EIDÖGENSSISCHE TECHNISCHE HOCHSCHULE LAUSANNE
POLITECNICO FEDERALE DI LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY LAUSANNE

SCHOOL OF BASIC SCIENCES | MATHEMATICS INSTITUTE FOR ANALYSIS AND APPLICATIONS | Chair of Mathematical Statistics

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Statistical Inference in Positron Emission Tomography

SEMESTER PROJECT (Spring 2014)
Department of Mathematics

Author:

Matthieu SIMEONI, MA-MA2, EPFL.

matthieu.simeoni@epfl.ch.

Under the supervision of:

Mikael KUUSELA, SMAT, EPFL.

Prof. Victor PANARETOS, SMAT, EPFL.

Last modified on: June 15, 2014.

Table of Contents

1

Introduction

Positron Emission Tomography (PET) [7, 10, 13] is a medical diagnostic technique that enables a physician to study blood flow in and metabolic activity in an organ in a visual manner (see Figure 1.1). This technique, introduced by David Khul in the late 1950s, has since become a very central visualization tool in various fields of medicine: oncology, neuroimaging, cardiology, pharmacokinetics... PET scans of brains of people suffering from schizophrenia have for example revealed very distinctive metabolic patterns associated with this disorder. The study of these metabolic portraits suggested new treatment strategies and facilitated the diagnosis of this disease. A more detailed discussion on the subject as well as on the numerous other possible applications of PET imagery can be found in [10].

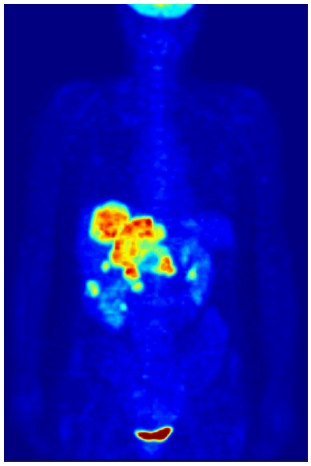


Figure 1.1: Whole-body PET scan of a woman in the context of a tumor diagnosis: besides normal accumulation of the tracer in the heart, bladder, kidneys and brain, liver metastases of a colorectal tumor are clearly visible within the abdominal region of the image. Source: [19].

As many noninvasive techniques, positron emission tomography belongs to the vast family of **reconstruction problems**: the metabolic activity is not directly observed, but inferred from the action of *probes* on the organ of interest. More specifically, a biochemical metabolite labeled with a *positron emitting* radioactive substance is introduced into the organ of interest and the radioactive emissions are then counted[1] using a **PET scanner** (machine consisting in possibly many rings of detectors surrounding the patient's body; see Figure 1.2). The choice of the biochemical and radioactive tracer strongly depends on the organ of interest: for example, labeled glucose is often used in PET scans of the brain, as glucose is the primary source of energy of the brain.

Figure 1.2: Image of a typical PET facility. Source: [20].

Then, recovering the metabolic activity of an organ based on the scanner records is an instance

[1] In reality, we cannot directly count the radioactive emissions, we can only record gamma rays emanating from the encounter between a positron and an electron.

of an **inverse problem** [3,5,6], and as such is very sensitive to the mathematical algorithm chosen for the reconstruction process: in mathematical terms, we face an ill-posed problem. In this report, we take inspiration from the physics of PET to design a mathematical model tailored to the problem. We think of positron emissions as an output of an indirectly observed Poisson process and formulate the link between the emissions and the scanner records through the Radon transform. This model allows us to express the image reconstruction in terms of a standard problem in **statistical estimation from incomplete data**. Then, we investigate different algorithms as well as stopping criterion, and compare their relative efficiency.

In all this report we will consider the brain as the organ to be studied.

1 The Physics of PET

We give in this section a slightly more detailed description of the physics of PET. A better comprehension of the underlying physical phenomenon will in fact be very useful while elaborating the mathematical model.

This short introduction to the physics of PET follows the one presented in [7]. The idea of PET is the following: as primary source of the brain's energy, glucose will concentrate in the various regions of the brain in quantities proportional to the brain's metabolic activity. Then, if we label this glucose with a radioactive substance, the positron emissions will also be directly proportional to the glucose consumption and therefore to the brain's metabolic activity. Thus, if we were able to record the locations of each emission, we could produce a portrait of the brain's metabolic activity.

Unfortunately, we *cannot identify the exact location of positron emissions*. However, we can determine a cylindrical volume in which the emission occurred. In fact, when a positron is emitted, it quickly annihilates with an electron, naturally present in the brain's medium. This annihilation generates two gamma rays, flying off in nearly opposite random directions. Therefore, by positioning a ring of **scintillation detectors** around the head of the patient, one can detect those gamma rays, that hit in coincidence a pair of detectors (see Figure 1.3). This pair of detector defines then a cylindrical volume (to be referred to as a **detector tube**), which provides us with a partial information on the location of

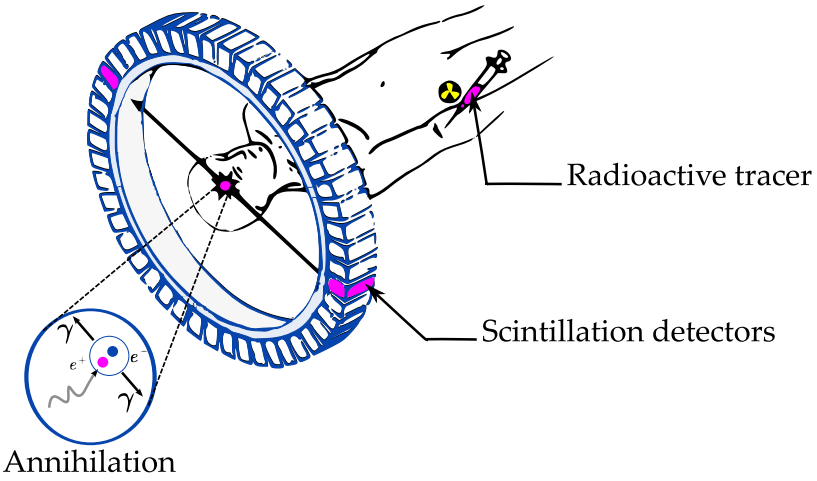


Figure 1.3: The physics of PET. Glucose is labeled with a radioactive tracer, and deposited in the brain in quantities proportional to the glucose consumption. When a positron is emitted, it quickly annihilates with an electron. This annihilation generates two gamma rays, which are detected in coincidence by a pair of scintillation detectors. Adapted from [15].

the positron emission. During a PET acquisition, the scanner will record each of these coincidences and count the total number of them for each detector tube.

2

Preliminary Concepts

1 Basic Theory of Point Processes

In this chapter, we first provide some basic results on point processes and Poisson processes. Then, we present briefly the Radon transform, that we show to be of central importance in positron emission tomography.

Most of the material presented in this section is inspired from the sections I.1 and II.7 of [1], from which we selected only results relevant to the context of PET.

A point process on a space $\mathcal{S}$ is a stochastic rule for the occurrence and position of point events. Those are ideal models for the study of random samples

$$X_1, X_2, \ldots, X_\xi,$$

where $X_1, X_2, X_3, \ldots$ are independent, identically distributed (i.i.d.) random elements in the **state space** $\mathcal{S}$, and ξ is a **random sample size independent** of the X_i's. Let $(x_1, \ldots, x_k)$ be a realization of such a random sample. Then, the examination of this observation can be carried out by counting the number of of points in subsets B of $\mathcal{S}$. This leads to the concept of **point measures**.

Example 1.1 We have represented on Figure 2.1 earthquakes in the Pacific ocean, for a period spanning from February 2014 to March 2014. A model based on point processes could be proposed to calculate the probability of a certain number of events within a specific subarea of the state space. In this particular case, the state space is multi-dimensional (five dimensions: longitude, latitude, depth,magnitude and time of occurrence). ■

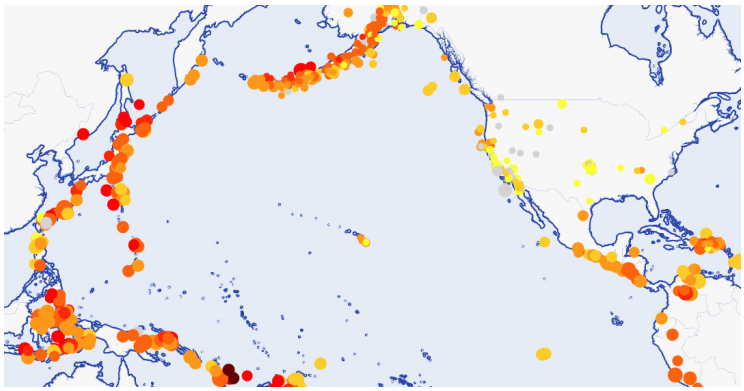


Figure 2.1: Earthquakes of magnitude greater than 2.5 from February 2014 to March 2014, in the Pacific Ocean. Circles' size indicates the magnitude of the earthquakes, while the color indicates the depth of the epicenter (Data collected by [17]).

1.1 Point Measures

An output $(x_1, ..., x_k)$ of a random sample can be represented by a point measure.

We endow the state space $\mathcal{S}$ with a σ-field $\mathcal{B}$. Let $(x_1, \ldots, x_k)$ be an outcome of the random experiment previously described. Then, we propose to represent the observed points x_i by a discrete measure μ on the σ-field $\mathcal{B}$, that we call a **point measure**:

Definition 1.1 — *Point Measure.* *Let $\{x_i : i \in I\} \subset \mathcal{S}$ a countable set of points and $\mathcal{B}$ a σ-field on $\mathcal{S}$. Then, the following discrete measure μ on $\mathcal{B}$ is called a **point measure**:*

$$\mu = \sum_{i \in I} \epsilon_{x_i},$$

*where $\epsilon_x(B) := \mathbb{1}_B(x),\ \forall B \in \mathcal{B}$, a **Dirac measure** with mass 1 at x. The **space of point measures** on $\mathcal{B}$ is denoted by*

$$\mathbb{M} = \mathbb{M}(\mathcal{S}, \mathcal{B}).$$

Remark 1.1 *If $I = \emptyset$ then μ is the null measure (that is, $\mu(B) = 0, \forall B \in \mathcal{B}$).*

Point measures μ can be identified with their truncations on the borel sets $B \in \mathcal{B}$. Point processes will inherit this useful property, which will simplify their study.

Since μ is a Borel-measure, we can identify it with $(\mu(B))_{B \in \mathcal{B}}$,

$$\mu(B) = \sum_{i \in I} \epsilon_{x_i}(B).$$

This legitimates the introduction of the **projection** mapping:

$$\pi_B : \begin{cases} \mathbb{M} \to \mathbb{N} \cup \{\infty\}, \\ \mu \mapsto \mu(B). \end{cases}$$

Then, $\mathbb{M}(\mathcal{S}, \mathcal{B})$ is endowed with the σ-field $\mathcal{M}(\mathcal{S}, \mathcal{B})$ such that the projections π_B, $B \in \mathcal{B}$ are measurable.

1.2 Point Processes

Point processes are random point measures.

Point measures extend naturally to **point processes**, which are simply random point measures on $(\mathcal{S}, \mathcal{B})$:

Definition 1.2 — *Point Processes.* *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, we call a **point process** on $(\mathcal{S}, \mathcal{B})$ a measurable mapping:*

$$N : \begin{cases} (\Omega, \mathcal{F}) & \to (\mathbb{M}(\mathcal{S}, \mathcal{B}), \mathcal{M}(\mathcal{S}, \mathcal{B})), \\ \omega & \mapsto N^\omega. \end{cases}$$

As point processes are random analogs of point measures, we would like to be able to manipulate them in the same fashion. To this end, we introduce the mapping $N(B) = \pi_B \circ N$, composition of the projection π_B and the point process N:

$$N(B) : \begin{cases} \Omega \to \mathbb{N} \cup \{\infty\}, \\ \omega \mapsto N^{\omega}(B). \end{cases}$$

Vocabulary 1.1 *We call the mapping $N(B)$ a **one-dimensional marginal** of N with index $B \in \mathcal{B}$.*

The following criterion provides us with the desired identification of a point process with its one-dimensional marginals:

A point process can be identified with its one-dimensional marginals. See section I.1 of [1] for a proof of Criterion 1.1.

Criterion 1.1 *The following two assertions are equivalent:*

- *$N : \Omega \to \mathbb{M}$ is a point process,*
- *$N(B) : \Omega \to \mathbb{N} \cup \{\infty\}$ is $(\mathcal{F}, \mathcal{P}(\mathbb{N} \cup \{\infty\}))-$ measurable for each $B \in \mathcal{B}$ (with $\mathcal{P}(\mathbb{N} \cup \{\infty\})$ the power set).*

Before continuing our developments, we present simple but yet fundamental examples of point processes.

Example 1.2 — *Fundamental point processes.* Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables from Ω to $\mathcal{S}$.

- **Empirical Processes:** This collection of random variables can be represented by the following point process (where the information about the order of appearance is lost) :

  $$N_n = \sum_{i=1}^{n} \epsilon_{X_i}.$$

  From the Criterion 1.1 we can affirm that N_n is indeed a point process as $N_n(B) = \sum_{i=1}^{n} \mathbb{1}_B(X_i)$ is obviously measurable. Moreover, we observe that $N_n(B)$ is binomial for every $B \in \mathcal{B}$.
- **Mixed empirical processes:** We now consider a random sample size ξ, independent from the X_i. Then, the following point process is called a mixed empirical point process :

  $$N = \sum_{i=1}^{\xi} \epsilon_{X_i}.$$

  One more time, we can invoke Criterion 1.1 to verify that this is indeed a point process.

■

We now have a useful characterization of a point process in terms of its one-dimensional marginals. However, in a modeling context, we are interested in the distribution of a point process, in order to compute probabilities of events of interest.

The following theorem, uniquely determines the distribution of a point process by its finite dimensional marginals.

The distribution of a point process can be seen as the collection of probabilities $\mathbb{P}\{N(B_1) = n_1, ..., N(B_k) = n_k\}$ for all possible choices of Borel sets $B_1, ..., B_k \in \mathcal{B}$ and all $k \in \mathbb{N}$. See section I.1 of [1] for a proof of Theorem 1.2.

Theorem 1.2 — *Uniqueness Theorem.* Let N_0 and N_1 be point processes on $(\mathcal{S}, \mathcal{B})$. The following two assertions are equivalent:

- $N_0 \overset{d}{=} N_1$,
- For every $k \in \mathbb{N}$ and $B_1, \dots, B_k \in \mathcal{B}$:

$$(N_0(B_1), \dots, N_0(B_k)) \overset{d}{=} (N_1(B_1), \dots, N_1(B_k)),$$

where $\overset{d}{=}$ denotes the equality in distribution.

Finally, we wish to design summary features of a point process. Point processes being random by essence, it is natural to consider the expected value of these random measures. This leads to the concept of **intensity measure** (or **mean measure**):

Definition 1.3 *Let N be a point process on $(\mathcal{S}, \mathcal{B})$. The **intensity measure** Λ of N is defined by the expectations:*

$$\Lambda(B) = \mathbb{E}N(B), \quad \forall B \in \mathcal{B}.$$

$\Lambda(B)$ is the expected number of observations in B.

It is quite easy to show that Λ is indeed a measure, convincing ourselves about the legitimity of the appellation. In the Euclidean case where $\mathcal{S} = \mathbb{R}^k$ and $B = [b_1, x_1[\times \cdots \times [b_k, x_k[$, then the **intensity** (or **density**) of Λ (provided that Λ is absolutely continuous), is given by the derivative:

$$\lambda(x_1, \dots, x_k) = \frac{\partial^k \Lambda(B)}{\partial x_1 \cdots \partial x_k}.$$

Intensity measures play a key role in the characterization of point processes, role that will even strengthen up in the particular case of Poisson processes

Even if there is no one-to-one correspondence between distributions of point processes and their intensity measure, we can still notice that two point processes with the same distribution will have the same intensity measure. The converse, which does not hold for general point processes, will appear to be true in the special case of Poisson processes, where we will be able to fully characterize the distribution of a Poisson process by the knowledge of its intensity measure.

