at different angles.

The mathematical algorithms for tomographic reconstructions are based on the interaction between the radiation and the material of which the object is composed, producing projection data.

Such interactions can be formally described with line integrals of some characteristic of the object: for many applications these curved paths can be modeled by straight lines but in general the beam of photons can be absorbed by atoms or can be scattered away, deviating from this linear behavior. For many important practical applications, approximation of these curved paths by straight lines is acceptable. [1]

In CT scan, the X-rays pass through the object and the attenuation of the intensity of the radiation exiting the object is measured. The intensity of the measured beam is proportional to N_{in}, the number of photons entering the object; this reduces due to absorption as:

$$N_{out} = N_{in} e^{-\int_{ray} w(x,y) ds} \tag{1}$$

where N_{out} is the number of photons exiting the object and $w(x, y)$ is the attenuation coefficient of the object and depends on two spatial coordinates x, y: this is a strong assumption because in general this coefficient is a function of photon energy. If the X-ray beam is monochromatic, the monoenergetic photons have all the same energy: so we can assume the spatial dependence of the attenuation coefficient mentioned above.

A different approach for tomographic imaging is the algebraic one, where we assume that the cross section is an array of unknowns, a discretized grid. This approach seems more suitable in the presence of diffracting sources, as refraction or diffraction. [2]

If the object is homogeneous, the intensity I_0 of the monochromatic beam is reduced due to absorption as:

$$I = I_0 e^{-wT} \tag{2}$$

where w is the attenuation coefficient of the object and T is the distance traveled by the beam inside the object.

In a more realistic scenario the object is inhomogeneous, so the decay of the intensity of the beam must take into account the different possible tissues of the object:

$$I = I_0 e^{-w_1 T_1} e^{-w_2 T_2} ... e^{-w_N T_N} \tag{3}$$

where the image to be reconstructed is discretized in a set of N pixels.

Taking the logarithms and considering M measurements, the problem of reconstructing the image is mapped to a linear estimation problem, with M equations in N unknown; each equation will be of the type:

$$y_m = log\frac{I_0}{I} = \sum_{i=1}^{N} w_i T_{mi} \tag{4}$$

where y_m is the log-ratio between the intensities of the incoming and outgoing beams, and this single ray-sum equation is an approximation of the line integral of X-ray attenuation. [3]

So using the attenuation coefficient w_i of the tissue it is possible to infer diagnostic medical information, namely the discretized cross-section of the object. In the next chapter we will denote this discretized grid of N pixels with $\boldsymbol{x} = (x_1, ..., x_N)^T \in \mathbb{R}^N$ and define a general setup for linear estimation problem.

The perfect (and ideal) reconstruction would be possible with an infinite number of rays from any possible angle. Clearly this ideal scenario is not possible due to the dangerous radiation dose, so in image reconstruction it is important to reduce the number of measurements and to design accurate image reconstruction techniques.

1.1 State of the art

There are two main categories of reconstruction algorithms, analytical reconstruction and iterative reconstruction (IR) : iterative methods iteratively optimize an objective function, while analytical procedures directly reconstruct the images in one step. The first ones are more efficient but they are also computationally more expensive with respect to the analytical procedures.

The most used analytical technique is the back-filtered-projection (BFP) , where the density's image is recovered by using the inverse Radon transform. Let us define the tomographic image as a function $f(x, y)$ defined on the plane; we need to exploit the information obtained through the sinogram, a set of projected data under different angles. From the mathematical point of view, the inverse Radon transform in 2D corresponds to the integral transform of a function Rf, defined on the space of straight lines (the output scan of each ray) into a function $f(x, y)$ on a plane, the tomographic image.

So the inverse Radon transform can be used in image reconstruction with BFP: we can obtain directly the reconstructed image in one single iteration step. In this case the tomographic image is defined as a continuous function on the plane; iterative reconstructions instead use an algebraic approach, where the image we want to reconstruct is discretized in a grid of discrete variables, the pixel intensities.

The ART (algebraic-reconstruction techniques) algorithm was introduced in image reconstruction in [4] and it is based on this algebraic approach. The main improvement over filtered-back projection is the possibility to include more easily a prior knowledge in the image reconstruction process.

In this algorithm at each iteration there is a linear estimation problem, where we know some observations and a linear operator generating them and we solve a set of linear equations.

In a general context, standard linear estimation consists of solving this system:

$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta} = \boldsymbol{p} \tag{5}$$

where $\boldsymbol{p} = (p_1, ..., p_M)^T \in \mathbb{R}^M$ is the vector of measurements affected by the possible presence of noise $\boldsymbol{\eta} \in R^M$, $\boldsymbol{A} \in \mathbb{R}^{M\text{xN}}$ is a known linear operator and $\boldsymbol{x} \in \mathbb{R}^N$ is an unknown vector collecting the pixel intensities.

ART algorithm is an iterative solver of the system of linear equations without noise:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{p} \tag{6}$$

In this framework $\boldsymbol{x} = (x_1, ..., x_N)^T \in \mathbb{R}^N$is the reconstructed image and $\boldsymbol{p} = (p_1, ..., p_M)^T \in \mathbb{R}^M$is the measurements vector of projected data, while the linear operator $\boldsymbol{A} = (f_{ij}) \in \mathbb{R}^{M\text{xN}}$is the projection-matrix (each entry of the matrix $\boldsymbol{A}$ represents the intensity of the intersection between the i-th projection ray and the j-th pixel). In other words each row of the projection-matrix represents the intersections of a given ray with all pixels. At each iteration ART algorithms minimizes the l_2 error defined as: $\boldsymbol{x}^* = argmin_{\boldsymbol{x}} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{p}||_2$.

Another algebraic approach is that of the simultaneous iterations reconstruction technique (SIRT) algorithms [5]: the difference is that ART algorithms sequentially solve one single ray-sum equation after another, while one iteration of SIRT algorithms simultaneously consider all equations of the system to be solved (SIRT is iterative in the sense that we can consider the same set of equations at each iteration).

SIRT algorithms do not take into account the order of the equations to be solved, therefore the concept of "ray in a specific direction" is absent. The idea behind this work [2] is therefore to improve the performances of the ART algorithms by considering simultaneously a subset of equations corresponding to a given projection direction. The SART (simultaneous algebraic reconstruction technique) algorithm tries to obtain the advantages of the algebraic methods described so far.

Whatever algebraic algorithm, the system of linear equations of our interest is in the limited data regime, because the number of measurements is smaller than the number of variables $M < N$ (also called underdetermined system of linear equations) : without noise this underdetermined system has an infinite number of solutions, so we need to add further information, like some regularization term to reduce the solution space. Notice that in a noisy system there is no solution because the presence of noise brings to an inconsistent set of equations.

Tomographic images are not sparse, they are, instead, quite constant in extended spaces (as within an organ) and they have rapid changes only at their boundaries. As a consequence, neighboring pixels are more likely to assume the same values: the discrete gradient of a tomographic image could be quite sparse, even if the original image is not.

It therefore seems reasonable inserting also the l_1 norm of the discrete gradient image in the objective function to be minimized :

$$\boldsymbol{x}^* = argmin_{\boldsymbol{x}} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{p}||_2 + \lambda ||\nabla_{img}\boldsymbol{x}||_1 \tag{7}$$

where we define $(\nabla_{img}\boldsymbol{x})_i = \left(x_{i_x} - x_i, x_{i_y} - x_i\right)$ with $i_x$and $i_y$are the neighboring pixels to the right and below i, respectively, and λ is a parameter to weight the contribute of the image gradient regularization.

An iterative image reconstruction algorithm based on the minimization of the image total variation (TV) is presented here [6].

1.2 Aim of the thesis

This thesis is organized as follows: after this introductory chapter about Computed X-ray tomography and the general image reconstruction methods, in Chapter 2.1 we will discuss the Bayesian approach to this problem, that is maximizing the posterior distribution considering the likelihood and priors. As we will see, in this framework it is important to retrieve any possible information in priors, and this can generate distributions that are difficult to treat from an analytic point of view.

Expectation Propagation allows us to handle with these distributions thanks to its approximation scheme. We will explain a basic version of EP in Chapter 2.2.

To make the most of the potential of EP we will introduce an other set of variables (the auxiliary ones, e.g difference variables), obtained through some linear transformation of the pixel variables. The EP algorithm with the introduction of this kind of variables is explained in section 2.3.

In chapter 3 we explain how it is possible to infer the functional form and parameters of the functions that describe the statistics of auxiliary variables obtained through a linear transformation of the pixels: in this way we obtain the so called "empirical priors". Then we show the Haar transform case and one possible application in data compression of the auxiliary variables obtained with this linear transform.

The results on the image reconstructions are shown in chapter 3.4: we have compared some methods that differ in the auxiliary variables implemented in the EP algorithm. The best two implementations were compared with the TV method and the SART one.

2 Methods

2.1 Bayesian Inference approach

Statistical inference is the process of using data to deduce properties of the distribution that generated the data. Among the different techniques of statistical inference in this work we will use Bayesian Inference, that takes advantage of Bayes' theorem to investigate the problem in a probabilistic framework. Thanks to this theorem we can design a posterior probability distribution for the variables $\boldsymbol{x}$ given a set of measurements $\boldsymbol{p}$.

According to Bayes' theorem, the posterior probability $\mathscr{P}(\boldsymbol{x}|\boldsymbol{p})$ is computed considering the probability of a hypothesis before observing the data (prior term $\mathscr{P}_0(\boldsymbol{x})$) and the probability of the observed data conditioned to the pixel values, the likelihood term $\mathscr{P}(\boldsymbol{p}|\boldsymbol{x})$:

$$\mathscr{P}(\boldsymbol{x}|\boldsymbol{p}) = \frac{\mathscr{P}(\boldsymbol{p}|\boldsymbol{x})\,\mathscr{P}_0(\boldsymbol{x})}{\mathscr{P}(\boldsymbol{p})} \tag{8}$$

The problem of image reconstruction can be formulated through a Bayesan Inference approach, in which the reconstructed image is the maximum a posteriori (MAP) estimation:

$$\boldsymbol{x}^* = argmax_{\boldsymbol{x}}\mathscr{P}(\boldsymbol{x}|\boldsymbol{p}) \tag{9}$$

where $\mathscr{P}(\boldsymbol{x}|\boldsymbol{p})$ is the posterior distribution of images $\boldsymbol{x}$ given the vector of measurements $\boldsymbol{p}$.

In many cases the MAP estimate may be not accurate, in particular when the posterior distribution is not concentrated: this problem arises because MAP is a point estimate, whereas Bayesian techniques use distributions to summarize data. Another approach in Bayesian inference is the minimum mean square error (MMSE), that corresponds to compute the mean value of the variables with respect to the posterior distribution (one can prove that minimizing the mean square error between the true value and the reconstructed one is the same as computing the expectation value with respect to the marginal of the posterior distribution, $x_i^* =< x_i >_{\mathscr{P}(x_i|\boldsymbol{p})}$). whereas Bayesian techniques use distributions to summarize data

The likelihood term $\mathscr{P}(\boldsymbol{p}|\boldsymbol{x})$ enforces the linear constraints of the set of equations: for the noiseless case it will be $\mathscr{P}(\boldsymbol{p}|\boldsymbol{x}) = \delta(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{p})$, but we can also add Gaussian noise, leading to this likelihood term: $\mathscr{P}(\boldsymbol{p}|\boldsymbol{x}) \propto e^{-\frac{\beta}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{p})^2}$, where β is the inverse variance of the noise.

The prior term $\mathscr{P}_0(\boldsymbol{x})$ is the crucial one: to find a solution of the underdetermined system, we need other information that may express beliefs about these quantities before some evidence is considered. This information can concern each single pixel, as a $l_1$or l_2 regularization, $\mathscr{P}_0^{(single)}(x) \propto e^{-\lambda||\boldsymbol{x}||_1}$or

$\mathscr{P}_0^{(single)}(x) \propto e^{-\lambda||\boldsymbol{x}||_2}$ respectively. Here we want to include also an a priori knowledge on some function of pixels, in order to exploit the intrinsic correlated structure of images, as smoothness between neighboring pixels.

So let us consider this form for the prior term:

$$\mathscr{P}_0(\boldsymbol{x}) \propto \mathscr{P}_0^{(single)}(\boldsymbol{x})\, \mathscr{P}_0^{(pair)}(\boldsymbol{x}) \tag{10}$$

where all priors with a probability distribution factorized over pairs of variables are included in the pair term $\mathscr{P}_0^{(pair)}$. A standard choice for this term is:

$$\mathscr{P}_0{}^{(pair)}(\boldsymbol{x}) \propto e^{-\frac{J}{2}\boldsymbol{x}^T L \boldsymbol{x}} \propto e^{-\frac{J}{2}\sum_{i=1}^{N}\sum_{j\in\partial i}(x_i - x_j)^2} \tag{11}$$

where $\boldsymbol{L}$ is the Laplacian matrix of the nearest-pixels adjacency graph, J is a weight parameter and the sum $j \in \partial i$ is over all the neighbors of pixel i.

The posterior distribution can be treated analytically with standard convex optimization techniques if we insert a $l_1$or l_2 regularization.

However, an analysis of computed tomography scans shows that difference variables between neighboring pixels have non-convex empirical distributions, with a peak in zero, as shown in figure (1) in [7] . Intuitively, in fact, if we take all the pairs of neighboring pixels, the difference variables will often take on values around zero because it is reasonable to think that very often they will have similar values (except for example in the case of the boundary between two different organs).

So a bimodal distribution seems more suitable to describe this non-convex behavior; we will see possible functional forms (e.g. the Bernoulli-Gaussian) in the chapter about EP with auxiliary variables. In this case, both MAP or MMSE estimation lead to a non-convex optimization problem.

These problems are NP-hard because it is possible to find a local minimum in polynomial-time algorithms but the proof of optimality is not (it requires an exponential-time algorithm).

So the main drawback of Bayesian Inference with non-convex prior is the massive computational cost, but thanks to recent applications of statistical mechanics there are algorithms that make these non convex optimization problems computationally feasible through approximations. Examples of these techniques are Belief Propagation algorithms [8], that uses a message-passing procedure reformulating the problem in a graphical way.

These algorithms are kind of ancestors of those based on Expectation Propagation (EP). We will explain more precisely the general case of EP in the next section.

In this work [7] an inference model has been introduced to consider the information obtained through the empirical distributions of the pixel difference variables, and in particular their non convex shape.

