



Data harmonization in PET imaging

PhD thesis

PhD Candidate

Nicola Alchera

Advisors

Giovanni Mettivier
Università degli studi di Napoli
Piernicola Oliva
Università degli studi di Sassari

Tutor

Andrea Chincarini
Istituto Nazionale di Fisica Nucleare (INFN)
Università degli Studi di Genova

Università degli Studi di Genova

Ciclo XXXIII – A. A. 2017-2020

Preface

Medical imaging physics has advanced a lot in recent years, providing clinicians and researchers with increasingly detailed images that are well suited to be analyzed with a quantitative approach typical of hard sciences, based on measurements and analysis of clinical interest quantities extracted from images themselves. Such an approach is placed in the context of quantitative imaging.

The possibility of sharing data quickly, the development of machine learning and data mining techniques, the increasing availability of computational power and digital data storage which characterize this age constitute a great opportunity for quantitative imaging studies.

The interest in large multicentric databases that gather images from single research centers is growing year after year. Big datasets offer very interesting research perspectives, primarily because they allow to increase statistical power of studies. At the same time, they raised a compatibility issue between data themselves. Indeed images acquired with different scanners and protocols could be very different about quality and measures extracted from images with different quality might be not compatible with each other.

Harmonization techniques have been developed to circumvent this problem. Harmonization refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies. Harmonization can be done before acquiring data, by choosing a-priori appropriate acquisition protocols through a preliminary joint effort between research centers, or it can be done a-posteriori i.e. images are grouped into a single dataset and then any effects on measures caused by technical acquisition factors are removed.

Although the a-priori harmonization guarantees best results, it is not often used for practical and/or technical reasons. In this thesis I will focus on a-posteriori harmonization.

It is important to note that when we consider multicentric studies, in addition to the technical variability related to scanners and acquisition protocols, there may be a demographic variability that makes single centers samples not statistically equivalent to each other. The wide individual variability that characterize human beings, even more pronounced when patients are enrolled from very different geographical areas, can certainly exacerbate this issue. In addition, we must consider that biological processes are complex phenomena: quantitative imaging measures can be affected by numerous confounding demographic variables even apparently unrelated to measures themselves.

A good harmonization method should be able to preserve inter-individual variability

and remove at the same time all the effects due acquisition technical factors. Heterogeneity in acquisition together with a great inter-individual variability make harmonization very hard to achieve.

Harmonization methods currently used in literature are able to preserve only the inter-subjects variability described by a set of known confounding variables, while all the unknown confounding variables are wrongly removed. This might lead to incorrect harmonization, especially if the unknown confounders play an important role. This issue is emphasized in practice, as sometimes happens that demographic variables that are known to play a major role are unknown.

The final goal of my thesis is a proposal for an harmonization method developed in the context of amyloid Positron Emission Tomography (PET) which aim to remove the effects of variability induced by technical factors and at the same time are able to keep all the inter-individual differences. Since knowing all the demographic confounders is almost impossible, both practically and a theoretically, my proposal does not require the knowledge of these variables.

The main point is to characterize image quality through a set of quality measures evaluated in regions of interest (ROIs) which are required to be as independent as possible from anatomical and clinical variability in order to exclusively highlight the effect of technical factors on images texture. Ideally, this allows to decouple the between-subjects variability from the technical ones: the latter can be directly removed while the former is automatically preserved.

Specifically, I defined and validated 3 quality measures based on images texture properties. In addition I used a quality metric already existing, and I considered the reconstruction matrix dimension to take into account image resolution.

My work has been performed using a multicentric dataset consisting of 1001 amyloid PET images. Before dealing specifically with harmonization, I handled some important issues: I built a relational database to organize and manage data and then I developed an automated algorithm for images pre-processing to achieve registration and quantification.

This work might also be used in other imaging contexts: in particular I believe it could be applied in fluorodeoxyglucose (FDG) PET and tau PET. The consequences of harmonization I developed have been explored at a preliminary level. My proposal should be considered as a starting point as I mainly dealt with the issues of quality measures, while the harmonization of the variables in itself was done with a linear regression model. Although harmonization through linear models is often used, more sophisticated techniques are present in literature. It would be interesting to combine them with my work. Further investigations would be desirable in future.

Contents

1	Research context	7
1.1	Prospective and retrospective harmonization	8
1.2	Big data and harmonization issues in medical imaging physics	9
2	Neuroimaging	12
2.1	Neuroimaging overview	12
2.2	Quantitative imaging	13
2.2.1	Biomarker	13
2.2.2	Radiotracers	14
2.3	PET physical principle	15
2.4	Data acquisition	16
2.4.1	Detection	16
2.4.2	PET detectors	18
2.4.3	Factors affecting acquired data	19
2.5	Image Reconstruction	25
2.5.1	The image reconstruction problem	25
2.5.2	Analytic reconstruction methods	26
2.5.3	Iterative reconstruction methods	30
2.5.4	3D reconstruction methods: an overview	35
2.6	Resolution modeling	36
2.7	DICOM files and NIfTI images	38
2.8	Performance characteristics of PET scanners and image quality measures .	39
2.9	Final considerations	40
3	Methods	42
3.1	Machine Learning overview	42
3.1.1	Supervised Learning	43
3.1.2	Supervised Learning and small size datasets	48
3.1.3	Unsupervised Learning	52
3.2	Resampling techniques	53
3.3	Decision tree based ML algorithms	55
3.3.1	Decision Trees	55
3.3.2	Classification and regresion trees (CART)	56
3.3.3	Bootstrap aggregating	61
3.3.4	Random Forest	62
3.4	Principal Component Analysis	63
3.5	Statistical evaluators	66
3.5.1	Statistical test of parametric hypothesis	66

3.5.2	Akaike Information Criterion and Bayesian Information Criterion	69
3.5.3	Receiver Operating Characteristics	70
3.6	Image Registration	72
3.6.1	ANTs software	76
3.7	Image semi-quantification	78
3.7.1	Standardized Uptake Value ratio (SUVR)	79
3.7.2	Evaluation of Brain Amyloidosis (ELBA)	80
3.8	Computing	83
3.9	Database	85
3.9.1	Database building and managing	85
4	Materials	88
4.1	AmyDB Database	88
4.2	Image preprocessing	94
4.2.1	Image preprocessing overview	94
4.2.2	Registration pipeline	97
4.2.3	Quantification pipeline	99
5	Data Harmonization in PET imaging	101
5.1	Multicentric and monocentric studies	102
5.2	Prospective and retrospective harmonization in PET imaging	104
5.3	Fixing the notation	106
5.4	PET retrospective harmonization methods discussion	108
6	Image quality estimation and data harmonization: proposal for a novel approach	111
6.1	Batch effect estimation using quality measures directly extracted from the PET images	111
6.2	ROIs selection	113
6.3	Quality measures	113
6.3.1	Watershed	113
6.3.2	Delta Contrast	115
6.3.3	Acutance	117
6.3.4	Natural Image Quality Evaluator (NIQE)	119
6.3.5	Matrix Dimension	120
6.4	Quality measures extraction and post-processing	120
6.5	Quality Measures Validation	121
6.5.1	Independence between quality measures and clinical profiles	121
6.5.2	Visual validation	122
6.5.3	Testing the ability of data provenance reconstruction	122
6.6	Harmonization of quantification values	125
6.7	Estimation of AIC and BIC for different models	126
6.8	Discussion	128
	Bibliography	132

Chapter 1

Research context

We live in an era in which knowledge is increasingly based on the analysis of large amounts of data.

More and more advanced technologies and the possibility of sharing digital information thanks to the World Wide Web, has allowed the creation of huge databases. In parallel, the development of machine learning and data mining techniques, combined with the development of computing technology, give us the opportunity to analyze and investigate data as never before has been possible.

When databases reach very large volumes, when there is a continuous and rapid flow of new data, when data comes from various sources and are acquired with different criteria and sometimes in different formats, then it is legitimate to speak of big data. There is no precise definition of what the minimum volume is to be considered big data; according to Magoulas et al [100] we are dealing with big data when the size and performance requirements for data management become significant design and decision factors for implementing a data management and analysis system. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

Large databases along with machine learning and data mining methods represent a great and relatively new opportunity for advancement of science in many different fields. The former increase statistical power of studies, while the latter allow to investigate data space to obtain significant results and complex predictions even when there is no theoretical model that drives studies.

However, studying big databases has not only benefits; one of the main drawbacks is the potential heterogeneity of data from different sources.

Large databases are usually built through a joint data sharing effort of several research centers. If data are collected without using a well-defined a priori shared protocol, then important steps such as the choice of sample, the data acquisition process as well as data manipulation may depend, even strongly, on the research center. The potential incompatibility between data with different provenance is considered as the main limitation of multicentric studies. Data provenance refers to information about the origin, the process and the methodologies by which data were produced. [66].

Multiple effects could be involved in data provenance and take them all into account is often a very hard challenge as each operation performed on data will modify data themselves in a specific way. It is important to emphasize that the raw data, before being analyzed, are typically subjected to more or less complex workflows which allow to map them in a suitable format for analysis. In conclusion, data provenance leaves a mark on data: the greater is the differences in provenance, the greater is the data heterogeneity and hence the more difficult is to analyze data gathered in single multicentric studies.

Analyzing heterogeneous data without taking into account the provenance issue could lead to incorrect and misleading results.

In this framework, the so-called data harmonization plays a crucial role: data harmonization refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies [70]. Harmonization includes all practices which enable the pooling of data from multiple cohorts at a level of precision that is scientifically adequate. The key challenge of harmonization is remove, or at least mitigate, the data provenance heterogeneity, making possible to adequately combine data from different studies in order to increase the sample size, and hence statistical power and generalizability of the results.

1.1 Prospective and retrospective harmonization

Harmonization can be divided into two categories: prospective (also called a-priori) and retrospective harmonization (also called a-posteriori) [61]. In the former researchers would agree in advance on a series of practices to collect data in such a way as to directly enable pooled analysis, while the latter refers to all the techniques used to make compatible data after they have been collected.

Prospective harmonization can in turn be divided into two categories: *stringent prospective harmonization* (also called standardization) and *flexible prospective harmonization* [76].

Standardization refers to the implementation of uniform processes for prospective collection, storage and transformation of data. Standardization implies that precisely the same methods, protocols and standard operating procedures are used in every study or study phase contributing to the analyses [76].

In particular harmonization is considered as stringent if data are collected across different studies using identical data collection tools (e.g. identical measuring instruments) and standard operating procedures [76].

Using standardized methods across multiple studies greatly facilitates analyses of datasets from separate research centers. However, imposing identical procedures is a very hard challenging in practice: differences in measuring instruments, in staff training, in participant characteristics, in financial resources as well as differences in legislation on ethical issues (if studies involve different countries) could make standardization difficult to achieve [59].

A balance can be struck between the use of precisely uniform measures and procedures that render data synthesis straightforward and the acceptance of some flexibility that may be appropriate and more realistic in a collaborative context: such an approach is called flexible prospective harmonization [60].

It is important to note that the goal of harmonization is to obtain data from different research centers that can be integrated into a single study. The use of a precisely identical data acquisition and data processing protocols (i.e. standardization) is a sufficient condition, but not necessary to achieve this aim: flexible harmonization permits the utilization of different data collection tools and procedures, however it is required that data acquired and pre-processed in individual research centers are sufficiently compatible to each other in order to allow a valid integration of the data themselves. This can be achieved by designating studies a priori with a concerted effort: data acquisition, data preprocessing tools and protocols are required to be defined ab-initio in order to obtain comparable data.

However, a-priori harmonization, whether flexible or stringent, is in practice difficult to achieve because it requires a joint effort that is rarely made. A further limitation of prospective harmonization is the impossibility of using existing data that have not been acquired according to criteria chosen a-priori, or whose provenance is unknown (or only partially known).

Retrospective harmonization targets synthesis of information already collected by existing studies. The ability to retrospectively harmonize data from existing studies facilitates the rapid generation of new scientific knowledge: harmonization can make use of existing data, hence the construction of a synthesized dataset can be achieved relatively rapidly. [60].

While the a-priori harmonization is based on a well-defined scheme, depending in detail on the topic considered, but basically based on a common agreement between research centers during the design phase of the study, a-posteriori harmonization cannot be defined by a common general scheme. In other words, addressing the problem of a posteriori harmonization requires the definition of a specific problem.

1.2 Big data and harmonization issues in medical imaging physics

Interestingly, the introductory discussion on harmonization was very general. Indeed, analysis of large datasets, problem of data provenance, and strategies related to harmonization constitute cross-cutting topics that involve many fields of science, such as medicine, neuroscience, physics, economics, chemistry, biology, sociology.

In physics, for example, instrument calibration could be viewed as an a-priori harmonization. Typically, instruments are calibrated in the absence of signal in order to characterize the noise. This simple approach can sometimes be very complex to achieve. For example, a straightforward calibration of an instrument that detects gravitational waves would imply characterizing the instrument simulating the absence of gravity: this is very hard to achieve since it is not possible to shield a gravitational field.

Big data are particularly present in particle physics framework: the CERN LHCs

(Large Hadron Collider) sensors record hundreds of millions of collisions between particles. Clearly, this generates a huge amount of data, the LHC alone generates around 90 petabytes of information a year ¹. However, since these data come from a single source (the LHC), the harmonization problem is essentially not present.

A field where harmonization is particularly important is medical imaging physics: it is well known that when medical images acquired with different instruments and different acquisition protocols are pooled in a single study, measures extracted from images themselves usually affected by these technical factors. In the context of medical imaging, data provenance consists of all the technical factors involved in the reconstruction and acquisition of the data: the most important are the properties of the scanner and the image acquisition and reconstruction protocols. These technical factors can have a great impact on image quality, as we will discuss in chapter 2, and clinical measures extracted from images could be sensitive to image quality. Data harmonization is often an essential step to use medical imaging multicentric datasets [113, 156, 90, 2].

Research in medical physics is in some respects very different from research in fundamental physics. In medical physics there are ethical and legal aspects that constrains studies and data collection: patients cannot be subjected to invasive examinations that are not necessary, and this typically creates a bias in data acquisition for which there is no availability of young and healthy subjects images. Furthermore, data often come from naturalistic populations, and not from ad-hoc designed studies. A naturalistic database is a collection of clinical images of patients who have undergone examinations.

Studies conducted on naturalistic databases can be affected by a sampling bias: a naturalistic population is unlikely to constitute a representative sample of the whole population. A further important issue is linked to the equivalence of the statistical units considered, i.e patients. In fundamental physics the experiments are conducted on samples in which the variability is controlled and limited to the purposes of the experiment itself. This does not happen in medical studies: the inter-patient variability can be very wide and in practice is impossible to characterize it completely, as there are too many variables potentially involved (for example, lifestyle, income, quality of night rest, unknown pathologies, genetic differences etc...).

Therefore, in investigating a relation between a given input and a given output of interest, a large number of unknown confounding variables may be involved, and this can lead to misleading results [93, 158]. This issue is enhanced by the complexity of the biological mechanisms: this complexity means that there may be many confounding variables at play, even apparently not related with measures of interest.

Furthermore, the samples size in studies involving medical imaging studies is enormously lower than that typical samples size which characterize fundamental physic experiments: many monocentric studies in neuroimaging use samples made of about 50 subjects [142]. For this reason there is an increasing interest in multicentric studies. i.e. studies that involve data (i.e. medical images) from many clinical and research centers [41, 103, 2, 90]. The price to pay for having a more powerful statistic is that of having heterogeneous data. This aspect together with the problem of sampling bias constitutes a main challenge for medical imaging studies. Heterogeneity in acquisition together with

¹<https://home.cern/science/computing/storage>

a great inter-individual variability make postreconstruction harmonization very hard to achieve, as a good harmonization method should be able to preserve inter-individual variability and remove at the same time all the effects due acquisition technical factors.