We finish this section by presenting the computation of the intensity measures of the two point processes introduced in Example 1.2:

Example 1.3 — *Intensity measures.* The intensity measure of an empirical process is given by

$$\Lambda(B) = n\mathbb{P}\{X_1 \in B\}, \quad \forall B \in \mathcal{B}.$$

For a mixed empirical process, the intensity measure is given by

$$\Lambda(B) = \mathbb{E}N(B) = \mathbb{E}\left[\mathbb{E}\left[N(B)|\xi\right]\right] = \sum_{m=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{m} \epsilon_{X_i}(B)\Big|\xi = m\right] \mathbb{P}\{\xi = m\},$$

$$= \mathbb{P}(X_1 \in B) \sum_{m=1}^{\infty} m\mathbb{P}\{\xi = m\} = \mathbb{P}(X_1 \in B)\mathbb{E}\xi.$$

■

1.3 Poisson Processes

We introduce here the **Poisson process** with a finite intensity measure ($\Lambda(B) < \infty \; \forall B \in \mathcal{B}$), which plays a similar central role as the Gaussian distribution. In fact, in some situations, Poisson processes arise as the limit of many sequences of point processes. Therefore, in applications involving point processes models, we often restrict our attention to the class of Poisson processes for the estimation of the underlying stochastic process.

For a Poisson process, the number of points in a given subset follows a Poisson distribution, and the number of points occurring in separate subsets are mutually independent. Therefore, Poisson processes are ideal candidate models for random scatter. However, there are physical phenomena for which Poisson processes appear to be poor models (e.g. phenomena where there is a natural spacing or clustering).

In our context, it is even more legitimate to restrict our attention to the Poisson process, as it can be shown that the positron emission process is a Poisson process.

Definition 1.4 — *Poisson Process.* *Let Λ be a finite measure ($\Lambda(B) < \infty \; \forall B \in \mathcal{B}$). A point process $N : \Omega \to \mathbb{M}(\mathcal{S}, \mathcal{B})$ is a **Poisson process** if the following properties are satisfied:*

- *For every $B \in \mathcal{B}$,*

$$N(B) \overset{d}{\sim} Poisson(\Lambda(B)),$$

*with $Poisson(\Lambda(B))$ the **Poisson distribution** with parameter $\Lambda(B) \geq 0$. Recall that a discrete random variable $X \in \mathbb{R}$ is said to have a Poisson distribution with parameter $\lambda \geq 0$ if its density function is given by*

$$\mathbb{P}\{X = k\} = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \mathbb{N},$$

with the convention $0^0 = 1$ when $\lambda = k = 0$.

- *For every $k \in \mathbb{N}$, and all disjoint sets $B_1, \ldots, B_k \in \mathcal{B}$, the random variables $N(B_1), \ldots, N(B_k)$ are independent.*

Remark 1.2 *Observe that $\Lambda(B)$ is the intensity measure of N. In fact, from the definition of a Poisson process and well-known properties of the Poisson distribution we have $\mathbb{E}N(B) = \Lambda(B), \quad \forall B \in \mathcal{B}$. Moreover, from Criterion 1.1 and the above definition, it is straightforward to see that Poisson processes with equal finite intensity measure Λ are equal in distribution.*

Poisson processes can be explicitly represented using mixed empirical processes introduced in Example 1.2. More precisely, the Poisson process may be seen as a mixed empirical process with the sample size ξ being a Poisson random variable.

Proposition 1.3 — *Construction of Poisson Processes.* *An explicit representation of a Poisson Process N on $(\mathcal{S}, \mathcal{B})$ with finite intensity measure Λ is given by*

$$N = \sum_{i=1}^{\xi} \epsilon_{X_i},$$

where $\xi, X_1, X_2, X_3, \ldots$ are independent random variables such that

- $\xi \overset{d}{\sim} \mathcal{P}(\Lambda(\mathcal{S}))$,
- $X_i \overset{d}{\sim} \Lambda/\Lambda(\mathcal{S}), \quad \forall i \in \mathbb{N}$.

We can show that the point process defined above is indeed a Poisson process with intensity measure Λ but we will not provide the proof here.

Example 1.4 — *Homogeneous Poisson Process.* Let $\Lambda = c\mu$ with $c \in \mathbb{R}_+$ and μ the Lebesgue measure. Then the homogeneous Poisson process on $[0, T]$ is obtained by $N = \sum_{i=1}^{\xi} \epsilon_{X_i}$, where $\xi \overset{d}{\sim} \mathcal{P}(cT)$, and $X_i \overset{d}{\sim} \mathcal{U}(0, T)$. ■

1.4 Indirectly Observed Poisson Processes

In the context of positron emission tomography, we do not directly observe the positron emission process: we register impacts of gamma rays on the detectors resulting from the encounter of a positron with an electron, naturally present in the brain medium. This is an indirect observation of the stochastic process, leading to the concept of indirectly observed Poisson processes.

Assume we are interested in a Poisson process $N_1 = \sum_{i=1}^{\xi} \epsilon_{X_i}$, in some space $(\mathcal{S}, \mathcal{B})$, with ξ and the X_is as in Proposition 1.3. Imagine now that only *indirect* information on this Poisson process is available to us, through another Poisson process $N_2 = \sum_{i=1}^{\xi} \epsilon_{Y_i}$ in a measurable space $(\mathcal{T}, \mathcal{C})$, where $Y_1, Y_2, \ldots$ are independent and identically distributed random variables such that

$$\mathbb{P}_Y = \int_S K(x) d\mathbb{P}_X(x), \tag{2.1}$$

with $K : (\mathcal{S}, \mathcal{B}) \to (\mathbb{M}(\mathcal{T}, \mathcal{C}), \mathcal{M}(\mathcal{T}, \mathcal{C}))$ a **Markov kernel**:

Definition 1.5 — *Markov Kernel.* *Let $(\mathcal{S}, \mathcal{B})$ and $(\mathcal{T}, \mathcal{C})$ be two measurable spaces. A **Markov Kernel** with source $(\mathcal{S}, \mathcal{B})$ and target $(\mathcal{T}, \mathcal{C})$ is a map*

$$K : \begin{cases} (\mathcal{S}, \mathcal{B}) \to \mathbb{M}(\mathcal{T}, \mathcal{C}), \\ x \mapsto K(x), \end{cases}$$

that associates to each point $x \in \mathcal{S}$ a probability measure $K(x) \in \mathbb{M}(\mathcal{T}, \mathcal{C})$ such that $\forall C \in \mathcal{C}$, the map $x \mapsto K(x)(C)$ is measurable with respect to the σ-algebra $\mathcal{B}$.

Then, from Example 1.3 and (2.1), we obtain

$$\Lambda_2 = \mathbb{E}\xi\mathbb{P}_Y, = \int_S K(x)\mathbb{E}\xi d\mathbb{P}_X(x), = \int_S K(x) d\Lambda_1(x), \tag{2.2}$$

where Λ_1 and Λ_2 are the respective mean measures of N_1 and N_2. If we assume existence of densities for Λ_1, Λ_2 and $K(x)$, we can re-express (2.2) in terms of

densities

$$\lambda_2(y) = \int_S k(y;x)\lambda_1(x)dx,$$

with λ_1, λ_2 the respective densities of Λ_1 and Λ_2, and $k(\cdot;x)$ the density of $K(x)$. From this, we see that the kernel k also relates the intensities of the two Poisson processes. In such a case, we call N_1 an **indirectly observed Poisson process**.

2 Introduction to the Radon Transform

The main results presented in this section are adapted from the Chapters 1,2, 3 and 5 of [2]. We recommend an interested reader to consult this reference for a much more detailed discussion on the Radon transform and its properties, as well as its relationships with the Fourier transform and other classical transforms.

The problem of estimating the density of an indirectly observed Poisson process is a special instance of a larger class of problems, called **reconstruction problems**. These problems all share a common framework: we are interested in some internal distribution of an object (in our case the intensity map of the brain's metabolic activity) that we cannot directly access, often to avoid damaging the object. But we can still **act** on the latter with **probes** (X-rays, gamma rays, micro-waves...), whose action provides a projection or **profile** of the distribution of interest (see Figure 2.2). Then, the problem is to recover the internal distribution based on a collection of such profiles. The **Radon transform** is then a convenient mathematical tool that encapsulates most of the reconstruction problems in the same framework. This operator, denoted by $\mathcal{R}$, maps the distribution of interest (approximated by some density function f) to the projected distribution, or profile $\check{f}$:

PET is a reconstruction problem: we wish to determine the internal structure of an object without dissecting or damaging it. Then, we use internal or external probes that act on the object, providing a collection of profiles of it, from which we seek to estimate properties of interest of the object. The appropriate unifying framework for this type of problems is the Radon transform.

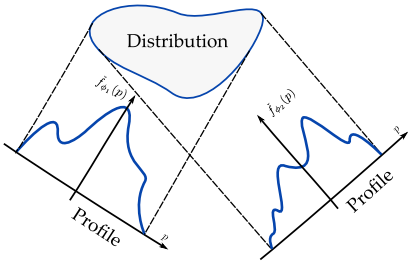


Figure 2.2: Internal distribution and two profiles. Each profile is a projection of the internal distribution (see Vocabulary 2.1 for a rigorous definition).

$$\check{f} = \mathcal{R}f.$$

The Radon transform can be thought of as the formal mathematical representation of the action of the probe onto the distribution (see Figure 2.3).

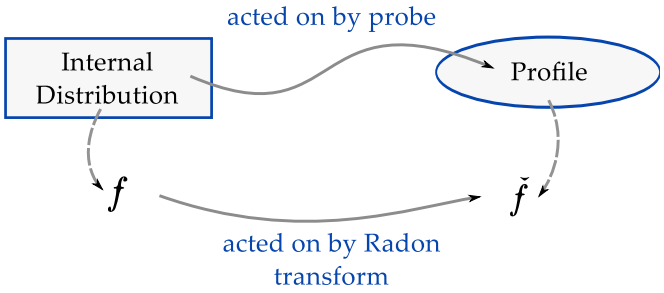


Figure 2.3: Schematic illustration of the mathematical framework. The internal distribution is approximated by a function f, the profile by a function $\check{f}$, and the action of the probe is modeled by the Radon transform, that maps f to $\check{f}$.

2.1 Definition of the Radon Transform

Two dimensions

We will first define the Radon transform in the two dimensional case. We will propose different equivalent definitions: the first one, more intuitive, will help us having a better geometrical understanding of the transform, while the second, a bit more formal, will appear to be particularly well suited to generalize the transform to higher dimensions in a straightforward manner.

Let $(x, y) \in \mathbb{R}^2$, and consider an arbitrary function f defined on some domain $D \subset \mathbb{R}^2$. If we call $L \subset \mathbb{R}^2$ any line in the plane, then the mapping defined by the **projection** or line integral of f along all possible lines L is the **(two-dimensional) Radon transform** of f provided the integral exists.

Then, depending on the chosen parametrization of the line L, the explicit definition of the transform will take different form. A first natural parametrization for L is

$$L = \{(x, y) \in \mathbb{R}^2 : x\cos(\phi) + y\sin(\phi) = p\},$$

with[1] p the distance of the line to the origin, and ϕ the angle between the x-axis and the unit vector ξ normal to the line (see Figure 2.4). Then, the Radon transform $\check{f}$ depends on the values of p and ϕ:

$$\check{f}(p, \phi) = \int_L f(x, y)dl. \tag{2.3}$$

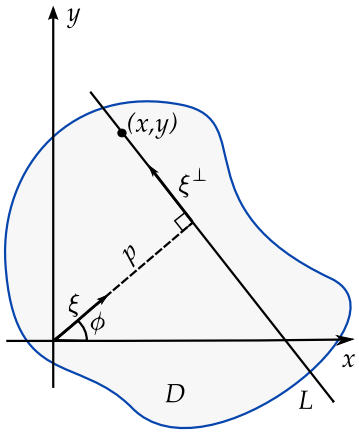


Figure 2.4: Different parametrizations of a line L in the domain D. The distance p to the origin is allowed to take negative values, and $\phi \in [0, \pi]$.

Using the fact that a point $\mathbf{v} = (x, y) \in L$ admits the parametrization $\mathbf{v} = p\xi + t\xi^{\perp}$, with $t \in \mathbb{R}$ and $\xi = (\cos(\phi), \sin(\phi))$, we can re-write (2.3) in a more explicit form:

$$\check{f}(p, \phi) = \int_{-\infty}^{\infty} f\left(p\cos(\phi) - t\sin(\phi), p\sin(\phi) + t\cos(\phi)\right) dt.$$

In practice, because of the discrete nature of the detectors, we only have access to a sample of the Radon transform.

If $\check{f}(p, \phi)$ is known for each p and ϕ, then $\check{f} = \mathcal{R}f$ is the two-dimensional **Radon transform** of f. When $\check{f}$ is only known for certain values of p and ϕ , then we say that $\check{f}$ is a **sample** of the Radon transform.

Vocabulary 2.1 — *Profile.* *Suppose that $\phi = \phi_0$ is held constant as p varies. Then, the points (p, ϕ_0) define a line on the half-cylinder $[0, \pi]\times]-\infty, \infty[$. For each point along this line, we associate a real number $\check{f}_{\phi_0}(p) = \check{f}(p, \phi_0)$. Then, the set of points $\check{f}_{\phi_0}(p)$ is called a **profile** of the function f (see Figure 2.2 and 2.5b).*

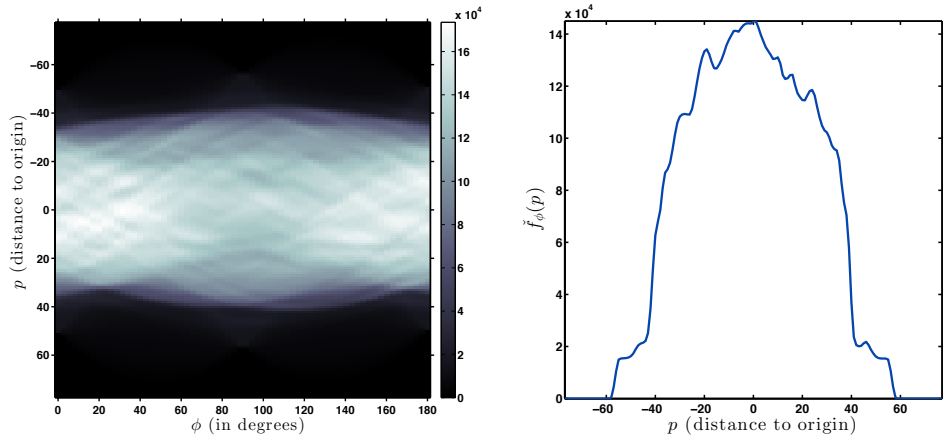
Vocabulary 2.2 — *Sinogram.* *In medical imagery, we call **sinogram** the intensity map of the raw data $\check{f}(p, \phi)$ collected by the detectors (see Figure 2.5a).*

Alternatively, we could also choose another parametrization for the line L. Let ξ and p be as in Figure 2.4. Then, we can express L as:

$$L = \{\mathbf{x} \in \mathbb{R}^2 : p - \xi \cdot \mathbf{x} = 0\}.$$

Using the Dirac delta function $\delta(\cdot)$, we can then write the Radon transform as

[1] in this report, to conform with the convention used in the Matlab function `radon`, we allow p to take negative values and constrain ϕ to the range $[0, \pi]$.



(a) Sinogram. **(b)** Profile for $\phi = 90°$.

Figure 2.5: Example of a sinogram and the associated profile for $\phi = 90°$.

an integral over $\mathbb{R}^2$:

$$\check{f}(p, \xi) = \int_{\mathbb{R}^2} f(\mathbf{x})\delta(p - \xi \cdot \mathbf{x})d\mathbf{x}. \tag{2.4}$$

Higher dimensions

The generalization of (2.4) to higher dimensions is simply obtained by letting $\mathbf{x}$ and ξ lie in $\mathbb{R}^n$. We obtain the following definition:

Definition 2.1 — ***Extension of the Radon transform to*** $\mathbb{R}^n$***.*** *Let* $\mathbf{x} \in \mathbb{R}^n$ *and* $\xi \in \mathcal{S}^{n-1}$*, with* $\mathcal{S}^{n-1}$ *the* $n-1$ *dimensional generalized unit sphere. Let also* $f : \mathbb{R}^n \to \mathbb{R}$ *be an arbitrary function and* $p \in \mathbb{R}$*. Then, the* ***(n-dimensional) Radon transform*** *of* f*, designated by* $\check{f}$*, is obtained by integrating* f *over each hyperplane in the space, provided the integral exists. We note then* $\check{f} = \mathcal{R}f$ *and we can express* $\check{f}$ *in the explicit form:*

$$\check{f} : \begin{cases} \mathbb{R} \times \mathcal{S}^{n-1} \to \mathbb{R}, \\ (p, \xi) \mapsto \int_{\mathbb{R}^n} f(\mathbf{x})\delta(p - \xi \cdot \mathbf{x})d\mathbf{x}, \end{cases}$$

with $\delta(\cdot)$ *the* ***Dirac delta function****.*

Remark 2.1 *The geometrical interpretation developed in the two dimensional case can be extended to* $\mathbb{R}^n$*: the parameter* p *is the perpendicular distance from the origin to the hyperplane and the unit vector* ξ *defines the orientation of the hyperplane.*

We conclude this section by presenting the explicit calculation of the Radon transform of the two-dimensional Gaussian distribution.