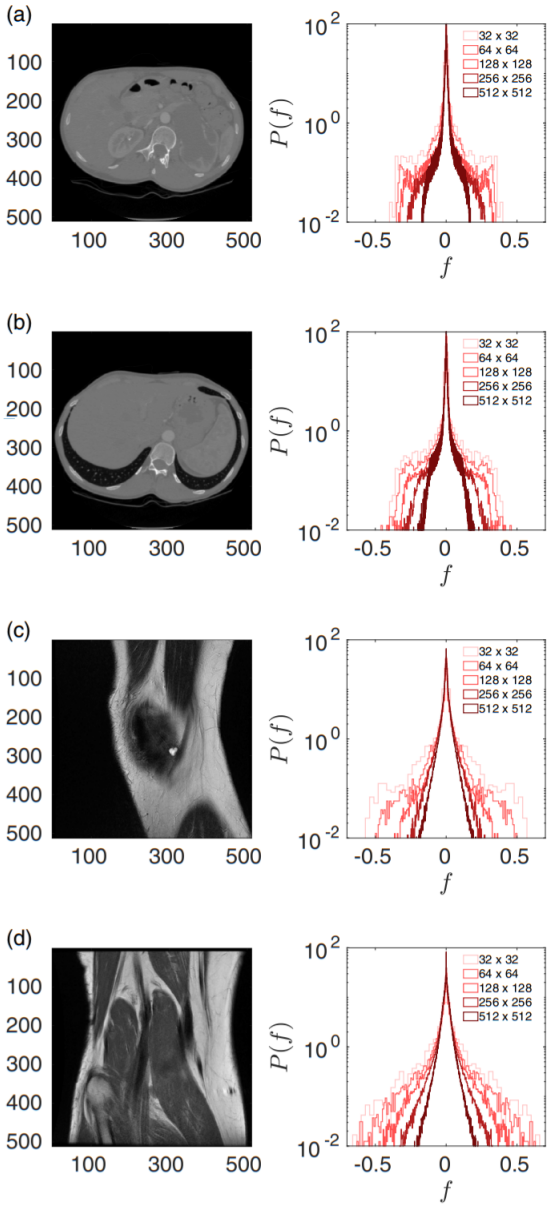

Figure 1: Computed tomography scans (left panels) and relative empirical distributions (right) of two different viewpoints of the abdomen (in (a) and (b) panels) or of a knee, as in (c) and (d). From the histogram it can be seen that by increasing the resolution of the image, from 64x64 to 512x512 pixels, the distribution has an increasingly narrow shape. Indeed, the statistics do not depend much on which organ is analyzed but on the coarse-graining of the image. Image taken from [7].

Rewriting the likelihood term $\mathscr{P}(\boldsymbol{p}|\boldsymbol{x}) \propto e^{-\frac{\beta}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{p})^2}$ in the standard form of a multivariate Gaussian, the posterior distribution reads:

$$\mathscr{P}(\boldsymbol{x}|\boldsymbol{p}) \propto \frac{1}{Z} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \prod_{i \in X \cup Y} \psi_i(z_i) \tag{12}$$

with precision matrix $\boldsymbol{\Sigma}^{-1} = \beta \boldsymbol{A}^T \boldsymbol{A}$ and mean $\boldsymbol{\mu} = \boldsymbol{\Sigma}\beta \boldsymbol{A}^T \boldsymbol{p}$. All priors $\psi_i(z_i)$ are both for the variables belonging to the set of pixel intensity variables $x_i \in X$and the difference ones $y_i \in Y$. If we deal with N pixel intensity variables, if $i < N$ the variables belong to the set of pixel intensity variables X, while if $i > N$ we are considering the set Y of difference variables.

The results of this work allows us to reduce the number of measurements to few projections and to increase the accuracy of the reconstructions: they also demonstrate the possibility of introducing empirical priors into the EP algorithm, opening up a new scenario. [9]

In fact the same reasoning could be used for other auxiliary variables, in some way connected with the intensity variables of the pixels through some linear transformation (we will see the Haar Transform case); so we can study the empirical behavior of these variables from a training set and impose prior with inferred parameters.

The difference we want to emphasize with respect to the case of difference variables is the way these priors are designated: in [9], all priors of all pairs of neighboring pixels have the same functional form and the same parameters, while in this work we will deal with specific priors for each auxiliary variable. We create a training dataset in order to infer the functional forms and the parameters associated with each prior.

We will discuss in detail how to proceed with empirical priors in Chapter 3.

2.2 EP method

The problem of reconstructing images belongs to the larger family of linear estimation problems: as already explained above (5), in these problems we have to solve a set of M linear equations in N unknowns, where M is the length of the vector of measurements.

We have seen that reformulating a linear estimation problem with a Bayesian approach certainly brings advantages; the drawback of Bayesian Inference is its excessive computational cost, especially in non-convex optimization problems.

A particularly powerful and flexible method inspired by statistical physics able to incorporate non convex prior information is Expectation Propagation (EP), developed for Bayesian inference problems in [10, 11, 12, 13].

Expectation Propagation is an iterative approach that approximates intractable posterior probability distributions. This efficient technique finds applications not only in image reconstruction, but in various linear estimation problems, as in this work [14], where EP was used to approximate the feasible space of cellular metabolic fluxes.

Let us briefly explain how EP generally works.

The "true" posterior distribution that we want to approximate is a multivariate Gaussian times the product of all uni-variate priors:

$$\mathscr{P}\left(\boldsymbol{x}|\boldsymbol{p}\right)=\frac{1}{Z}e^{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}\prod_{j}\psi_{j}\left(x_{j}\right) \tag{13}$$

where the only difference with respect to (12) is that we now only consider pixel intensity variables, just for the sake of simplicity.

Notice that in the multivariate Gaussian we can collect different terms, like the likelihood or a pair prior. There are different choices for the uni-variate prior terms $\psi_{j}\left(x_{j}\right)$, as for example the interval prior, the sparse prior or the binary one. These priors on the single pixel intensity value have to impose independent local constraints.

With the interval prior we assume an uniform measure on a given support $\left[x_{i}^{min},x_{i}^{max}\right]$:

$$\mathscr{P}_{0,int}^{(single)}\left(x_{i}\right)=\frac{\mathbb{I}\left[x_{i}^{min}\leq x_{i}\leq x_{i}^{max}\right]}{x_{i}^{max}-x_{i}^{min}} \tag{14}$$

where $\mathbb{I}$ is the indicator function which takes value 1 when x_i belongs to the interval $\left[x_{i}^{min},x_{i}^{max}\right]$ and value 0 otherwise.

An alternative to this prior is the sparse prior, useful for reconstruct images with monochromatic backgrounds:

$$\mathscr{P}_{0,sparse}^{(single)}\left(x_{i}\right)=\left[s\delta\left(x_{i}\right)+\left(1-s\right)\frac{\mathbb{I}\left[x_{i}^{min}\leq x_{i}\leq x_{i}^{max}\right]}{x_{i}^{max}-x_{i}^{min}}\right] \tag{15}$$

where the δ-function $\delta\left(x_{i}\right)$ is weighted by the sparseness parameter $s\in\left(0,1\right)$. This s is equal to the average fraction of background pixels in the image.

A third choice for the single variable prior is the binary prior, when only black and white colors are available (as in binary tomography) :

$$\mathscr{P}_{0,binary}^{(single)}\left(x_{i}\right)=\left[s\delta\left(x_{i}\right)+\left(1-s\right)\delta\left(x_{i}-1\right)\right] \tag{16}$$

Moreover, this approach can also be used when the image to be reconstructed contains three or more gray scales, as for example in this work [15].

For now let us consider generic single prior variable $\psi_{j}\left(x_{j}\right)$.

In EP we will consider the approximated posterior obtained replacing each single prior by a Gaussian distribution $\phi_{j}\left(x_{j}\right)$:

$$\mathscr{Q}\left(\boldsymbol{x}|\boldsymbol{p}\right)=\frac{1}{Z_{Q}}e^{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}\prod_{j}\phi_{j}\left(x_{j}\right) \tag{17}$$

where $\phi_{j}\left(x_{j}\right)=\frac{1}{\sqrt{2\pi b_{j}}}e^{-\frac{\left(x_{j}-a_{j}\right)^{2}}{2b_{j}}}$ and $\boldsymbol{a}=\left(a_{1},...,a_{N}\right)$ and $\boldsymbol{b}=\left(b_{1},...,b_{N}\right)$ are the mean and variance vectors of the uni-variate Gaussians of the approximated posterior.

Comparing this distribution with the "true" posterior in (13) we can see that $\mathscr{Q}(\mathbf{x}|\mathbf{p})$ is a Gaussian multivariate and therefore calculating the moments or the single variable marginals presents no particular problems, conversely to $\mathscr{P}(\boldsymbol{x}|\boldsymbol{p})$ that is analytically more difficult to deal with.

Now let us introduce the tilted-distribution for the i-th variable:

$$\mathscr{Q}^{(i)}(\boldsymbol{x}|\boldsymbol{p}) = \frac{1}{Z_{Q^{(i)}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \psi_i(x_i) \prod_{j \neq i} \phi_j(x_j) \tag{18}$$

The distribution with all Gaussian priors and the tilted one can be written in a similar form, where the only difference is in the i-th variable, because in the tilted distribution there is the exact prior $\psi_i(x_i)$, while in the fully approximated one there is the Gaussian prior $\phi_i(x_i)$:

$$\mathscr{Q}^{(i)}(\boldsymbol{x}|\boldsymbol{p}) \propto \left[e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \prod_{j \neq i} \phi_j(x_j) \right] \psi_i(x_i) \tag{19}$$

$$\mathscr{Q}(\boldsymbol{x}|\boldsymbol{p}) \propto \left[e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \prod_{j \neq i} \phi_j(x_j) \right] \phi_i(x_i) \tag{20}$$

In both formulas we can merge in a multivariate Gaussian all the priors except for the i-th variable, obtaining the so called "cavity Gaussian" $\mathscr{Q}^{\setminus i}$:

$$\mathscr{Q}^{\setminus i}(\boldsymbol{x}|\boldsymbol{p}) \propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \prod_{j \neq i} \phi_j(x_j) \propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T \boldsymbol{\Sigma}_{(i)}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} \tag{21}$$

where $\boldsymbol{\Sigma}_{(i)}^{-1}$ is the precision matrix and $\boldsymbol{\mu}^{(i)}$ is the vector of means.

Both quantities clearly depend on the i-th variable we are updating:

$$\boldsymbol{\Sigma}_{(i)}^{-1} = \beta \boldsymbol{A}^T \boldsymbol{A} + \boldsymbol{B}^{(i)} \tag{22}$$

$$\boldsymbol{\mu}^{(i)} = \boldsymbol{\Sigma}_{(i)} \left(\beta \boldsymbol{A}^T \boldsymbol{p} + \boldsymbol{B}^{(i)} \boldsymbol{a} \right) \tag{23}$$

where $\boldsymbol{B}^{(i)}$is a diagonal matrix whose elements are $B_{nn} = \frac{1}{b_n}$ for $n \neq i$ and $B_{ii} = 0$ (this zero entry is the "cavity").

We want to approximate $\psi_i(x_i)$ with the closest uni-variate Gaussian, so we need the values of a_i and b_i that best reproduce $\mathscr{P}(\boldsymbol{x}|\boldsymbol{p})$.

One possible approach would be minimizing the Kullback-Leibler distance between the exact prior ψ_i and the uni-variate Gaussian ϕ_i but this approach does not give in general good performances [14]; instead of approximating the prior itself, the EP algorithm approximates the effect of the prior on the full distribution.

To do this we minimize the Kullback-Leibler distance between two distributions that differ only in the i-th term: the approximated distribution $\mathscr{Q}(\boldsymbol{x}|\boldsymbol{p})$

with all uni-variate Gaussian priors (17) and the tilted one $\mathscr{Q}^{(i)}(\boldsymbol{x}|\boldsymbol{p})$, with only one non-Gaussian prior (18) .

$$a_i, b_i = argmin_{(a_i,b_i)} D_{KL}\left[\mathscr{Q}^{(i)}(\boldsymbol{x}|\boldsymbol{p})\,||\mathscr{Q}(\boldsymbol{x}|\boldsymbol{p})\right] \tag{24}$$

The equivalence between the minimization of the KL divergence and the following moment matching condition is explained in the appendix A:

$$< x_i >_{\mathscr{Q}^{(i)}(\mathbf{x})} = < x_i >_{\mathscr{Q}(\mathbf{x})} \tag{25}$$

$$< x_i^2 >_{\mathscr{Q}^{(i)}(\mathbf{x})} = < x_i^2 >_{\mathscr{Q}(\mathbf{x})} \tag{26}$$

where $< ... >_{\mathscr{Q}^{(i)}(\mathbf{x})}$ is the expectation value with respect to the tilted distribution and $< ... >_{\mathscr{Q}(\mathbf{x})}$ with respect to the Gaussian approximation.

The distributions (19) and (20), in the case of an interval prior (14) ,$\psi_i(x_i) = \frac{\mathbb{I}\left[x_i^{min} \leq x_i \leq x_i^{max}\right]}{x_i^{max} - x_i^{min}}$, would be respectively:

$$\mathscr{Q}(\boldsymbol{x}|\boldsymbol{p}) = \frac{1}{Z_Q} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T \boldsymbol{\Sigma}_{(i)}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} \frac{e^{-\frac{(x_i-a_i)^2}{2b_i}}}{\sqrt{2\pi b_i}} \tag{27}$$

$$\mathscr{Q}^{(i)}(\boldsymbol{x}|\boldsymbol{p}) = \frac{1}{Z_{Q^{(i)}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T \boldsymbol{\Sigma}_{(i)}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} \frac{\mathbb{I}\left[x_i^{min} \leq x_i \leq x_i^{max}\right]}{x_i^{max} - x_i^{min}} \tag{28}$$

The average with respect to the tilted distribution depends on the choice of the exact priors, while we can compute easily mean and variance of the full Gaussian distribution:

$$< x_i >_{\mathscr{Q}(\mathbf{x})} = \left(\frac{1}{\Sigma_{ii}^{(i)}} + \frac{1}{b_i}\right)^{-1} \left(\frac{\mu_i}{\Sigma_{ii}^{(i)}} + \frac{a_i}{b_i}\right) \tag{29}$$

$$< x_i^2 >_{\mathscr{Q}(\mathbf{x})} - < x_i >^2_{\mathscr{Q}(\mathbf{x})} = \left(\frac{1}{\Sigma_{ii}^{(i)}} + \frac{1}{b_i}\right)^{-1} \tag{30}$$

So if we apply the moment matching condition we will obtain the updating values of mean and variance of the Gaussian ϕ_i that approximates the exact prior ψ_i :

$$a_i = b_i \left[< x_i >_{\mathscr{Q}^{(i)}} \left(\frac{1}{b_i} + \frac{1}{\Sigma_{ii}}\right) - \frac{\mu_i}{\Sigma_{ii}}\right] \tag{31}$$

$$b_i = \left(\frac{1}{< x_i^2 >_{\mathscr{Q}^{(i)}} - < x_i >^2_{\mathscr{Q}^{(i)}}} - \frac{1}{\Sigma_{ii}}\right)^{-1} \tag{32}$$

Starting with initial values of a_i and b_i for each variable, these two equations are iterated until convergence, numerically reached when the error ε is smaller than a fixed threshold.