To conclude, I believe that a consideration on the concept of big data in medicine is appropriate. Typical medical imaging studies certainly does not reach the typical volumes of big data. However according to Eurostat reporting, each year, one person in ten in Europe undergoes computed tomography imaging (CT), one in 13 undergoes magnetic resonance imaging (MRI) and one in 200 positron emission tomography imaging (PET) [3]. This huge number of images combined with a research increasingly oriented to study large multicentric dataset, suggests that the future trend is certainly oriented towards a numerosity that can legitimately fall within the field of big data.

It should also be considered that each medical image is made up of a relatively large amount of data, which researchers would like to summarize in a few useful information. However when we consider one medical image we are actually considering about a 0.1 - 1 gigabyte of data. Furthermore the heterogeneity of medical imaging due to provenance as well as the relatively quickly flow of new data are typical characteristics of big data.

Nowadays, according to Smith et al. [142], we may consider an imaging study as "big" if it has 1000 or more subjects.

In the thesis I will deal with the issues of harmonization and data provenance in the context of Amyloid β ($A\beta$) PET neuroimaging. I will use a naturalistic database made up of 1001 subjects, thus falling within the limit of the big data category. The database I considered for my work is very heterogeneous both in terms of demography and in terms of acquisition methods. This fact certainly enhances the problem of data provenance.

Chapter 2

Neuroimaging

2.1 Neuroimaging overview

The terms neuroimaging refers to the use of various techniques to either directly or indirectly image the structure, function and pharmacology of the central nervous system.

Neuroimaging techniques can be classified into two categories: structural and functional. Structural neuroimaging (just imaging for the sake of simplicity) refers to approaches that are specialized for the visualization and analysis of anatomical properties of the brain, then it is particularly useful for detecting brain damage and abnormalities as well as for obtaining accurate knowledge of patients brain anatomy[79]. In this context,

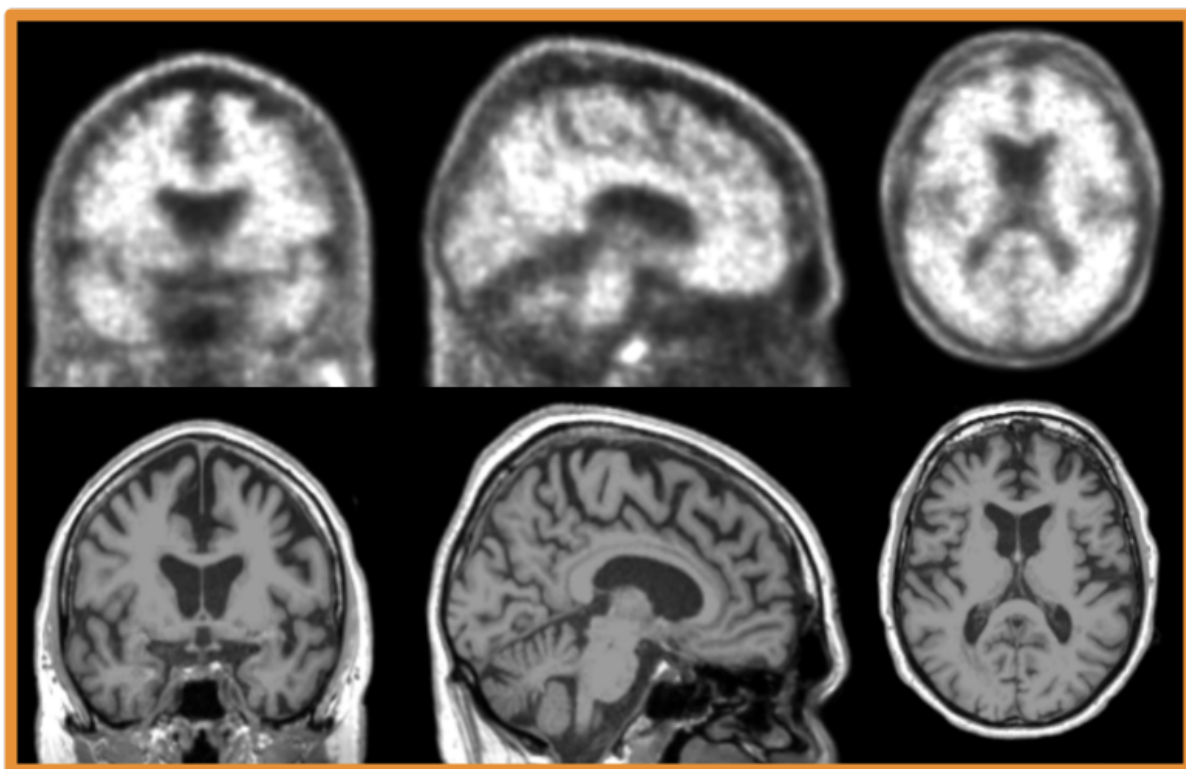


Figure 2.1: An example of functional (amyloid-PET, at the top) and structural (T1 MRI, at the bottom) are provided, respectively. From left to right are shown the same sagittal, coronal and axial section respectively. Both the images are acquired scanning the same patient.

structural imaging analyses can be used to quantify geometric structural properties such as the size and volume of a given structure [148] or the thickness of a cortical area (e.g., gray matter)[58]. Structural imaging include many techniques such as MRI, CT and Diffusion Tensor Imaging (DTI).

On the other side, functional imaging is used to identify brain areas and underlying brain processes that are associated with performing a particular cognitive or behavioral task [79]. Functional imaging consists in family of techniques widely used for detecting and measuring changes in metabolism, blood flow and regional chemical composition of tissues. Functional images can be acquired with a broad category of techniques such as functional Magnetic Resonance Image (fMRI) , Single-photon emission computed tomography (SPECT), and PET. In this thesis I will only discuss $A\beta$ PET imaging and I will especially focus on Amyloid-PET.

An example of structural and functional imaging can be found in figure 2.1.

2.2 Quantitative imaging

Quantitative imaging (QI) refers to the extraction and use of numerical/statistical features from medical images [74]. As a research field, QI includes the development, standardization, optimization, and application of anatomical, functional, and molecular imaging acquisition protocols, data analyses, display methods, and reporting structures, as well as the validation of QI results against relevant biological and clinical data[1].

The QI concept is closely tied to that of a biomarker, which will be defined in the next subsection.

2.2.1 Biomarker

There are different definitions of what a biomarker is, highlighting some aspects rather than others. One of the most used and universally accepted by scientific community is the following:

Biomarker definition: *any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease.*[13]

Biomarkers are something of very general in medicine and is not necessarily related to imaging, for example bone mineral density, blood pressure and blood glucose level are biomarkers.

Focusing on quantitative imaging, the Quantitative Imaging Biomarkers Alliance ¹ (QIBA), organized by the Radiological Society of North America (RSNA), has formally defined a QI biomarker as follows:

QI Biomarker definition: *an objective characteristic derived from an in vivo image measured on a ratio or interval scale as indicators of normal biological processes,*

¹<https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance>

pathogenic processes, or a response to a therapeutic intervention[1].

Biomarkers stand in contrast to medical symptoms, which are limited to those indications of health or illness perceived by patients themselves: they may but do not necessarily correlate with a patient's experience and sense of wellbeing.[145].

Going beyond the definitions, as a biomarker is a measurable quantity which aim to predict an output with regard to a subject's certain disease, two fundamental questions arise. Suppose we are studying a certain disease, thus

- what are the appropriate biomarkers as regard to the investigated disease? What is the more appropriate one?
- How can a given biomarker be measured?

The first question is implicitly based on having a medical theory that causally links the biomarker to the disease under investigation.

The second question opens up a vast scenario related to everything that concerns a measurement: reproducibility, calibration of the instrument, error theory, statistical analysis of the measured results, ability to distinguish signal from noise etc...

$A\beta$ PET images, which I worked on during my PhD, are involved in neurodegenerative disease studies; in particular brain cortical $A\beta$ burden represents a biomarker for Alzheimer's Disease (AD) [37, 153, 25, 164, 120].

2.2.2 Radiotracers

As quantitative neuroimaging is based on the measure of biomarkers extracted from medical images, a further important issue is related to the ability of a medical image to highlight what is to be measured.

For example, with regard to $A\beta$ PET imaging which I will discuss in this thesis, one may ask how is it possible to detect the presence of amyloid so that it can be measured through the so called quantification process.

The answer of this question relies both on radiotracers and PET working principles, which I will describe in this subsection and in the next section respectively.

A radiotracer (also called tracer for the sake of simplicity) is a chemical compound in which one or more atoms have been replaced by a radioactive isotope. Thus, a tracer can be considered as consisting of two key elements: a carrier molecule and a radioactive atom (radio-isotope) [64]. Carrier molecules are engineered to take part in a specific biological process related to the pathology under consideration. As an example, In $A\beta$ their role is to chemically bind, as specific as possible, to the red $A\beta$.

The most commonly used radiotracers for $A\beta$ -PET are Amyvid (Florbetapir), NeuraCeQ (Florbetaben) and Vyzamil (Flutemetamol), which are marked with the ^{18}F radioisotope (Figure 2.2). Another common radiotracer is Pittsburgh Compound-B (PIB) which is ^{11}C marked. ^{18}F and ^{11}C are largely used as they can be easily produced in cyclotrons and that their half-life is rather short: $\simeq 110$ min and $\simeq 20$ min respectively: this allows to avoid enduring radiation hazard for patients.

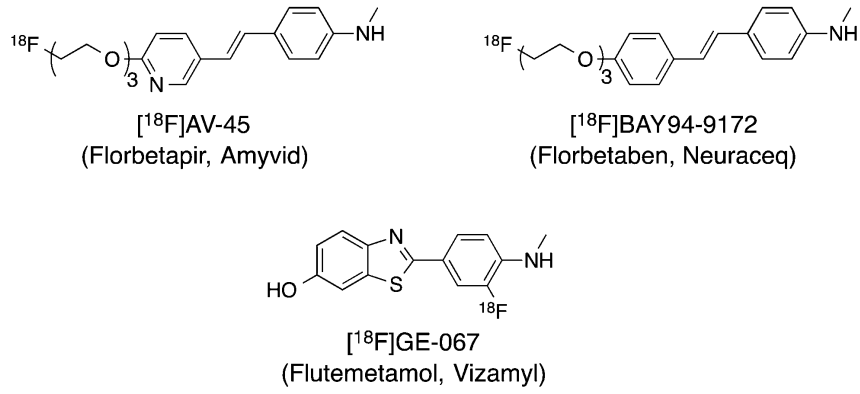


Figure 2.2: Examples of fluorinated radiotracers for $A\beta$ detecting

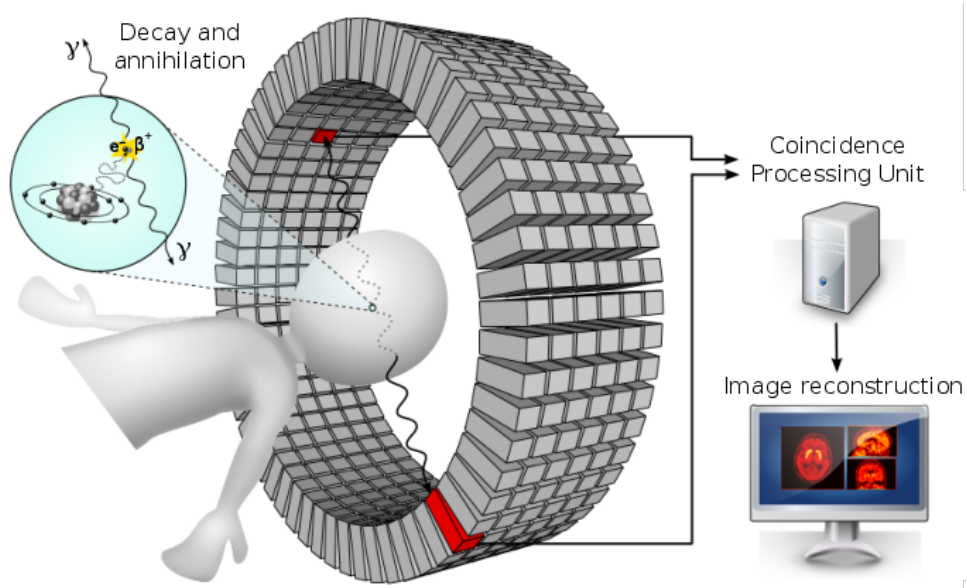


Figure 2.3: Pictorial representation of a typical PET detector, consisting of a ring of scintillators coupled to photomultipliers and a processing unit to retrieve the emitter position

2.3 PET physical principle

PET is an imaging technique used in nuclear medicine that allows non-invasive visualization and quantification of biological and physiological processes. This technique provides three-dimensional images of the anatomical district under examination by exploiting the decay processes of radioactive isotopes, which were previously introduced into the patients bloodstream by injection of a radiotracer [143].

PET operating principle relies on the decay processes of radioactive isotopes which mark the tracer. Tracer is administered intravenously and, after its administration, the patient is placed in the scanner for acquisition.

In this context, the measure consists in detecting the number of counts in a given volume, where the measure timings are short when compared to the half life of the radioisotope, and long with respect to the biochemistry kinetics that guide the tracer spread into the anatomical region under investigation.

The radioisotopes which mark the tracer decay β^+ emitting a positron e^+ and an

electron neutrino ν_e . The most common radioisotope is ^{18}F which decays as follows:

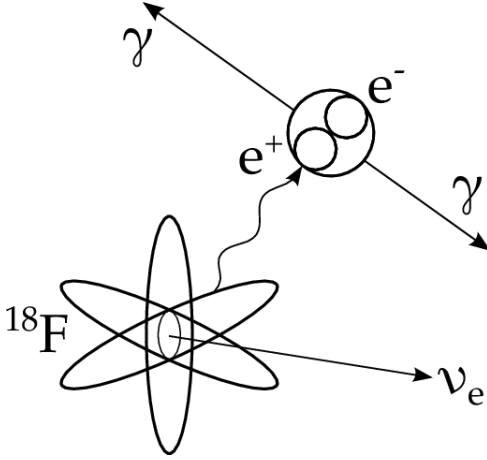
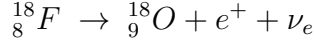


Figure 2.4: Decay of the radioisotope ^{18}F and annihilation of the positron with an electron resulting in the emission of two γ photons.

In organic tissue the emitted positron e^+ has a mean free path of about 1-2 mm. Then the positron annihilates with an electron and two collinear 511 KeV photons are emitted (as illustrated in figure 2.4). As the photon direction is random, the key point is to detect the coincidence of coplanar events.

A series of scintillators organized in a circular array around a central axis detects a signal only when two collinear events are detected within a time-difference of 6-20 ns [132], thus revealing the emission line of sight.

The raw signal which is obtained from all the revealed lines of sight are then processed in order to obtain all the emitter positions and thus to reconstruct the final three dimensional image.

The PET workflow is schematically illustrated in fig. 2.3

Two main steps are needed to obtain a three-dimensional image of the patient. The first is data acquisition which basically regards signal's detection and storage. The second step, called image reconstruction, allows to obtain a three dimensional images from the stored raw signal.

I will discuss both these steps in the next sections.