Example 2.1 — *Gaussian distribution.* Let $f(x,y) = e^{-\pi(x^2+y^2)}, \forall (x,y) \in \mathbb{R}^2$. Then,

$$\check{f}(p,\xi) = \int_{\infty}^{\infty}\int_{\infty}^{\infty} e^{-\pi(x^2+y^2)}\delta(p-\xi_1 x-\xi_2 y)dxdy, \tag{2.5}$$

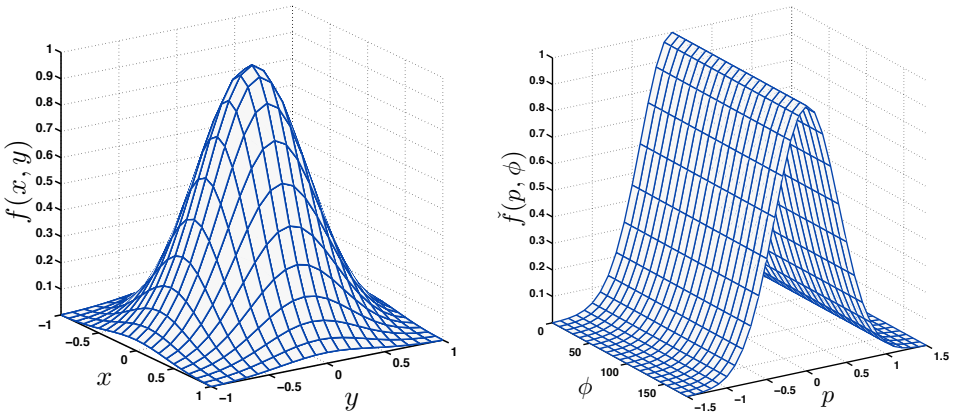
with $\xi = (\xi_1, \xi_2)$ a unit vector. Then, we perform the following orthogonal transformation:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \xi_1 & \xi_2 \\ -\xi_2 & \xi_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

We note that the columns of the matrix are respectively $\xi^{\perp}$ and ξ, two unit vectors perpendicular to each other. Therefore, the transformation is indeed orthogonal, and thus $u^2 + v^2 = x^2 + v^2$. Moreover, in this new basis, the equation of the line $L = \{(x,y) \in \mathbb{R}^2 : \xi_1 x + \xi_2 y = p\}$ becomes simply: $L = \{(u,v) \in \mathbb{R}^2 : u = p\}$ (see Figure 2.4) . Therefore, following this change of variable, (2.5) becomes:

$$\begin{aligned}\check{f}(p,\xi) &= \int_{\infty}^{\infty}\int_{\infty}^{\infty} e^{-\pi(u^2+v^2)}\delta(p-u)dudv, \\ &= e^{-\pi p^2}\int_{\infty}^{\infty} e^{-\pi v^2}dv = e^{-\pi p^2},\end{aligned}$$

where in the last step we performed the change of variable $\sqrt{\pi}v = t$ and the well known result $\int_{\infty}^{\infty} e^{-t^2}dt = \sqrt{\pi}$. Hence, we have: $\mathcal{R}\{e^{-\pi(x^2+y^2)}\} = e^{-\pi p^2}$ (see Figure 2.6). ■



(a) The Gaussian distribution.

(b) Radon transform of the Gaussian distribution.

Figure 2.6: The Gaussian distribution and its Radon transform.

2.2 Basic Properties of the Radon Transform

In this section, we continue our investigation of the Radon transform and present some of its basic properties. We will use all of these properties to compute the Radon transform of a test distribution, that will serve later as a reference to verify the correctness of our discrete Radon transform procedure.

Several elementary properties can be derived directly from the definition of the Radon transform. In all that follows, the functions are assumed smooth enough so that all the quantities involved exist.

Proposition 2.1 — *Linearity.* *Let f and g be two functions and $\alpha, \beta \in \mathbb{R}$. Then, the Radon transform is linear:*

$$\mathcal{R}\{\alpha f + \beta g\} = \alpha \mathcal{R} f + \beta \mathcal{R} g.$$

■ **Proof 2.1** We have $\mathcal{R}\{\alpha f + \beta g\} = \int (\alpha f(\mathbf{x}) + \beta g(\mathbf{x}))\, \delta(p - \xi \cdot \mathbf{x}) d\mathbf{x} = \alpha \check{f} + \beta \check{g}$, which proves the result. ■

Proposition 2.2 — *Transform of a linear transformation.* *Let f be a function, $\mathbf{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ an invertible matrix. Let $(p, \xi) \in \mathbb{R} \times \mathcal{S}^{n-1}$. Then we have:*

$$\mathcal{R}\{f(A\mathbf{x})\} = \frac{|det(B)|}{\|B^T \xi\|} \check{f}\left(\frac{p}{\|B^T \xi\|}, \frac{B^T \xi}{\|B^T \xi\|}\right),$$

with $B := A^{-1}$.

■ **Proof 2.2** By definition, we have:

$$\mathcal{R}\{f(A\mathbf{x})\} = \int f(A\mathbf{x}) \delta(p - \xi \cdot \mathbf{x}) d\mathbf{x}. \tag{2.6}$$

With the change of variable $\mathbf{y} = A\mathbf{x}$ and by defining $B := A^{-1}$, (2.6) becomes

$$\begin{aligned} \mathcal{R}\{f(A\mathbf{x})\} &= |det(B)| \int f(\mathbf{y}) \delta(p - \xi \cdot B\mathbf{y}) d\mathbf{y}, \\ &= |det(B)| \int f(\mathbf{y}) \delta(p - B^T \xi \cdot \mathbf{y}) d\mathbf{y}. \end{aligned} \tag{2.7}$$

It would be tempting to conclude from (2.7) that $\mathcal{R}\{f(A\mathbf{x})\} = |det(B)| \check{f}(p, B^T \xi)$. Unfortunately, $B^T \xi$ is not a unit vector in general, so we have to normalize it in order to avoid violating the domain of definition of $\check{f}$ (see Definition 2.1). To this end, we introduce the following change of variables: $\|B^T \xi\| \mathbf{y} = \mathbf{v}$. Then, (2.7) becomes

$$\mathcal{R}\{f(A\mathbf{x})\} = \frac{|det(B)|}{\|B^T \xi\|^n} \int f(\|B^T \xi\|^{-1} \mathbf{v}) \delta\left(p - \frac{B^T \xi}{\|B^T \xi\|} \cdot \mathbf{v}\right) d\mathbf{v}. \tag{2.8}$$

Let H_0 be the hyperplane $H_0 = \text{span}(\frac{B^T \xi}{\|B^T \xi\|})^\perp$ and $H = H_0 + p \frac{B^T \xi}{\|B^T \xi\|}$ the translated hyperplane selected by the Dirac delta function in (2.8). If $\mathcal{H}_0 = \{\mathbf{h}_1, \ldots, \mathbf{h}_{n-1}\}$ is a basis of H_0, then any $\mathbf{v} \in H$ can be written as

$$\mathbf{v} = p \frac{B^T \xi}{\|B^T \xi\|} + \sum_{i=1}^{n-1} \alpha_i \mathbf{h}_i, \quad \alpha := (\alpha_1, \ldots, \alpha_{n-1}) \in \mathbb{R}^{n-1}.$$

Then, we can express explicitly (2.8) as

$$\mathcal{R}\{f(A\mathbf{x})\} = \frac{|det(B)|}{\|B^T \xi\|^n} \int_{\mathbb{R}^{n-1}} f\left(\frac{p}{\|B^T \xi\|} \frac{B^T \xi}{\|B^T \xi\|} + \sum_{i=1}^{n-1} \frac{\alpha_i}{\|B^T \xi\|} \mathbf{h}_i\right) d\alpha.$$

Finally, with the change of variables $\beta = \alpha/\|B^T\xi\|$, we obtain

$$\begin{aligned}\mathcal{R}\{f(A\mathbf{x})\} &= \frac{|det(B)|}{\|B^T\xi\|}\int_{\mathbb{R}^{n-1}} f\left(\frac{p}{\|B^T\xi\|}\frac{B^T\xi}{\|B^T\xi\|} + \sum_{i=1}^{n-1}\beta_i\mathbf{h}_i\right)d\beta,\\ &= \frac{|det(B)|}{\|B^T\xi\|}\int f(\mathbf{v})\delta\left(\frac{p}{\|B^T\xi\|} - \frac{B^T\xi}{\|B^T\xi\|}\cdot\mathbf{v}\right)d\mathbf{v},\\ &= \frac{|det(B)|}{\|B^T\xi\|}\check{f}\left(\frac{p}{\|B^T\xi\|}, \frac{B^T\xi}{\|B^T\xi\|}\right),\end{aligned}$$

which is the desired result. ■

Proposition 2.3 — *Shifting property.* *Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$ and f a function. Then, the Radon transform of the shifted function $f(\mathbf{x}-\mathbf{a})$ is given by*

$$\mathcal{R}\{f(\mathbf{x}-\mathbf{a})\} = \check{f}(p - \xi\cdot\mathbf{a}, \xi),$$

with $(p,\xi) \in \mathbb{R}\times\mathcal{S}^{n-1}$.

■ **Proof 2.3** $\mathcal{R}\{f(\mathbf{x}-\mathbf{a})\} = \int f(\mathbf{x}-\mathbf{a})\delta(p-\xi\cdot\mathbf{x})d\mathbf{x} = \int f(\mathbf{y})\delta(p-\xi\cdot\mathbf{a}-\xi\cdot\mathbf{x})$. ■

The explicit computation of the Radon transform of the characteristic function of an ellipse will be very useful as the phantom image on which we will test our algorithms later will be composed of ellipses.

Example 2.2 — *Radon transform of the characteristic function of an ellipse.* Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases}1, & \left(\frac{x-x_0}{a}\right)^2 + \left(\frac{y-y_0}{b}\right)^2 \leq 1,\\ 0, & \text{otherwise},\end{cases}$$

with $a, b \in \mathbb{R}_+^*$. We are interested in finding the Radon transform of this function. For this, we observe first that the Radon transform of the characteristic function $u(x,y)$ of a unit circle centered at the origin is simply the length of a chord at a distance p from the center, given by

$$\check{u}(p,\phi) = \begin{cases}2\sqrt{1-p^2}, & |p| < 1,\\ 0, & \text{otherwise}.\end{cases}$$

Consider now the linear transformation defined by the matrix A:

$$A = \begin{pmatrix}\frac{1}{a} & 0\\ 0 & \frac{1}{b}\end{pmatrix}.$$

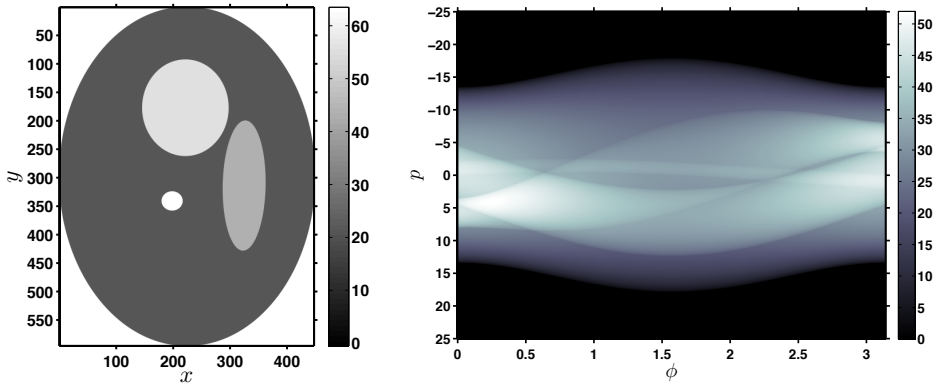
Then, with $\mathbf{x} = (x,y)$, we have: $u(A\mathbf{x}) = e(\mathbf{x})$, where $e(x,y)$ is the characteristic function of the ellipse implicitly defined by: $\{(x,y)\in\mathbb{R}^2 : \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 \leq 1\}$. Then, using the Proposition 2.2, we can express the Radon transform of e as:

$$\begin{aligned}\check{e}(p,\phi) &= \frac{|det(B)|}{\|B^T\xi\|}\check{u}\left(\frac{p}{\|B^T\xi\|}, \frac{B^T\xi}{\|B^T\xi\|}\right),\\ &= \begin{cases}\frac{2ab}{s}\sqrt{1-\left(\frac{p}{s}\right)^2}, & \left|\frac{p}{s}\right| < 1,\\ 0, & \text{otherwise},\end{cases}\end{aligned}$$

where $s = \sqrt{a^2\cos^2(\phi) + b^2\sin^2(\phi)}$ and $B := A^{-1}$. Finally, as $f(x, y) = e(x - x_0, y - y_0)$ we conclude by invoking Proposition 2.3 with $\mathbf{a} = (x_0, y_0)^T$. We obtain:

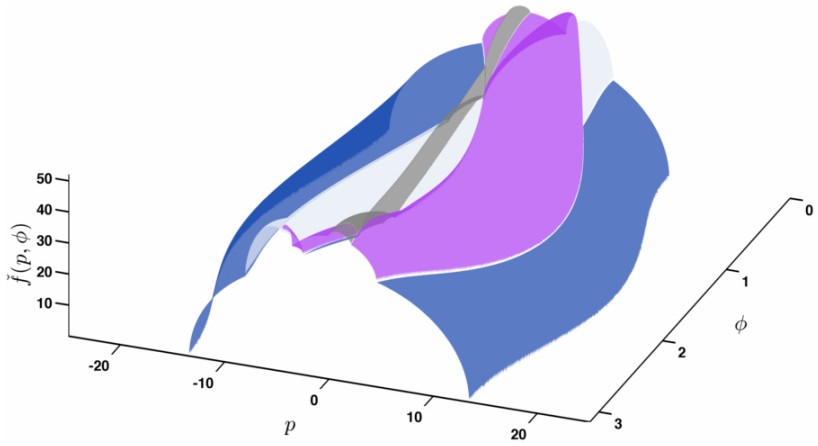
$$\check{f}(p,\phi) = \begin{cases} \frac{2ab}{s}\sqrt{1 - \left(\frac{p-p_0}{s}\right)^2}, & \left|\frac{p-p_0}{s}\right| < 1, \\ 0, \quad \text{otherwise}, \end{cases} \tag{2.9}$$

where $s = \sqrt{a^2\cos^2(\phi) + b^2\sin^2(\phi)}$ and $p_0 = x_0\cos(\phi) + y_0\sin(\phi)$. ∎



(a) Sum of ellipses' characteristic functions with heights given by the color of each ellipse.

(b) Sinogram.



(c) Radon transform. We can observe that the Radon transform is the sum of the four Radon transforms.

Figure 2.7: The Radon transform of a sum of ellipses' characteristic functions.

In Figure 2.7, we use this result together with the linearity of the Radon transform to compute the Radon transform of the sum of four ellipses' characteristic functions. Each characteristic function have been rescaled by a factor directly proportional to the grayscale image on Figure 2.7a.

2.3 Inversion of the Radon Transform

The problem we are facing is an inverse problem: we want to recover an internal distribution f from its Radon transform $\check{f}$, the only quantity we have access to. For this problem to admit a unique solution, it would be desirable that the Radon transform be an invertible operator.

Interestingly enough, an inversion formula for the Radon transform depends on the parity of the ambient space's dimension. Therefore, it will be convenient to consider the inversion formula in two parts, depending on wether $\mathbb{R}^n$ is odd or even. A detailed demonstration of the inversion formula is not in the scope of this work, and an interested reader can find more details in the Chapter 5 of [2]. In all that follows, we assume that the functions are smooth enough so that all the quantities involved have a well-defined meaning.