We define ε as the maximum of the sum of the differences between the first two moments:

$$\varepsilon = \max_i | < x_i >^{t+1}_{\mathcal{Q}^{(i)}} - < x_i >^{t}_{\mathcal{Q}^{(i)}} | + | < x_i^2 >^{t+1}_{\mathcal{Q}^{(i)}} - < x_i^2 >^{t}_{\mathcal{Q}^{(i)}} |. \tag{33}$$

This error ε tells us how the approximated distribution change in two consecutive iterations $t+1$ and t.

The general update expressions (31) and (32) are independent of the prior, which only affects the expectation values with respect to the tilted distribution $< ... >_{\mathcal{Q}^{(i)}(\mathbf{x})}$. Therefore the EP procedure presented so far is quite generic; we want to see now an example of how the update equations develop in the case of a specific prior.

Considering the distribution (28), with an interval prior for the pixel intensities, mean and variance will be:

$$< x_i >_{\mathcal{Q}^{(i)}} = \mu_i + \frac{\mathcal{N}\left(\frac{x_i^m-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right) - \mathcal{N}\left(\frac{x_i^M-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right)}{\Phi\left(\frac{x_i^M-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right) - \Phi\left(\frac{x_i^m-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right)} \Sigma_{ii}^{(i)} \tag{34}$$

$$< x_i^2 >_{\mathcal{Q}^{(i)}} - < x_i >^2_{\mathcal{Q}^{(i)}} = \Sigma_{ii}^{(i)} \left\{ 1 + \frac{\frac{x_i^m-\mu_i}{\Sigma_{ii}^{(i)}}\mathcal{N}\left(\frac{x_i^m-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right) - \frac{x_i^M-\mu_i}{\Sigma_{ii}^{(i)}}\mathcal{N}\left(\frac{x_i^M-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right)}{\Phi\left(\frac{x_i^M-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right) - \Phi\left(\frac{x_i^m-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right)} \right.$$

$$\left. - \left(\frac{\mathcal{N}\left(\frac{x_i^m-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right) - \mathcal{N}\left(\frac{x_i^M-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right)}{\Phi\left(\frac{x_i^M-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right) - \Phi\left(\frac{x_i^m-\mu_i}{\sqrt{\Sigma_{ii}^{(i)}}}\right)} \right)^2 \right\} \tag{35}$$

where $\mathcal{N}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $\Phi(x)$ is the cumulative function $\Phi(t;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} dx e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$.

We will see explicitly how these update equations become with other possible choices for the prior.

In the computation of the moments or the partition function of the tilted distribution it is useful to use the following simplification, regardless of the prior you choose.

Looking at (27) and (20) we can rewrite the tilted distribution as:

$$\mathcal{Q}^{(i)}(\boldsymbol{x}|\boldsymbol{p}) \propto \mathcal{Q}(\boldsymbol{x}|\boldsymbol{p}) \, \psi_i(x_i) \, \phi_i^{-1}(x_i) \tag{36}$$

in this way whenever we should perform integrals over the tilted distribution, we can exploit multivariate Gaussian properties for marginal distribution.

In fact in these integrals there is a multivariate Gaussian times a function of the i-th variable, so we can easily first integrate over all other variables except the i-th:

$$\begin{aligned} Z &= \int dx_i \int d\mathbf{x}_{/i} \mathcal{Q}\left(\boldsymbol{x}|\boldsymbol{p}\right) \psi_i\left(x_i\right) \phi_i^{-1}\left(x_i\right) \\ &= \int dx_i \psi_i\left(x_i\right) \phi_i^{-1}\left(x_i\right) \int d\mathbf{x}_{/i} \mathcal{Q}\left(\boldsymbol{x}|\boldsymbol{p}\right) \end{aligned} \tag{37}$$

The integral over all variables except the i-th is its marginal distribution.

We will denote with $\bar{\mu_i}$e $\overline{\Sigma}_{ii}$ the mean and variance of the i-th variable on which we are calculating the marginal according to the full Gaussian approximation. The normalization constant Z will be proportional to:

$$Z \propto \int dx_i \psi_i(x_i) e^{\frac{\left(x_i-a_i\right)^2}{2b_i}} e^{-\frac{1}{2}\frac{\left(x_i-\bar{\mu_i}\right)^2}{\overline{\Sigma}_{ii}}} \tag{38}$$

We can merge the exponential terms, obtaining:

$$Z \propto \int dx_i \psi_i\left(x_i\right) e^{-\frac{1}{2}\frac{\left(x_i-\mu_i\right)^2}{\Sigma_{ii}}} \tag{39}$$

where the entries of the covariance matrix and the mean are:

$$\Sigma_{ii} = \left(\frac{1}{\overline{\Sigma}_{ii}} - \frac{1}{b_i}\right)^{-1} \tag{40}$$

$$\mu_i = \left(\frac{1}{\overline{\Sigma}_{ii}} - \frac{1}{b_i}\right)^{-1} \left(\frac{\overline{\mu}_i}{\overline{\Sigma}_{ii}} - \frac{a_i}{b_i}\right) \tag{41}$$

we have greatly simplified the expression for the Z in (39) , as we now have only one Gaussian $e^{-\frac{1}{2}\frac{\left(x_i-\mu_i\right)^2}{\Sigma_{ii}}}$ (the cavity) times the exact prior $\psi_i\left(x_i\right)$.

This procedure that we have carried out for the Z can also be used in the computation of moments.

2.2.1 Computational cost of EP

The moment matching condition is present in every EP implementation. At each iteration step and for each pixel this condition requires inverting $\mathbf{\Sigma}_{(i)}^{-1}$, defined in (22). Indeed, as it can be seen from the update equations of the i-th element of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in (31) and (32), there is a direct dependence on the i-th element of the matrix $\mathbf{\Sigma}$.

Inverting a matrix $N \times N$ requires a number of operations that scale as $O(N^3)$. In EP algorithm we invert a matrix for each pixel, so the computational time scales as $O(N^4)$ per iteration step. One can reduce this cost to a single matrix inversion per iteration step using Eqs. (41) and (40), for the parameters of the cavity distributions.

2.3 EP with auxiliary variables

Until now we have considered EP with only the pixel intensity variables, so let us expand our analysis by including auxiliary variables $y_i \in Y$, obtained through some linear transformation $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$. Notice that the following computations are general and they not depend on the linear transformation $\boldsymbol{F}$.

As already mentioned for the general case, in each iteration step we need to compute the first two moments of the posterior distribution with all Gaussian priors and the tilted one.

In the case of the approximated posterior:

$$\mathscr{Q}\left(\mathbf{x}|\boldsymbol{p}\right) \propto \frac{1}{Z} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \delta\left(\boldsymbol{y}-\boldsymbol{F}\boldsymbol{x}\right) \prod_{j\in X} e^{-\frac{\left(x_j-a_j\right)^2}{2b_j}} \prod_{j\in Y} e^{-\frac{\left(\boldsymbol{e}_j^T \boldsymbol{F}\boldsymbol{x}-a_j\right)^2}{2b_j}} \tag{42}$$

where $\boldsymbol{e}_j \in \mathbb{R}^N$ is the vector of the standard basis, whose components are all zero, except the j-th that equals 1. This will boil down to compute mean and covariance of a multivariate Gaussian distribution, but we must distinguish the two sets of variables.

Rearranging the Gaussian prior for the auxiliary variables :

$$\frac{\left(\left(\boldsymbol{e}_j^T \boldsymbol{F}\right)\boldsymbol{x}-a_j\right)^2}{b_j} = \boldsymbol{x}^T \frac{1}{b_j}\left(\boldsymbol{F}^T \boldsymbol{e}_j \boldsymbol{e}_j^T \boldsymbol{F}\right)\boldsymbol{x} - \frac{2}{b_j}\left(a_j \boldsymbol{e}_j^T \boldsymbol{F}\right)\boldsymbol{x} + const \tag{43}$$

In this way we can generalize and consider pixel and auxiliary variables together:

$$\begin{aligned} -\sum_{j\in X\cup Y} \frac{\left(x_j-a_j\right)^2}{2b_j} = \frac{1}{2}\boldsymbol{x}^T \left(\sum_{j\in X} \frac{1}{b_j}\boldsymbol{e}_j\boldsymbol{e}_j^T + \sum_{j\in Y} \frac{1}{b_j}\left(\boldsymbol{F}^T \boldsymbol{e}_j \boldsymbol{e}_j^T \boldsymbol{F}\right)\right)\boldsymbol{x} + \\ + \boldsymbol{x}^T \left(\sum_{j\in X} \frac{a_j}{b_j}\boldsymbol{e}_j + \boldsymbol{F}^T \sum_{j\in Y} \frac{a_j}{b_j}\boldsymbol{e}_j\right) + const \end{aligned} \tag{44}$$

At each step of EP the equations will be updated according to:

$$\begin{cases} \boldsymbol{A}' = \boldsymbol{\Sigma}^{-1} + \sum_{j\in X} \frac{1}{b_j}\boldsymbol{e}_j\boldsymbol{e}_j^T + \sum_{j\in Y} \frac{1}{b_j}\left(\boldsymbol{F}^T\boldsymbol{e}_j\right)\left(\boldsymbol{F}^T\boldsymbol{e}_j\right)^T \\ \boldsymbol{c}' = \boldsymbol{c} + \sum_{j\in X} \frac{a_j}{b_j}\boldsymbol{e}_j + \boldsymbol{F}^T \sum_{j\in Y} \frac{a_j}{b_j}\boldsymbol{e}_j \\ \boldsymbol{\Sigma}' = \boldsymbol{A}'^{-1} \\ \boldsymbol{\mu} = \boldsymbol{\Sigma}'\boldsymbol{c}' \end{cases}$$

So we get the multivariate Gaussian:

$$\mathscr{Q}\left(\boldsymbol{x}|\boldsymbol{p}\right) \propto \frac{1}{Z} e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}'\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{c}'} \tag{45}$$

The average with respect to the tilted distribution depends on the choice of the exact priors, instead of mean and variance of the full Gaussian approximated distribution:

$$< x_i >_{\mathcal{Q}(\mathbf{x})} = \int d\boldsymbol{x} \mathcal{Q}(\boldsymbol{x}|\boldsymbol{p}) x_i \propto \int d\boldsymbol{x} e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}' \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{c}'} x_i \tag{46}$$

$$< x_i^2 >_{\mathcal{Q}(\mathbf{x})} \propto \int d\boldsymbol{x} e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}' \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{c}'} x_i^2 \tag{47}$$

So the moments of the approximated posterior distribution for the pixel intensity variables are:

$$< x_i^2 >_{\mathcal{Q}(\mathbf{x})} - < x_i >^2_{\mathcal{Q}(\mathbf{x})} = \boldsymbol{\Sigma}'_{ii} \tag{48}$$

$$< x_i >_{\mathcal{Q}(\mathbf{x})} = \boldsymbol{\mu}_i \tag{49}$$

while for the auxiliary variables:

$$< y_i^2 >_{\mathcal{Q}(\mathbf{x})} - < y_i >^2_{\mathcal{Q}(\mathbf{x})} = \boldsymbol{e}_i^T \boldsymbol{F} \boldsymbol{\Sigma}' \boldsymbol{F}^T \boldsymbol{e}_i \tag{50}$$

$$< y_i >_{\mathcal{Q}(\mathbf{x})} = \boldsymbol{e}_i^T \boldsymbol{F} \boldsymbol{\mu} \tag{51}$$

in this way the introduction of auxiliary variables does not make the EP algorithm slower because we can calculate the moments of the auxiliary variables from the statistics of the pixel variables.

As was shown in the general EP method with only pixel intensity variables, the key step of each iteration of the algorithm is the moment matching condition, when we impose the equality of moments of the tilted and the fully approximated distributions.

So now let us compute the moments of the tilted distribution, where the only difference with respect to (42) is one uni-variate Gaussian replaced by an exact prior.

The subsequent calculations depend on the choice of the exact prior $\psi_i(x_i)$, and obviously it is different as we deal with intensity or auxiliary variables.

In the next section, we show a possible choice for the auxiliary variables.

Let us introduce the difference variables: they are defined as the difference between the intensity of neighboring pixels, $y_{ij} = x_i - x_j$. The new variables take value in the interval $[x_{min} - x_{max}, x_{max} - x_{min}]$.

One can prove that the linear operator $\boldsymbol{F}$ generating these difference variables in $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$ is the incidence matrix $\boldsymbol{F} = \boldsymbol{R}$. We can imagine our image as an indirect graph with N vertices and N_e edges: in the column of edge e of the matrix $\boldsymbol{R}$, there is one 1 in the row corresponding to one vertex of e and one -1 in the row corresponding to the other vertex of e, and all other rows have 0.

The empirical distribution of these variables shows that they can be well fitted by a function of the form:

$$\mathscr{P}^{BG}(y_i) \propto \rho\delta(y_i) + (1-\rho)\frac{e^{-\frac{y_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \tag{52}$$

this is the so called Bernoulli-Gaussian prior where ϱ is the sparse parameter and σ^2 is the variance.

If we choose an interval prior (14) for the intensity variables and a Bernoulli-Gaussian for the auxiliary ones, the "true" posterior probability is:

$$\mathscr{P}(\mathbf{x}|\mathbf{p}) \propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}\prod_{i\in X}\mathbb{I}\left[x_i^{min}\leq x_i\leq x_i^{max}\right]\prod_{j\in Y}\left\{(1-\rho)\delta(y_j)+\rho\frac{e^{-\frac{y_j^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right\} \tag{53}$$

while we have already written the fully approximated distribution (42).

Since now the exact prior is different depending on the variable (intensity or auxiliary), we will deal with two tilted distributions:

$$Q^{(i)}(x_i|\boldsymbol{p}) \propto e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\Sigma_{ii}}}\frac{\mathbb{I}\left[x_i^{min}\leq x_i\leq x_i^{max}\right]}{x_i^{max}-x_i^{min}} \tag{54}$$

$$Q^{(i)}(y_i|\boldsymbol{p}) \propto e^{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\Sigma_{ii}}}\left\{(1-\rho)\delta(y_i)+\rho\frac{e^{-\frac{y_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right\} \tag{55}$$

We have already seen mean and variance in the case of pixel intensity variables in (34) and (35); now we will compute the moments of the tilted distribution in the case of auxiliary variables.