2.4 Data acquisition

2.4.1 Detection

The PET acquisition system is the PET scanner, which basically consists of a ring of scintillators coupled to photomultipliers.

PET working is based on the detection in coincidence of the two annihilation photons that originate from the β^+ emitting sources, which in clinical practice is the patient positioned within the ring.

The detection is obtained by means of scintillators, that convert high energy photons (generated by the annihilation of the positron) into photons of relatively low energy (in the visible spectrum). These photons reach the photocathode of the photomultipliers, which converts the light signal into an electrical signal.

Ideally, each annihilation event should be detected by the scintillator, while all events not due to the same annihilation should not be counted. To avoid counting photons not due to an annihilation event, it is necessary to select the events with respect to energy range and time window.

The energy range, which is also called energy window, is usually chosen as 350 - 650 keV [133]. About the time window, signal is detected through a coincidence filter, which selects only nearly simultaneous event. The time window Δt depends on the type of detector, but it is typically selected between $\Delta t = 6ns$ and $\Delta t = 20ns$ [132].

When two photons are detected both within the time window and within the energy window, they are considered as due to the same annihilation event, and then they are registered as a coincident count. Coincidence counts are also called *prompt events*.

The detection of a couple of photons related to the same annihilation event define an imaginary line connecting the two scintillators that recorded the event. This line is called line of response (LOR). It is important to emphasize that we do not know the position of the annihilation along the LOR: only the LOR is known and the event could be occurred anywhere along the LOR itself.

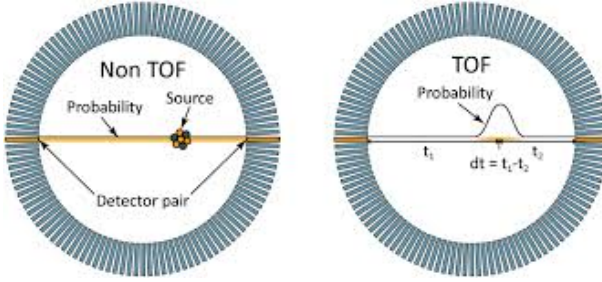


Figure 2.5: Schematic illustration of non-TOF (left) and TOF (right) photon detecting.

TOF vs NON-TOF detection The event may have occurred anywhere along a LOR, so, in the absence of further information it is assumed that the event is equiprobable along the whole LOR.

A considerable improvement relies on the ability of locating the annihilation event along the LOR itself; this is achieved thanks to the time of flight (TOF) technique. Thanks to the recent introduction of new scintillator crystals and electronics with better temporal resolution, it is now possible to obtain information on the position of the annihilation event along the LOR by measuring the small time difference in the detection of the two photons.

For an annihilation event placed at Δx from the center of the FOV, the Δt between the two detections will be:

$$\Delta t = 2 \frac{\Delta x}{c} \quad (2.1)$$

The time difference is really tiny: for a spatial offset of 9 cm the required temporal resolution is 600 ps.

This allows to go beyond the equiprobability assumption about the localization along the LOR, as the position of a given event can be exploited through a probability distribution definite on the LOR itself, as illustrated in figure 2.5.

TOF information considerably improves image quality, contrast and signal-to-noise ratio [154].

Digital data storing We notice that a LOR is uniquely defined by two polar coordinates (r, θ) . The radial coordinate r represents the distance between the center of the scanner and the LOR, while θ is the angular coordinate of the LOR with respect of a given axis, as illustrated in figure 2.6.

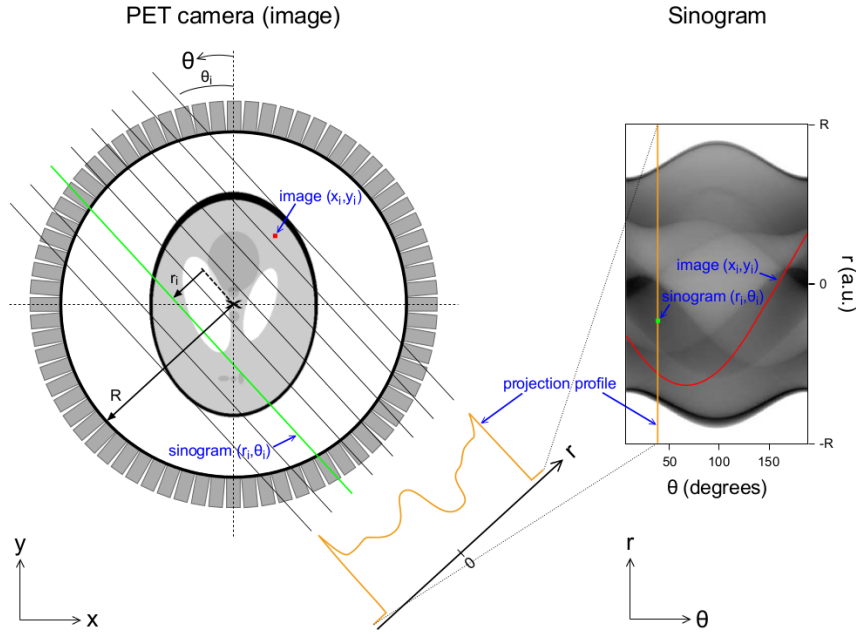


Figure 2.6: The image space (x,y) is represented in sinogram space (r,θ) . A group of LORs constitute a projection profile, corresponding to a column in the sinogram. A point in the image space corresponds a curve in the space of the sinogram, while a point in the space of the sinogram is a LOR in the image space (courtesy of [73])

Each event is uniquely identified by a LOR and it is registered as a count into an histogram, which is called sinogram. The sinogram is basically a 2-dimensional histogram of the LORs in the (r,θ) coordinates in a given detection plane. Thus, each LOR (and hence, detector pair) corresponds to a particular pixel (or element) in the sinogram, characterized by the coordinates r and θ . The relation between LORs and sinograms is further explained in figure 2.6. PET data are acquired directly into a sinogram which is a matrix of appropriate size in the computer memory. In the final sinogram the total counts in each sinogram's pixel represent the number of coincidence events detected during the counting time by the two detectors along the LOR. A typical sinogram can be found in figure 2.7.

2.4.2 PET detectors

As mentioned in the previous section, the key point of PET scan is the detection of coincidence events. Therefore, the properties of the PET scan detector is a really important issue which will be discussed in this section. Typically, the choice of a detector is based on the following characteristics:

- Stopping power of the detector for 511keV photons
- Scintillation decay time
- Light output per keV of photon energy
- Energy resolution of the detector

The stopping power of the detector determines the mean distance the photon travels until it stops after complete deposition of its energy, and depends on the density and effective

atomic number of the detector material.

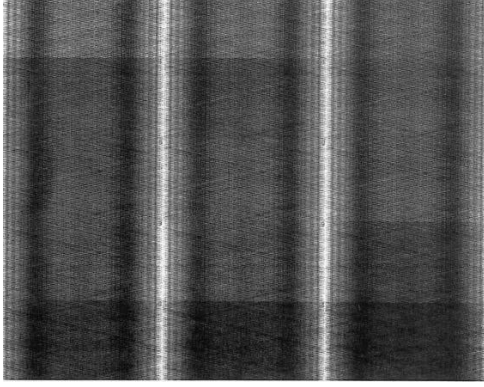


Figure 2.7: A typical sinogram

The scintillation decay time arises when a photon interacts with an atom of the detector material, and the atom is excited to a higher energy level, which later decays to the ground state, emitting visible light. This time of decay is called the scintillation decay time (given in nanoseconds) and it depends on the detector material. The shorter the decay time, the higher the efficiency of the detector at high count rates.

A high-light-output detector produces a well-defined pulse resulting in better energy resolution. The intrinsic energy resolution is affected by inhomogeneities in the crystal structure of the detector and random variations in the production of light in it. The energy resolutions at 511keV in different detectors vary from 6% to 20% [132], for routine integration time of pulse formation, which runs around a few microseconds. However, in PET imaging, the integration time is a few hundred nanoseconds in order to exclude random coincidences, and the number of photo-electrons collected for a pulse is small, thus degrading the energy resolution. Consequently, the detectors in PET scanners have relatively poorer energy resolution (10% to 25%) [132].

The detection efficiency of a detector is another important property in PET technology. Since it is desirable to have shorter scan times and low tracer activity for administration, the detector must detect as many of the emitted photons as possible. The 511 keV photons interact with detector material by either photoelectric absorption or Compton scattering. Thus, the photons are attenuated (absorbed and scattered) in the detector, and the fraction of incident photons that are attenuated is determined by a linear attenuation coefficient which gives the detection efficiency.

The most commonly used detectors are the BGO (bismuth germanate) and LSO (lutetium oxyorthosilicate).

2.4.3 Factors affecting acquired data

The data acquired in the form of sinograms are affected by a number of factors, namely variations in detector efficiencies between detector pairs, random coincidences, scattered coincidences, photon attenuation, dead time, radial elongation and acquisition modality (2-D vs 3-D). Each of these factors contributes to the sinogram to a varying degree. In this subsection many of the factors which affect the acquired data will be discussed.

Photon attenuation factor After the annihilation event, the photons will travel in the patients tissues in opposite directions; in this phase, different types of interaction occur between the photon and the medium, which attenuate or block the intensity. The modulation property of the medium depends on its characteristics (such as the density),

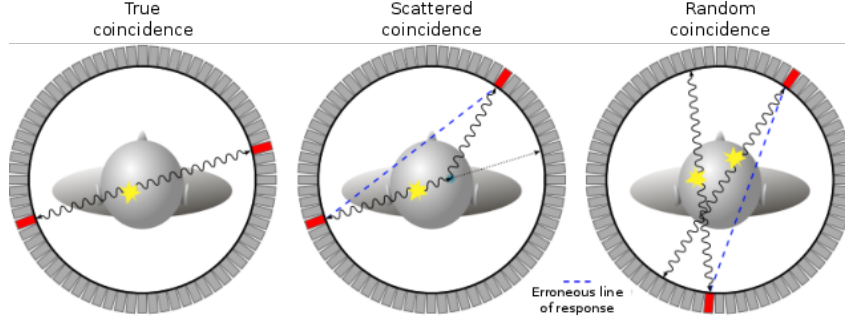


Figure 2.8: Types of coincidences in a PET scanner (courtesy of [73])

and is described by a linear attenuation coefficient μ . The fraction of photons that crosses a thickness x of matter with attenuation coefficient μ is expressed by the following relation:

$$\Gamma = e^{-\mu x} \quad (2.2)$$

which extends naturally to a sequence of materials with thickness x_i and coefficient μ_i as:

$$\Gamma = e^{-\sum_i \mu_i x_i} \quad (2.3)$$

The attenuation phenomenon in human tissues can be rather significant and can lead to nonuniformities in the images, because of the loss of relatively more coincidence events from the central tissues than the peripheral tissues of an organ and also because the two photons may transverse different organs along the LOR. Therefore it is necessary to know the magnitude of this effect to obtain the real distribution of the tracer within the patient.

The solution of this problem relies on computed tomography imaging. In modern scanners, PET is combined with CT, which delivers morphological and tissue properties information. The CT image is converted into an attenuation map and used to correct the intensities in the image [143, 96].

Thus, if the attenuation map is known from CT, the photons counts can be corrected applying the attenuation factors provided by (2.3) to all individual LOR counts in the sinogram.

Scatter and random coincidences Ideally, the only events that should be recorded are those associated with the actual annihilation of positrons, called *true coincidence*; however, a series of fringe events which satisfy the coincidence detection criteria are still considered as annihilation events, resulting in noise, artifacts, and degradation of spatial resolution.

Most of the spurious coincidence events are classifiable in random and scatter coincidences (figure 2.8).

A *scattered coincidence* occurs when one or both photons undergo Compton scattering. The Compton interaction causes both energy loss and change of direction, resulting in an incorrect location of annihilation, and ultimately in image degradation, background noise increase and reduction in contrast.

Since both scattered and true coincidence rates vary linearly with the administered activity, the scatter-to-true ratio does not change with the activity. Also, this ratio does not

change with the width of the time window, because scatter events arise from the same annihilation event and the two similar photons arrive at the two detectors almost at the same time. In 2-D acquisition, as previously said, the use of septa collimators removes additional scattered events, whereas in 3-D acquisition, they become problematic. Typically, the scatter fraction ranges from 15% in 2-D mode to more than 40% in 3-D mode in modern PET scanners. In practice, the correction for scatter is made by taking the counts just outside the field of view, where no true coincidence counts are expected. The outside counts contain both random and scatter events. After subtracting random counts, the scatter counts are subtracted from the prompt counts across the field of view to give true coincidence counts.

A *random coincidence* occurs when photons associated with two distinct annihilations are seen by the detection system as coming from a common annihilation event. Random events add to the background causing artifacts and loss of image contrast and are more problematic in low-efficiency detectors and in 3-D acquisition modality.

The rate of random coincidences is given by

$$R = C_1 C_2 \Delta t \quad (2.4)$$

where Δt is the time window in nanoseconds for the system and where C_1 and C_2 are the single count rates in counts/sec on each of the two detectors on the LOR. Random events increase with the square of the administered activity whereas the true coincidence events increase linearly with the administered activity.

Furthermore random coincidences are proportional to the time and energy window.

A common method of correcting for random events is to employ two coincidence circuits one with the standard time window of Δt ns and another with a delayed time window (e.g. from 50 ns to 50 Δt ns) using the same energy window. The counts in the standard time window include both the randoms plus trues, whereas the delayed time window contains only the randoms. For a given source, the random events in both time windows are the same within statistical variations. Thus, correction for random coincidences is made by subtracting the delayed window counts from the standard window counts.

2D vs 3D acquisition Axially, PET scanners consist of several rings of detector elements that may or may not be separated by thin annular septa of photon-absorptive material that provide collimation. With collimation, all data is acquired in 2-dimensional slices between the septa (2.9).

The annular septa are usually made of tungsten and their thickness is of ~ 1 mm while their radial width is of about 7-10cm.

2D acquisition mode is characterized by the presence of septa which collimate photons, while 3D mode works without any septa.

Detector pairs connected in coincidence in the same ring give the direct plane event. In 2D mode most of the random and scattered 511 keV photons from outside the ring are prevented by the septa to reach the detectors, leaving the true coincidences to be recorded. However, considering direct planes coincidences only typically leads to a very small sensitivity (i.e. low counts). Then, to improve sensitivity in 2-D acquisition, detector pairs in two adjacent rings are connected in a coincidence circuit.