Theorem 2.4 — *Inversion formula.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth enough function (let's say $\mathcal{C}^\infty$ with compact support), and $\check{f}$ its Radon transform: $\check{f} = \mathcal{R}f$. Then, with the same notations as before, $\Delta_{\mathbf{x}}$ the **Laplacian operator**, and $\mathcal{C}_n := \frac{1}{2}(2\pi i)^{-(n-1)}$,we have

- **Odd dimension ($n \geq 3$):**

$$f(\mathbf{x}) = \mathcal{C}_n \Delta_{\mathbf{x}}^{(n-1)/2} \int_{|\xi|=1} \check{f}\left(\xi \cdot \mathbf{x}, \xi\right) d\xi, \quad \forall \mathbf{x} \in \mathbb{R}^n, \tag{2.10}$$

- **Even dimension ($n \geq 2$):**

$$f(\mathbf{x}) = \frac{\mathcal{C}_n}{i\pi} \int_{|\xi|=1} \int_{-\infty}^{\infty} \frac{\left(\frac{\partial}{\partial p}\right)^{n-1} \check{f}\,(p, \xi)}{p - \xi \cdot \mathbf{x}} dp\, d\xi, \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{2.11}$$

Remark 2.2 *It is interesting to note that the inversion problem is changing of nature depending on the evenness of the dimension. Indeed, for $n \geq 3$ odd we see in (2.10) that the problem is* local*, in the sense that we only need to know the Radon transform $\check{f}$ for $p = \xi \cdot \mathbf{x}$ (i.e. on planes that go through x). On the other hand, for $n \geq 2$ even, we see in (2.11) that the knowledge of $\check{f}$ for $p = \xi \cdot \mathbf{x}$ is not sufficient anymore: the problem becomes* global *as we integrate $\check{f}$ for all $p \in \mathbb{R}$ (i.e. for all planes) . Considering this intriguing behavior, one could argue on the physical admissibility of the Radon transform to model the action of the probes onto the distribution of interest. However, as surprising as this phenomenon might be, there are examples of physical problems that behave in the same fashion: on page 80 of [Evans] a similar behavior is noticed for the solution of the wave equation. Therefore, despite its exotic behavior, the Radon transform might still be a legitimate framework for reconstruction problems.*

As every inverse problem, the PET reconstruction problem is intrinsically ill-posed. Therefore, we have to expect instabilities in practice, complicating the task of reconstructing the metabolic activity from the sinogram.

Remark 2.3 *The inversibility of the Radon transform means that if we were able to measure $\check{f}$ with an infinite precision, then we could exactly recover f. Unfortunately in practice we cannot reach arbitrary precision in measurements.*

If the invertibility of the Radon transform is a necessary condition for the feasibility of our reconstruction problem, it is unfortunately not sufficient to ensure the well-posedness of it. As a matter of fact, the study of the singular value decomposition of the Radon transform made by Johnstone et al. [9] and Natterer [3] reveals that the reconstruction problem is a **midly ill-posed problem**: the spectrum of the Radon transform operator *decays polynomially*.

This rather gentle ill-posedness at the continuous level could worsen a lot during the discretization[2] of the problem, resulting in a numerically badly behaved problem.

3 The EM Algorithm

In this section, we describe the so-called EM algorithm, a general-purpose algorithm for maximum likelihood estimation. This algorithm is applicable in a wide variety of situations, referred to as incomplete-data problems. We will see that the PET reconstruction problem can be expressed as such.

In this section, we follow the approach of the Chapter 1 of [4], which outlines a unified and complete treatment of the theory and methodology of the EM algorithm and its extensions, as well as applications. The EM algorithm, or the **expectation-maximization algorithm** is one of the most important algorithm in modern statistics, as it allows to conveniently compute the maximum likelihood estimator in a wide variety of situations, where the maximization of the likelihood function would be otherwise very challenging. Problems where the EM algorithm has been successfully applied are referred to as *incomplete-data problems*. This very general term encapsulates not only evidently incomplete-data situation, such as problems with missing data, truncated distributions, censored or grouped observations, but also more general situations, where the incompleteness of the data is far from being obvious. The idea behind the EM algorithm is quite simple, and therefore it has been discovered and re-discovered many times in various forms and in several different fields of science. However, Dempster, Laird and Rubin [11] were the first to propose a general formulation of the algorithm and to derive its main properties. Following this celebrated paper, a whole variety of applications of the algorithm have appeared in the literature.

Assume that we observe an output $\mathbf{y}$ of a random vector $\boldsymbol{Y} \in \mathbb{R}^n$ and we know that its intensity function $f(\mathbf{y}|\boldsymbol{\theta})$ depends on an unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. The ultimate goal is to find the MLE of $\boldsymbol{\theta}$ by maximizing the likelihood function $L(\boldsymbol{\theta}; \mathbf{y})$, (or equivalently the log-likelihood function). Unfortunately, it might occur that this maximization is very difficult to perform, because of some properties of the likelihood function, but almost straightforward with the knowledge of some additional unobserved data. Then, regarding $\boldsymbol{Y}$ as an incomplete version of some complete data random vector $\boldsymbol{X}$ with density $f_c(\mathbf{x}|\boldsymbol{\theta})$ might simplify a lot the apparently challenging task of finding the MLE. Formally, we have $\boldsymbol{Y} = g(\boldsymbol{X})$ for some non-injective function g. Often, we choose g to be a projection on some linear subspace. In this context, we call $\mathbf{x}$ the vector containing the **augmented data** $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, with $\mathbf{z}$ the vector containing the **additional unobservable data** (also referred to as **latent variables**). Then, g is the projection of $\mathbf{x}$ onto the $\mathbf{y}$ component, and the incomplete-data likelihood is given by

$$L(\boldsymbol{\theta}; \mathbf{y}) = \int f_c(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z} = \int L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) d\boldsymbol{z},$$

[2] In practice, the discretization of the problem is enforced by the detectors, that cannot reach arbitrary precision. Therefore, the continuous case is only a theoretical framework, but full of insights regarding the origin of the ill-posedness of the problem.

with $L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z})$ the complete-data likelihood. Often, the complete-data likelihood is in a nice form that greatly simplifies its maximization, relatively to the maximization of the incomplete-data likelihood function. The idea of the EM algorithm is then to exploit this reduced complexity in order to indirectly solve the original maximization problem of the incomplete-data likelihood function.

At this stage, the only problem is that as nice as the complete-data likelihood might be, we are unable to maximize it, as we cannot observe $\mathbf{x}$. Therefore, we compute the expectation of the complete-data likelihood function $L(\boldsymbol{\theta}; \mathbf{x})$ (or equivalently of the complete-data log-likelihood $l(\boldsymbol{\theta}, \mathbf{x})$), given the observed data $\mathbf{y}$ and the current estimate of $\boldsymbol{\theta}$. More precisely, we compute

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \mathbb{E}[l(\boldsymbol{\theta}; \mathbf{x}) | \boldsymbol{y}, \boldsymbol{\theta}^{(k)}], \tag{2.12}$$

with $\boldsymbol{\theta}^{(k)}$ the current estimate of the parameter $\boldsymbol{\theta}$. This step is called the **expectation step (or E-step)**. Then, we update the estimate $\boldsymbol{\theta}^{(k)}$ during the **maximization step (or M-step)**. The new estimate $\boldsymbol{\theta}^{(k+1)}$ is chosen as the maximizer of (2.12), namely

$$\boldsymbol{\theta}^{(k+1)} = \text{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}).$$

Then, we repeat iteratively these two steps on the new estimate $\boldsymbol{\theta}^{(k+1)}$, until we detect the convergence of the algorithm with some stopping criterion. To summarize, the EM algorithm is given by:

► The EM Algorithm

Input: An initial estimate $\boldsymbol{\theta}^{(0)}$ of θ and observations $\boldsymbol{y}$.
Output: The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$.

1. k:=0;
2. **while** the stopping criterion is not met **do**;
   - **E-step:** Compute $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) := \mathbb{E}[l(\boldsymbol{\theta}; \mathbf{x}) | \mathbf{y}, \boldsymbol{\theta}^{(k)}]$;
   - **M-step:** $\boldsymbol{\theta}^{(k+1)} := \text{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$;
   - k:=k+1;
   - Compute stopping criterion;

   endwhile;
3. $\hat{\boldsymbol{\theta}}_{ML} := \boldsymbol{\theta}^{(k)}$;

Remark 3.1 *It might occur that the maximum in the M-step is not unique during the iteration. In this case, it is sufficient to choose any of the $\boldsymbol{\theta}$'s maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$. The choice of the stopping criterion depends on the considered problem. A common criterion is to stop the iteration when the improvement in the log-likelihood function is less than a certain threshold ϵ. In the context of PET, we will see that such a naive stopping criterion is not sufficient, and will design a goodness-of-fit stopping criterion.*

We conclude this section, with an important result, proved by Dempster, Laird and Rubin [11], that tells us that after each iteration of the EM algorithm, the incomplete-data likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ has either increased or remains unchanged.

Theorem 3.1 The EM sequence of likelihoods $\{L(\boldsymbol{\theta}^{(k)}; \mathbf{y})\}_{k\in\mathbb{N}}$ increases monotically, that is

$$L(\boldsymbol{\theta}^{(k+1)}; \mathbf{y}) \geq L(\theta^{(k)}; \mathbf{y}), \ \forall k = 0, 1, 2, \ldots$$

■ **Proof 3.1** See [4] Section 3.2. ■

An immediate consequence of this theorem is that for bounded likelihood functions $L(\theta, \mathbf{y})$ the sequence $\{L(\boldsymbol{\theta}^{(k)}; \mathbf{y})\}_{k\in\mathbb{N}}$ converges to a limit L^*. Care must be taken however not to overinterpret this theorem. First, the convergence of the sequence does not imply the existence of a $\boldsymbol{\theta}^*$ such that $L^* = L(\boldsymbol{\theta}^*; \mathbf{y})$, nor does it imply that L^* is a maximum of $L(\boldsymbol{\theta}, \mathbf{y})$. Moreover, even if under some regularity assumptions we can show that there exists a stationary point $\boldsymbol{\theta}^*$ of $L(\boldsymbol{\theta}, \mathbf{y})$ such that $L(\boldsymbol{\theta}^{(k)}, \mathbf{y}) \to L(\boldsymbol{\theta}^*, \mathbf{y})$, nothing tells us that the iterates $\boldsymbol{\theta}^{(k)}$ actually converge to the stationary point $\boldsymbol{\theta}^*$.

Hopefully in the particular case of PET, we will be able to show that these desirable results hold.

3

A Statistical Model

The statistical model derived in this chapter is inspired from the developments in [7]. However, our approach differs a little: while they chose to model the problem in terms of a thinned Poisson process, with each positron emission having a certain probability to be detected by a given detector, we preferred to reformulate everything in terms of two Poisson processes, with intensities relating through the Radon transform. Even if the two formulations are still very similar, we believe that the Radon transform framework provides a better understanding on the origin of the instabilities encountered during the image reconstruction process.

1 Construction of a Mathematical Model

We begin our investigation by proposing a mathematical model tailored to the physics of the positron emission process. To facilitate the task, let us first make the unrealistic assumption of a perfect knowledge of the location of each positron emission within the brain. Under this assumption, a collection of positron emissions occurring during a certain fixed period of time has every aspect of a *random scatter* (see Figure 3.1). In Chapter 2 we introduced point processes and particularly Poisson processes, that revealed to be ideal candidate models for the study of such random samples. It would then seem legitimate to try and model the stochastic rule governing the positron emissions with a Poisson process. But such a choice of model is not as straightforward as we would like it to be, and we should not rush headlong into it. In fact, there are situations in which Poisson processes appear to be very poor models. This is particularly true when the studied phenomenon exhibits a natural clustering[1], as it seems to be the case in Figure 3.1. Hopefully, in our context we can invoke empirical arguments to decide on the validity of the model. Indeed, we have observed experimentally that the

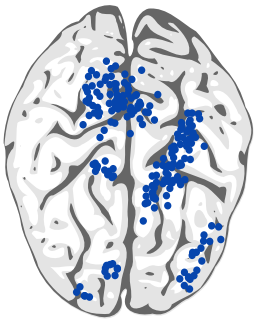


Figure 3.1: Schematic illustration of a random scatter of positron emissions within the brain.

[1] Because of the property of independence in Definition 1.4 , Poisson processes are unable to capture spatial dependencies between observations.

positron emissions occur independently from one another (or at least there is no strong evidence for the contrary). Thus, the clusters observed in Figure 3.1 do not result from a dependency among emissions, but simply from the metabolic activity of the brain, that will naturally concentrate the radioactive tracer in areas stimulated by the cerebral activity of the patient.Another way to formulate this is to say that, conditional on the metabolic activity, the positron emissions are observations from independent random variables.

The empirically observed independence of positron emissions allow us to investigate Poisson processes as possible models for this phenomenon.

This empirical observation preventing us from one of the main pitfall of Poisson processes, we can now embrace this model with a bit more confidence. Let us first re-express all the quantities involved in terms of the vocabulary of point processes, introduced in Chapter 2.

▶ PET in the Poisson process framework

- We consider as the **state space** $\mathcal{S}$ the brain, that we see as a compact subset of $\mathbb{R}^2$, endowed with the Borel $\sigma-$algebra $\mathcal{B}$.
- For every $B \in \mathcal{B}$, we consider positron emissions as the output of many independent **Bernoulli trials**: for each radioactive nucleus contained in a subregion B of the brain, we have two possible outcomes, "success" if it decayed and emitted a positron during the fixed period of observation, "failure" if not (see Figure 3.2).
- Finally, let N be a spatial point process in $(\mathcal{S}, \mathcal{B})$ with finite mean measure Λ, according to which the positron emissions occur.

Under the above modeling assumptions, we can show that the point process N is approximately a Poisson process. In fact, for every $B \in \mathcal{B}$, the number of positron emissions $N(B)$ is a sum of many independent Bernoulli trials. Therefore, the exact distribution of $N(B)$ is a binomial distribution, but as the number of radioactive nuclei is extremely large in practice and the decay probability is very small, we can use asymptotic results and approximate this distribution with a Poisson distribution, with rate $\Lambda(B)$ (this is a famous result known as the *law of rare events*). Finally, the second property of Definition 1.4 follows directly from the independence of positron emissions. This achieves

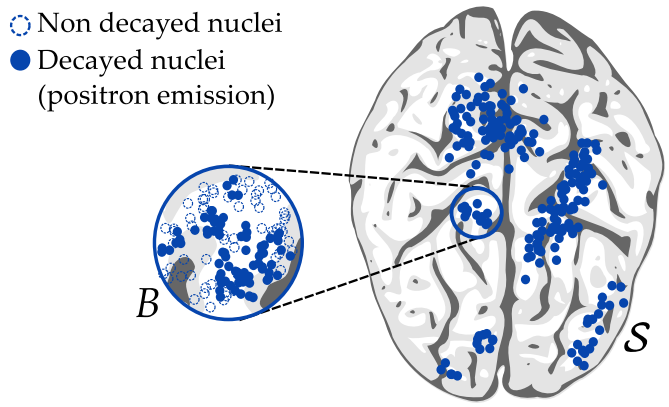


Figure 3.2: Positron emissions can be seen as the output of many Brenoulli trials.

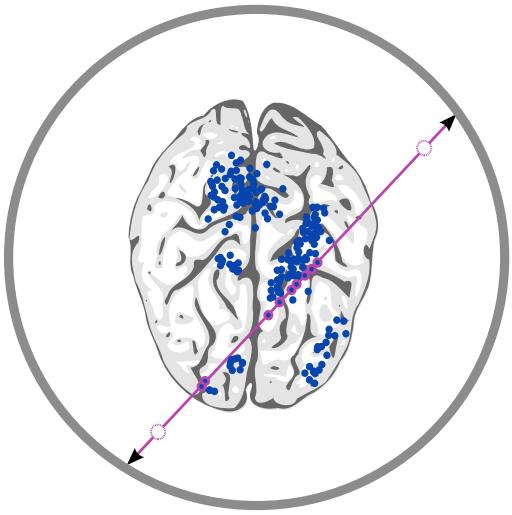


Figure 3.3: By placing a ring of detectors around the head of the patient, we can record gamma rays resulting from the annihilation of a positron with an electron. These gamma rays are recorded in coincidence by a pair of detectors. In the ideal case of infinitely many detectors, a pair of detectors can be identified by a chord on the detector ring.

to convince us that the Poisson process is a reasonable model for the positron emissions process.

Then, if we still assume the knowledge of the locations of each positron emission, we could very easily exploit this model in order to assess the metabolic activity of the brain, that would simply be given by an **estimate of the mean measure**[2] Λ , or equivalently of its **intensity function** λ. Such an estimate could be obtained for example by a 2D histogram of the emissions' locations.