We report the formulas for the Bernoulli-Gaussian prior calculated in this work [9] :

$$< y_i >_{\mathscr{Q}^{(i)}} = \frac{1}{Z_{Q^{(i)}}}\int dy_i y_i e^{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\Sigma_{ii}}}\left\{(1-\rho)\delta(y_i)+\rho\frac{e^{-\frac{y_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right\} \tag{56}$$

$$< y_i >_{\mathscr{Q}^{(i)}} = \frac{1}{Z_{Q^{(i)}}}\varrho\sqrt{\frac{\Sigma_{ii}}{\Sigma_{ii}+\lambda}}\frac{\lambda\mu_i}{\Sigma_{ii}+\lambda} \tag{57}$$

$$< y_i^2 >_{\mathscr{Q}^{(i)}} = \frac{1}{Z_{Q^{(i)}}}\int dy_i y_i^2 e^{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\Sigma_{ii}}}\left\{(1-\rho)\delta(y_i)+\rho\frac{e^{-\frac{y_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right\} \tag{58}$$

$$< y_i^2 >_{\mathscr{Q}^{(i)}} = \frac{1}{Z_{Q^{(i)}}}\varrho\sqrt{\frac{\Sigma_{ii}}{\Sigma_{ii}+\lambda}}\left[\frac{\lambda\Sigma_{ii}}{\Sigma_{ii}+\lambda}+\left(\frac{\lambda\mu_i}{\Sigma_{ii}+\lambda}\right)^2\right] \tag{59}$$

And the partition function Z is:

$$Z_{Q^{(i)}} = (1-\varrho)e^{\frac{-\lambda\mu_i^2}{2\Sigma_{ii}(\Sigma_{ii}+\lambda)}} + \varrho\sqrt{\frac{\Sigma_{ii}}{\Sigma_{ii}+\lambda}} \tag{60}$$

3 Results

3.1 Empirical priors

In the chapter of EP with auxiliary variables we have seen an example of a linear function of the pixel intensity variables, namely the difference between neighboring pixels. In [7] the empirical distribution of these variables has been analyzed from a set of tomographic images of a certain portion of the body. Then the prior (52) was designed estimating the parameters ϱ and σ with an "expectation-maximization" [16] step in the EP algorithm, as explained in [7].

The results obtained have demonstrated the effectiveness of introducing auxiliary variables and the possibility of exploiting new information from some function of the pixel variables.

In a similar way to what was done with the difference variables, now we want to use other auxiliary variables, in some way connected with the intensity variables through a linear transformation. The main improvement is that we do not want to estimate the same parameters (or also the same functional form) for all the priors of the auxiliary variables.

In this work we create a training dataset and, after applying a linear operator, we infer both the functional form and the parameters of each prior of the auxiliary variables.

We need to create an image dataset because it is not easy to use EP algorithm with large images, as the real-world tomographic images (large in the sense of high resolution). In fact we have to invert matrices with large dimensions and the computational cost of EP would increase significantly.

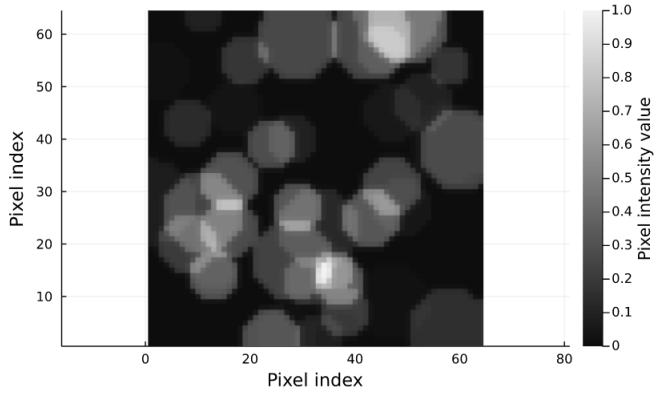

Figure 2: Dataset image example in gray-scale with linear size $L = 64$.

To this purpose, a training dataset of 20.000 images was created: we must think of each of these images as a grid of numbers (a matrix) where each number corresponds to a shade of gray; therefore, each entry varies from zero (black) to

one (white). We try to reproduce the typical features of tomographic images, which present the tissues of the organs in grayscale on a black background. The dataset images are created by inserting a given number of circles on a black background. We set the number of these circles equal to $n_{circles} = 40$: for each of them the color is sampled from a uniform distribution in $[0, 1]$, corresponding to the grayscale range, while the radius is sampled from a Gaussian distribution of zero mean and standard deviation $\sigma = \frac{L}{unif(4,6,7,10)*\sqrt{n_{circles}}}$.

Finally, we set the linear size L of the images (e.g. 64 or 128) and we create the training dataset of 20.000 images. If we set to $L = 64$ the linear size of the image, we deal with $N = L \times L = 4096$ pixel variables.

The auxiliary variables are obtained through a linear transformation:

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} \tag{61}$$

where $\boldsymbol{F}$ is the transform matrix, $\boldsymbol{x} = vec(\boldsymbol{X})$ is the vector containing the pixel intensities of an image and $\boldsymbol{y} = vec(\boldsymbol{Y})$ is the corresponding vector of auxiliary variables.

Then, we observe the behavior of the histograms of the auxiliary variables and we infer the functional form and the parameters for each transformed variable, fitting the data. We will show more in detail the fitting in the following subsections, after defining the linear operator we have used, the Haar transform matrix.

This approach is not very different from that used for the difference variables, in fact we can also define a linear transformation $\boldsymbol{F}$ for these auxiliary variables $\boldsymbol{F} = \boldsymbol{R}$, where $\boldsymbol{R}$ is the incidence matrix.

3.2 Haar case

The Haar transform is the first discrete wavelet transform, invented in 1909 by Alfred Haar. The wavelet transform is the representation of a signal through orthonormal series generated by a wavelet, similar to what happens in Fourier analysis, where, however, the orthonormal series are trigonometric functions. The main difference is that wavelets are localized in both time and frequency while the standard Fourier transform is localized only in frequency.

We cannot localize sharply a signal in both time and frequency domains: at a fixed time, we can measure an original signal and obtain the exact amplitude but zero information about the frequency, while the Fourier transform give us all information about the frequency and maximum uncertainty about time.

The wavelet transform is a sort of intermediate case between the two limit situations just described: wavelets analysis give us information both in time and frequency domain, but with less accuracy in the frequency spectrum with respect to the Fourier analysis.

Let us introduce the Haar functions $h_k(z)$ defined for $z \in [0, 1]$, where k is the row index and can be decomposed uniquely in two integers p, q:

$$k = 2^p + q - 1 \tag{62}$$

with $k = 0, 1, ..., L-1$, $L = 2^n$, and where $0 \leq p \leq n-1$, $0 \leq q \leq 2^p$for $p \neq 0$ and $q = 0$ or $q = 1$ for $p = 0$.

The Haar functions are:

$$h_0(z) \equiv h_{00}(z) = \frac{1}{\sqrt{L}}, \tag{63}$$

$$h_k(z) \equiv h_{pq}(z) = \frac{1}{\sqrt{L}} \begin{cases} 2^{\frac{p}{2}} & \frac{q-1}{2^p} \leq z < \frac{q-\frac{1}{2}}{2^p} \\ -2^{\frac{p}{2}} & \frac{q-\frac{1}{2}}{2^p} \leq z < \frac{q}{2^p} \\ 0 & otherwise \end{cases} \tag{64}$$

The Haar transform matrix of order L consists of rows made up of $h_k(z)$; for example if $L = 8$ the Haar transform matrix consists of 8 rows: $h_0(z), h_1(z), ..., h_8(z)$:

$$\boldsymbol{H} = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix} \tag{65}$$

$\boldsymbol{H}$ is orthogonal, meaning that $\boldsymbol{H}^{-1} = \boldsymbol{H}^T$. In this work we set $L = 64$ in order to obtain a suitable Haar transform matrix $\boldsymbol{H} \in \mathbb{R}^{LxL}$ for our dataset images: in the figure 3 we visualize an heatmap of the transform matrix to get an idea of the 2D arrangement of the Haar coefficients.

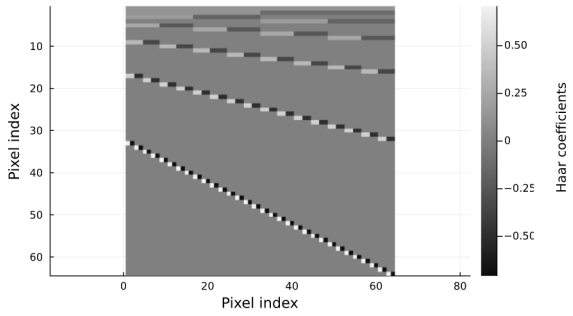

Figure 3: Heatmap of the Haar transform matrix for L=64.

The Haar transform is applied to each pixel of the image as:

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X}\boldsymbol{H}^T \tag{66}$$

where $\boldsymbol{x} = vec(\boldsymbol{X})$ is the vector containing all pixel intensities and $\boldsymbol{y} = vec(\boldsymbol{Y})$ is the corresponding vector of transformed variables.

We want to define the linear operator that can be applied to the vector $\boldsymbol{x}$:

$$\boldsymbol{y} = (\boldsymbol{H} \otimes \boldsymbol{H})\, \boldsymbol{x} \tag{67}$$

where $\otimes$ is the Kronecker product.

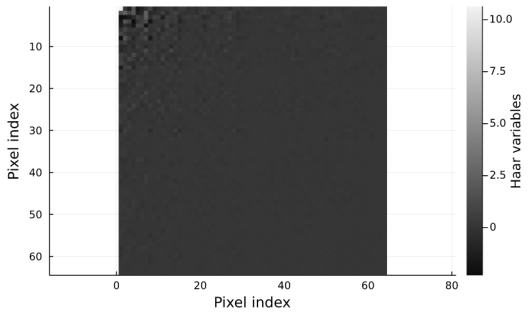

Figure 4: Dataset image after applying the Haar transform; the values of many auxiliary variables are close to zero.

Looking at this image 4, it is quite easy to see the sparsity of the auxiliary variables , the Haar coefficients. A huge number of these variables thus obtained assume values close to zero, as confirmed by the peak present in the histogram in figure 5 (here the quantities in the y-axis represent empirical frequencies).

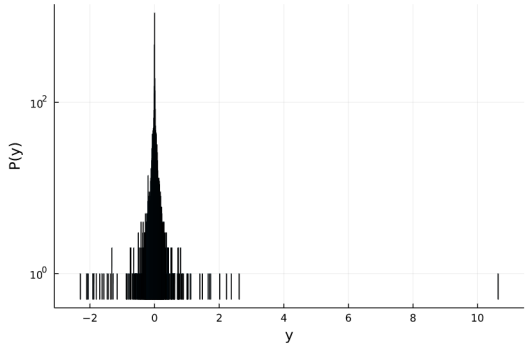

Figure 5: Log-scale histogram of the Haar auxiliary variables of a single image: $P(y)$ is the absolute frequency, with no normalization.

Moreover, comparing different transformed images it has been noticed that the informative part of each image resides in the upper left corner, while the opposite corner is the one where there are more transformed variables equal to zero.

We will show in the next section a possible application in data compression of the Haar transform to show that we can, after applying the transform on the pixel variables, set a large number of auxiliary variables equal to zero and obtain good reconstruction results by doing the inverse transform.

3.2.1 The Haar transform to compress data

We take an image of the dataset and apply the Haar transform: then we set some of these auxiliary variables equal to zero and apply the inverse Haar transform to return an approximation of the image. We decimate the Haar coefficients according to three criteria specified in the following and we study the behavior of the reconstructed image as a function of the number of non-zero coefficients used within the decompression.

In order to clarify the problem, we pass to a representation in terms of vectors; if $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$ is the vector of the auxiliary variables, we need the inverse of the linear operator $\boldsymbol{F} = \boldsymbol{H} \otimes \boldsymbol{H}$.

Thanks to the property of the Kronecker product, it is not difficult to prove that the inverse of $\boldsymbol{F}$ is its transpose.

So after having set some variables equal to zero, we return to the starting image with:

$$\bar{\boldsymbol{x}} = \boldsymbol{F}^T \bar{\boldsymbol{y}} \tag{68}$$

where we denote with $\bar{\boldsymbol{y}}$ the vector of auxiliary variables with some of them set to zero and $\bar{\boldsymbol{x}}$ the vector of the image obtained in this way.

Clearly if we do not set to zero any variable, we obtain exactly the starting image $\boldsymbol{x}$.

In the following we will explain the three selection methods that were used to choose the order in which the variables were set equal to zero:

1. the *index-based* selection starts from the variables of the lower right corner up to the upper left corner, where there are the variables that are further away from zero, i.e. the most informative ones.

2. in the *coefficient-based* method variables closest to zero are cut out and the others gradually.

3. *Prior based.* As regards the third selection method, however, it is necessary to consider not only the single reference image of the dataset, but the entire training dataset. In fact we collect the 20,000 values of each auxiliary variable in a histogram: if we normalize by sum of weights only, we obtain a discrete probability function for each bin (their relative frequency), as shown in figure 6. Then we consider the height of the bin in zero as the value of the probability of having zero: the selection order will

first "silence" the variables with the largest values of this probability and then all the others in order up to its smallest value.

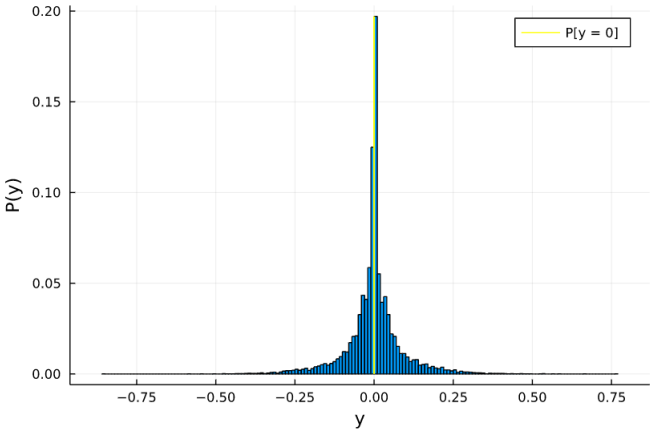

Figure 6: Empirical histogram of an auxiliary variable (index 500): the normalization is by sum of weights only, so here $P(y)$ represents the discrete probability of each bin, with its value at zero highlighted.

We define the error with which we compare the reconstructed images after having obscured a certain number of variables:

$$\epsilon = \frac{|\boldsymbol{x} - \bar{\boldsymbol{x}}|}{N} \tag{69}$$

where $N = 4096$ is the number of auxiliary variables and $\bar{\boldsymbol{x}}$ is the vector of the image obtained after having set equal to zero some auxiliary variables.

In figure 7 we plot the error ϵ as a function of K, the fraction of auxiliary variables used to reconstruct the starting image.

When we decrease the amount of auxiliary variables, and so the information per pixel used, the reconstruction error with the prior-based method increases with less slope than the other two methods.

So, the prior-based method obtain better results: in particular with this method it is possible to reduce the number of auxiliary variables used to reconstruct the original image and to obtain however a good resolution in the image, as can be seen in the image reconstructions in figure 8.

In table 1 we show the fraction K of components used and the relative errors ϵ of the four reconstructions of image 8. The auxiliary variables used are sorted with the prior-based method. The results show that an acceptable reconstruction error (of order $\sim 10^{-4}$) is obtained even for a large number of variables set equal to zero: this is a particular feature of the Haar variables.