Coincidence events from a detector pair in this arrangement are detected also on the so-called cross plane that falls midway between two adjacent detector rings. Direct plane

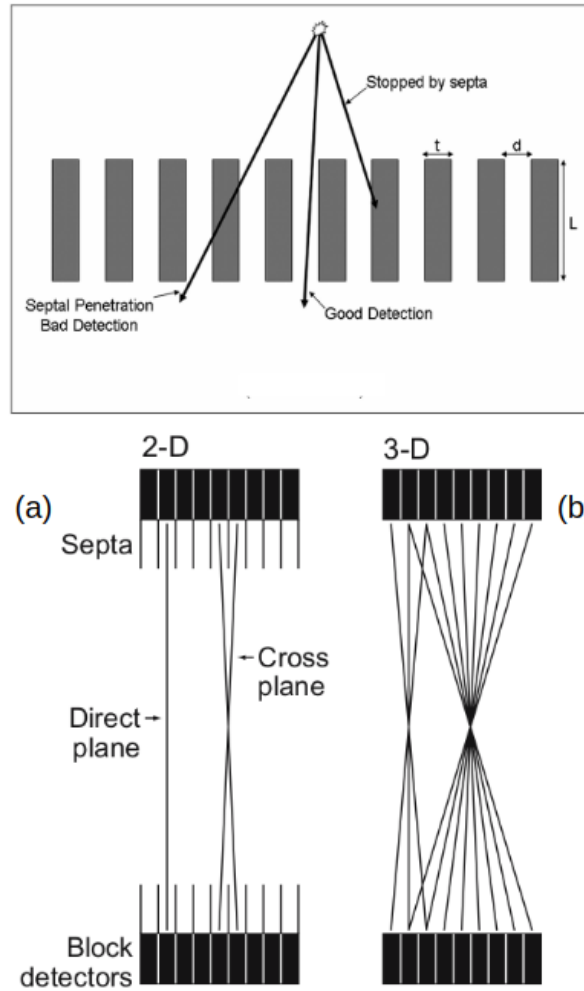


Figure 2.9: Top: Septa allow to detect only photons which lie in accepted planes only. The septa geometry, i.e. L , d and t in figure, defines the accepted planes. Bottom (a): 2D data acquisition with the septa placed between. Detectors connected in the same ring give direct plane events. Detectors connected in adjacent rings give cross plane events. Bottom (b): When septa are removed, the 3-D data acquisition takes place.

and cross plane coincidences are illustrated in figure 2.9. Moreover, instead of two adjacent rings, such cross planes can be obtained from other nearby rings that are connected in coincidence. The maximum acceptable ring difference is usually of ± 5 rings [132], i.e. a maximum of 5 rings across can be interconnected in coincidence. Coincidences between detectors in a connected n system-rings neighboring rings are summed or rebinned to produce of $2n - 1$ sinograms, (n from direct planes, $n - 1$ from cross planes). These sinograms may be reconstructed into images using standard 2D techniques, which will be discussed in the following section. Obviously, increasing the number of cross planes increases the sensitivity as well as increases number of spurious coincidences. 5

When a scanner is operated without collimation (i.e. no septa), coincidences from all axial angles in the FOV will be accepted. Data storage, correction, and image reconstruction is considerably more complex in the 3D case. Modern PET scanners are able to operate in either 2D-only or 3D-only mode, or in a 2D/3D mode for those with retractable septa. Figure 2.9 shows the effect that collimation has on the acquisition of coincidence counts: the septa block a fairly large number of true coincidences from ever reaching the detector surface, decreasing sensitivity. However, they also reduce the

noise: in 3D acquisition spurious coincidences fractions are 30%-40%, while in the 2D acquisition are $\sim 15\%$ [132]. In this context, of special importance are accidental counts that partly originate from outside of the area between the detector surfaces (true FOV), because without collimation, the scanner is sensitive to activity from a very large area outside the true FOV [16], as illustrated in figure 2.10.

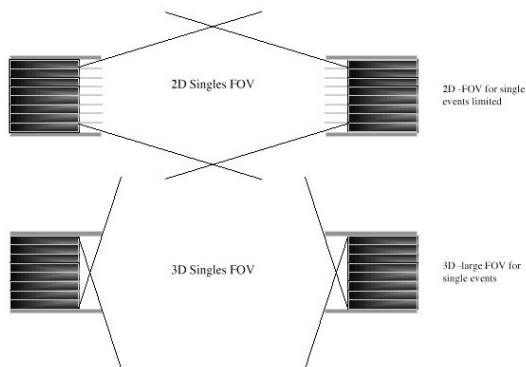


Figure 2.10: Effect of septa removal on sensitivity to coincidences which occur out of the true FOV.

The decision on whether to use 2D or fully 3D acquisitions is still under debate, weighing the reduction of background counts against the loss of sensitivity. [135].

To summarize, 3D acquisition has a higher sensitivity than the 2D one. However 2D acquisition is simple to reconstruct and less sensitive to spurious coincidences, then image contrast and quality could be better in 2D despite its lower sensitivity. At the end of the story, there is no definitive answer, as the 2D vs 3D choice can

be considered as choosing the best trade-off between noise and sensitivity, and this best trade-off depends on the investigated issues and can not be defined in general.

Dead Time For detection systems that record discrete events, such as particle and nuclear detectors, the dead time is the time after each event during which the system is not able to record another event [92].

Thus, all the events which occur during the dead time are not recorded by the detector. The dead-time represents a serious problem at high count rates and varies with different PET systems. The loss of coincidence events due to dead-time can be reduced by using detectors with shorter scintillation decay time and faster electronics components in the PET scanners. Dead time correction is made by empirical measurement of observed count rates as a function of increasing concentrations of activity. From these data, the dead time is calculated and a correction is applied to compensate for the dead time loss.

Detector Size One factor that greatly affects the spatial resolution is the intrinsic resolution of the scintillation detectors used in the PET scanner. For multidetector PET scanners, the intrinsic resolution R_i is related to the detector size d . R_i is normally given by $d/2$ on the scanner axis at midposition between the two detectors and by d at the face of either detector [132]. Thus it is best at the center of the FOV and deteriorates toward the edge of the FOV. For a 6mm detector, the R_i value is about 3mm at the center of the FOV and about 6mm toward the edge of the FOV.

0.7

Positron range Energetic positrons travel a distance in tissue, losing most of their energy. Then they annihilate and produce a couple of photons. As a consequence, the

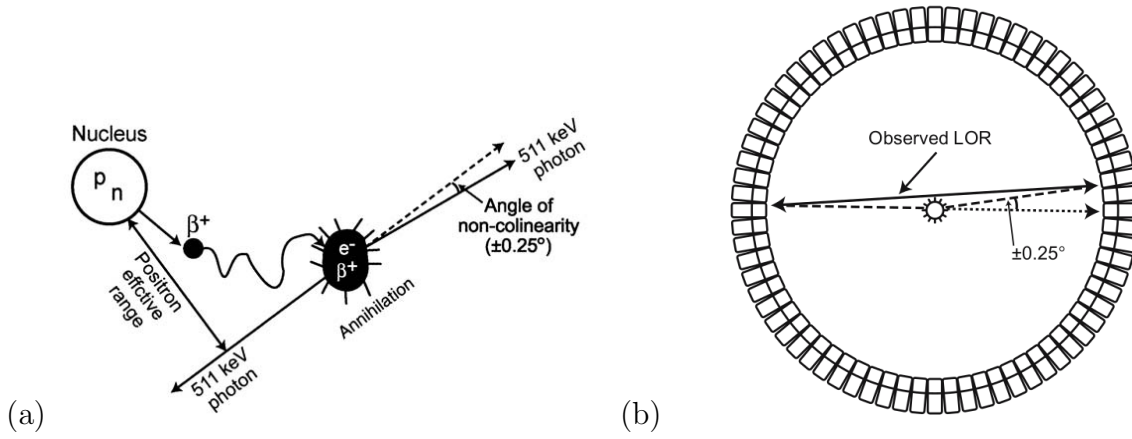


Figure 2.11: A pictorial representation of positron range (a) and non-collinearity (b)

site of e^+ emission differs from the site of annihilation, as illustrated in figure 2.11 a.

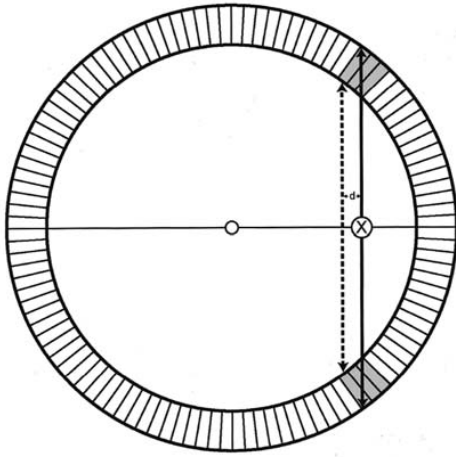


Figure 2.12: An illustration of radial elongation. An off-center event (solid line) strikes the back of the detector pair tangentially. The dashed LOR defined by the detector positions is a distance d away from the actual location of the positron annihilation

The distance traveled by the positron increases with its energy, but decreases with the tissue density. Since the positrons are emitted with a spectrum of energy, the positron range is essentially an effective range, which is given by the shortest distance from the emitting nucleus to the positron annihilation line. The typical positron range in tissue is of about 1 mm for ^{18}F [83].

Non-collinearity Another factor of concern is the non-collinearity that arises from the deviation of the two annihilation photons from the exact 180 degrees position. That is, two 511keV photons are not emitted at exactly 180 degrees after the annihilation process (see figure 2.11 a), because of some small residual momentum of the positron at the end of the positron range. The maximum deviation from the direction is ± 0.25 degrees. Thus, the observed LOR between the two detectors does not intersect the point of annihilation, but is somewhat displaced from it, as illustrated in figure 2.11 b.

The contribution from non-collinearity worsens with larger diameter of the ring, and it amounts to 1.8 to 2mm for currently available 80-cm to 90-cm PET scanners.

Radial Elongation Radial elongation is caused by the photons penetrating into the detector ring coupled with the fact that the detector module does not determine the interaction point, but the interaction crystal.

The off-centered 511 keV photons can strike tangentially at the backside of the detector pair and form a coincidence event. As seen in figure 2.12, the LOR defined by the detectors

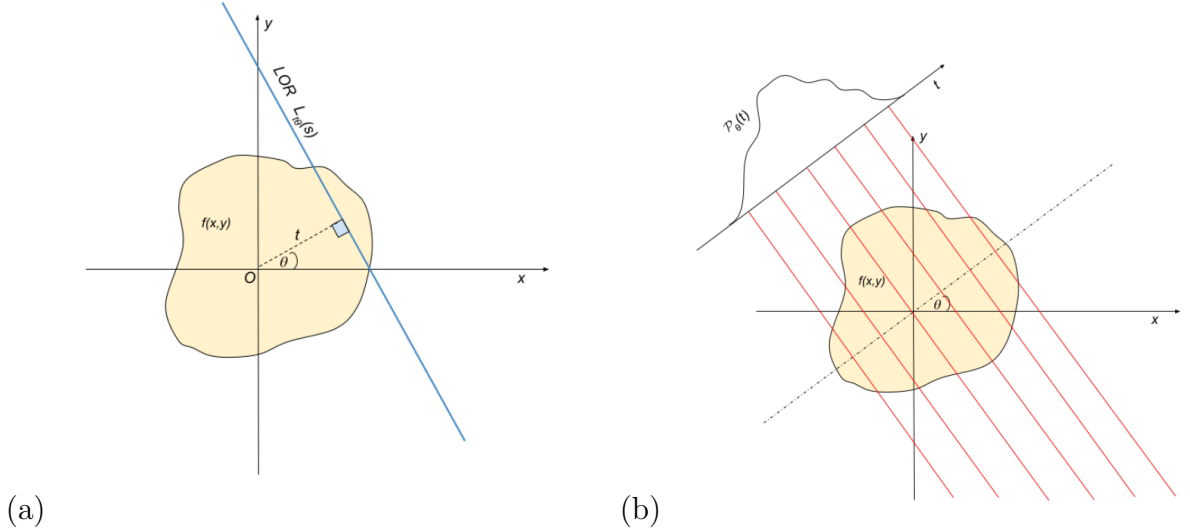


Figure 2.13: In both figure $f(x, y)$ represent the tracer distribution. On the left: the blue line represent the LOR $L_{t,\theta}$, defined by t and θ . We move on the LOR by varying the parameter s . On the right: the projection $\mathcal{P}_\theta(t)$ is obtained by integrating all the LORs defined by θ (the red lines). We move on the projection varying t , while choosing another θ allow us to change projection.

(dashed line) is at a distance d away from the actual LOR (solid line), resulting in the blurring of the image due to unknown depth of interaction in the detector material.

2.5 Image Reconstruction

2.5.1 The image reconstruction problem

The goal of image reconstruction step is to recover the radiotracer distribution starting from the sinograms. The radiotracer distribution can be considered as a function $F(x, y, z)$ where x, y are the transaxial orthogonal coordinates and z is the axial coordinate with respect to the PET detectors rings. The axis origin will be supposed to be located on the symmetry axis of the rings.

As explained in the last section, there are two main ways to acquire images: the 2D and the 3D acquisition methods, which respectively lead to a 2D and 3D image reconstruction techniques. In this section I will discuss the most important 2D reconstruction methods and then I will give a brief overview about the 3D reconstruction techniques.

For this purpose, we will generically consider a bi-dimensional transaxial slice of the radiotracer distribution $f(x, y) = F(x, y, z = z_0)$, as illustrated in figure 2.13. Before continuing, I want to point out that both the reconstructed images and sinograms are not continuous object, as they are stored in digital array. However we will treat them as if they were continuous object, and we will discretize them only when strictly needed (typically in the iterative methods subsections).

Let now consider a generical LOR $L_{t,\theta}$. As illustrated in figure 2.13 we notice that $L_{t,\theta}$ is uniquely defined by two coordinates (t, θ) , where $t \in \mathbb{R}$ and $\theta \in [0, 2\pi)$: (t, θ) are coordinates for the space of all lines of \mathbb{R}^2 . Furthermore, we notice that the versor $(\cos \theta, \sin \theta)$ represents the direction of the distance $\overline{OL_{t,\theta}}$, while $(-\sin \theta, \cos \theta)$ is the

versor of $L_{t,\theta}$.

Therefore we can parameterize a given line $L_{t,\theta}$ in terms of a real number s as follows:

$$L_{t,\theta}(s) = (x(s), y(s)) = (t \cos \theta - s \sin \theta, t \sin \theta + s \cos \theta) \quad \forall s \in \mathbb{R} \quad (2.5)$$

We notice that by varying t we move across parallel LORs, while by fixing $t = t_0$ and varying s we move along the specific LOR $L_{t_0,\theta}$.

Radon Transform Given a function $f(x, y)$ defined on \mathbb{R}^2 with compact support, the Radon transform \mathcal{R} of the function f is defined as follows

$$(\mathcal{R}f)(t, \theta) = \int_{L_{t,\theta}} f(x, y) ds = \int_{-\infty}^{+\infty} f(t \cos \theta - s \sin \theta, t \sin \theta + s \cos \theta) ds \quad (2.6)$$

where $t \in \mathbb{R}$ and $\theta \in [0, 2\pi)$. To better understand the meaning of Radon Transform, let fix $t = t_0$ and $\theta = \theta_0$. The Radon Transform $\mathcal{R}f(t_0, \theta_0)$ is the line integral of $f(x, y)$ along the line (t_0, θ_0) , thus it determines the total amount of tracer along the line (t_0, θ_0) itself. Furthermore, if we consider $\mathcal{R}f(t, \theta_0)$, we obtain the total amount of tracer (i.e. the projection of $f(x, y)$) along θ_0 direction (see figure 2.13). Therefore, we denote the projection \mathcal{P} of a function f along θ_0 as follows

$$(\mathcal{P}f)_{\theta_0}(t) = (\mathcal{R}f)(t, \theta_0) \quad (2.7)$$

It is important to notice that $(\mathcal{R}f)(t, \theta)$ represents the continuous version of the sinogram, which is the data we actually measure. We notice that $(\mathcal{P}f)_{\theta_0}(t)$ is the $\theta = \theta_0$ sinogram column, as illustrate in figure 2.6.

At the end of the story, the 2D image reconstruction problem can be explicitly formalized for each slice as follows: given $(\mathcal{R}f)(t, \theta)$, what is the function $f(x, y)$ which originates $(\mathcal{R}f)(t, \theta)$?

There are two main ways to answer to this question: the analytical one, in which the mathematical model is analytically inverted, and iterative one in which the tomographic image is reconstructed using iterative statistical methods [143, 82].