Unfortunately, things are not so simple in practice. In fact, as explained in the Section 1 of the Introduction, we *cannot identify the exact location of positron emissions*. The best we can do is to determine a cylindrical volume in which the emission occurred, by recording coincidences of gamma rays resulting from the annihilation of an emitted positron with an electron (see Section 1 of the Introduction for a more detailed explanation). To do so, a ring of detectors is positioned around the head of the patient. In the ideal case of an infinite number of detectors, a pair of detectors can be associated with a chord on the detector ring (see Figure 3.3). Then, instead of looking at a random sample of positron emissions in the *brain space*, we observe a random sample of coincidences in the *detector space*, the set of all possible chords on the detector ring. This new random sample inherits the independence property of the positron emissions and therefore can be modeled by another Poisson process N^*. Of course, these two Poisson processes are linked together and we can make this link rigorous by the use of indirectly observed Poisson processes, introduced in Section 1.4, and the Radon transform, introduced in Section 2. If λ^* is the intensity function of N^* and λ is the intensity function of N, then we have:

The stochastic process governing positron emissions can be modeled as an indirectly observed Poisson process.

$$\lambda^* = \mathcal{R}\lambda. \tag{3.1}$$

In this formulation, N is said to be an **indirectly observed Poisson process**. It

[2]Recall that for every $B \in \mathcal{B}$, $\Lambda(B)$ is the expected number of emissions in B, proportional to the consumption of glucose and therefore indicator of the metabolic activity in B.

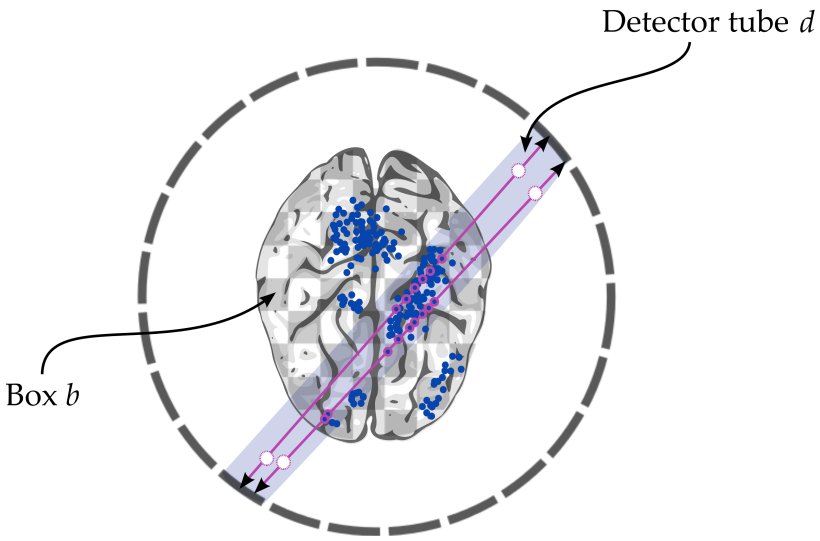


Figure 3.4: In practice, the number of detectors is necessary finite. Pairs of detectors form a detector tube. We count the number of coincidences in each tube. Finally, to simplify the mathematics, we assume the existence of a grid of B boxes covering the brain, such that λ is constant on each box.

is possible to provide some intuition about this modeling approach. In fact, if we parametrize chord on the detector ring by two coordinates $(p, \xi) \in \mathbb{R} \times \mathbb{S}^1$, we can re-write (3.1) as

$$\lambda^*(p, \xi) = \int_{\mathbb{R}^2} \lambda(\mathbf{x}) \delta(p - \xi \cdot \mathbf{x}) d\mathbf{x}, \quad \forall (p, \xi) \in \mathbb{R} \times \mathbb{S}^1,$$

with p, ξ as in Figure 2.4. Then, λ^* can be interpreted as the intensity function of the expected number of counts recorded by a detector aimed along a line defined by p and ξ (see Figure 3.3).

Of course, in practice the number of detector cells is *necessary finite*, and thus we can only have access to a **sample of the Radon transform**. In this situation, a pair of detectors no longer form a line but a cylindrical volume, to be referred to as a **detector tube** (see Figure 3.4). Moreover, if D is the number of detector tubes, we have $\forall d \in \{1, \dots, D\}$,

$$\lambda^*(d) = \int_{\text{tube}_d} \lambda^*(p, \xi) d\xi dp = \int_{\text{tube}_d} \left(\int_{\mathbb{R}^2} \lambda(\mathbf{x}) \delta(p - \xi \cdot \mathbf{x}) d\mathbf{x} \right) d\xi dp, \tag{3.2}$$

with $\lambda^*(d)$ the expected number of counts in tube d.

The goal is now to estimate $\lambda^*(1), \dots, \lambda^*(D)$ from an output of the random vector $(N^*(1), \dots, N^*(D))$ where $N^*(d) \sim \text{Poisson}(\lambda^*(d))$ is the number of coincidences detected in the tube d. Then, based on these estimates, we want to recover λ by exploiting (3.2). But at this stage, this is an *infinite dimensional problem*: we have to estimate a continuum of parameters on the basis of finite data. Therefore, to simplify the problem, we will assume that there exists a grid of B boxes covering the brain and that in the bth box the value of $\lambda(\mathbf{x})$ is a constant. Thus, the emission density λ is a **step function** that can be written as

To simplify the problem, we look for λ in the class of step functions.

follows:

$$\lambda(\mathbf{x}) = \sum_{b=1}^{B} \frac{\lambda(b)}{V(b)} \mathbb{1}_{\text{box}_b}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{S} \subset \mathbb{R}^2, \tag{3.3}$$

where $\lambda(b),\ b = 1, \ldots, B$ is the expected number of emissions in box b, $V(b)$ is the volume of the bth box and $\mathbb{1}_{\text{box}_b}$ denotes the indicator function of the bth box. Finally, substituting (3.3) in (3.2) yields

$$\begin{aligned} \lambda^*(d) &= \sum_{b=1}^{B} \lambda(b) \overbrace{\frac{1}{V(b)} \int_{\text{tube}_d} \left(\mathcal{R}\mathbb{1}_{\text{box}_b}\right)(p, \xi) dp d\xi}^{:=R(d,b)}, \\ &= \sum_{b=1}^{B} R(d, b)\lambda(b). \end{aligned} \tag{3.4}$$

If we fill a matrix $R \in \mathbb{R}^{D \times B}$ with the coefficients $R(d, b)$, we can re-write (3.4) as

$$\boldsymbol{\lambda}^* = R\boldsymbol{\lambda}, \tag{3.5}$$

with $\boldsymbol{\lambda}^* = (\lambda^*(1), \ldots, \lambda^*(D))^T \in \mathbb{R}^D$ and $\boldsymbol{\lambda} = (\lambda(1), \ldots, \lambda(B))^T \in \mathbb{R}^B$. We will refer to R as the **Radon matrix**. The goal is now to recover λ from an estimate of λ^* and the knowledge of R.

2 Computing the Radon Matrix

To compute the discrete Radon transform R, we used the Matlab function `radon`. This function takes two inputs (see Figure 3.5):

- A gray scale image, of which we wish to compute the discrete radon transform. The dimensions of this image naturally define the grid in the brain space: a box correspond to a pixel. Moreover, the expected number of counts $\lambda(b)$ at each box b is given by the grayscale intensity of the corresponding pixel.
- A range of angles (in degrees), that determines the spacing between each detector cell on the detector ring.

Then, the function returns the discrete radon transform of the image as a **sinogram** image (see Vocabulary 2.2 for a definition of a sinogram). The dimensions of this image are determined from the range of angles provided to the function and the dimensions of the input image (more precisely the diagonal of the input image, see Figure 3.5). Finally, the grayscale intensity of this image corresponds to the intensity function λ^*.

The `radon` function does not provide directly the Radon matrix R as an output. However, we can form this matrix very easily for images with given dimensions $N = m \times n$. First, we consider the image as a vector in $\mathbb{R}^N$. Then, we compute the discrete Radon transform of the basis vectors $e_i = (\delta_{ij})_{j=1,\ldots,N} \in \mathbb{R}^N$ with the function `radon`. From a classic result of linear algebra, the image of

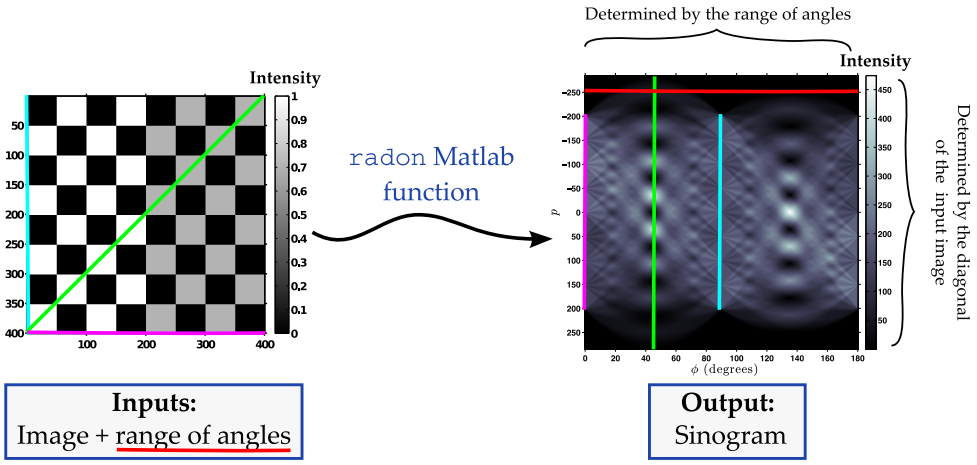


Figure 3.5: A typical output of the Matlab function `radon`. The dimensions of the output image are determined by the range of angles provided as an input and the diagonal of the input image. The intensities of the input and output images are to be interpreted respectively as the densities λ and λ^*.

each of these basis vectors e_i, $i = 1, \ldots, N$, will form the ith column of the matrix R.

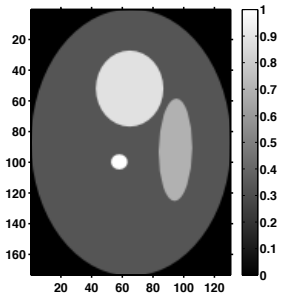
2.1 Validation of the Procedure

In order to validate the preceding procedure, we will use the phantom image[3] from Figure 2.7 as a testbed, and compare the discrete Radon transform with the exact Radon transform (calculated in Example 2.2). A qualitative comparison of the two sinograms is presented in Figure 3.6, where we chose 120 equally spaced detectors cells to compute the discrete Radon transform. We observe that the two sinograms appear very similar both in shape and intensity: the only apparent effect of the discretization process is a slight pixelation of the discrete sinogram Figure 3.6b. Of course this pixelation becomes stronger as the number of detectors decreases, but this is an expected behavior.

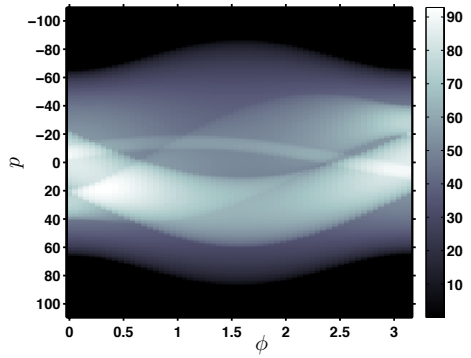
3 Computer Simulation of a PET Scan

In order to test the relative efficiencies of different estimators of λ, we need to simulate the output of a PET scan. For this, we first draw a **phantom image** (see Figure 3.7) that describes the intensity λ from which we wish to generate the data and that we wish to recover. This phantom image is made of eight ellipses, that are meant to provide a simplification of both the brain's geometry and metabolic activity: the skull, ventricles and tumors metabolize at different rates, between 0.1 and 2.0. This phantom image is widely used among researchers in tomography to test the numerical accuracy of reconstruction algorithms (see for example [7]). The rates in Figure 3.7 have been converted to grayscale values, that are naturally understood as intensities by the procedure described in Section 2. We then resize this image so that its dimensions are 100×128 pixels. As explained in Section 2 resizing the image is equivalent as choosing a bi-dimensional grid of size

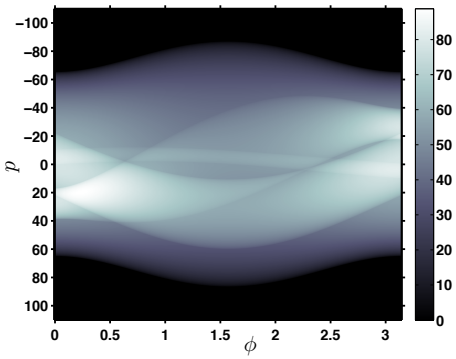
[3] Test images such as the one in Figure 3.6a are usually referred to as **phantom images** by researchers in tomography.



(a) Phantom image.



(b) Sinogram computed with the discrete Radon transform.



(c) Exact Sinogram.

Figure 3.6: Qualitative comparison of the discrete Radon transform **(b)** and the exact Radon transform **(c)** of the same phantom image **(a)**. The grid in the brain space is determined by the pixels of the phantom image **(a)**. To compute the discrete Radon transform in **(b)** we chose 120 equally spaced detectors cells.

$B = 100 \times 128$ that covers the brain. In order to provide physical interpretability to each of the values $\lambda(b)$, $b = 1, \ldots, B$, we rescale them by a constant factor, so that the total number of positron emissions is equal to 10^5 (approximately the number of coincidences we observe during a PET scan).

In the detector space now, we choose a sector ring composed of 60 equally spaced detectors, which gives us $D = 3540$ detector tubes. Thanks to the procedure previously described, we then compute the discrete Radon transform λ^* of the image, which gives us the sinogram in Figure 3.8a. Then, for each $d = 1, \ldots, D$, we simulate a Poisson random variable with mean $\lambda^*(d)$. We obtain this way a new sinogram (see Figure 3.8b) that we interpret as the output of a simulated PET scan (number of coincidences detected in each detector tube), and that we will use in order to recover λ.

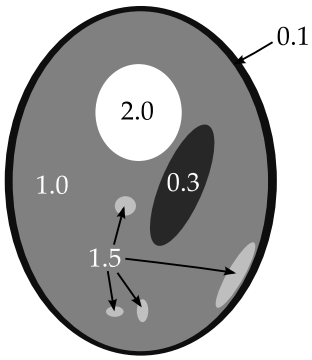
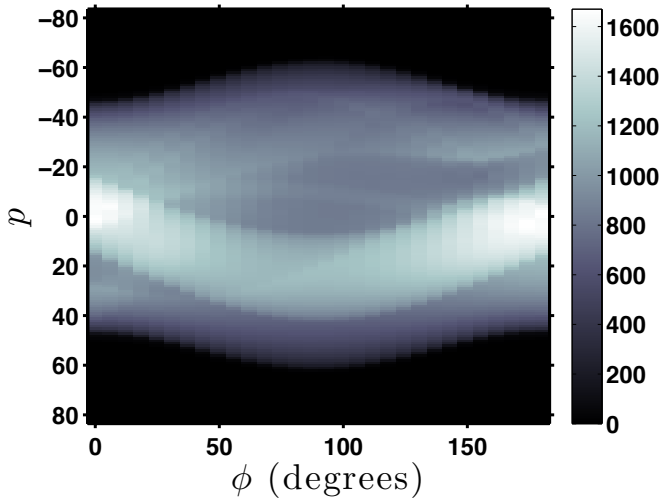
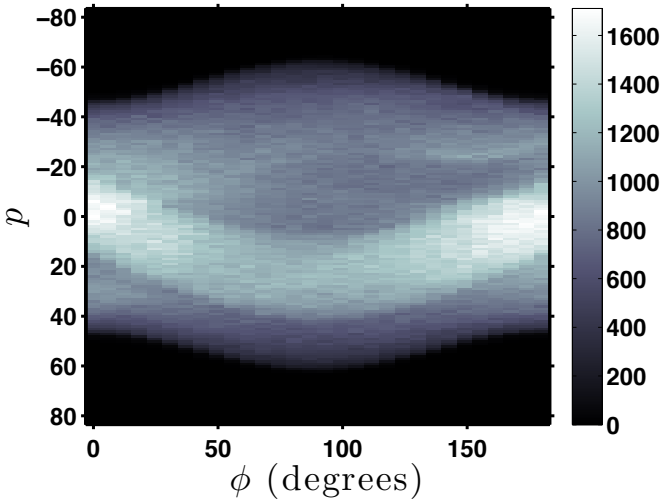


Figure 3.7: Phantom image used in the computer simulation of the PET scan. This image is a simplification of the brain's geometry and metabolic activity. Smaller ellipses represent features in the brain (ventricles, tumors, etc...) metabolizing at different rates. This image was reproduced from the phantom image proposed in [7].