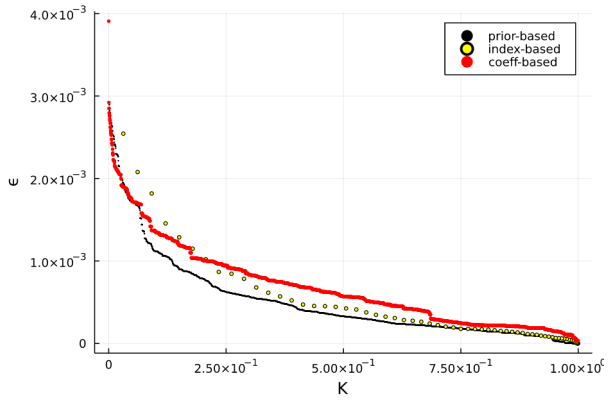

Figure 7: The error ϵ as a function of K, the fraction of components used to reconstruct the image, for the three methods described above: prior-based, index-based and coefficient-based.

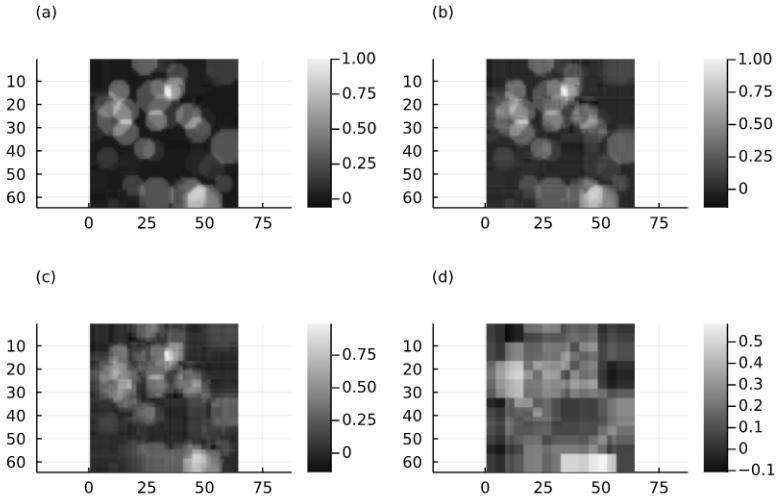

Figure 8: Image reconstructions using the inverse of the transform matrix after silencing a certain number of auxiliary variables, sorted with the prior-based method. In the (a) panel we have a very good reconstruction as we used enough auxiliary variables. Also in the (b) panel the error is acceptable even if we used about half of the auxiliary variables. Note the (c) panel with an error of order $\sim 10^{-4}$ using only a fraction $K = 0.26$ of auxiliary variables. The (d) panel is the only one with an error of the order 10^{-3}, as can be clearly seen from the inaccurate reconstruction.

	K	ϵ
(a)	0.75	2×10^{-4}
(b)	0.51	5×10^{-4}
(c)	0.26	9×10^{-4}
(d)	0.02	2×10^{-3}

Table 1: Reconstruction error ϵ and fraction of auxiliary variables used K in the four images in figure 8.

3.3 Fitting

Curve fitting corresponds to find the functional forms and the relative parameters of a mathematical function ("mapping function") that best fits a series of data.

After applying the Haar transform to all the 20.000 images of the training set, we have to fit a suitable function to each histogram of the auxiliary variables. We normalize all histograms by sum of weights and bin sizes: in this way the histogram represents a probabilistic density function (PDF).

This means that we have to use probability distributions as mapping functions: in the following we will define the distributions used in order to fit the data.

Gaussian functions are widely used in statistics, so if we choose this function:

$$\mathscr{P}^{GA}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \tag{70}$$

we need to estimate two parameters, the mean μ and the variance σ^2.

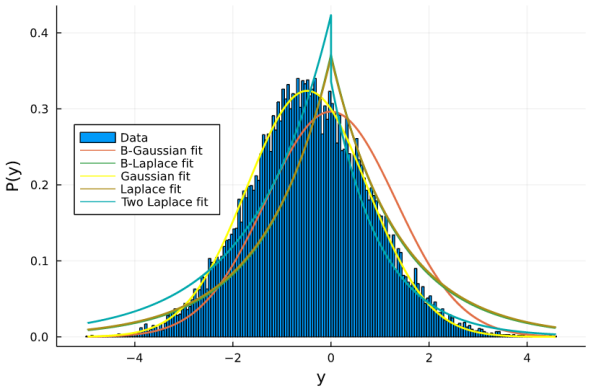

Figure 9: Empirical histogram of an auxiliary variable: the normalization is given by summing the weights and bins size, so here $P(y)$ represents a probabilistic density function (PDF), that can be fitted by distributions. The function that best fits this data is the Gaussian distribution.

Another possible mapping function is the Laplace distribution:

$$\mathscr{P}^{LA}(y) = \frac{1}{2}\lambda e^{-\lambda|y|} \tag{71}$$

this distribution can be seen as two exponential distributions joined together back-to-back. In the case of zero mean there is only the parameter λ to infer and the distribution is symmetrical with respect to the y axis. An example of Gaussian fitting is displayed in figure 9, while a Laplace one is shown in figure 10: in the empirical histograms of both auxiliary variables there is no peak, so we can use these unimodal distributions to fit the data.

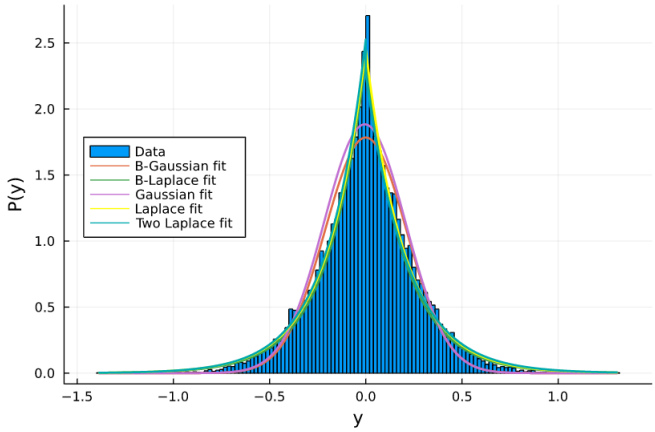

Figure 10: Empirical histogram of an auxiliary variable, where $P(y)$ represents a probabilistic density function (PDF): the best curve fitting is given by the Laplace distribution.

To take into account the possibility of asymmetric behaviors, let us define the asymmetric Laplace distribution:

$$\mathscr{P}^{2L}(y) = \mathbb{I}[y > 0]\,\varrho e^{-\lambda_1|y|} + \mathbb{I}[y < 0]\,(1-\varrho)\,e^{-\lambda_2|y|} \tag{72}$$

the two parameters λ_1 and λ_2 to be inferred allow us to modulate different trends for positive or negative arguments, weighted by the ϱ parameter. An example of an auxiliary variables with this behavior is shown in figure 11.

All three distributions written above are uni-modal, there is only one value that appears with the maximum frequency. In the next we will consider two bimodal distributions, made up by a linear combination with a delta function in zero and an uni-modal distribution (Gaussian or Laplace). The delta function can modulate really sharp distributions.

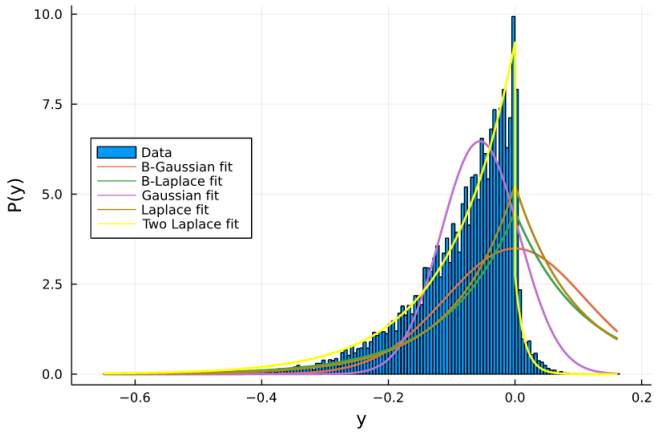

Figure 11: Empirical histogram of an auxiliary variable, where $P(y)$ represents a probabilistic density function (PDF): the curve the best approximates this data is the asymmetric Laplace distribution.

In the Bernoulli-Gaussian distribution, the parameters ϱ and σ will be estimated for each auxiliary variable. So we will use:

$$\mathscr{P}^{BG}(y) \propto \rho\delta(y) + (1-\rho)\frac{e^{-\frac{y^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \tag{73}$$

We have also fit the data with a combination of the delta function and the Laplace distribution (Bernoulli-Laplace prior), and the parameters to estimate are ϱ and λ :

$$\mathscr{P}^{BL}(y) \propto \rho\delta(y) + (1-\rho)\frac{1}{2}\lambda e^{-\lambda|y|} \tag{74}$$

Notice that in our Haar transform case we deal with $N = 4096$ intensity variables and for each of them the histogram of the relative auxiliary one takes into account information collected in the whole training set (20.000 images). An example of these histograms is shown in figure 12, with the relative curve fitting.

For these five distributions we fit each parameter and then we compute the norm of the these functions with respect to the observed data, in order to compare their accuracy. In some cases, however, comparing bimodal and unimodal distributions using the norm can produce inaccurate results in the choice of priors.

So we introduce also another parameter κ which indicate the presence of a peak in the histogram of the data: this is defined as the absolute value of the difference between the histogram weight in zero and its neighbors. According to this value we can distinguish between uni-modal (Gaussian or Laplace) and bimodal distributions (Bernoulli-Gaussian and Bernoulli-Laplace). For example

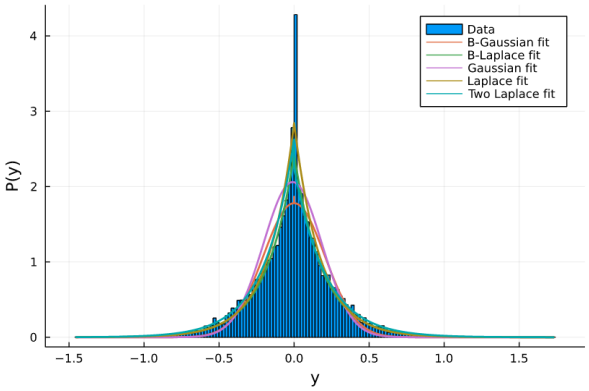

Figure 12: Empirical histogram of an auxiliary variable: the normalization is given by summing the weights and bins size, so here $P(y)$ represents a probabilistic density function (PDF), that can be fitted by distributions. Unlike the previous plots, now the distributions that best fit the data are the bimodal ones, and in particular in this case the Bernoulli-Laplace.

in the figure 12 the curve fitting of an auxiliary variable is displayed: the distribution with the smallest norm would be a unimodal distribution, in contrast to the presence of the peak clearly visible from the plot. With the κ parameter discrimination it will be chosen instead a bimodal distribution (in this case the Bernoulli-Laplace one).

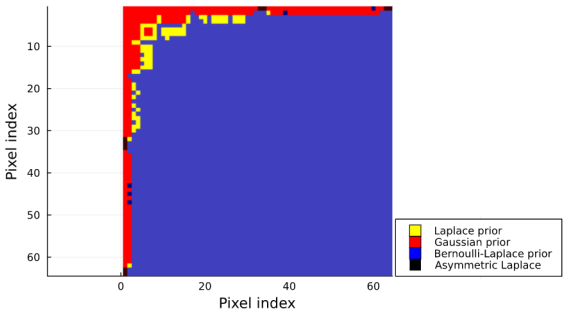

Figure 13: Priors heatmap of auxiliary variables. The informative angle is well described by Gaussian priors (in red) : moving away from this, the empirical distributions become more peaked and are well described by Laplace prior (in yellow). At a certain point only bimodal distributions (blue) can describe the trends of the steeper auxiliary variables.

We plot in figure 13 an heatmap with the prior chosen for each auxiliary variable. In this way it is easy to see that uni-modal auxiliary variables that do

not have a peak are found mainly in the upper left corner: these are fitted by Gaussian and Laplace distributions. Moving away from this angle, on the other hand, the variables increase more and more with a peak of values around zero, and they need Spike and Slab bimodal distributions.

There is also a particular case of asymmetric histograms, fitted by the asymmetric Laplace distribution.

In appendix B we report the calculations of the moments of the tilted distribution with respect to the various Laplace priors: the formulas thus computed have been implemented in the EP algorithm, but they are unstable from a numeric point of view because very often the normalization of the tilted distributions (which contain the cumulative density function of a Gaussian density) go to very small numbers. We tried to solve this problem by replacing the cumulative distribution with an expansion but EP algorithm with these priors continues to be numerically unstable.

To avoid this problem we unfortunately need to replace the Laplace, asymmetric Laplace and Bernoulli-Laplace priors. In the figure 14 we can see the new heatmap of priors we will use in the reconstructions.

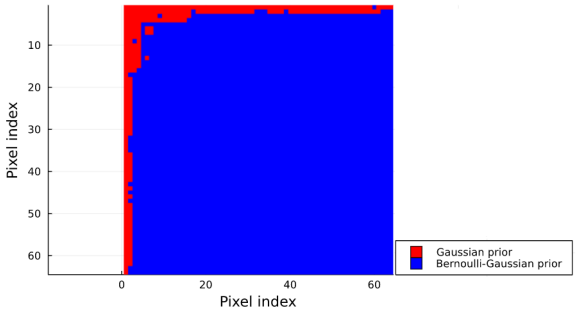

Figure 14: Priors heatmap excluding the Laplace priors, with red Gaussian priors and blue Bernoulli-Gaussian priors.

The asymmetric Laplace has been replace by a Gaussian prior. Moreover the Bernoulli-Laplace has been replaced by the Bernoulli-Gaussian prior, because the error slightly differs between bimodal distributions.

We substitute the Laplace priors comparing the norm of the Gaussian and BG priors with respect to the observed data.

3.4 Reconstructions

In this section we report the results of some reconstructions carried out using the EP algorithm: after describing the measurement process and the implementations of the algorithm, we compare the performances of the different techniques used. In fact, we implement several approaches depending on the auxiliary variables used and whether the parameters of the relative priors are estimated

through the use of the training set or not. Therefore in a first analysis we report the performances of these EP-based algorithms differing in the use of the auxiliary variables.

Then we compare our best performers against other two methods: total variation (TV) and simultaneous iterations reconstruction technique (SIRT).