2.5.2 Analytic reconstruction methods

Analytic methods provide a direct solution for the formation of the image, based on a line-integral model and the Fourier theory. Although they have low requirements for computational resources and they are relatively fast, the reconstructed images suffer from high noise levels and streak artefacts. This is a result of the approximation that along a projection line, the number of counts is linearly proportional to the integral of the tracer density.

The goal of analytical methods is to find some type of inversion formula for the Radon transform that will allow us to recover our starting function f . In this subsection we discuss the Filter Back Projection (FBP) reconstruction, which are the widely used analytical reconstruction method. Before explaining FBP, it will be convenient to introduce the central slice theorem and the back-projection operator.

Central Slice Theorem The central slice theorem (also known as the projection slice theorem) is the most important relationship in analytic image reconstruction. In this paragraph we derive the two-dimensional version. To do that, we require the imaging process be shift invariant, which allows the use of Fourier transforms. Shift-invariance means that if we scan a shifted object, the projections are also shifted but are otherwise identical to the projections of the unshifted object. Shift-invariance is a natural property of two-dimensional imaging.

Central Slice Theorem

Let denote with \mathcal{F}_2 and \mathcal{F} the 2-dimensional and 1-dimensional Fourier transform operator respectively. For a function f defined on \mathbb{R}^2 the central slice theorem states that

$$(\mathcal{F}_2 f)(K \cos \theta, K \sin \theta) = \{\mathcal{F}(\mathcal{R}f)\}(K, \theta) \quad (2.8)$$

for all $K \in \mathbb{R}$ and $\theta \in [0, 2\pi)$.

Proof

We first recall that the two dimensional Fourier transform of $f(x, y)$ is given by

$$(\mathcal{F}_2 f)(k_x, k_y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) e^{-i(xk_x + yk_y)} dx dy \quad (2.9)$$

The generical point (k_x, k_y) in the spatial frequencies domain can be expressed in polar coordinates $(k_x = K \cos \theta, k_y = K \sin \theta)$. Therefore we obtain

$$(\mathcal{F}_2 f)(K \cos \theta, K \sin \theta) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) e^{-iK(x \cos \theta + y \sin \theta)} dx dy \quad (2.10)$$

We perform the following coordinates change suggested from (2.5) :

$$x(s, t) = t \cos \theta - s \sin \theta \quad (2.11)$$

$$y(s, t) = t \sin \theta + s \cos \theta \quad (2.12)$$

from which we have $t = x \cos \theta + y \sin \theta$. We notice that the determinant of the Jacobian is 1, thus $dx dy = ds dt$. Thus (2.9) became

$$(\mathcal{F}_2 f)(K \cos \theta, K \sin \theta) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(t \cos \theta - s \sin \theta, t \sin \theta + s \cos \theta) e^{-iKt} ds dt \quad (2.13)$$

Because e^{-iKt} has no dependence on s , we are able to rearrange the above integral as follows:

$$(\mathcal{F}_2 f)(K \cos \theta, K \sin \theta) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f(t \cos \theta - s \sin \theta, t \sin \theta + s \cos \theta) ds \right) e^{-iKt} dt \quad (2.14)$$

where the inner integral is the Radon transform (2.6). Therefore, we finally obtain

$$(\mathcal{F}_2 f)(K \cos \theta, K \sin \theta) = \int_{-\infty}^{+\infty} \mathcal{R}f(t, \theta) e^{-iKt} dt = \{\mathcal{F}\mathcal{R}f\}(K, \theta) \quad (2.15)$$

and then the theorem is proofed.

We shown that the bi-dimensional Fourier transform of f evaluated along a given direction is equivalent to the Fourier transform of the projection of f into the same direction. A pictorial representation of central slice theorem is given in figure 2.14.

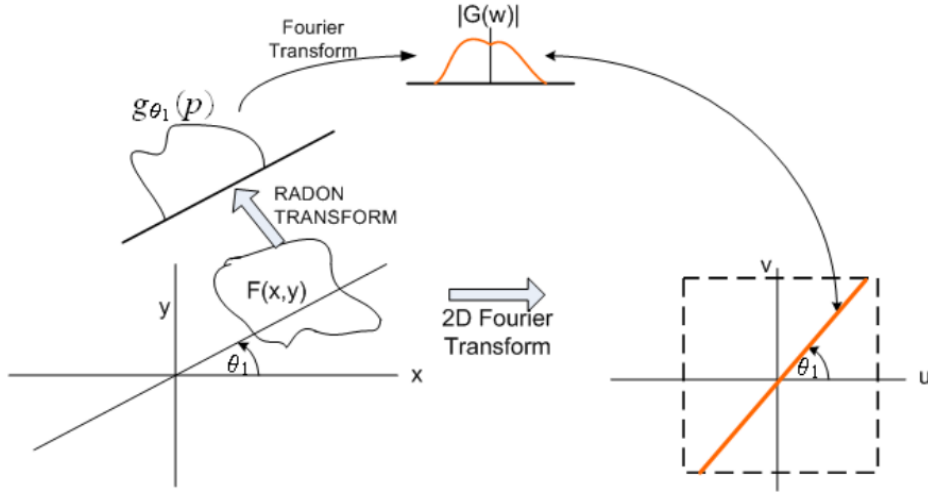


Figure 2.14: A pictorial representation of the central slice theorem

Backprojection In this paragraph we introduce the backprojection operator. This is an essential step in image reconstruction. Conceptually the backprojection can be described as placing a value of $\mathcal{R}f(t, \theta)$ back into an image array along the appropriate LOR $L_{t\theta}$, but because the knowledge of where the values came from was lost in the integration with respect to s (see (2.7)), the best we can do is place a constant value given by $\mathcal{R}f(t, \theta)$ along each LOR $L_{t,\theta}$.

This can be done using the projection (2.7)

$$f(t \cos \theta - s \sin \theta, t \sin \theta + s \cos \theta) = (\mathcal{P}_\theta f)(t) \quad (2.16)$$

We notice that (2.16) gives a function f made up of constant values along the chosen θ direction, or saying it in other words, which is constant along each LOR (i.e. with respect to s).

Let now consider all the projections contributions: we integrate respect to θ the equation (2.16). This leads to introduce the backprojection operator \mathcal{B} which maps functions from image space to projection space. Given a function $h(t, \theta)$ defined in the projection space, we define the backprojection $\mathcal{B}h$ at a point (x, y) of the image space as

$$(\mathcal{B}h)(x, y) = \frac{1}{\pi} \int_0^\pi h(x \cos \theta + y \sin \theta, \theta) d\theta \quad (2.17)$$

where we used the relation $x \cos \theta + y \sin \theta = t$ obtained from the inversion of (2.11). Now applying the backprojection operator to equation (2.16) we obtain

$$(\mathcal{B}\mathcal{R}f)(x, y) = \frac{1}{\pi} \int_0^\pi (\mathcal{R}f)(x \cos \theta + y \sin \theta, \theta) d\theta \quad (2.18)$$

where $(\mathcal{B}\mathcal{R}f)(x, y)$ is a function in the image space.

The function $(\mathcal{B}\mathcal{R}f)(x, y)$ is the so called simple backprojected (SB) image reconstructed.

As is clear from (2.18), the backprojection operator is not the inverse operator of Radon transform. Thus, even if we use many projections to recover the original image, we get a result which is an approximation of the original f and which is really very poor

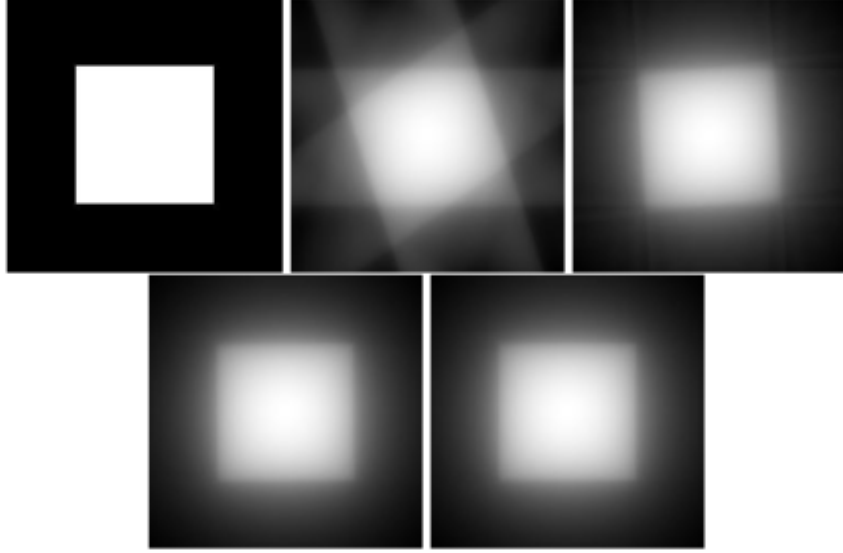


Figure 2.15: Simple Backprojection of a square in 5, 25, 100, and 1000 directions (courtesy of [20]).

about quality and not very similar to the original image. In particular the SB has the problem of star pattern artifacts (as you can see in figure 2.15), which is a direct consequences of assigning a constant value on each LOR (see equation (2.16)) and summing up these constant values .

Filter Backprojection In this paragraph we go beyond the SB and we discuss the Filter Backprojection (FBP) method, which allows to explicitly inverse the Radon transform in order to exactly recover the original f . FBP is an based on SB theory and gives better result than SB, however even if an inversion formula for the Radon transform exist, recovering the original f is not possible, as will be explained below.

We will now inverse the radon transform. Let consider $f(x, y)$: we can write the following identity

$$f(x, y) = (\mathcal{F}_2^{-1} \mathcal{F}_2) f(x, y) = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\mathcal{F}_2 f)(k_x, k_y) e^{i(xk_x + yk_y)} dk_x dk_y \quad (2.19)$$

where \mathcal{F}_2^{-1} is the bi-dimensional inverse transform. We now change the integration variables from Cartesian (k_x, k_y) to polar coordinates (K, θ) , where $k_x = K \cos \theta$ and $k_y = K \sin \theta$, with $K \in \mathbb{R}$ and $\theta \in [0, \pi]$. This gives the Jacobian determinant $|\det \mathbf{J}| = |K|$. Thus the (2.19) became

$$f(x, y) = \frac{1}{4\pi^2} \int_0^\pi \int_{-\infty}^{+\infty} (\mathcal{F}_2 f)(K \cos \theta, K \sin \theta) e^{iK(x \cos \theta + y \sin \theta)} |K| dK d\theta \quad (2.20)$$

Applying the central slice theorem (2.8) to the integrand function, we obtain

$$f(x, y) = \frac{1}{4\pi^2} \int_0^\pi \int_{-\infty}^{+\infty} (\mathcal{FR} f)(K, \theta) e^{iK(x \cos \theta + y \sin \theta)} |K| dK d\theta \quad (2.21)$$

We notice that the inner integral is equal to the 1-dimensional Fourier inverse transform:

$$\begin{aligned} & \int_{-\infty}^{+\infty} (\mathcal{FR}f)(K, \theta) e^{iK(x \cos \theta + y \sin \theta)} |K| dK t = \\ & = 2\pi \mathcal{F}^{-1} \left(|K| (\mathcal{FR}f)(K, \theta) \right) (x \cos \theta + y \sin \theta, \theta) \end{aligned} \quad (2.22)$$

Inserting the (2.22) into the (2.21) we obtain

$$f(x, y) = \frac{1}{2\pi} \int_0^\pi \mathcal{F}^{-1} \left(|K| (\mathcal{FR}f)(K, \theta) \right) (x \cos \theta + y \sin \theta, \theta) d\theta \quad (2.23)$$

Using (2.17), the integral can be rewrite as a backprojection

$$f(x, y) = \frac{1}{2} \mathcal{B} \left\{ \mathcal{F}^{-1} \left(|K| \mathcal{FR}f \right) \right\} (x, y) \quad (2.24)$$

where the dependence from coordinate K, θ has not been made explicit for the sake of simplicity.

The important factor in this formula is the $|K|$ multiplier that occurs between the Fourier transform and its inverse. This additional $|K|$ can be interpreted as a filter (K is in the Fourier space) we must apply to the projected data in the Fourier domain (i.e. to $\mathcal{FR}f$) in order to recover the original image f . This required filtering operation gives its name to the filtered backprojection formula.

However the formula (2.24) is clearly awkward to handle. Recalling the convolution theorem

$$\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g) \quad (2.25)$$

and rewriting $|K|$ as $\mathcal{FF}^{-1}|K|$ in (2.24) we obtain an handful version of the analytical inversion formula

$$f(x, y) = \frac{1}{2} \mathcal{B} \left\{ (\mathcal{F}^{-1}|K|) * (\mathcal{R}f) \right\} (x, y) \quad (2.26)$$

The ideal reconstruction requires that the projection data to be filtered with by a ramp filter $|K|$ before backprojecting, as show in (2.24) (or in physical coordinates, to be convoluted with $\mathcal{F}^{-1}|K|$ as shown in (2.26)).

Such a ramp filter in practice is impossible to build as it has infinite in length. Moreover, even if it were possible to build such a filter, it would not really be convenient: in real images noise is always present and an ideal ramp filter would amplify high frequency noise. In practice, a variety of different filters may be used, such as the ramp band limited filter, the Hanning filter and the Butterworth filter [163, 159, 99, 17]. These filters all have trade-offs among resolution and the presence of artifacts. Once the projections have been filtered, they are back-projected as described above. Exapmles of PET images FBP reconstructed are given in figure 2.16.

2.5.3 Iterative reconstruction methods

Iterative methods estimate a series of tentative radioactivity distributions and compares the respective projections with those actually acquired, refining the former at each iteration until correspondence is satisfactory [160]. While computationally more expensive, it allows to modulate and account for the statistical fluctuations associated with noise, both on the reconstructed images and on the raw data side [78].

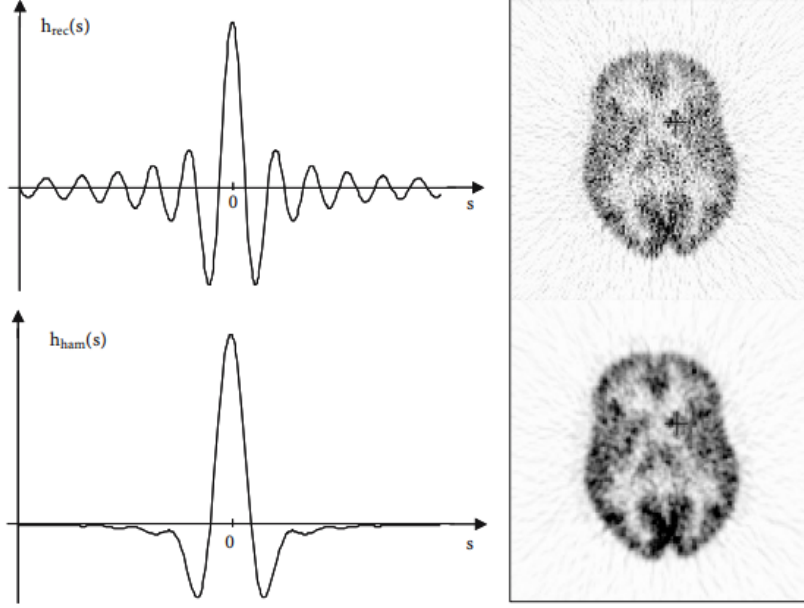


Figure 2.16: On the left: convolution kernels corresponding to a generic rectangular filter window (top) and to a generic Hamming filter window (bottom). On the right: A transaxial slice of an PET brain scan reconstructed using FBP with these two windows. (Courtesy of [45]).