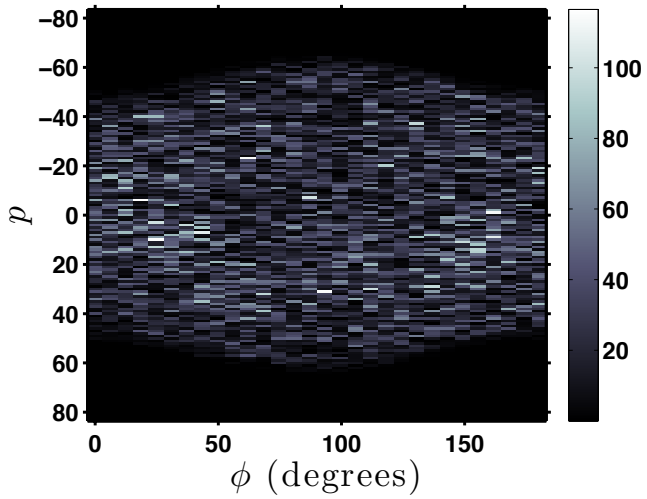
Figure 3.8: Simulation of the output of a PET scan **(b)** by generating Poisson noise with densities given by the sinogram **(a)** of the phantom image in Figure 3.7.



(a) Sinogram of the phantom image in Figure 3.7.



(b) Poisson generated sinogram with densities given by **(a)**.



(c) Difference between the two sinograms.

4

Estimation of the Poisson Intensity

In this chapter we investigate and compare the relative performance of various estimators of the intensity λ, proposed by Vardi et al. in [7]. We begin naively our investigation with the classic least squares estimator, that will, as it could have been expected by the nature of the problem, perform very badly. To try and provide a cure to the bad performances of this estimator, we propose to constrain it positively, as proposed in [7]. This positively constrained least squares estimator will provide much more satisfying results than the unconstrained one, but still not entirely satisfactory (especially in terms of computational cost of the estimation). Finally, we investigate the maximum likelihood estimator, that we compute by means of the classic EM algorithm, after having re-expressed the problem in terms of a standard problem in statistical estimation for incomplete data. As we will show, the quality of the last two estimators is highly dependent on the choice of a good stopping criterion. We propose here a *goodness-of-fit* stopping criterion, similarly at what has been proposed by Veklerov et al. in [8]. However, we design a different statistical test than the on proposed in [8] , that we believe provides a interesting mathematical interpretation of the problem of overfitting that occurs when letting the EM algorithm run for too long.

1 Least Squares Estimator

A sensible estimate of $\boldsymbol{\lambda}^*$ is given by the moment estimator. Indeed, for every $d = 1, \dots, D$, $\lambda^*(d)$ is the expected number of coincidences falling into the dth pair of detectors. As the typical output of a PET experiment is a single observation $\mathbf{n}^* = (n^*(1), \dots, n^*(D))^T \in \mathbb{R}^D$ from N^*, the empirical mean is simply the observed data and the moment estimator[1] $\hat{\boldsymbol{\lambda}}^*$ of $\boldsymbol{\lambda}^*$ is given by

$$\hat{\boldsymbol{\lambda}}^* = \mathbf{n}^*.$$

Using (3.5) we can obtain an estimate of $\boldsymbol{\lambda}$ by solving

$$\mathbf{n}^* = R\hat{\boldsymbol{\lambda}}. \tag{4.1}$$

[1] which is also the maximum likelihood estimator in this case

If we try and solve system (4.1) three problems can arise:

1. the linear system might have no solution ($\mathbf{n}^*$ does not belong to the range of R),
2. the solution to this system might not be unique,
3. the solution might be very sensitive to errors in the data, because of the bad conditioning of the matrix R.

In practice, the first issue often occurs when R is not full row-rank (for example when the number of detector tubes is greater than the number of pixels). Then, $\mathbf{n}^*$ might fall far enough from $\boldsymbol{\lambda}^*$ and therefore not belong to the range of R. Such a behavior is the rule rather than the exception in PET, because of the high number of detector tubes with low counts rate. These detector tubes having low densities $\lambda^*(d)$, the standard deviation $\sqrt{\lambda^*(d)}$ of the Poisson random variable $N^*(d)$ is comparable with $\lambda^*(d)$ and therefore the observations $n^*(d)$ might fall far away from $\lambda^*(d)$. To avoid this undesirable situation, we can alternatively try and solve the following **normal equations**:

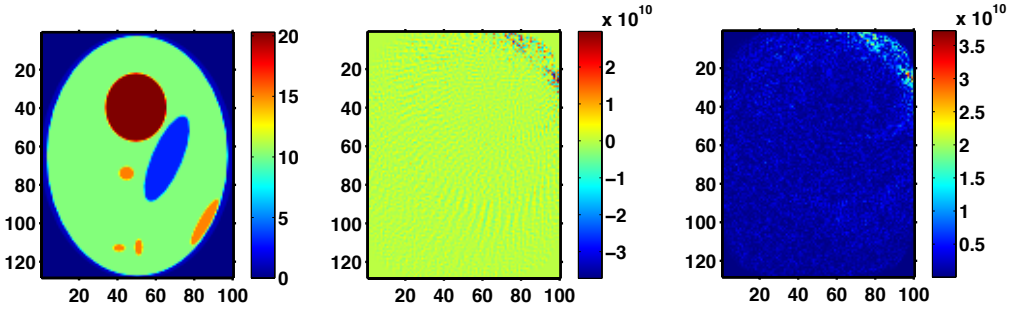
The moment estimator (4.1) of $\boldsymbol{\lambda}$ generally does not exist in practice, because of the bad conditioning of the problem. The least squares estimator (4.3) always exist, but might (and will very likely) lead to physically inadmissible estimates.

$$R^T\mathbf{n}^* = R^T R\hat{\boldsymbol{\lambda}},$$

which always admit a solution. It can be shown (see [Kaipo] Chapter I) that this is equivalent as minimizing the quantity $\|\mathbf{n}^* - R\boldsymbol{\lambda}\|^2$. Therefore an estimate of $\boldsymbol{\lambda}$ is given by the **least squares estimator**

$$\hat{\boldsymbol{\lambda}}_{LS} = \text{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}^B}\|\mathbf{n}^* - R\boldsymbol{\lambda}\|^2. \tag{4.2}$$

Even if this estimator always exist, there is no guarantee that the resulting density $\hat{\boldsymbol{\lambda}}_{LS}$ be positive, which is an absolute requirement to maintain physical interpretability. Moreover, the solution to (4.3) might not be unique. In this case, we can select among the solutions the one with minimal norm. Such a solution is provided by the **Moore-Penrose pseudoinverse** of R. In Figure 4.1 we present the reconstruction process of the phantom image described in Section 3 by means of the least squares estimator $\hat{\boldsymbol{\lambda}}_{LS}$ (to solve (4.3) we used the Matlab function `pinv`). We observe that the estimated density (see Figure 4.1b) is dramatically far from the true density (see Figure 4.1c). The reconstruction is so bad that we cannot even begin to guess the shapes of the ellipses constituting the phantom image. Moreover, the density $\hat{\boldsymbol{\lambda}}_{LS}$ is physically inadmissible, as it takes large



(a) True density $\boldsymbol{\lambda}$. **(b)** Estimated density $\hat{\boldsymbol{\lambda}}_{LS}$. **(c)** Error $|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_{LS}|$.

Figure 4.1: Reconstruction of the brain's metabolic activity by using the least squares estimator $\hat{\boldsymbol{\lambda}}_{LS}$.

negative values.

We need thus to be more careful in the optimization process, in order for the final estimate to be physically acceptable. For this, we propose to solve a constrained version of (4.3):

In order to avoid physically inadmissible estimates, we need to constrain positively the least squares optimization problem.

$$\hat{\boldsymbol{\lambda}}_{LS,C} = \text{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}^B, \boldsymbol{\lambda}\geq 0} \|\mathbf{n}^* - R\boldsymbol{\lambda}\|^2, \tag{4.3}$$

where we only look for positive $\boldsymbol{\lambda}$ minimizing $\|\mathbf{n}^* - R\boldsymbol{\lambda}\|^2$. We solve this constrained optimization problem with the function `fmincon` of Matlab, an iterative method based on the **interior point algorithm** (see [12]). On Figure 4.3a, we can observe the evolution of the estimate $\hat{\boldsymbol{\lambda}}^{(k)}_{LS,C}$ for selected iterations of the `fmincon` function. Surprisingly, it seems that the quality of the estimate does not improves monotically with the number of iterations. This is confirmed by Figure 4.2 where we plotted the **squared error (SE)** $\|\hat{\boldsymbol{\lambda}}^{(k)}_{LS,C} - \boldsymbol{\lambda}\|^2$ versus the number of iterations: the SE first decreases, reaches a minimum around iteration 14, and then increases. It seems that we face a typical bias/variance tradeoff: the estimate is less and less biased as the number of iterations increases, but suffers from an increasing variance, that creates an undesirable noise on the reconstructed density for high iterations. This phenomenon is even more apparent on Figure 4.3b, where the ellipses' boundaries progressively fade away as the bias decreases and an increasing noise spreads in the domain.

We will come back on this peculiar behavior when analyzing the EM algorithm, and try to provide some intuition on why we could have expected this result. However, we can already affirm in the light of theses observations that the stopping criterion will have a significant influence on the quality of the reconstruction.

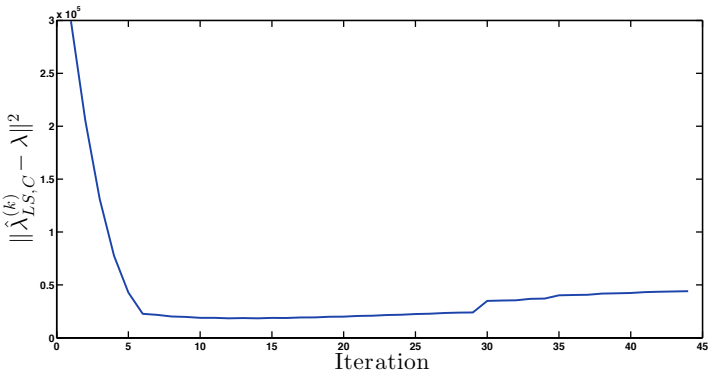
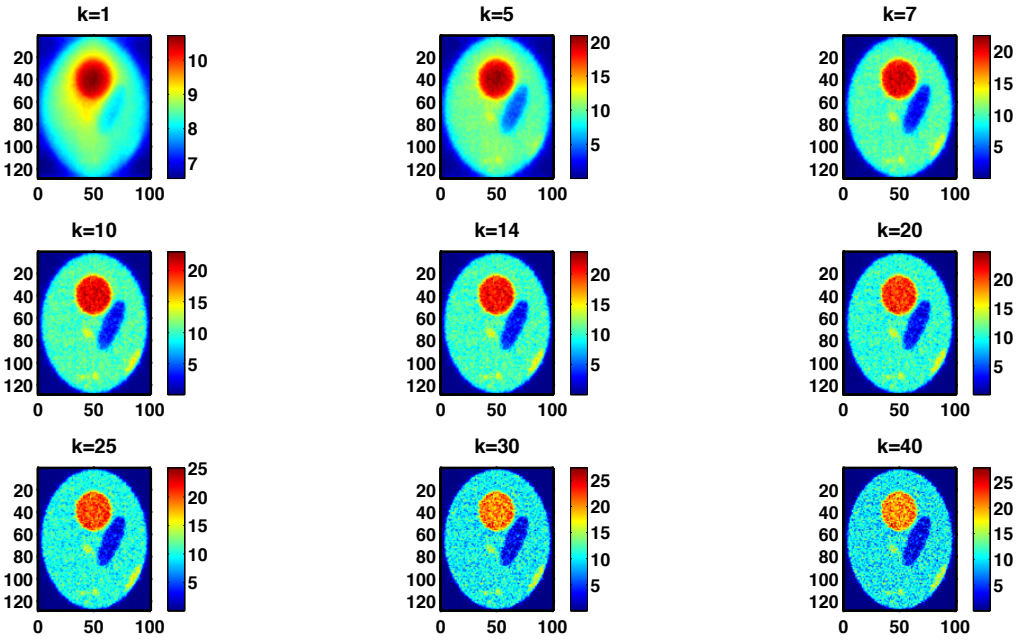


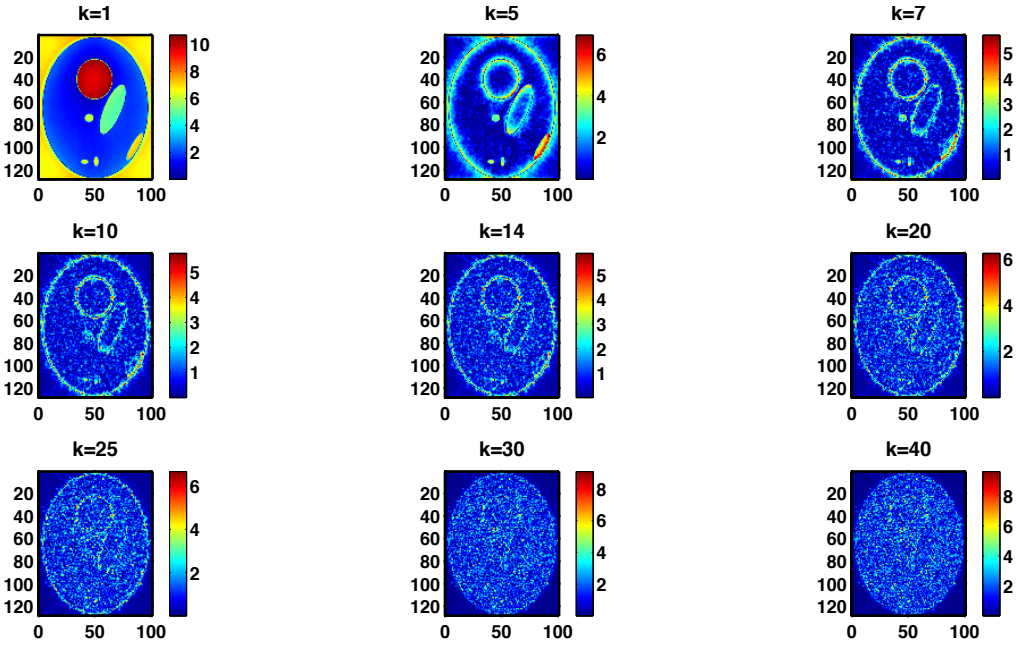
Figure 4.2: Squared error (SE) $\|\hat{\boldsymbol{\lambda}}^{(k)}_{LS,C} - \boldsymbol{\lambda}\|^2$ versus the number of iterations. We observe that the SE first decreases up to iteration 14, and then increases.

2 The Maximum Likelihood Estimator

We continue our investigation with the **maximum likelihood (ML)** estimator, originally proposed by Vardi et al. [7]. From Definition 1.4, we know that the random variables $N^*(1), \ldots, N^*(D)$ are independent Poisson variables with



(a) Estimate $\hat{\boldsymbol{\lambda}}_{LS,C}^{(k)}$ at iteration k.



(b) Error $|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_{LS,C}^{(k)}|$ at iteration k.

Figure 4.3: Iterations of the `fmincon` function used to compute the positively constrained least squares estimator $\hat{\boldsymbol{\lambda}}_{LS,C}$. We observe that the quality (in terms of squared error) of the estimate does not necessarily improve as the number of iterations grows: we have a bias/variance tradeoff.

respective means $\lambda^*(d),\ d = 1, \dots, D$. Therefore, the likelihood of the observed data $\mathbf{n}^*$ is given by

$$L(\boldsymbol{\lambda}) = \prod_{d=1}^{D} e^{-\lambda^*(d)} \frac{\lambda^*(d)^{n^*(d)}}{n^*(d)!},$$
$$\overset{(3.4)}{=} \prod_{d=1}^{D} e^{-\left(\sum_{b=1}^{B} R(d,b)\lambda(b)\right)} \frac{\left(\sum_{b=1}^{B} R(d,b)\lambda(b)\right)^{n^*(d)}}{n^*(d)!}, \tag{4.4}$$

with R the discrete Radon matrix, and B the number of pixels composing the brain's grid. By taking the log of (4.4) we obtain the so-called **log-likelihood function** $l(\boldsymbol{\lambda}) = \log L(\boldsymbol{\lambda})$. Then, the first and second derivatives of $l(\boldsymbol{\lambda})$ are given by:

$$\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b_0)} = -\sum_{d=1}^{D} R(d, b_0) + \sum_{d=1}^{D} \frac{n^*(d) R(d, b_0)}{\sum_{b=1}^{B} \lambda(b) R(d, b)},$$

and

$$\frac{\partial^2 l(\boldsymbol{\lambda})}{\partial \lambda(b_0) \partial \lambda(b_1)} = -\sum_{d=1}^{D} \frac{n^*(d) R(d, b_0) R(d, b_1)}{\left(\sum_{b=1}^{B} \lambda(b) R(d, b)\right)^2}.$$

We can show that the matrix of second derivatives in negative semidefinite (see [7]), and therefore $l(\boldsymbol{\lambda})$ is concave. Hence, all its maxima are global maxima.