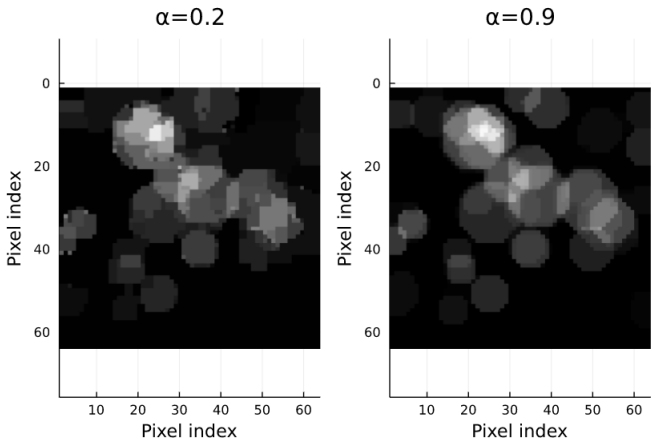

Figure 15: Image reconstructions of EP algorithm in the two measurement regimes: the sampling rate is $\alpha = 0.2$ in the left panel, while in the right one it is equal to $\alpha = 0.9$. You can see the good accuracy of the reconstruction in the high measurements regime.

We run EP algorithm in different measurements regimes, identified by the parameter $\alpha = \frac{M}{N}$, the sampling rate. In this image 15 we report the reconstructions of the same image in the two extreme situations: the lowest and highest measurement regimes, showing the clearly different performances of the EP algorithm. For these reconstructions we have used *EP-diff*, explained later in the sub-chapter about the different implementations of the EP algorithm.

3.4.1 Measurement process

The measurement process is the mathematical description of the process of acquiring the information obtained from the scanners device. We want to design suitable projection matrix $\boldsymbol{A}$ that stores these information.

Clearly, the characteristics of the scanners device must be matched by the matrices. There are different types of detectors with different functionality and acquisition methods but a complete survey of this topic is beyond the scope of this thesis (see [1] for details).

We will use two different methods to construct the projection matrices: for the first experiments the projection matrix $\boldsymbol{A}$ is built using single ray projections with uniformly chosen random directions in $[0, 2\pi]$. We recall that $\boldsymbol{A} \in \mathbb{R}^{M \mathrm{x} \mathrm{N}}$

and each element A_{ij} is the length of the portion of ray i passing through pixel j. Instead to compare the performance of EP with the other methods (TV and SIRT) we will reconstruct the images using a projection matrix representing a 2-D parallel beam geometry, as can be in the figure 16. In this projection matrix we can tune the distances between detectors (changing the number of parallel X-rays) and the projection angles (influencing the total number of X-rays).

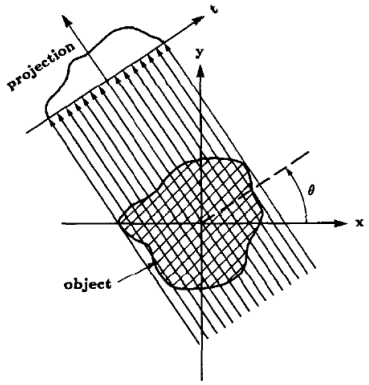

Figure 16: 2-D parallel beam geometry: the main difference with respect to the projection matrix used in the first experiment is that we have a certain number of parallel X-rays for a given angle. Image taken from [17].

Each row in this matrix represents a single ray and the number of rows M in the projection matrix depends on the sampling rate α, so it will change according to which measurement regime we are considering for that given reconstruction. If we sum all rows of a given pixel (namely a given column index of this matrix), we can get an idea of the total intensity of X-rays passing through each pixel for the two different projection matrices, as can be seen in figure 17. When the projection matrix represents a parallel-beam geometry we have more information in the central part of the image, where the rays at various angles intersect more with each other with respect to when the direction of rays is chosen uniformly in the range $[0, 2\pi]$.

3.4.2 EP implementations

In this section we describe the six different implementations of the EP algorithm used in the reconstructions.

- *EP-int* uses only the pixel intensity variables as described in section 2.2, each of them described with a flat interval prior $\psi_j(x_j)$ (14). The following methods are distinguished by the various auxiliary variables used. The difference therefore lies in the term of the priors of these variables, since all methods have an interval prior over the intensity variables. In the chapter 2.2 "EP method" we explained how EP works in this case in order

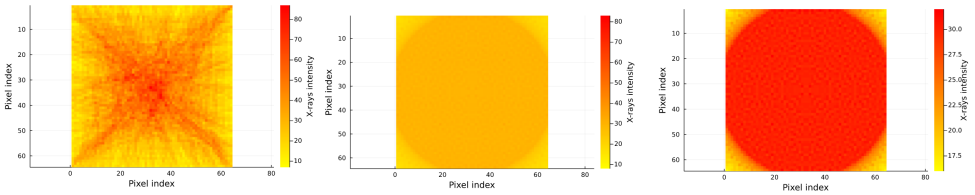

Figure 17: Heatmap showing the total information per pixel of different projection matrices: the left panel shows a superposition of rays whose direction is chosen uniformly at random in the range $[0, 2\pi]$, the center and right panel depict the overlap of multiple parallel rays. In the right plot we reduce the range of the colorbar to get an idea of this parallel beam geometry.

to approximate this "true" posterior distribution:

$$\mathscr{P}(\boldsymbol{x}|\boldsymbol{p}) = \frac{1}{Z} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \prod_j \psi_j(x_j) \tag{75}$$

where $\boldsymbol{\Sigma}^{-1} = \beta \boldsymbol{A}^T \boldsymbol{A} + J\boldsymbol{L}$ and mean $\boldsymbol{\mu} = \boldsymbol{\Sigma}\beta \boldsymbol{A}^T \boldsymbol{p}$. Note we included the pair prior in the multivariate Gaussian, as defined in (11).

- *EP-diff* introduces the use of empirical priors and auxiliary variables, in particular the difference ones; they are obtained applying the linear operator $\boldsymbol{F}$ on the vector containing all pixel intensity variables, where $\boldsymbol{F}$ is the oriented incidence matrix. The estimation of the parameters ϱ and σ of the difference-variable prior (the Bernoulli-Gaussian prior already defined in (52) are estimated in this work [7] using the Gradient Descent method over an approximate free energy. So in this case the length of the priors vector is $N_{tot} = N + N_e$, because we deal with $N = 4096$ interval priors for the pixel intensity variables and N_e Bernoulli-Gaussian priors for the difference ones. The number of edges N_e between neighboring pixel is computed using : $\sum_{v \in V} deg(v) = 2N_e$, where we sum all the degrees of the vertices. The degree $deg(v)$ of a vertex v is defined as the number of edges that have v as an endpoint. The total number of edges of a square grid of linear size $L = 64$ is $N_e = 8064$, so the length of the priors vector considering both variables is $N_{tot} = N + N_e$.

- We use two versions of *EP-Haar:* in both versions the auxiliary variables obtained through the Haar transform are used, with the linear operator $\boldsymbol{F} = \boldsymbol{H} \otimes \boldsymbol{H}$ of (67). The difference lies in the priors used for the auxiliary variables.

  1. in *EP-Haar-nofit* the vector of the priors of the auxiliary variables is composed of Bernoulli-Gaussian priors with the parameters estimated with EP, as done for the difference variables.

2. in *EP-Haar-fit* instead we design for each auxiliary variable a specific prior, whose functional form and relative parameters are estimated using the training dataset, as explained in the chapter about Empirical priors.

- The last two EP implementations use both the difference auxiliary variables and those obtained through the Haar transform. So in this case the matrix $\boldsymbol{F} \in R^{M\text{xN}}$ is composed by two blocks:

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{H} \otimes \boldsymbol{H} \end{bmatrix}$$

where the number of auxiliary variables is $M = N_e + N_{haar}$, giving a total amount of variables $N_{tot} = N_{int} + N_e + N_{haar} = 16256$.

Also in this situation the difference lies in choosing the priors: in both we use the Spike-and-Slab prior for the difference variables, as in the *EP-diff* case.

1. in *EP-both-nofit* we use the Bernoulli-Gaussian prior also for the Haar auxiliary variables.

2. *in EP-both-fit* instead we design a suitable prior, for each Haar variable, according to the information of the dataset, as explained for the *EP-Haar-fit* method.

3.4.3 Results

In a first analysis the different implementations of the EP algorithm were compared.

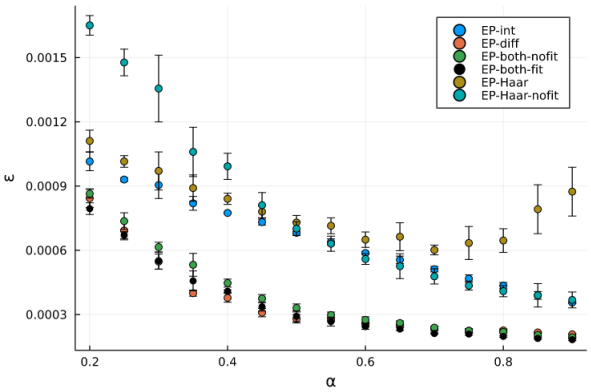

Figure 18: The error ε of reconstructed images as a function of the sampling rate α for the six EP methods presented, with an error bar representing the standard deviation.The projection matrix $\boldsymbol{A}$ is different for a given sampling rate, while we aim to reconstruct the same image.

We quantify the performances of the reconstructed images comparing the reconstruction error ε, defined as the average l_2 norm of the difference between the original image $\boldsymbol{x}$ and its reconstruction $\boldsymbol{x}^*$:

$$\varepsilon = \frac{||\boldsymbol{x}^* - \boldsymbol{x}||_2}{N} \tag{76}$$

where N is the length of the pixel intensity vector.

The error ε is plotted as a function of the sampling rate $\alpha = \frac{M}{N}$, which takes value in $\alpha \in [0.2, 0.9]$.

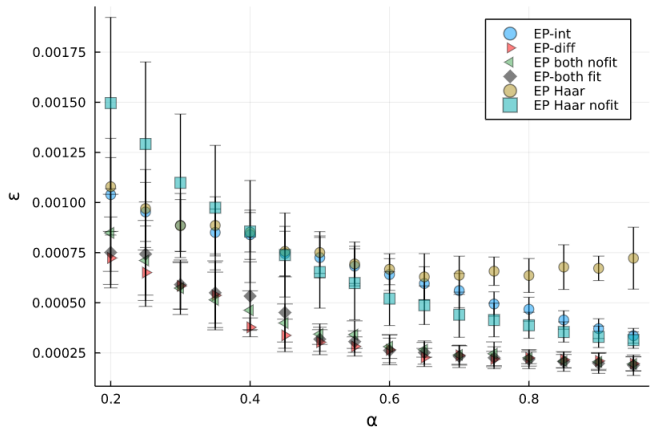

Figure 19: The error ε of reconstructed images as a function of the sampling rate α for the six EP methods presented: the projection matrix is the same for a given sampling rate, while images are different. From the error bar we can see more variance in the results with respect to the plot in 18.

In figure 18 we report the results comparing the mean error of five reconstructions for each algorithm implementation under study in the different measurement regimes: fixed the sampling rate α, we use five different projection matrix $\boldsymbol{A}$ and we aim to reconstruct the same image. We recall that in these matrices each row represents a single-ray projection in a random direction.

From the other hand we can also reconstruct different images using the same projection matrix$\boldsymbol{A}$. In a similar way to the previous plot we show the results in figure 19: in this case the results show higher variance, specially in the low measurement regime, but this is reasonable because now we are reconstructing different images.

Apart from the case of EP-Haar, whose performance at some point no longer improves, we can see in the graphs a similar trend for the other EP methods, with an error in the reconstructed images that decreases as the sampling rate increases (therefore as the measurements increase).

From both plots we can see that the key step consists of including as auxiliary variables the pixel differences: in fact in all measurements regimes we obtain

better performances with all the three implementations of EP algorithm with difference variables.

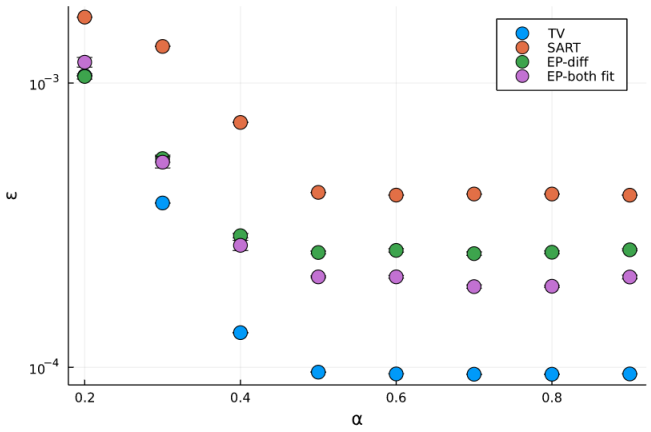

Figure 20: The error ε of reconstructed images as a function of the sampling rate α comparing the best two EP implementations with the TV method and the SART one.

Including also the Haar variables this error slightly decreases, specially in the high measurement regimes.

So far we have compared the performances of EP algorithms with each other: now we extend our analysis by comparing the two EP implementations showing better results with other methods commonly used in literature, the total variation method (TV) and the simultaneous algebraic reconstruction technique (SART).

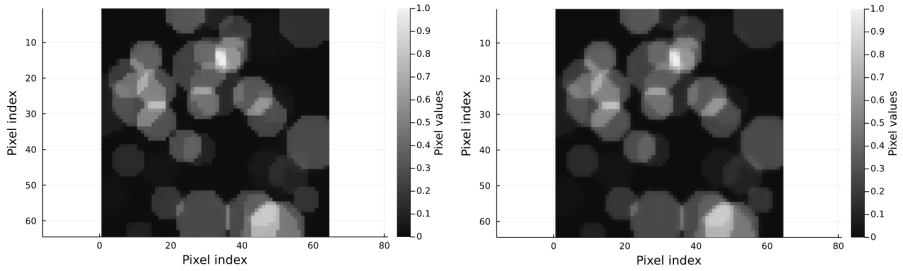

Figure 21: Image reconstructions: in the left panel we used *EP-both-fit*, while in the right one the image is reconstructed using the TV method. In both of them we are in the high measurement regime, with $\alpha = 0.7$. The error using TV (of order $\backsim 9 * 10^{-5}$) is smaller with respect to the error of *EP-both-fit* (of order $\backsim 2 * 10^{-4}$) but the difference in the image resolution is not visible to the naked eye.

In this work we used TomoForward.jl and XfromProjections.jl packages, [2, 18].

In this case, the projection matrix represents a 2-D parallel beam geometry (see paragraph 3.4.1 on measurement processes). The EP methods employed in this analysis are *EP-diff,* with as auxiliary variables only the difference ones, and *EP-both-fit,* which in addition to these EP also exploits the auxiliary Haar variables.

In figure 20 we report the results of the four methods described: for each of them we plot the mean of five reconstructions for different images, with the same projection matrix for a given sampling rate. Compared to the previous plots, the error bar representing the standard deviation is much lower: this is reasonable, as now we have a matrix of the measures constructed in a different way, which does not allow much variance as in the results shown above.

The reconstruction errors displayed show the best results obtained by TV method compared to the other techniques *(EP-both-fit*, *EP-diff* or SART). However, as can be seen in the figure 21, errors smaller than 10^{-4}do not improve the quality of the reconstructed images: the two images, reconstructed using *EP-both-fit* and TV, are indistinguishable to the naked eye.