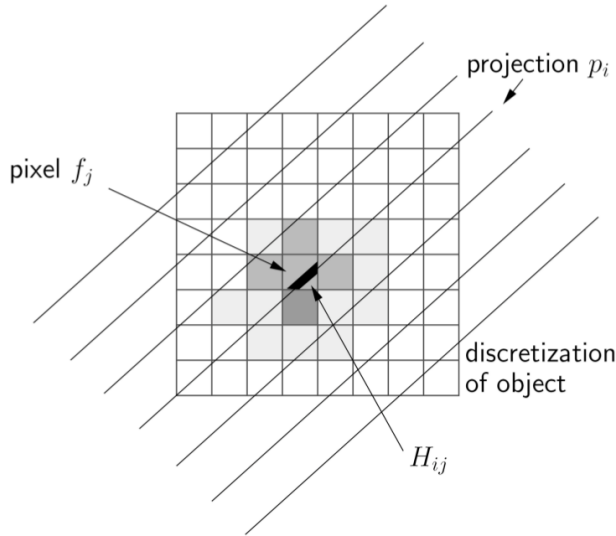


Figure 2.17: Pictorial representation of (2.27)(courtesy of [161])

The tracer distribution we aim to recover is a continuous function in the spatial domain, while the measured data acquired from the PET scanner are discretized among several sinogram bins. For this reason, in reality the function we aim to recover is the discrete version of the tracer distribution. Focusing on a $z = \text{constant}$ slice, the tracer distribution is given by $f(x, y)$. Therefore we approximate the continuous f choosing a n -by- n grid and assuming that $f(x, y)$ takes constant values on each grid element, also called pixel. As the discrete version of $f(x, y)$ is made up of $N = n^2$ elements, it will be considered as a vector $\mathbf{f} = (f_1, \dots, f_N)^T$

The grid dimension should be appropriate: too small N will lead to poor spatial resolution (i.e. an ability to resolve two adjacent high-contrast objects), while too large N will lead to high computational costs.

Note that the projection space is also discrete, with the projection data represented by the column vector \mathbf{p} . It is worth to remark that \mathbf{p} is a vectorial rearrangement of a sinogram, as sinogram is matrix: the element \mathbf{p}_i represents all the counts detected on the i -th LOR. The elements of \mathbf{p} are also called projection bins.

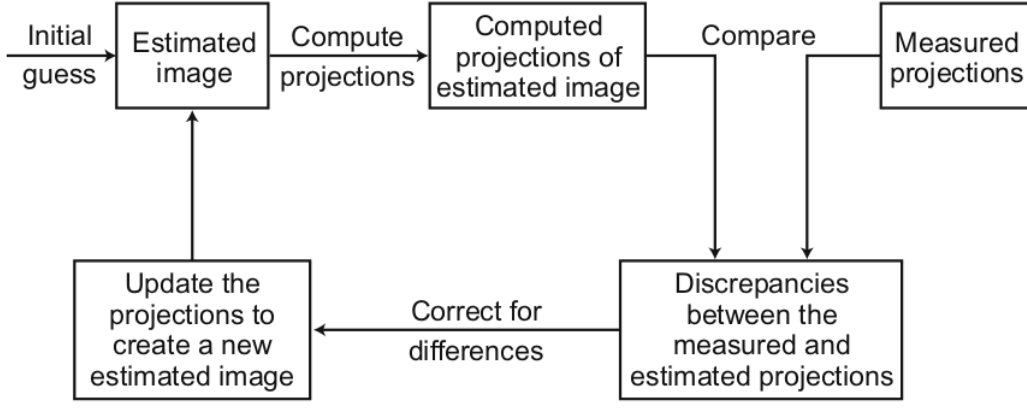


Figure 2.19: Flowchart of a generic iterative reconstruction algorithm.

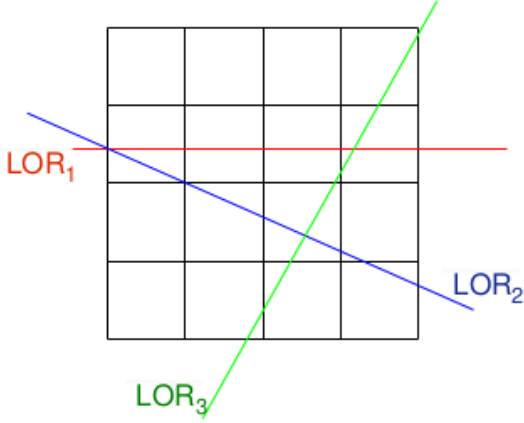


Figure 2.18: A simple example of ray-tracing for three lines of response (LOR) on a small 2-dimensional image grid.

as well as in the introduction of noise and artifacts.

\mathbf{H} describes the imaging process, and can represent attenuation and any linear blurring mechanisms. Each element of \mathbf{H} denoted by H_{ij} represents the mean contribution of the j -th pixel in the image space to the i -th bin in the projection space.

The matrix elements can be defined so that a projection bin receives contributions only from pixels that are intersected by a given line, while the contributions of pixels that do not intersect the line are set to zero, as illustrated in figure 2.18.

Iterative methods working principle Here we outline the common features of most iterative reconstruction algorithms and discuss some of their general properties. Let $\mathbf{m} \in \mathbb{R}^M$ the measured data in the projection space (i.e. the sinogram).

Most iterative reconstruction algorithms fit the general model shown in figure 2.19. First of all, the process begin with some initial estimate $\mathbf{f}^{(0)}$ of the pixel intensity values

System matrix All the iterative methods is based on a system matrix (also called projection forward operator). The system matrix is a linear operator $\mathbf{H} \in \mathbb{M}(M, N, \mathbb{R})$ which map elements of the image space into elements of the projection space through the following relation:

$$\mathbf{p} = \mathbf{H}\mathbf{f} \quad (2.27)$$

The quality of the reconstructed images directly depends on the accuracy of the system matrix \mathbf{H} and therefore accurate modelling of the photon detection process is imperative. An inaccurate projection operator may result in wrong assignment of the acquired data in the reconstructed images,

in the image space (typically $\mathbf{f}^{(0)}$ constant).

Let now consider the generic n -th step of an iterative methods.

- A projection step is applied to the current image estimate $\mathbf{f}^{(n)}$
- Then we obtain a set of projection values $\mathbf{p}^{(n)}$ that would be expected if $\mathbf{f}^{(n)}$ were the true image
- The predicted projections $\mathbf{p}^{(n)}$ are then compared with the actual measured data \mathbf{m} to create a set of projection-space error values $\mathbf{e}_p^{(n)}$.
- The error $\mathbf{e}_p^{(n)}$ are mapped back to the image space through a back-projection operation to produce image-space error values $\mathbf{e}_f^{(n)}$ that are used to update the image estimate, which becomes the new estimate $\mathbf{f}^{(n+1)}$.

This process just described is repeated again and again until the iteration stops automatically or is terminated by the user. Each of these repetitions is called an iteration. At the conclusion of the process, the current image estimate is considered to be the final solution. A schematic representation of the generic iterative algorithm is given in figure 2.19.

The differences between iterative reconstruction algorithms are due to the details of forward-projection, comparison, back-projection, and update steps. Note that direct reconstruction methods such as FBP use only the backprojection portion of the loop, so that there is no feedback about whether the image estimate, when projected, is consistent with the measured data. The power of iterative methods lies in the use of this feedback loop to refine the reconstructed image.

MLEM Maximum Likelihood Expectation Maximization (MLEM) [47, 28] is one of the most important algorithms, not so much because of its use in practice (it is computationally inefficient), but because it lays the foundations for faster algorithms such as OSEM (Order Subset Expectation Maximization).

The basic idea of MLEM is to maximize a likelihood function at each step in order to obtain the best estimation of the tracer distribution. Maximum likelihood estimators are advantageous because they offer unbiased, minimum variance estimates as the number of measurements increases towards infinity. This means that as the number of measurements or projections becomes large the expected value of the image estimate approaches the true image [161].

MLEM is based on the fact that the process of both radioactive disintegration and photon detection are described by the Poisson distribution [82]. Thus, the measured data \mathbf{m} in projection space (i.e. the sinogram) can be considered as a possible realization of a stochastic Poissonian process. According to that, using the system matrix in (2.27), given a tracer distribution \mathbf{f} we expect to obtain the following projected data p_i for each bin

$$p_i = \sum_{j=1}^N H_{ij} f_j + n_i \quad (2.28)$$

$$E(m_i) = p_i \quad (2.29)$$

where n_i models the noise contribution (e.g. scatter and random coincidences) to the i -th bin and where E is the expected value. As explained by (2.29), we are assuming that the

photon counts m_i measured in the i -th bin is a realization of a stochastic process with expected value p_i defined by the system matrix (2.28).

Therefore, assuming that each m_i is an independent identically distributed Poissonian variable, the probability to observe the measured data \mathbf{m} , given the unknown tracer distribution \mathbf{f} is given by

$$\mathcal{L}(\mathbf{f}|\mathbf{m}) = \prod_{i=1}^M \frac{p_i^{m_i} e^{-p_i}}{m_i!} \quad (2.30)$$

where the choice of indicating the probability with \mathcal{L} is due to the fact that (2.30) is precisely the likelihood function.

The maximum likelihood expectation method consists in finding the best tracer distribution \mathbf{f}^* for which the measured data would have had the greatest likelihood, so the problem can be mathematically formulated as

$$\mathbf{f}^* = \arg \max_{\mathbf{f}} \mathcal{L}(\mathbf{f}|\mathbf{m}) \quad (2.31)$$

Estimation of the ML solution suffers from increased variance (i.e. unstable solutions) especially for low-counts images. This often leads to the use of suitable regularization procedures, which restrict the acceptable solutions, ideally by exploiting prior knowledge of the unknown tracer distribution \mathbf{f} . This prior knowledge can be either based on anatomical information acquired from other modalities (e.g. CT or MRI) [95, 65] and the regularization can be practically achieved by adding a term to the objective function $\mathcal{L}(\mathbf{m}|\mathbf{f})$.

Estimation of the maximum likelihood (ML) solution is achieved by solving the (2.31). In particular, MLEM is based on finding the best solution \mathbf{f}^* through K iterations (steps): for each step the objective function (i.e. the likelihood) increases until a fixed point is reached. More in detail, the MLEM solution of (2.30) (neglecting noise n_i) gives the iterative equation

$$f_j^{(k+1)} = \frac{f_j^{(k)}}{\sum_i H_{ij}} \sum_i H_{ij} \left(\frac{p_i}{\sum_{j'} H_{ij'} f_{j'}^{(k)}} \right) \quad (2.32)$$

where $f_j^{(k)}$ is value of the j -th pixel of the image at the k -th iteration. The EM algorithm converges monotonically to the solution \mathbf{f}^* , thus at each iteration the updated image will increase the value of the likelihood function.

OSEM The MLEM algorithm is a simple algorithm, with attractive properties, but unfortunately it has a very slow convergence rate. Since MLEM requires a projection and a rear projection at each iteration, the total processing time is considerably greater than that required by any analytical algorithm. To improve on this, current standards employ the ordered-subset version (OSEM) of the MLEM. The OSEM algorithm has become a major workhorse in today's scanner [167].

In MLEM the comparison is made between estimated and measured projections each iteration: this requires considerable computation as the projections have to be calculated at typically 64 or 128 angles, calculation of each projection taking at least as long as a complete FBP reconstruction.

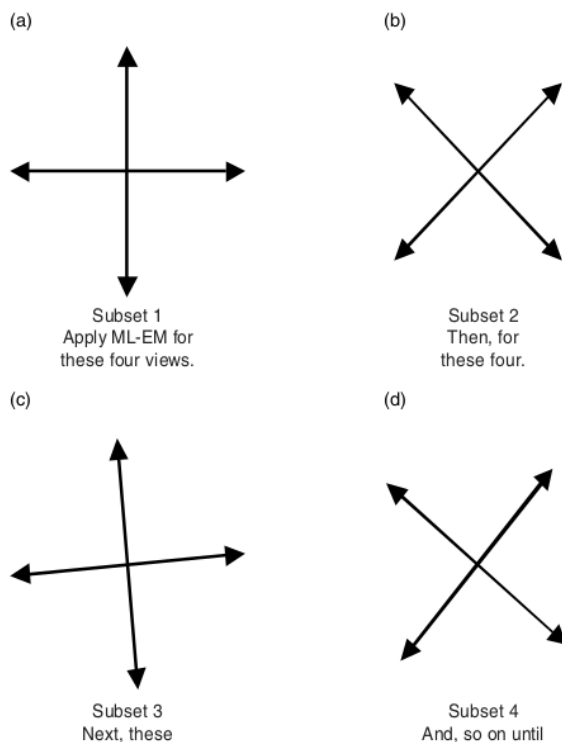


Figure 2.20: The process of using subset updates in a reconstruction algorithm. (a) Subset 1: Apply ML-EM for these four views. (b) Subset 2: Then apply the algorithm for these four. (c) Subset 3: Then apply the algorithm for these four. (d) Subset 4: Continue until all data are used.

OSEM method is a slight modification of MLEM method: the measured data are divided into several disjoint partitions, called *subsets*, so that each partition contains an equal integer number of projection angles (see figure 2.20).

Using OSEM the comparison and update steps are based on only a small number of the projections each iteration, progressively using other projections in each further iteration.

For this reason, OSEM ensure a rapid convergence of the algorithm [82].

Both MLEM and OSEM algorithms produce high variance, especially for large iteration numbers.

The best absolute choice of subsets and iterations number does not exist, as they are context and problem dependent. However, we can make some general considerations.

The image quality initially improves as the number of iterations increases. However, if the iterations continue after a certain point, the resulting images show high variance. [10, 130]. This variance, which is demonstrated as salt-and-pepper noise in

the reconstructed images, can be alleviated either by premature termination of the reconstruction process (lower iteration) or by post-smoothing the images using a Gaussian kernel [12]. An example Post-smoothing reconstruction is provided in figure 2.21.

With regard to the number of subsets, increasing the number of subsets resulted in increased noise as well as subtle shape artifacts in these images, especially for the highest numbers of subsets [139, 106].

An example of image reconstruction using different number of iterations and subsets is provided in figure 2.21.

2.5.4 3D reconstruction methods: an overview

Reconstruction of images from 3-D data is complicated by a very huge volume of data. However, both iterative and analytical methods discussed in the previous sections have their 3-D counterpart [160].

The filtered backprojection can be applied to 3-D image reconstruction with some manipulations which basically consist in considering the 3-D data sinograms as consisting of a set of 2-D parallel projections, and the FBP is applied to these projections.

The iteration methods can also be generally applied to the 3-D data. However, the complexity, large volume, and incomplete sampling of the data due to the finite axial length of the scanner are some of the factors that limit the use of the FBP and iterative methods directly in the fully 3-D reconstruction.