The maximum likelihood estimate $\hat{\boldsymbol{\lambda}}_{ML}$ of $\boldsymbol{\lambda}$ is then by definition maximizing the likelihood function (or equivalently the log-likelihood function):

$$\hat{\boldsymbol{\lambda}}_{ML} = \text{argmax}_{\boldsymbol{\lambda} \in \mathbb{R}^B, \boldsymbol{\lambda} \geq 0}\ l(\boldsymbol{\lambda}). \tag{4.5}$$

Because of the concavity of the log-likelihood function, sufficient conditions for $\hat{\boldsymbol{\lambda}}_{ML}$ to be a solution of Equation (4.5) are given by the **Karush Kuhn Tucker conditions** (see [Zangwill Theorem 2.19])

Proposition 2.1 — *Karush Kuhn Tucker conditions.* *$\hat{\boldsymbol{\lambda}}_{ML}$ is a solution of (4.5) if the **Karush Kuhn Tucker conditions** hold:*

$$\lambda(b) \frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\Big|_{\hat{\boldsymbol{\lambda}}_{ML}} = 0, \quad \textit{and} \quad \frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\Big|_{\hat{\boldsymbol{\lambda}}_{ML}} \leq 0 \ \textit{ if } \ \hat{\lambda}(b) = 0, \quad b = 1, \dots, B. \tag{4.6}$$

Remark 2.1 *Figure 4.4 provides an illustration of the Karush Kuhn Tucker conditions. The first order condition $\lambda(b) \frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\Big|_{\hat{\boldsymbol{\lambda}}_{ML}} = 0$ tells us that either $\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\Big|_{\hat{\boldsymbol{\lambda}}_{ML}} = 0$ (the $\lambda(b)$ maximizing the likelihood is positive), or $\hat{\lambda}(b) = 0$ (the $\lambda(b)$ maximizing the likelihood*

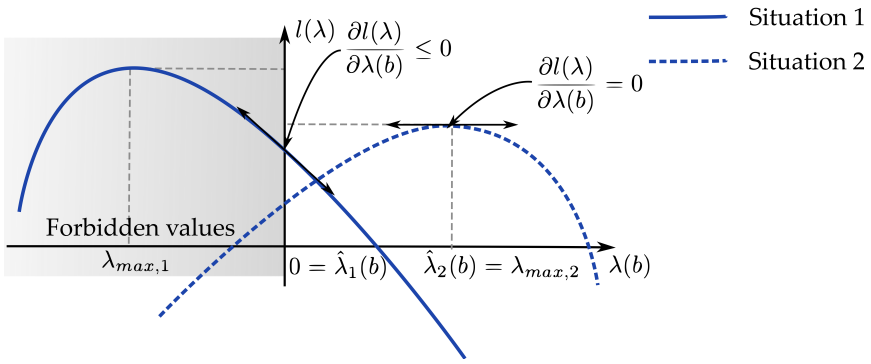


Figure 4.4: Schematic illustration of the Karush Kuhn Tucker conditions.

is negative so we choose 0, as negative values are unadmissible). The supplementary condition $\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda(b)}\Big|_{\hat{\boldsymbol{\lambda}}_{ML}} \leq 0$ *if* $\hat{\lambda}(b) = 0$ *is making sure that whenever* $\hat{\lambda}(b) = 0$ *the likelihood function does not improve in the positive direction, so that the maximum of the likelihood is indeed negative (remember that the likelihood function is concave).*

3 The EM Algorithm in the Context of PET

We now describe how the EM algorithm can be used to efficiently compute the MLE, solution of (4.5). To derive the EM algorithm in the context of PET we take inspiration from the developments in [7], that present a derivation tailored to this very specific problem. Here we consider a bit more general derivation, that follows more closely the recipe proposed in [11]. Our treatment is very similar to the one in [16] that applies the same methodology in the context of high energy physics.

In order to apply the EM algorithm to (4.5), we first need to re-express the problem in the framework of statistical estimation from *incomplete-data*. To this end we define the random quantities $N(b,d)$, $b = 1,\ldots,B$ and $d = 1,\ldots,D$ as the number of gamma rays originating from the pixel b and falling in the detector tube d. From the statistical model constructed in Chapter 3, we know that these random variables are independent Poisson variables with means $\lambda(b,d) = \lambda(b)R(d,b)$ (see (3.4)) :

$$N(b,d) \sim \text{Poisson}(\lambda(b)R(d,b)), \quad d = 1,\ldots,D,\ b = 1,\ldots,B.$$

Then, we choose $\mathbf{n}_c = \{n(b,d); b = 1,\ldots,B,\ d = 1,\ldots,D\}$ as the *complete data*, and $\mathbf{n}^* = \{n^*(d); d = 1,\ldots,D\}$ as the *incomplete data*. We have

$$n^*(d) = \sum_{b=1}^{B} n(b,d), \quad \forall d = 1,\ldots,D.$$

As $N(b,d)$ are independent Poisson variables, the complete-data log-likelihood is given by

$$l(\boldsymbol{\lambda}; \mathbf{n}_c) = \sum_{b,d} n(b,d) \log(\lambda(b)R(d,b)) - \log n(b,d)! - \lambda(b)R(d,b). \tag{4.7}$$

Now we need to apply the E-step of the EM algorithm and compute the conditional expectation of the complete-data log-likelihood (4.7) given the observations $\mathbf{n}^*$ and the current estimate $\boldsymbol{\lambda}^{(k)}$ of $\boldsymbol{\lambda}$. This yields:

$$\begin{aligned} Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k)}) &= \mathbb{E}[l(\boldsymbol{\lambda}; \mathbf{x}) | \mathbf{n}^*, \boldsymbol{\lambda}^{(k)}], \\ &\propto \sum_{b,d} \mathbb{E}[n(b,d) | \mathbf{n}^*, \boldsymbol{\lambda}^{(k)}] \log(\lambda(b)R(d,b)) - \lambda(b)R(d,b) \end{aligned} \tag{4.8}$$

We need to compute $\mathbb{E}[n(b,d)|\mathbf{n}^*, \boldsymbol{\lambda}^{(k)}]$:

$$\begin{aligned} \mathbb{E}[n(b,d)|\mathbf{n}^*, \boldsymbol{\lambda}^{(k)}] &= \mathbb{E}[n(b,d)|n^*(d), \boldsymbol{\lambda}^{(k)}], \\ &= \frac{n^*(d)\lambda^{(k)}(b)R(d,b)}{\sum_{b'=1}^{B} \lambda^{(k)}(b')R(d,b')}. \end{aligned} \tag{4.9}$$

In the first equality of (4.9), we used the mutual independence of the $n^*(d)$'s and in the second equality, we used the fact that if X_i are independent Poisson variables with means θ_i then the conditional distribution of X_j knowing that $\sum X_i = x^*$ is Binomial$(x^*, \theta_j / \sum \theta_i)$ and so $\mathbb{E}[X_j | \sum X_i = x^*] = x^* \theta_j / \sum \theta_i$.

Substituting (4.9) in (4.8) yields

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k)}) \propto \sum_{b,d} \frac{n^*(d)\lambda^{(k)}(b)R(d,b)}{\sum_{b'=1}^{B} \lambda^{(k)}(b')R(d,b')} \log(\lambda(b)R(d,b)) - \lambda(b)R(d,b). \tag{4.10}$$

Finally, we apply the M-step and maximize (4.10) with respect to $\boldsymbol{\lambda}$. Differentiating with respect to each component of $\boldsymbol{\lambda}$ and equating to zero leads to the following update equation:

$$\lambda^{(k+1)}(b) = \lambda^{(k)}(b) \sum_{d=1}^{D} \frac{n^*(d)R(d,b)}{\sum_{b'=1}^{B} \lambda^{(k)}(b')R(d,b')}, \quad b = 1, \ldots, B. \tag{4.11}$$

Therefore, in the special case of PET the EM algorithm is given by:

► The EM Algorithm in the Context of PET

Input: An initial estimate $\boldsymbol{\lambda}^{(0)} > 0$ of $\boldsymbol{\lambda}$.
Output: The maximum likelihood estimator $\hat{\boldsymbol{\lambda}}_{ML}$ of $\boldsymbol{\lambda}$.

1. k:=0;
2. **while** the stopping criterion is not met **do**;
   - $\lambda^{(k+1)}(b) := \lambda^{(k)}(b) \sum_{d=1}^{D} \frac{n^*(d)R(d,b)}{\sum_{b'=1}^{B} \lambda^{(k)}(b')R(d,b')}, \quad b = 1, \ldots, B;$
   - $k := k + 1;$
   - Compute stopping criterion;

   endwhile;
3. $\hat{\boldsymbol{\lambda}}_{ML} := \boldsymbol{\lambda}^{(k)};$

We note that the estimate $\hat{\boldsymbol{\lambda}}_{ML}$ provided by the EM algorithm is positive as long as the initial guess $\boldsymbol{\lambda}^{(0)}$ is. In fact, if $\boldsymbol{\lambda}^{(0)} > 0$, then all the quantities involved in the update equation (4.11) are positive and we can show by induction that all the iterates will also be positive. Moreover, in the special case of PET, Shepp and Vardi asserted in [13] that the sequence of estimates produced by the EM algorithm converges to a maximum, as $k \to \infty$. The proof they provided in [13] had a gap but the assertion was correct, and was proven by Csiszàr and Tusnàdy in [14].

Theorem 3.1 The sequence of estimates $\{\boldsymbol{\lambda}^{(k)}\}_{k=0,1,2,\ldots}$ provided by the EM algorithm converges to a limit point $\hat{\boldsymbol{\lambda}}_{ML}$ that maximizes $l(\boldsymbol{\lambda})$.

■ **Proof 3.1** See Appendix of [7]. ■

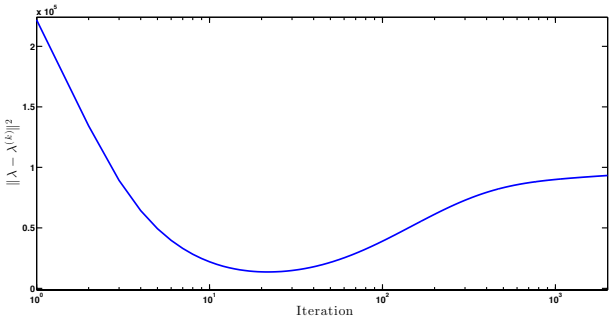
We now run the EM algorithm on our testbed, and analyze the evolution of the estimate with the number of iterations. We observe on Figure 4.5 a behavior very similar to the one of the positively constrained least squares estimator. Indeed the estimate $\boldsymbol{\lambda}^{(k)}$ does not necessarily improve with the number of iterations:

it first decreases, reaches a minimum aroun $k = 15$ and then starts to increase from this point. Finally, it seems that we once again face a bias/variance tradeoff: the estimate is less and less biased as the number of iterations increases, but suffers from an increasing variance, that creates an undesirable noise on the reconstructed density for large iterations.

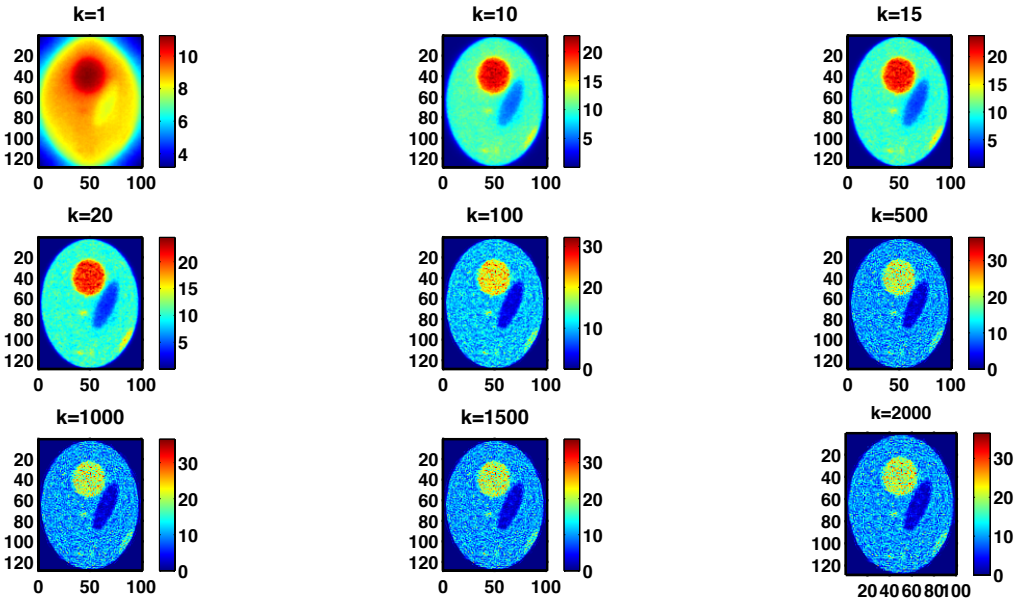
4 A Goodness-of-fit Stopping Criterion

We now investigate a goodness-of-fit stopping criterion, similarly at what has been proposed by Veklerov et al. in [8].

For both the least squares and maximum likelihood estimators, we encountered a notable difficulty while trying to compute them with iterative algorithms. The quality of the estimates is not always improving with the number of iterations: after some point, the estimates become considerably noisy. This is a quite unusual behavior that stresses the need for a good stopping criterion: running the algorithm for a very long time will not necessarily produce



(a) Squared error (SE) $\|\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}\|^2$ versus the number of iterations.



(b) Estimate $\boldsymbol{\lambda}^{(k)}$ at iteration k.

Figure 4.5: Iterations of the EM algorithm used to compute the maximum likelihood estimator $\hat{\boldsymbol{\lambda}}_{ML}$. We observe that the quality (in terms of squared error) of the estimate does not always improve as the number of iterations grows: we have a bias/variance tradeoff.

a good estimate, we need to know when to stop to avoid the deterioration previously observed. In order to design such a stopping criterion, we first need to better appreciate the origin of this odd phenomenon, and why it could have been expected. The answer to this inquiry lies in the nature of the problem itself: because of the extremely bad conditioning of the Radon matrix, the slight perturbation introduced by the Poisson noise in the detector space gets amplified when mapped back in the brain space during the inversion process, resulting in estimates $\hat{\boldsymbol{\lambda}}_{MLE}$ and $\hat{\boldsymbol{\lambda}}_{LS}$ far away from the true intensity. This is a very common situation in inverse problems (see Figure 4.7). In the light of this observation, we understand that we actually do not want the algorithm to converge, because the resulting intensity would be far away from the true intensity. However, as bad as the MLE or LS estimates might be, we can still extract some

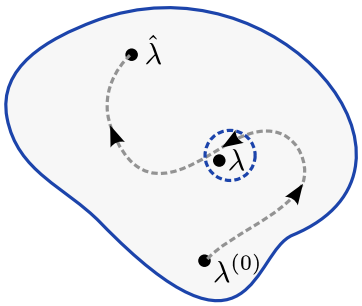


Figure 4.6: We use an estimator $\hat{\boldsymbol{\lambda}}$ as a target for our iterative algorithm and stop prematurely the procedure when the estimate pass close enough to the real intensity $\boldsymbol{\lambda}$.