4 Conclusion

Looking at the results obtained both by comparing the EP methods with each other and by comparing some of these with other image reconstruction algorithms, we confirm the importance of introducing the auxiliary difference variables in the EP algorithm to reconstruct tomographic images. Furthermore, EP algorithms have shown that they can include information via non-convex priors, which is difficult or even not possible with standard optimization tools.

The idea behind this work was to exploit the advantages by introducing another set of auxiliary variables, the Haar ones. To do this we created a training dataset that tried to reproduce the characteristics of the tomographic images. Studying the statistics of the new variables obtained through the Haar linear transform we inferred the form and the parameters of the functions to be inserted through the priors. By inserting this additional information, several reconstructions have been launched to compare the results but the attempt made through this work shows a slightly improvement with respect to *EP-diff* in some regimes.

It is worth noting that the fitting performed can be improved by making it possible to implement the family of Laplace priors, whose calculations have been performed in the appendix B. These priors in fact make EP algorithms unstable from a computational point of view.

However, the approach used in this thesis can also be adopted for other auxiliary variables connected through a linear transform to the pixel intensity variables, trying to obtain lower reconstruction errors and algorithms that require a lower number of measurements.

A Appendix

A.1 Kullback–Leibler divergence

In mathematical statistics, the Kullback–Leibler divergence$D_{\mathrm{KL}}(P \parallel Q)$ is a measure of the difference between two probability distribution; in particular how much information lost when Q is used to approximate P. Typically P represents the "true" distribution of data, while Q typically represents a model or approximation of P.

For discrete probability distributions $P$and Q defined on the same probability space X, the KL is defined as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in X} P(x) \log \left[\frac{P(x)}{Q(x)} \right]. \tag{77}$$

The KL is non-negative and it is zero only if $P = Q$ everywhere, thus recalling the concept of distance.

A.2 Moments matching condition

In this section let's show how the moment matching condition is the equivalent of minimizing the KL divergence with respect to the parameters of the distribution we want to use to describe our data.

Recalling (19) and (20), let's write explicitly the partition function of both distributions:

$$\widetilde{Z}_{Q^{(i)}} = \int d^N \boldsymbol{x} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T \boldsymbol{\Sigma}_{(i)}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} \psi_i(x_i) \tag{78}$$

$$\widetilde{Z}_Q\left(a_i, b_i\right) = \int d^N \boldsymbol{x} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T \boldsymbol{\Sigma}_{(i)}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} e^{-\frac{(x_i-a_i)^2}{2b_i}} \tag{79}$$

where $\phi_i(x_i) = \frac{1}{\sqrt{2\pi b_i}} e^{-\frac{(x_i-a_i)^2}{2b_i}}$ and $\psi_i(x_i)$ is the exact prior we want approximate.

Inserting these results in the definition of the KL divergence:

$$D_{KL}\left[Q^{(i)} || Q\right] = \int d^N \boldsymbol{x} Q^{(i)}\left(\boldsymbol{x}|\boldsymbol{y}\right) log\left[\frac{\psi_i(x_i)\widetilde{Z}_Q\left(a_i, b_i\right)}{\phi_i(x_i)\widetilde{Z}_{Q^{(i)}}}\right]$$

$$= \int d^N \boldsymbol{x} Q^{(i)}\left(\boldsymbol{x}|\boldsymbol{y}\right)\left[\frac{\left(x_i - a_i\right)^2}{2b_i} + log\widetilde{Z}_Q\left(a_i, b_i\right)\right] + cost$$

$$\frac{< \left(x_i - a_i\right)^2 >_{Q^{(i)}}}{2b_i} + log\widetilde{Z}_Q\left(a_i, b_i\right) + cost$$

where inside *cost* there all terms not depending on the parameters $a_i$or b_i.

Imposing the minimization with respect to these parameters we obtain:

$$\frac{\partial D_{KL}\left[Q^{(i)}||Q\right]}{\partial a_i} = \frac{- < x_i >_{Q^{(i)}} + a_i}{b_i} + \frac{1}{\widetilde{Z}_Q}\frac{\partial \widetilde{Z}_Q}{\partial a_i} \tag{80}$$

$$\frac{\partial D_{KL}\left[Q^{(i)}||Q\right]}{\partial b_i} = -\frac{< (x_i - a_i)^2 >_{Q^{(i)}}}{2b_i^2} + \frac{1}{\widetilde{Z}_Q}\frac{\partial \widetilde{Z}_Q}{\partial b_i} \tag{81}$$

Now we insert the derivatives inside the integrals of the partition functions:

$$\frac{1}{\widetilde{Z}_Q}\frac{\partial \widetilde{Z}_Q}{\partial a_i} = \frac{1}{\widetilde{Z}_Q}\int d^N\boldsymbol{x} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T\boldsymbol{\Sigma}_{(i)}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} e^{-\frac{(x_i-a_i)^2}{2b_i}}\left(\frac{x_i - a_i}{b_i}\right) =< \frac{x_i - a_i}{b_i} >_Q$$

$$\frac{1}{\widetilde{Z}_Q}\frac{\partial \widetilde{Z}_Q}{\partial b_i} = \frac{1}{\widetilde{Z}_Q}\int d^N\boldsymbol{x} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T\boldsymbol{\Sigma}_{(i)}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})} e^{-\frac{(x_i-a_i)^2}{2b_i}}\left(\frac{(x_i - a_i)^2}{2b_i^2}\right) = \frac{< (x_i - a_i)^2 >_Q}{2b_i^2}$$

Multiplying the multivariate Gaussian times the uni-variate that is missing in the cavity we obtain the distribution with all Gaussian priors $Q(\boldsymbol{x}|\boldsymbol{p})$, defined in (17).

Let's set equal to zero the derivatives in (80) and (81) :

$$0 = \frac{- < x_i >_{Q^{(i)}} + a_i}{b_i} + < \frac{x_i - a_i}{b_i} >_Q$$

$$0 = -\frac{< (x_i - a_i)^2 >_{Q^{(i)}}}{2b_i^2} + \frac{< (x_i - a_i)^2 >_Q}{2b_i^2}$$

Assuming $b_i \neq 0$, we obtain the moment matching condition:

$$< x_i >_{\mathcal{Q}^{(i)}(\mathbf{x})} =< x_i >_{\mathcal{Q}(\mathbf{x})} \tag{82}$$

$$< x_i^2 >_{\mathcal{Q}^{(i)}(\mathbf{x})} =< x_i^2 >_{\mathcal{Q}(\mathbf{x})} \tag{83}$$

B Appendix

B.1 Tilted distribution moments

In the following we compute the moments (mean and variance) of the tilted distribution with the Laplace prior, the Asymmetric Laplace distribution and the Bernoulli-Laplace.

Let's introduce some useful results that we will use in the next subsections to compute these moments.

First of all, the expectation values calculated with respect to the tilted distribution will be denoted with $< x >$ instead of $< x >_{Q^{(i)}}$ to simplify the notation (this is done also for the variance).

The cumulative distribution function $F(t, m, \Sigma)$ (CDF) of a random variable x, evaluated at a given value t, is the probability that the distribution will take a value less than or equal to t:

$$F(t; m, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} \int_{-\infty}^{t} dx e^{-\frac{(x-m)^2}{2\Sigma}} \tag{84}$$

$$1 - F(t) = \frac{1}{\sqrt{2\pi\Sigma}} \int_{t}^{+\infty} dx e^{-\frac{1}{2}\frac{(x-m)^2}{\Sigma}}$$

The truncated normal distribution is the probability distribution derived from that of a normally distributed random variable by bounding the random variable from either below or above (or both).

For the interval $[a, b]$, let's define $\alpha = \frac{(a-m)}{\sqrt{\Sigma}}$ and $\beta = \frac{(b-m)}{\sqrt{\Sigma}}$.

We have used this formula of truncated Gaussian distributions:

$$\frac{1}{\sqrt{2\pi\Sigma_{ii}}} \int_{a}^{b} dx x e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} = m_1 + \sqrt{\Sigma_{ii}} \frac{\Phi(\alpha) - \Phi(\beta)}{F(\beta) - F(\alpha)} \tag{85}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$ is the CDF.

Notice that $\Phi(\infty) = \Phi(-\infty) = 0$ and $F(\infty) = 1$, $F(-\infty) = 0$.

So, for $[0, +\infty]$:

$$\mathrm{E}\left(X|X>0\right) = \mu_i + \sqrt{\Sigma_{ii}}\left[\frac{\Phi\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}{1-F\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}\right] \tag{86}$$

$$\mathrm{Var}\left(X|X>0\right) = \Sigma_{ii}\left[1-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\left[\frac{\Phi\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}{1-F\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}\right] - \left(\frac{\Phi\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}{1-F\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}\right)^2\right] \tag{87}$$

And for $[-\infty, 0]$:

$$\mathrm{E}\left(X|X<0\right) = \mu_i - \sqrt{\Sigma_{ii}}\frac{\Phi\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}{F\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)} \tag{88}$$

$$\mathrm{Var}\left(\mathrm{X}|\mathrm{X}<0\right) = \Sigma_{ii}\left[1+\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\frac{\Phi\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}{F\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)} - \left(\frac{\Phi\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}{F\left(-\frac{\mu_i}{\sqrt{\Sigma_{ii}}}\right)}\right)^2\right] \tag{89}$$

B.2 Laplace prior

Considering the partition function of the tilted distribution with the Laplace prior:

$$Z = \frac{1}{\sqrt{2\pi\Sigma_{ii}}}\int_{-\infty}^{+\infty} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}}\left\{\frac{1}{2}\lambda e^{-\lambda|x_i|}\right\}$$

$$= \frac{\lambda}{2\sqrt{2\pi\Sigma_{ii}}}\left(\int_0^{+\infty} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}-\lambda x_i} + \int_{-\infty}^{0} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}+\lambda x_i}\right)$$

Rewriting the first quadratic expression in the form $-\frac{1}{2}Ax^2+bx+c$:

$$-\tfrac{1}{2}\tfrac{(x_i-\mu_i)^2}{\Sigma_{ii}}-\lambda x_i = -\tfrac{1}{2}\tfrac{x_i^2+\mu_i^2-2x_i\mu_i}{\Sigma_{ii}}-\lambda x_i = -\tfrac{1}{2\Sigma_{ii}}\left(x_i^2+\mu_i^2+x_i\left(-2\mu_i+2\lambda\Sigma_{ii}\right)\right)$$

Mean and variance will be:

$$\sigma^2 = \frac{1}{A} = \Sigma_{ii}$$

$$m_1 = \sigma^2 b = \Sigma_{ii} - \frac{1}{2}\left[\frac{(-2\mu_i+2\lambda\Sigma_{ii})}{\Sigma_{ii}}\right] = \mu_i - \lambda\Sigma_{ii}$$

Proceeding in a similar way for the second quadratic form we obtain:

$$\sigma^2 = \Sigma_{ii}$$

$$m_2 = \mu_i + \lambda\Sigma_{ii}$$

In order to obtain an equivalent expression we need to add a constant term; finally we obtain in the partition function:

$$Z = \frac{\lambda\rho}{2\sqrt{2\pi\Sigma_{ii}}}\left(\int_0^{+\infty} dx_i e^{-\frac{(x_i-m_1)^2}{2\Sigma_{ii}}+\frac{1}{2}\lambda^2\Sigma_{ii}-\lambda\mu_i} + \int_{-\infty}^{0} dx_i e^{-\frac{(x_i-m_2)^2}{2\Sigma_{ii}}+\frac{1}{2}\lambda^2\Sigma_{ii}+\lambda\mu_i}\right)$$

Using the CDF (84) :

$$Z = \frac{\lambda}{2}e^{\frac{1}{2}\lambda^2\Sigma_{ii}}\left\{e^{-\lambda\mu_i}\left[1-F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}}\right)\right] + e^{\lambda\mu_i}F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}}\right)\right\} \tag{90}$$

Let's compute mean and variance of the tilted distribution with the Laplace prior:

$$< x_i > \;=\; \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}}\int_{-\infty}^{+\infty} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}}\left\{\frac{\lambda}{2}e^{-\lambda|x_i|}\right\} = \frac{\lambda}{2Z\sqrt{2\pi\Sigma_{ii}}}\int_{-\infty}^{+\infty} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}-\lambda|x_i|}$$

Rearranging the exponential as written above:

$$= \frac{\lambda}{2Z\sqrt{2\pi\Sigma_{ii}}}e^{\frac{1}{2}\lambda^2\Sigma_{ii}}\left\{e^{-\lambda\mu_i}\int_0^{+\infty} dx_i x_i e^{-\frac{(x_i-m_1)^2}{2\Sigma_{ii}}} + e^{\lambda\mu_i}\int_{-\infty}^{0} dx_i x_i e^{-\frac{(x_i-m_2)^2}{2\Sigma_{ii}}}\right\}$$

Using (86) and (88) :

$$< x_i >= \frac{\lambda}{2Z}e^{\frac{1}{2}\lambda^2\Sigma_{ii}}\left\{e^{-\lambda\mu_i}\left[m_1+\sqrt{\Sigma_{ii}}\frac{\Phi\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}}\right)}{1-F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}}\right)}\right] + e^{\lambda\mu_i}\left[m_2-\sqrt{\Sigma_{ii}}\frac{\Phi\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}}\right)}{F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}}\right)}\right]\right\}$$

Inserting Z :

$$< x_i >= \frac{e^{-\lambda\mu_i}\left[m_1+\sqrt{\Sigma_{ii}}\frac{\Phi\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}}\right)}{1-F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}}\right)}\right] + e^{\lambda\mu_i}\left[m_2-\sqrt{\Sigma_{ii}}\frac{\Phi\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}}\right)}{F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}}\right)}\right]}{e^{-\lambda\mu_i}\left[1-F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}}\right)\right] + e^{\lambda\mu_i}F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}}\right)} \tag{91}$$

Let's compute the second moment:

$$\left\langle (x_i - \langle x_i \rangle)^2 \right\rangle = \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i \, (x_i - \langle x_i \rangle)^2 \, e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ \frac{1}{2}\lambda e^{-\lambda|x_i|} \right\}$$

As in the previous case, we separate the integrals and we apply (87) and (89) :

$$= \frac{\lambda}{2Z\sqrt{2\pi\Sigma_{ii}}} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \int_0^{+\infty} dx_i \, (x_i - \langle x_i \rangle)^2 \, e^{-\frac{(x_i-m_1)^2}{2\Sigma_{ii}}} + e^{\lambda\mu_i} \int_{-\infty}^{0} dx_i \, (x_i - \langle x_i \rangle)^2 \, e^{-\frac{(x_i-m_2)^2}{2\Sigma_{ii}}} \right\}$$

$$\left\langle (x_i - \langle x_i \rangle)^2 \right\rangle = \frac{e^{-\lambda\mu_i}\Sigma_{ii} \left\{ 1 + h_1 \left[\frac{\Phi(h_1)}{1-F(h_1)} \right] - \left[\frac{\Phi(h_1)}{1-F(h_1)} \right]^2 \right\} + e^{\lambda\mu_i}\Sigma_{ii} \left\{ 1 - h_2 \frac{\Phi(h_2)}{F(h_2)} - \left[\frac{\Phi(h_2)}{F(h_2)} \right]^2 \right\}}{e^{-\lambda\mu_i} \left[1 - F\left(h_1\right) \right] + e^{\lambda\mu_i} F\left(h_2\right)} \tag{92}$$