To circumvent these difficulties, a modified method of handling 3-D data is commonly

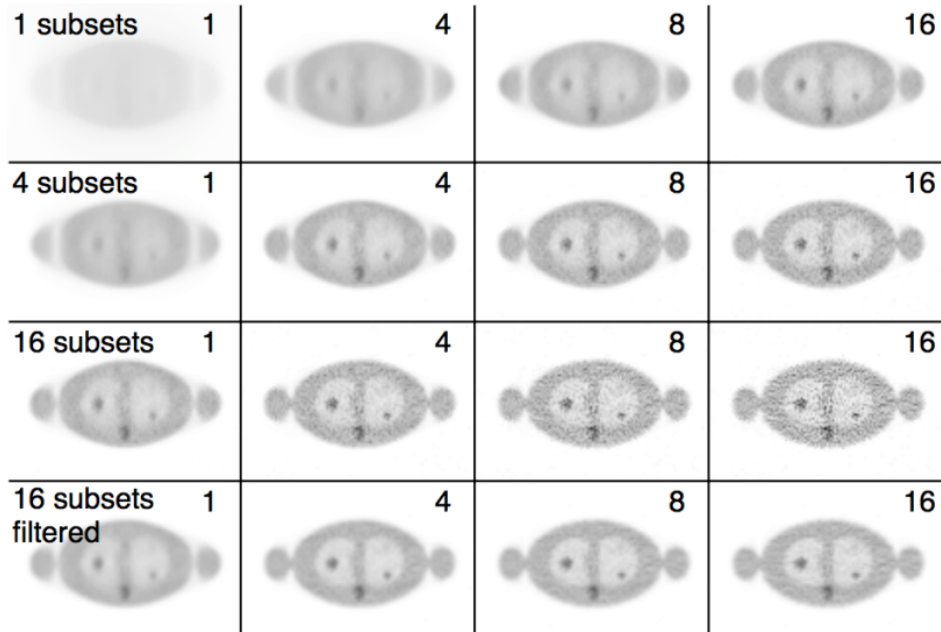


Figure 2.21: From left to right, image reconstruction using 1,4,8,16 iterations. From top to bottom, image reconstruction using 1,4,16 subsets. The bottom images is a 16 subsets gaussian post filtered reconstruction. We notice that increasing subsets allows the algorithm to converge more quickly.

used, which is described below.

A method of 3-D reconstruction involves the rebinning of the 3-D acquisition data into a set of 2-D equivalent projections. Rebinning is achieved by assigning axially tilted LORs to transaxial planes intersecting them at their axial midpoints. This is equivalent to collecting data in a multiring scanner in 2-D mode, and is called the single-slice rebinning algorithm (SSRB). This method works well along the central axis of the scanner, but steadily becomes worse with increasing radial distance. In another method, called the Fourier rebinning (FORE) algorithm, rebinning is performed by applying the 2-D Fourier method to each oblique sinogram in the frequency domain. This method is more accurate than the SSRB method because of the more accurate estimate of the source axial location. After rebinning of 3-D data into 2-D data sets, the FBP or iterative method is applied.

2.6 Resolution modeling

Resolution can be loosely defined as the level of reproduction of spatial detail in the imaging system. Many factors discussed in the section 2.4.3 contribute to the effective spatial resolution of a PET scanner: detectors size, positron range, non-collinearity, radial elongation and image reconstruction method are recognized as factors which affect PET spatial resolution. Furthermore, patients motion is to be considered as a degrading factor for spatial resolution. Depending on the PET scanner properties, on the reconstruction method and on the location within the ring, the PET spatial resolution typically goes from 2.5 mm to 5 mm [107, 87, 44, 68].

The effective spatial resolution is often characterized through the point spread function (PSF). Imagine a single point source is placed in the system and reconstructed into an image: the reconstructed image is a spread out version of the point. In PET, the PSF is frequently characterized by the transaxial and axial Full Width at Half Maximum

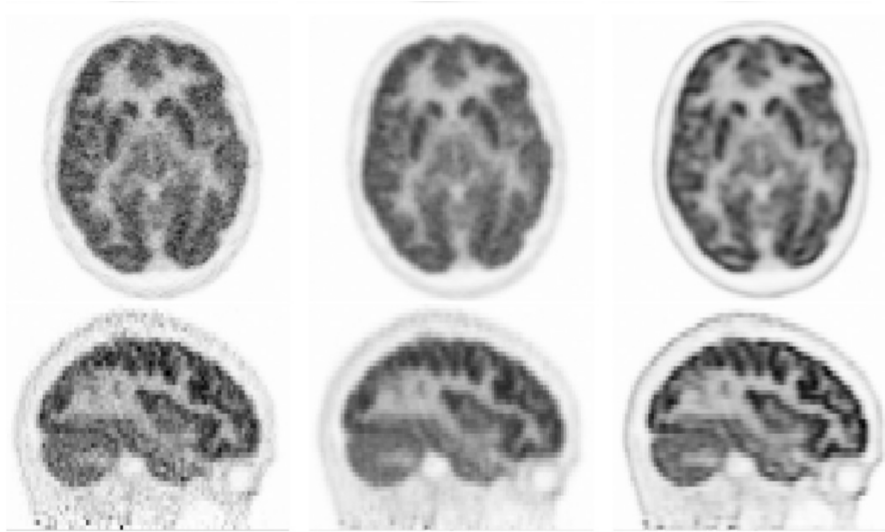


Figure 2.23: Transaxial (top, middle) and sagittal (bottom) images reconstructed with no PSF and no post-processing filter (left), no PSF post-filtered with a FWHM = 3 mm Gaussian filter (center) and PSF (right). Courtesy of [126]

(FWHM) term, which describes the width at half of the maximum value of the PSF of a Gaussian shaped function fitted to the PSF (in the transaxial or axial direction).

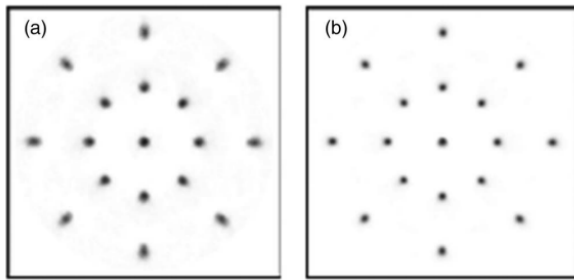


Figure 2.22: Transverse sections of point source images reconstructed without (a) and with (b) resolution modeling. Courtesy of [124]

However, the real PSF is not actually purely Gaussian: Gaussian PSF do not model the presence of varying inter-crystal blurring that is known to lead to the parallax effect, the degradation of resolution as one moves away from the center of the field of view. [124]. Generalizations of Gaussian PSF can be achieved by used space-varying blurring kernels (which can be anisotropic and asymmetric) in order to to better model the PSF.

Resolution model, which consists in characterizing the PSF, can be incorporate within the iterative reconstruction algorithm in the system matrix. As including PSF in reconstruction algorithm provides improvements in contrast recovery, resolution and quantification, particularly for small regions, the PSF modeling issue has received increased attention over the recent years [124]. Several methods have been proposed to estimate the PSF, which can be generally divided into three categories: Monte Carlo simulations [52], analytical modelization [123, 144] and experimental measurements [9, 115].

Including PSF model in reconstruction step contributes to recover the symmetry of the profiles, to obtain a more uniform resolution across the reconstruction FOV and to recover similar peak heights across the FOV [126].

Even though resolution modeling methods are very commonly seen to lead to improved resolution, at the same time it is important to notice that resolution modeling can significantly modify the image noise structure [124].

2.7 DICOM files and NIfTI images

In this chapter we discussed how PET images are reconstructed from sinograms. Reconstructed images are typically saved in a digital format as DICOM (Digital Imaging and Communications in Medicine) files [24], generally with one DICOM file per transaxial image section (also called slice).

In this section I will briefly describe the content of a DICOM file, then I will show how to obtain the final 3-D tracer distribution patient's image in a useful format.

A DICOM file consists of a header and an image data sets, all packed into a single file. The image data sets consists of voxel intensities written in the form of 2-byte integers [116]. The header contains metadata which provide important information such as matrix dimensions, slice thickness, spacing between slices, and the modality used to create the DICOM file.

Spacing between slices is a parameter (expressed in mm) that gives the distance between two adjacent slices.

The matrix dimension is the n -by- n grid used to reconstruct each sinogram from the projection space to the image space.

The slice thickness is a parameter that can be selected and represent the thickness of each slice in mm. Increasing the slice thickness leads to decreasing resolution, as slice thickness refers to the axial resolution of the scan.

Ideally a complete DICOM file should contain detailed information about the acquisition process, such as 2D or 3D acquisition, acquisition type (e.g. iterative or analytical), acquisition details (e.g. number subsets and iterations) and details about the scanner (manufacturer, model, detector properties etc...).

However, only some of this information is reported in practice: sometimes we deal with more detailed files, sometimes less.

Furthermore DICOM files usually contain patient's demographic information, such as age and sex, and the PET examination date.

Typically the 3D images are created by converting the DICOM files into the NIfTI (Neuroimaging Informatics Technology Initiative) format [55]. NIfTI images obtained from DICOM files provide 3D representations of the brain are called *raw images*.

This representation is actually a 3D n -by- n -by- p matrix whose elements are called *voxels* and whose dimensions n and p are respectively the matrix dimension and the number of slice.

This 3D n -by- n -by- p grid can be considered as a discrete space in which the raw image is embedded, and this called the *native space*.

The conversion from DICOM files to NIfTI images is achieved using dedicated software, such as *dcm2nii*².

3D NIfTI images can be visualized using appropriate tools (such as *MRICron*³), that

²<https://people.cas.sc.edu/rorden/mricron/dcm2nii.html>

³<https://www.nitrc.org/projects/mricron>

allows to visualize different 3D image's section image by varying the coordinates.

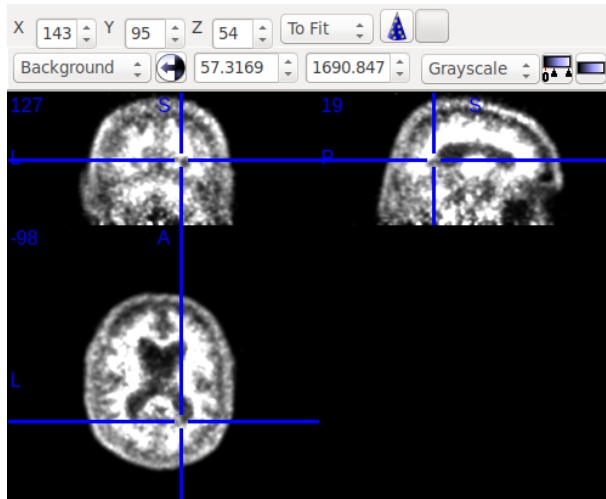


Figure 2.24: Example of a raw PET image in NIfTI format visualized using MRICron. Sagittal (top left), coronal (top right) and axial (bottom) sections are showed for a given triplet of matrix coordinate.

An example of a raw PET image visualized with MRICron can be found in figure 2.24.

2.8 Performance characteristics of PET scanners and image quality measures

A major goal of PET images acquisition is to obtain a good quality and detailed image of an object by the PET scanner, and so it depends on how well the scanner performs in image formation. The definition of image quality in nuclear medicine (and hence that of "best" image) is still rather open. An objective comparison of image quality is often difficult and can only be performed in the context of a specific application, or task.

Phantom studies Before going on, it is important to point out the following fact:

almost all the techniques currently used in literature to asses images quality and scanners performance are based on phantom studies. Phantoms are objects built for the purpose of calibrating and assessing performance of PET scanners and acquisition protocols. Phantoms have a known geometry and they typically consist of *hot regions* and *cold regions*: the former are usually of different sizes and filled with a precise amount of radiotracer, while the latter is filled of non-radioactive fluid (e.g. water) to simulating human tissue not involved in radiotracer uptake. An illustration of a typical phantom is provided in figure 2.25.

Since the phantom geometry is always known, the exact radiotracer distribution is available. Therefore, estimated and actual tracer distributions can be compared. In this context, one of the main criteria for assessing image quality relies on the ability to best distinguish between hot and cold regions. Phantoms must comply with the National Electrical Manufacturers Association (NEMA) standards[46].

Image quality measures In the context of phantom studies, image quality can be assessed via many measures extracted from the acquired phantom images, which I briefly describe below.

Measures usually used in assessing image quality are based on ratio between true coincidences and spurious coincidences as well as on signal to noise ratio (SNR) and contrast detection (CNR).

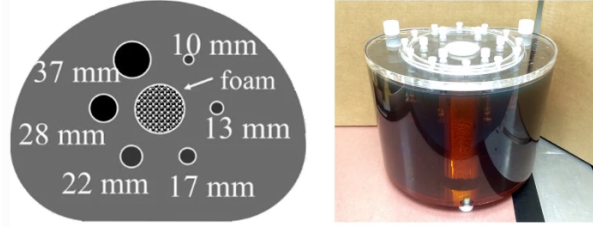


Figure 2.25: A typical phantom

One of the most used measure to estimate ratio between true coincidences and spurious coincidences is the noise equivalent count rate (NECR) which essentially describes the relative numbers of trues, scatters and randoms coincidences. NECR is defined as follows [132]

$$NECR = \frac{T^2}{T + R + S} \quad (2.33)$$

where T , R , and S are the true, random, and scatter coincidence count rates, respectively. We remind that scatter and random events can be measured as explained in section 2.4.3. The true events T are determined by subtracting scatter S and random R events from all the prompt events. From the knowledge of T , R and S , the NECR is calculated by (2.33). Image noise can be minimized by maximizing NECR.

The CNR measures are based on comparing intensity values in phantom hot regions against the background (i.e. phantom cold regions). Well contrasted images have the ability to better detect the difference between hot and cold regions. SNR measurements (extracted from images) are based on estimation of intensity and variance of uniform regions (hot or cold).

SNR and CNR can be defined in different ways in the imaging framework, however the most used definition are the following

$$SNR = \frac{\mu_B}{\sigma_B} \quad (2.34)$$

$$CNR = \frac{\mu_H - \mu_B}{\sigma_B} \quad (2.35)$$

where μ is the mean, σ is the variance, and the subscripts B and H are referred to the background (i.e. cold region) and to the hot region respectively.

Many studies compare image quality of different PET scanner and/or different image reconstruction methods acquiring phantom images and assessing quality using NECR measure as well as SNR and CNR measures [81, 111, 46, 168, 129].

2.9 Final considerations

In this chapter we have shown how PET scan acquisition and reconstruction is a complex issue which involves many factors: some of them can be directly controlled (e.g. the choice of a reconstruction method and the setting of its parameters (sections 2.5.3 and 2.5.2), the setting of time and energy windows in counts detection), while others that you can't control directly (e.g. factors that come into play in the acquisition phase discussed in section 2.4.3), are typically taken into account in a more or less accurate way.

We have also discuss technologies that considerably improve the quality of the reconstructed images, such as the use of TOF in the detection of annihilation and PSF in image reconstruction. TOF and PSF often play a predominant role in image quality

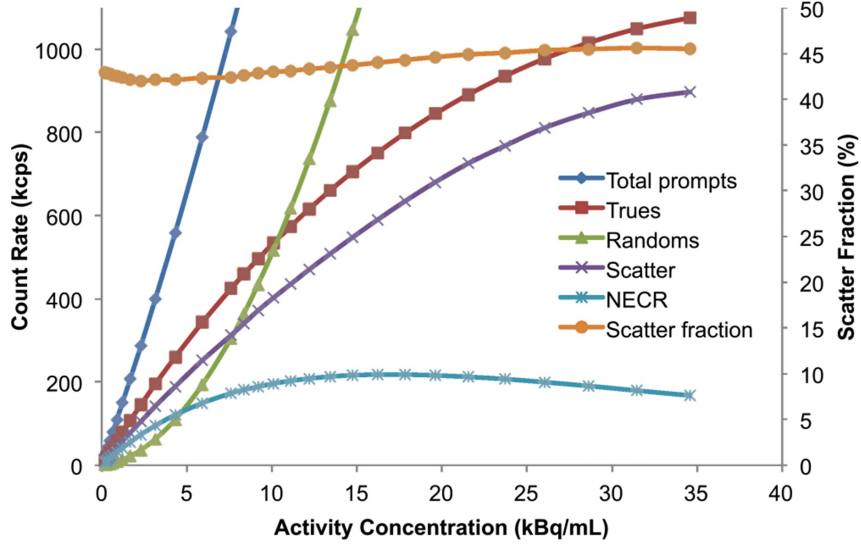


Figure 2.26: Graphic which shows the typical trend of true, scatter and random coincidences as well as of NECR and Scatter fraction. Scatter fraction represent the ratio between scatter and true coincidences (courtesy of [71]).

determination [6, 11, 39, 7, 22].