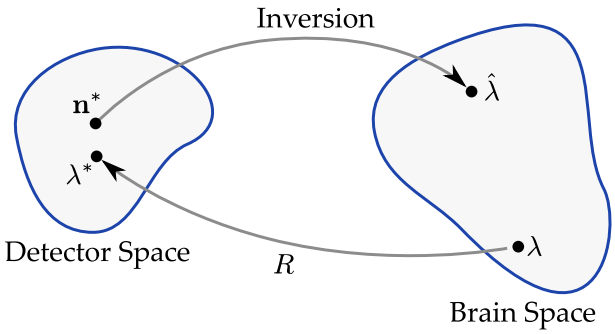


Figure 4.7: Because of the extremely bad conditioning of the Radon matrix, a slight perturbation introduced by some Poisson noise in the detector space gets amplified when mapped back in the brain space during the inversion process, resulting in an estimate $\hat{\boldsymbol{\lambda}}$ far away from the true intensity $\boldsymbol{\lambda}$.

information out of them, and use them wisely in order to get to a good estimate of the true intensity. The idea is the following: we use these estimators as targets for our iterative algorithms and stop prematurely the procedure when the estimate is "good enough". Roughly speaking, the hope is that the iterative algorithms will pass close enough to the true density on their way to the MLE or LS estimators, so that we can "recognize" it (in a sense that will be made precise later) and stop near to it (see Figure 4.6). Mathematically speaking, we perform a statistical test at each iteration and test wether or not the current estimate $\boldsymbol{\lambda}^{(k)}$ could have produced the observed data $\mathbf{n}^* = (n^*(1), \dots, n^*(D))^T$. More precisely we test:

$$H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)} \quad \text{vs.} \quad H_1 : \boldsymbol{\lambda} \neq \boldsymbol{\lambda}^{(k)},$$

or in the detector space:

$$H_0 : \boldsymbol{\lambda}^* = \boldsymbol{\lambda}^*_{(k)} \quad \text{vs.} \quad H_1 : \boldsymbol{\lambda}^* \neq \boldsymbol{\lambda}^*_{(k)}.$$

For this, we introduce the following test statistic:

$$C = \sum_{d=1}^{D} \frac{\left(n^*(d) - \lambda^*_{(k)}(d)\right)^2}{\lambda^*_{(k)}(d)}, \tag{4.12}$$

with $\boldsymbol{\lambda}^*_{(k)} := R\boldsymbol{\lambda}^{(k)}$ and $k \in \mathbb{N}$ the index of the iteration. Note that the test is performed in the detector space (where the observed data lie) and not directly in the brain space. Roughly speaking, we indirectly test the intensities in the brain space by comparing two sinograms (i.e. intensities in the detector space). Under the null hypothesis and our modeling assumptions, $n^*(d) \overset{d}{\sim} \text{Poisson}(\lambda^*_{(k)}(d)),\ \forall d = 1, \dots, D$. Therefore, from properties of the Poisson distribution, we have $\mathbb{E}[n^*(d)] = \text{Var}(n^*(d)) = \lambda^*_{(k)}(d)$. Moreover, as the number of detected coincidences in each detector tubes is usually very high[2] in PET experiments, the counts rate $\lambda^*_{(k)}(d),\ d = 1, \dots, D$ will also be very high and we can thus approximate the Poisson distribution by a Gaussian distribution:

[2] in our setup we have on average 400 coincidences per detector tube.

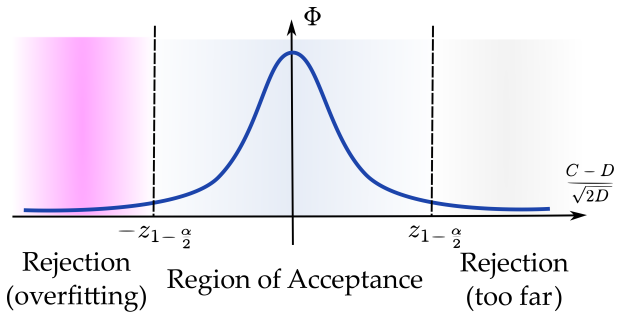


Figure 4.8: Rejection and acceptance regions for the test.

$$\frac{n^*(d) - \lambda^*_{(k)}(d)}{\sqrt{\lambda^*_{(k)}(d)}} \overset{d}{\sim} \mathcal{N}(0,1).$$

Then, it naturally follows that

$$\frac{\left(n^*(d) - \lambda^*_{(k)}(d)\right)^2}{\lambda^*_{(k)}(d)} \overset{d}{\sim} \chi^2(1),$$

and finally

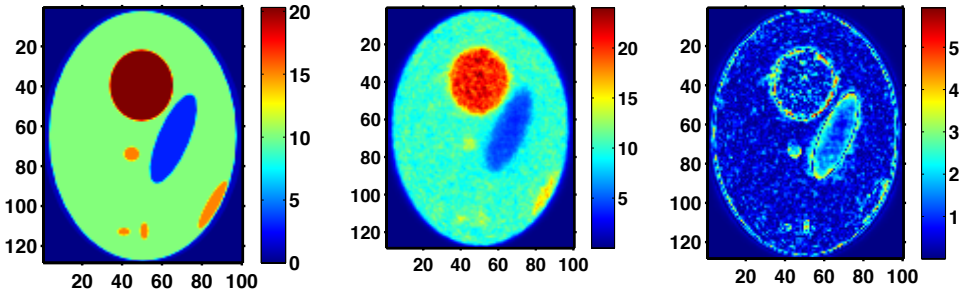
$$C = \sum_{d=1}^{D} \frac{\left(n^*(d) - \lambda^*_{(k)}(d)\right)^2}{\lambda^*_{(k)}(d)} \overset{d}{\sim} \chi^2(D).$$

We recognize here an instance of the classic **Pearson's chi-squared test**, commonly used as a measure of *goodness of fit*. A similar goodness-of-fit criterion for stopping the EM iteration is proposed in [8]. Finally, because the number of detector tubes is very big, we can approximate the $\chi^2(D)$ distribution by a normal distribution and we obtain

$$\frac{C - D}{\sqrt{2D}} \overset{d}{\sim} \mathcal{N}(0,1).$$

If we now choose a significance level $\alpha = 0.05$ we can define regions of acceptance and rejection for the test (see Figure 4.8). The right tail of the distribution $[z_{1-\alpha/2}, +\infty[$ (with $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2) = 1.96$) in Figure 4.8 corresponds to intensities too far away from the observed data to have possibly generated it. The acceptance region $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$ corresponds to intensities that could have likely enough produced the data. Finally, the left tail of the distribution $]-\infty, z_{1-\alpha/2}]$ corresponds to the region of *overfitting*: the intensities that fall in this region are so close from the observed data that it becomes very unlikely that they actually produced the data (indeed it is really unlikely than thousands of Poisson random variables all generate an output very close to their mean).

Then, we will stop the iterative algorithm as soon as one iterate falls into the acceptance region of the test previously designed. In Figure 4.9 we can see the estimated intensity based on the only iterate $\boldsymbol{\lambda}^{(18)}$ of the EM algorithm



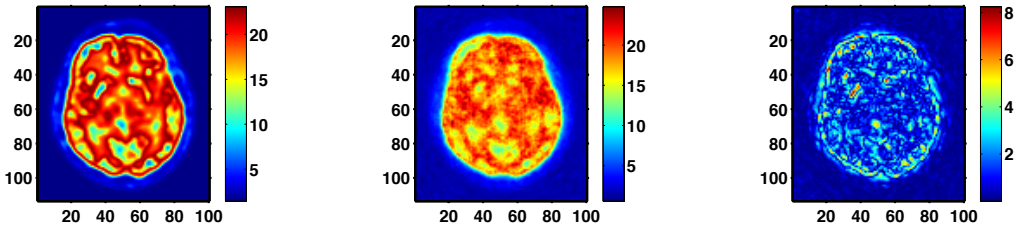
(a) True density $\boldsymbol{\lambda}$. **(b)** Estimated density $\hat{\boldsymbol{\lambda}}_{MLE}$. **(c)** Error $|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_{MLE}|$.

Figure 4.9: Reconstruction of the brain's metabolic activity by using the EM-algorithm together with a goodness-of-fit stopping criterion .

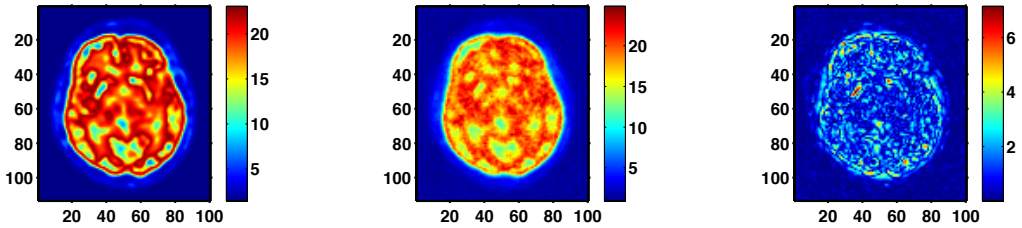
falling in the acceptance region. For our setup, we had been lucky enough so that only one iterate falls into the region of acceptance. But it could occur that some specific setups lead to many iterates falling into the region of acceptance or on the contrary not any. Then, we could still select an estimate in these situations by choosing the iterate falling the closest to 0, the middle of the segment $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$.

5 Test of the Procedure on More Realistic Data

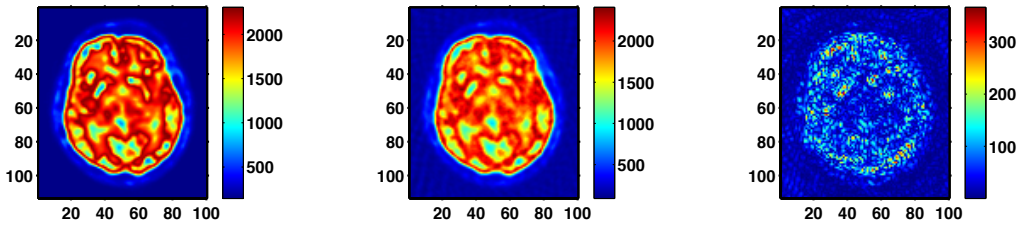
We conclude this study by analyzing the behavior of the reconstruction procedure previously described on a more complex and realistic intensity. More specifically, we use as a phantom image an actual PET scan output, provided by [18]. Then we test the reconstruction process with different setups, and observe the evolution of the reconstructed intensity, estimated with the EM algorithm together with the goodness-of-fit stopping criterion. In the different setups, we vary the number of detector cells c and the expected total number of observed coincidences n. The results are available on Figure 4.10. We can see that the estimated intensity improves much more with the number of observations n than with the number of detector cells c. Moreover, we also notice that the "convergence" of the EM algorithm happens much later as the number of observations increases. In comparison, the convergence of the algorithm seems less affected by the number of detector cells. Finally, the reconstructed intensity obtained on Figure 4.10d is very similar to the real intensity, which is very satisfying.



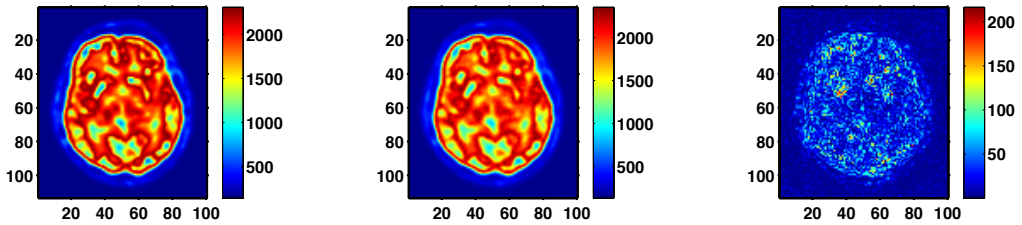
(a) Setup: $n = 10^5$ and $c = 30$. The EM algorithm stopped at iteration $k = 18$.



(b) Setup: $n = 10^5$ and $c = 60$. The EM algorithm stopped at iteration $k = 24$.



(c) Setup: $n = 10^7$ and $c = 30$. The EM algorithm stopped at iteration $k = 84$.



(d) Setup: $n = 10^7$ and $c = 60$. The EM algorithm stopped at iteration $k = 110$.

Figure 4.10: Reconstruction of the brain's metabolic activity in different experimental setups. For the reconstruction, we used the EM algorithm together with a goodness-of-fit stopping criterion. On each figure we have, from left to right: phantom image, estimated intensity $\hat{\boldsymbol{\lambda}}_{MLE}^{(k)}$, and error $|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_{MLE}^{(k)}|$.

5

Conclusion

In conclusion, our knowledge from Poisson point processes and the Radon transform from Chapter 2 allowed us to construct a statistical model tailored to the physics of PET. Based on this model, we investigated different estimators of the brain's metabolic activity, which both experienced an unusual behavior when ran long enough: the estimated intensity started to deteriorate after some point. While such a behavior is very peculiar for an algorithm, we have been able to provide some intuition on the nature of this odd phenomenon by looking at the reconstruction problem as an inverse problem. This discussion allowed us to design a goodness-of-fit stopping criterion, aimed at stopping prematurely the iterative algorithm before the deterioration of the estimate occurs. However, this methodology, while coherent with the interpretation we developed, can appear a bit ad-hoc: indeed, the whole success of the estimation process lies in the hope that the iterative algorithm will pass close enough to the true intensity while traveling to the MLE or LS estimators. But there is no guarantee that this will indeed happen and therefore it could be wise to try and construct some other estimators, such as penalized likelihood or Bayesian estimators.

Acknowledgments

I wish to acknowledge the Prof. Victor Panaretos for having proposed me this project, and Mr. Mikael Kuusela for having followed me all along its realization. His advices and support have always been very helpful, and I learnt a lot under his supervision. Moreover he has always been very dedicated to this task and very available, which I sincerely appreciated.

Bibliography

Books

[1] R.-D. Reiss, *A Course on Point Processes*, Springer Series in Statistics, Second edition, 1993.

[2] S. Deans, *The Radon Transform and Some of Its Applications*, John Wiley & Sons, 1983.

[3] F. Natterer, *The Mathematics of Computerized Tomography*, Society for Industrial and Applied Mathematics, 2001.

[4] G. J. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, Second Edition, 2008.

[5] H, W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, 1996.

[6] J. Kaipio, E. Somersalo, *Statistical and Computational Inverse Problems*, Applied Mathematical Sciences, Springer, 2005.

Articles

[7] Y. Vardi, L. A. Shepp, L. Kaufman, *A Statistical Model for Positron Emission Tomography*, Journal of the American Statistical Association, American Statistical Association, Vol. 80, No. 389, 2005.

[8] E. Veklerov, J. Llacer, *Stopping Rule for the MLE Algorithm Based on Statistical Hypothesis Testing*, IEEE Transactions on Medical Imaging, Vol. MI-6, No. 4, 1987.

[9] I. M. Johnstone, B.-W. Silverman, *Speed of Estimation in Positron Emission Tomography and Realted Inverse Problems*, The Annals of Statistics, Vol. 18, No. 1, 1990.

[10] G. L. Brownell, T. F. Budinger, P. C. Lauterbur, P. L. Johnstone, B. W. Silverman, P. L. McGeer, *Positron Tomography and Nuclear Magnetic Resonance Imaging*, Science, No. 215, 1982.

[11] A. P. Dempster, N. M. Laird, D. B. Rubin, *Maximum Likelihood From Incomplete Data via the EM Algortihm*, Journal of the Royal Statistical Society, No. 39, 1977.

[12] R. H. Byrd, J. C. Gilbert, J. Nocedal, *A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming*, Mathematical Programming, Vol 89, No. 1, 2000.

[13] Y. Vardi, L. A. Shepp, *Maximum Likelihood Reconstruction in Positron Emission Tomography*, IEEE Transactions on Medical Imaging, Vol. 1, 1982.

[14] I. Csiszàr, G. Tusnàdy, *Information Geometry and Alternating Minimization Procedures*, Technical Report, Mathematical Institute of the Hungarian Academy of Science, 1982.

[15] C. M. L. West, T. Jones, P. Price, *The potential of positron-emission tomography to study anticancer-drug resistance*, Nature Reviews Cancer, Vol. 4, 2004.

Other

[16] M. Kuusela, *Statistical Issues in Unfolding Methods for High Energy Physics*, Master's Thesis, Aalto University, 2012.

[17] United States Geological Survey (USGS), *Earthquake Hazards Program*, http://earthquake.usgs.gov/earthquakes/feed/v1.0/, Consulted in June 2014.

[18] J. Langner, *PET scan of the human brain*, Wikimedia Commons, http://en.wikipedia.org/wiki/Positron_emission_tomography#mediaviewer/File:PET-image.jpg, Uploaded on 2010, Consulted on 2014.

[19] J. Langner, *Whole-body PET scan using 18F-FDG*, Wikimedia Commons, http://en.wikipedia.org/wiki/Positron_emission_tomography#mediaviewer/File:PET-MIPS-anim.gif, Uploaded in 2006, Consulted in 2014.

[20] A. Praefcke, Wikimedia Commons, http://commons.wikimedia.org/wiki/File:KH_St_Elisabeth_RV_2013_Radiologie_PET-CT.jpg?uselang=fr, Uploaded in 2013, Consulted in 2014.