B.3 Asymmetric Laplace prior

Sometimes it might be convenient to have a Laplace distribution that behaves differently depending on whether the argument is positive or negative; so let's introduce the distribution and the corresponding tilted distribution, with partition function:

$$Z = \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ \rho\delta\left(x_i > 0\right) \lambda_1 e^{-\lambda_1|x_i|} + \left(1-\rho\right)\delta\left(x_i < 0\right)\lambda_2 e^{-\lambda_2|x_i|} \right\}$$

We proceed as in the Laplace prior, separating the integrals and apply (84):

$$= \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \left(\rho\lambda_1 \int_0^{+\infty} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} - \lambda_1 x_i} + \left(1-\rho\right)\lambda_2 \int_{-\infty}^{0} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} + \lambda_2 x_i} \right)$$

$$= \quad \rho\lambda_1 e^{\frac{1}{2}\lambda_1^2\Sigma_{ii} - \lambda_1\mu_i} \left[1 - F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}} \right) \right] + \left(1-\rho\right)\lambda_2 e^{\frac{1}{2}\lambda_2^2\Sigma_{ii} + \lambda_2\mu_i} F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}} \right)$$

For the mean we use (86) and (88):

$$< x_i >= \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ \rho\delta\left(x_i > 0\right) \lambda_1 e^{-\lambda_1|x_i|} + \left(1-\rho\right)\delta\left(x_i < 0\right)\lambda_2 e^{-\lambda_2|x_i|} \right\}$$

$$= \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}} \left(\rho\lambda_1 \int_0^{+\infty} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} - \lambda_1 x_i} + (1-\rho)\,\lambda_2 \int_{-\infty}^{0} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} + \lambda_2 x_i} \right)$$

$$= \frac{\left\{ \rho\lambda_1 e^{\frac{1}{2}\lambda_1^2\Sigma_{ii} - \lambda_1\mu_i} \left[m_1 + \sqrt{\Sigma_{ii}} \frac{\Phi(h_1)}{1-F(h_1)} \right] + (1-\varrho)\,\lambda_2 e^{\frac{1}{2}\lambda_2^2\Sigma_{ii} + \lambda_2\mu_i} \left[m_2 - \sqrt{\Sigma_{ii}} \frac{\Phi(h_2)}{F(h_2)} \right] \right\}}{\rho\lambda_1 e^{\frac{1}{2}\lambda_1^2\Sigma_{ii} - \lambda_1\mu_i} \left[1 - F\left(h_1\right)\right] + (1-\rho)\,\lambda_2 e^{\frac{1}{2}\lambda_2^2\Sigma_{ii} + \lambda_2\mu_i} F\left(h_2\right)}$$

Defining $c_1 = \rho\lambda_1 e^{\frac{1}{2}\lambda_1^2\Sigma_{ii} - \lambda_1\mu_i}\Sigma_{ii}$ and $c_2 = (1-\varrho)\,\lambda_2 e^{\frac{1}{2}\lambda_2^2\Sigma_{ii} + \lambda_2\mu_i}$, we can rewrite:

$$< x_i >= \frac{\left\{ c_1 \left[m_1 + \sqrt{\Sigma_{ii}} \frac{\Phi(h_1)}{1-F(h_1)} \right] + c_2 \left[m_2 - \sqrt{\Sigma_{ii}} \frac{\Phi(h_2)}{F(h_2)} \right] \right\}}{c_1 \left[1 - F\left(h_1\right)\right] + c_2 F\left(h_2\right)}$$

While for the variance we use (87) and (89):

$$Var(x_i) = \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i \left(x_i - \langle x_i \rangle\right)^2 e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ \rho\delta\left(x_i > 0\right)\lambda_1 e^{-\lambda_1|x_i|} + (1-\rho)\,\delta\left(x_i < 0\right)\lambda_2 e^{-\lambda_2|x_i|} \right\}$$

$$= \frac{c_1 \left\{ 1 + h_1 \left[\frac{\Phi(h_1)}{1-F(h_1)} \right] - \left[\frac{\Phi(h_1)}{1-F(h_1)} \right]^2 \right\} + c_2 \left\{ 1 - h_2 \frac{\Phi(h_2)}{F(h_2)} - \left[\frac{\Phi(h_2)}{F(h_2)} \right]^2 \right\}}{c_1 \left[1 - F\left(h_1\right)\right] + c_2 F\left(h_2\right)}$$

B.4 Bernoulli-Laplace prior

Let's compute the first two moments of the tilted distribution with the Bernoulli-Laplace prior:

$$Z = \frac{1}{\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ (1-\rho)\,\delta(x_i) + \rho\frac{1}{2}\lambda e^{-\lambda|x_i|} \right\}$$

$$= \frac{1}{\sqrt{2\pi\Sigma_{ii}}} (1-\rho)\, e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}}} + \frac{\lambda\rho}{2\sqrt{2\pi\Sigma_{ii}}} \left(\int_0^{+\infty} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} - \lambda x_i} + \int_{-\infty}^{0} dx_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} + \lambda x_i} \right)$$

Also in this case we can rearrange the exponential:

$$Z = \frac{(1-\rho)\, e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}}}}{\sqrt{2\pi\Sigma_{ii}}} + \frac{\lambda\rho}{2\sqrt{2\pi\Sigma_{ii}}} \left(\int_0^{+\infty} dx_i e^{-\frac{(x_i-m_1)^2}{2\Sigma_{ii}} + \frac{1}{2}\lambda^2\Sigma_{ii} - \lambda\mu_i} + \int_{-\infty}^{0} dx_i e^{-\frac{(x_i-m_2)^2}{2\Sigma_{ii}} + \frac{1}{2}\lambda^2\Sigma_{ii} + \lambda\mu_i} \right)$$

Using 84:

$$= \frac{(1-\rho)\, e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}}}}{\sqrt{2\pi\Sigma_{ii}}} + \frac{\lambda\rho}{2} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \left[1 - F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}} \right) \right] + e^{\lambda\mu_i} F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}} \right) \right\}$$

Now let's compute the mean and variance:

$$< x_i >= \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ (1-\rho)\, \delta(x_i) + \frac{\lambda\rho}{2} e^{-\lambda|x_i|} \right\}$$

$$= \frac{\lambda\rho}{2Z\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i x_i e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}} - \lambda|x_i|}$$

Inserting the already obtained results:

$$< x_i >= \frac{\lambda\rho}{2Z} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \left[m_1 + \sqrt{\Sigma_{ii}} \frac{\Phi\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}} \right)}{1 - F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}} \right)} \right] + e^{\lambda\mu_i} \left[m_2 - \sqrt{\Sigma_{ii}} \frac{\Phi\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}} \right)}{F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}} \right)} \right] \right\}$$

Recalling the expression of Z:

$$\begin{aligned} Z &= \frac{(1-\rho)\, e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}}}}{\sqrt{2\pi\Sigma_{ii}}} + \frac{\lambda\rho}{2} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \left[1 - F\left(-\frac{m_1}{\sqrt{\Sigma_{ii}}} \right) \right] + e^{\lambda\mu_i} F\left(-\frac{m_2}{\sqrt{\Sigma_{ii}}} \right) \right\} \\ &= \frac{\lambda\rho}{2} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \left[1 - F\left(h_1 \right) \right] + e^{\lambda\mu_i} F\left(h_2 \right) + \frac{(1-\rho)\, e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}}}}{\sqrt{2\pi\Sigma_{ii}} \frac{\lambda\rho}{2} e^{\frac{1}{2}\lambda^2\Sigma_{ii}}} \right\} \\ &= \frac{\lambda\rho}{2} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \left[1 - F\left(h_1 \right) \right] + e^{\lambda\mu_i} F\left(h_2 \right) + \frac{2\,(1-\rho)}{\lambda\rho\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}} - \frac{1}{2}\lambda^2\Sigma_{ii}} \right\} \end{aligned}$$

Finally:

$$< x_i >= \frac{e^{-\lambda\mu_i} \left[m_1 + \sqrt{\Sigma_{ii}} \frac{\Phi(h_1)}{1-F(h_1)} \right] + e^{\lambda\mu_i} \left[m_2 - \sqrt{\Sigma_{ii}} \frac{\Phi(h_2)}{F(h_2)} \right]}{e^{-\lambda\mu_i} \left[1 - F\left(h_1 \right) \right] + e^{\lambda\mu_i} F\left(h_2 \right) + \frac{2(1-\rho)}{\lambda\rho\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{{\mu_i}^2}{2\Sigma_{ii}} - \frac{1}{2}\lambda^2\Sigma_{ii}}}$$

Let's compute the second moment:

$$\left\langle \left(x_i - \langle x_i \rangle\right)^2 \right\rangle = \frac{1}{Z\sqrt{2\pi\Sigma_{ii}}} \int_{-\infty}^{+\infty} dx_i \left(x_i - \langle x_i \rangle\right)^2 e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}} \left\{ (1-\rho)\,\delta(x_i) + \rho \frac{1}{2} \lambda e^{-\lambda |x_i|} \right\}$$

As in the previous case:

$$= \frac{\lambda\rho}{2Z\sqrt{2\pi\Sigma_{ii}}} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left\{ e^{-\lambda\mu_i} \int_{0}^{+\infty} dx_i \left(x_i - \langle x_i \rangle\right)^2 e^{-\frac{(x_i-m_1)^2}{2\Sigma_{ii}}} + e^{\lambda\mu_i} \int_{-\infty}^{0} dx_i \left(x_i - \langle x_i \rangle\right)^2 e^{-\frac{(x_i-m_2)^2}{2\Sigma_{ii}}} \right\}$$

Using (87) and (89)

$$= \frac{\lambda\rho}{2Z} e^{\frac{1}{2}\lambda^2\Sigma_{ii}} \left[e^{-\lambda\mu_i}\Sigma_{ii} \left\{ 1 + h_1 \left[\frac{\Phi(h_1)}{1-F(h_1)} \right] - \left[\frac{\Phi(h_1)}{1-F(h_1)} \right]^2 \right\} + e^{\lambda\mu_i}\Sigma_{ii} \left\{ 1 - h_2 \frac{\Phi(h_2)}{F(h_2)} - \left[\frac{\Phi(h_2)}{F(h_2)} \right]^2 \right\} \right]$$

$$\left\langle \left(x_i - \langle x_i \rangle\right)^2 \right\rangle = \frac{e^{-\lambda\mu_i}\Sigma_{ii} \left\{ 1 + h_1 \left[\frac{\Phi(h_1)}{1-F(h_1)} \right] - \left[\frac{\Phi(h_1)}{1-F(h_1)} \right]^2 \right\} + e^{\lambda\mu_i}\Sigma_{ii} \left\{ 1 - h_2 \frac{\Phi(h_2)}{F(h_2)} - \left[\frac{\Phi(h_2)}{F(h_2)} \right]^2 \right\}}{e^{-\lambda\mu_i}\left[1 - F(h_1)\right] + e^{\lambda\mu_i} F(h_2) + \frac{2(1-\rho)}{\lambda\rho\sqrt{\Sigma_{ii}}} e^{-\frac{\mu_i^2}{2\Sigma_{ii}} - \frac{1}{2}\lambda^2\Sigma_{ii}}}$$

References

[1] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. IEEE Press, 1988.

[2] A. Andersen and A. Kak, "Simultaneous Algebraic Reconstruction Technique (SART): a Superior Implementation of the ART Algorithm," *Ultrasonic imaging*, vol. 6, pp. 81–94, 02 1984.

[3] M. Pieropan, "Expectation Propagation methods for approximate inference in linear estimation problems," Ph.D. dissertation, Politecnico di Torino, 2021.

[4] R. Gordon and G. Herman, "Algebraic Reconstruction Technique (ART) for three-dimensional electron microscopy and X-ray photography," *Journal of theoretical biology*, vol. 29, pp. 471–81, 01 1971.

[5] P. Gilbert, "Iterative methods for the three-dimensional reconstruction of an object from projections," *Journal of Theoretical Biology*, vol. 36, no. 1, pp. 105–117, 1972. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022519372901804

[6] E. Y. Sidky, C.-M. Kao, and X. Pan, "Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT," *Journal of X-ray Science and Technology*, vol. 14, pp. 119–139, 2006.

[7] A. P. Muntoni, R. D. H. Rojas, A. Braunstein, A. Pagnani, and I. P. Castillo, "Nonconvex image reconstruction via expectation propagation," *Physical Review E*, vol. 100, no. 3, Sep 2019. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.100.032134

[8] E. Gouillart, F. Krzakala, M. Mezard, and L. Zdeborova, "Belief Propagation Reconstruction for Discrete Tomography," *Inverse Problems*, vol. 29, 11 2012.

[9] A. Muntoni, "Statistical mechanics approaches to optimization and inference," Ph.D. dissertation, Politecnico di Torino, 2017.

[10] T. P. Minka, "Expectation Propagation for approximate Bayesian inference," *CoRR*, vol. abs/1301.2294, 2013. [Online]. Available: http://arxiv.org/abs/1301.2294

[11] M. Opper and O. Winther, "Expectation consistent free energies for approximate inference," in *Proceedings of the 17th International Conference on Neural Information Processing Systems*, ser. NIPS'04. Cambridge, MA, USA: MIT Press, 2004, pp. 1001–1008.

[12] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate inference techniques with expectation constraints," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 11,

p. P11015, nov 2005. [Online]. Available: https://doi.org/10.1088/1742-5468/2005/11/p11015

[13] M. Opper and O. Winther, "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling," *PRE*, vol. 64, no. 5, p. 056131, Oct. 2001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.64.056131

[14] A. Braunstein, A. P. Muntoni, and A. Pagnani, "An analytic approximation of the feasible space of metabolic networks," *Nature Communications*, vol. 8, no. 1, p. 14915, 2017. [Online]. Available: https://doi.org/10.1038/ncomms14915

[15] K. J. Batenburg and J. Sijbers, "DART: a practical reconstruction algorithm for discrete tomography." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 20, pp. 2542–53, Sep 2011.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/2984875

[17] S. Pan and A. Kak, "A computational study of reconstruction algorithms for diffraction tomography: Interpolation versus filtered-backpropagation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 5, pp. 1262–1275, 1983.

[18] A. Chambolle and T. Pock, "A First-Order Primal-Dual Algorithm for Convex Problems withÂ Applications to Imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011. [Online]. Available: https://doi.org/10.1007/s10851-010-0251-1