The great variability in the steps from PET scan to reconstructed ready-to-study image is a source of great variability about perceived image quality.

In the framework of quantitative imaging (see section 2.2), image quality plays a role which can not be ignored by researchers, as image quality could influence the results of the analysis performed on PET images.

We have also shown that image quality is currently assessed using ad-hoc phantom studies, as briefly discussed in section 2.8.

The relation between image quality and quantitative imaging analysis is the main issue of the PET harmonization.

PET data harmonization is the main issue of my thesis work and it will be detailed discusses in chapters 5 and 6.

Chapter 3

Methods

3.1 Machine Learning overview

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. The goal of ML is learn from data in order to finds generalizable predictive patterns: a well-constructed ML algorithm adapts well to the data from which it has learned, but at the same time has the ability to make predictions as correct as possible on new data.

The ML algorithms can be roughly categorized as either supervised or unsupervised. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures [18]. Examples of supervised learning algorithms are given by regression and classification problems, while examples of unsupervised learning are clustering and principal component analysis.

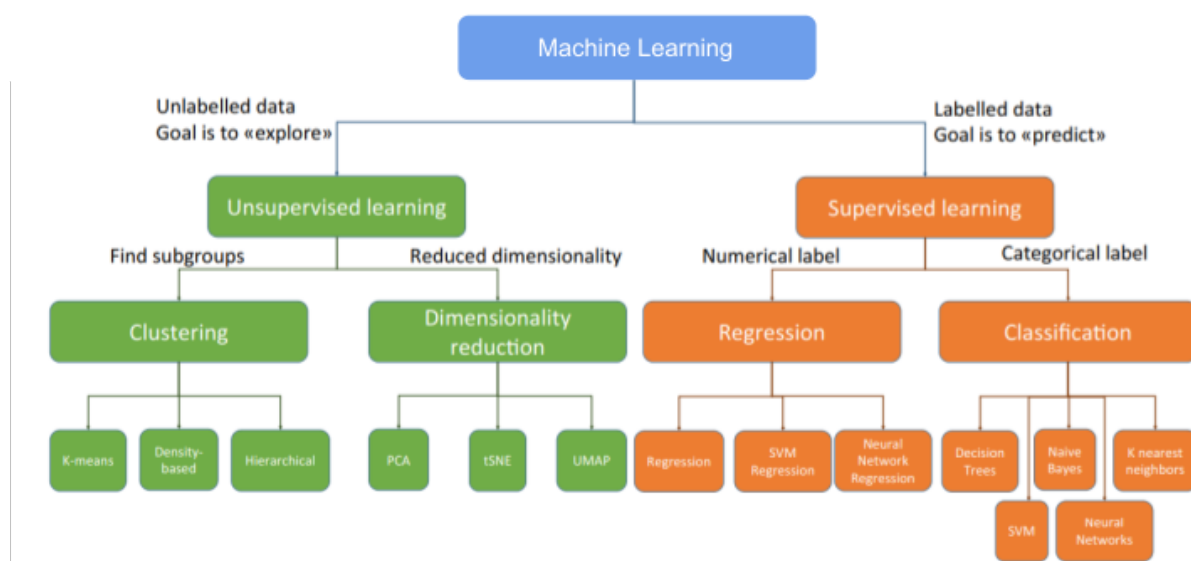


Figure 3.1: Taxonomy and overview of most used machine learning algorithms. Courtesy of [18].

3.1.1 Supervised Learning

Here we give a general review of how a supervised learning algorithm works.

Suppose we have a dataset D made of ordered couples of data $D = X \times Y = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where $X = \{x_1, \dots, x_N\}$ is the input set and where $Y = \{y_1, \dots, y_N\}$ the output set. The input set X belong to an input space \bar{X} , which typically is \mathbb{R}^k . The output set Y belong to an output space \bar{Y} . The output space usually can be

- a subset of \mathbb{R} . This represents a regression problem
- a finite set of element $\{l_1, \dots, l_p\}$ where the l_i are called labels. This represents a classification problem

Furthermore, we have to postulate the existence of a probability model for the data. The model should take into account the possible uncertainty in the task and in the data. We assume that exist a fixed unknown distribution $p(x, y)$ according to which the data are identically and independently sampled. The distribution $p(x, y)$ models different sources of uncertainty and it can be factorized as $p(x, y) = p_X(x)p(y|x)$; where the conditioned probability $p(y|x)$ can be seen as a form of noise in the output. For each input x there is a distribution of possible outputs $p(y|x)$, while the marginal distribution $p_X(x)$ models uncertainty in the sampling of the input points

Assuming a $p(y|x)$ distribution lead us to suppose the existence of a unknown *target function* f^* which relates the whole input and output space

$$f^* : \bar{X} \longrightarrow \bar{Y} \quad (3.1)$$

In regression problems $Y \subset \bar{Y}$, while in classification problems $Y = \bar{Y}$. Furthermore in regression problem we typically suppose the existence of a non-deterministic relation described as follows

$$y = f^*(x) + \epsilon \quad (3.2)$$

where $y \in \bar{Y}$ and $x \in \bar{X}$ and where ϵ is a noise term which represents the non-deterministic part of the relation. We suppose $\epsilon \sim \mathcal{N}(0, \sigma)$.

The goal of supervised ML is to obtain the best estimate of the function f^* using the elements of D , i.e. the input-output couples $(x_i, y_i) \in D$.

We will denote the estimator of $f^*(x)$ by $f_D(x)$. If there are no ambiguities, we will call $f_D(x)$ simply f .

We point out that we are looking for to a general relation which holds for all the possible new incoming data. As new data are not in D , the we are looking for must not be the best estimate relative to our dataset D because such a f_D would not lead to generalize for new data. Thus, the main point is the following: the f_D we are looking for must be a compromise between adapting well to data in D and at the same time predicting the output of new data as well as possible.

At this point we need to formalize the concept of best estimate of f^* . We need to fix a loss function. Let consider a couple $(x, y) \in D$, we define the loss function L as a point-wise error measure

$$L : Y \times \bar{Y} \ni (y, f_D(x)) \longrightarrow L(y, f_D(x)) \in [0, +\infty) \quad (3.3)$$

$$(3.4)$$

The loss function represents the cost of predicting $f_D(x)$ instead of the correct output y . The loss function for regression problem is usually the quadratic loss function. Considering a classification problem a common choice for the loss function is the zero-one loss function, where all misclassifications are charged a single unit. Many other choices are possible for L , both in classification and in regression problems.

We introduce now the expected loss (also called cost function):

$$\mathcal{E}(f_D) := E_{XY}[L(y, f_D(x))] = \int p(x, y) L(y, f_D(x)) dx dy \quad (3.5)$$

The $\mathcal{E}(f)$ can not be computed as $p(x, y)$ is unknown. However, given n data, we can estimate $\mathcal{E}(f)$ using an estimator. The best estimator for the expected value is the arithmetical mean. Thus the estimator $\hat{\mathcal{E}}$ of \mathcal{E} is given by

$$\hat{\mathcal{E}}(f_D) = \frac{1}{n} \sum_{i=1}^n L(y_i, f_D(x_i)) \quad (3.6)$$

where (x_i, y_i) are couples of D . If L is quadratic, the estimated cost function given by equation (3.6) became the mean square error

$$\hat{\mathcal{E}} = \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - f_D(x_i))^2 \quad (3.7)$$

An important point in ML is estimating the ability prediction of an algorithm. But if we minimize $\hat{\mathcal{E}}(f)$ using all the dataset D , the generalization ability can not be estimated, as we have no new data to predict in order to test the model.

To overcome this problem, ML provide the following solution: the whole dataset D is divided into two disjoint subsets: a training set \mathcal{T} and a test set \mathcal{S} .

We train the algorithm on the training set and we test the ability to generalize on the test set.

What does train an algorithm mean? Training a given algorithm means optimize the free parameters of the problem over the training set, i.e. finding the parameters which minimize the estimated expected loss given by equation (3.6) evaluated for data $\in \mathcal{T}$ set only .

This allows to define the training error and the test error. The training error is the average error that results from using a ML method to predict the response on the training set. Training error can be computed using (3.6) for the data in \mathcal{T} .

The test error is the average error that results from using a ML method to predict the response on new observations, i.e. observations not used in training the algorithm.

We are interested in test error estimation, as it provides an estimate of the prediction's ability of an algorithm.

As we just defined training and test sets, we will now introduce the concepts of overfitting, underfitting, capacity of a model, regularization and hyperparameters.

Overfitting and Underfitting If an algorithm works well on the training set but fails to generalize, we say it is overfitting. An overfitted model contains more parameters than can be justified by the data. The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e. the noise) as if that variation represented underlying model structure.

In other words, the model remembers a huge number of examples instead of learning to notice features.

The counterpart of overfitting is the underfitting. Underfitting occurs when a statistical model cannot adequately capture the underlying structure of the data. An under-fitted model is a model where some parameters or terms that would appear in a correctly specified model are missing. Under-fitting would occur, for example, when fitting a linear model to non-linear data. Such a model will tend to have poor predictive performance both on the dataset and on new incoming data.

Model capacity We can control whether a model is more likely to overfit or underfit by altering its capacity. Informally, a model's capacity is its ability to fit a wide variety of functions [40]. Models with low capacity may struggle to fit the data of the dataset. Models with high capacity can overfit by memorizing properties of the training set that do not serve them well on the test set.

One way to control the capacity of a learning algorithm is by choosing its hypothesis space, i.e. the set of functions that the learning algorithm is allowed to select as being the solution.

As an example, let consider the following polynomial regression

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_N x^N \quad (3.8)$$

The model capacity is controlled through the polynomial degree N . Increasing N the model increase its capacity, as the number of parameters (also called degrees of freedom) increases, leading to a lower training error. Adding too many degrees of freedom leads to overfitting. The extreme overfitting scenario occurs when the degrees of freedom N is greater or equal than the number of data: data are exactly fitted and the training error goes to zero. Obviously the generalization error will be very large. On the contrary, using a polynomial of too low a degree with respect to the complexity of the problem will lead to underfitting: for example, fitting an intrinsically quadratic relation with a polynomial of degree 1 will certainly lead to underfitting.

To summarize, ML algorithms will generally perform best when their capacity is appropriate for the true complexity of the problem and the amount of available training data. Models with insufficient capacity are unable to solve complex tasks (underfitting). Models with high capacity can solve complex tasks, but when their capacity is higher than needed to solve the present task, they may overfit. An illustration of overfitting, underfitting and model capacity can be found in figure 3.2.

Regularization Finding the optimal capacity of a model is not only matter of selection of a suitable set of functions: adding or removing functions from the hypothesis space of solutions is not the only way to control capacity.

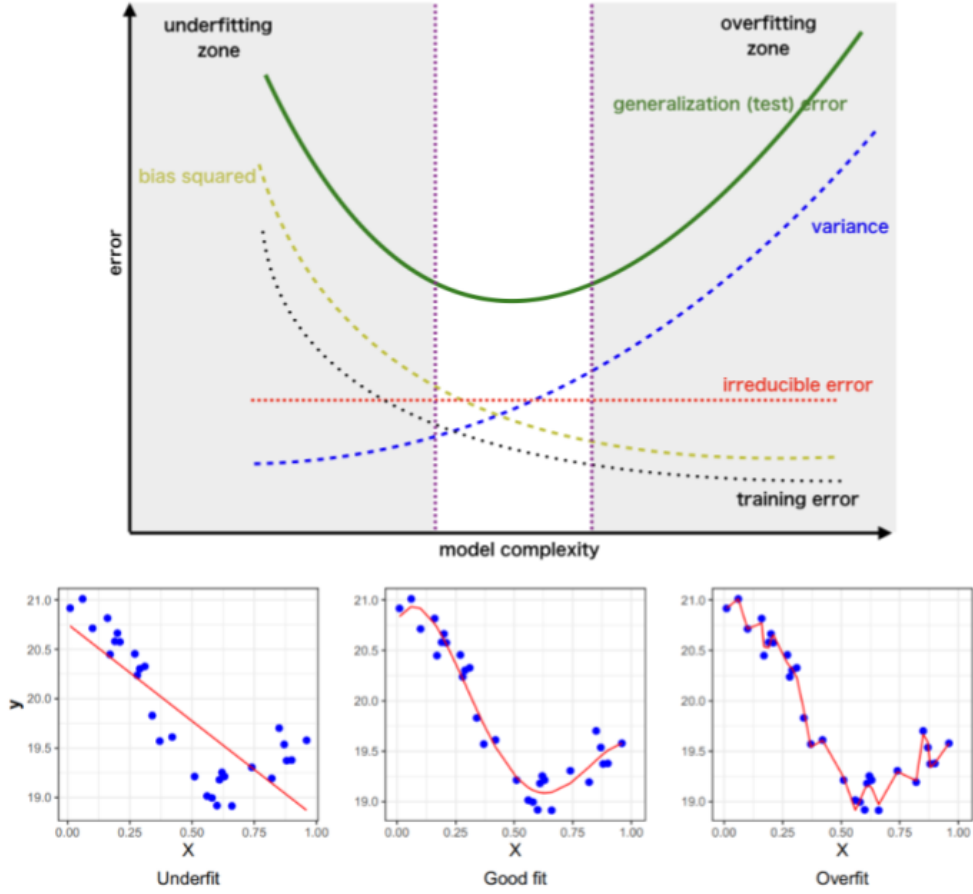


Figure 3.2: Top figure: typical trends of training error and generalization error versus model complexity. Bias and variance typical trends are furthermore provided. The left/right gray regions are the underfitting/overfitting regions, while the white one represent the region of good choices for model complexity. Bottom figure: Bidimensional representation of underfitting, good fitting and overfitting.

The behavior of our algorithm is strongly affected not just by how large we make the set of functions allowed in its hypothesis space, but also by the specific identity of those functions.

Fixed a set of functions, an algorithm can be trained to prefer a solution over another within the functions set. This means that both functions are eligible, but one is preferred. The unpreferred solution will be chosen only if it fits the training data significantly better than the preferred solution. How can we formalize this idea? This can be done by adding a penalty $\Omega(\mathbf{w})$, called regularizer, to the cost function. The penalty weight is controlled by a positive real number λ . Thus the equation to minimize became

$$J(f) = \hat{\mathcal{E}}(f) + \lambda\Omega(f) \quad (3.9)$$

$\Omega(f)$ is typically chosen to impose a penalty on the complexity of f . Concrete notions of complexity used include restrictions for smoothness (i.e. non rapidly oscillating solutions are preferred) and bounds on the vector space norm. Typical choices for Ω which penalize high vector norm are the L_2 and L_1 regularization, which we briefly describe in the next section.

The role of λ is crucial in underfitting/overfitting trade-off. Minimizing $J(f)$ results in a choice of weights that make a trade-off between fitting the training data and being

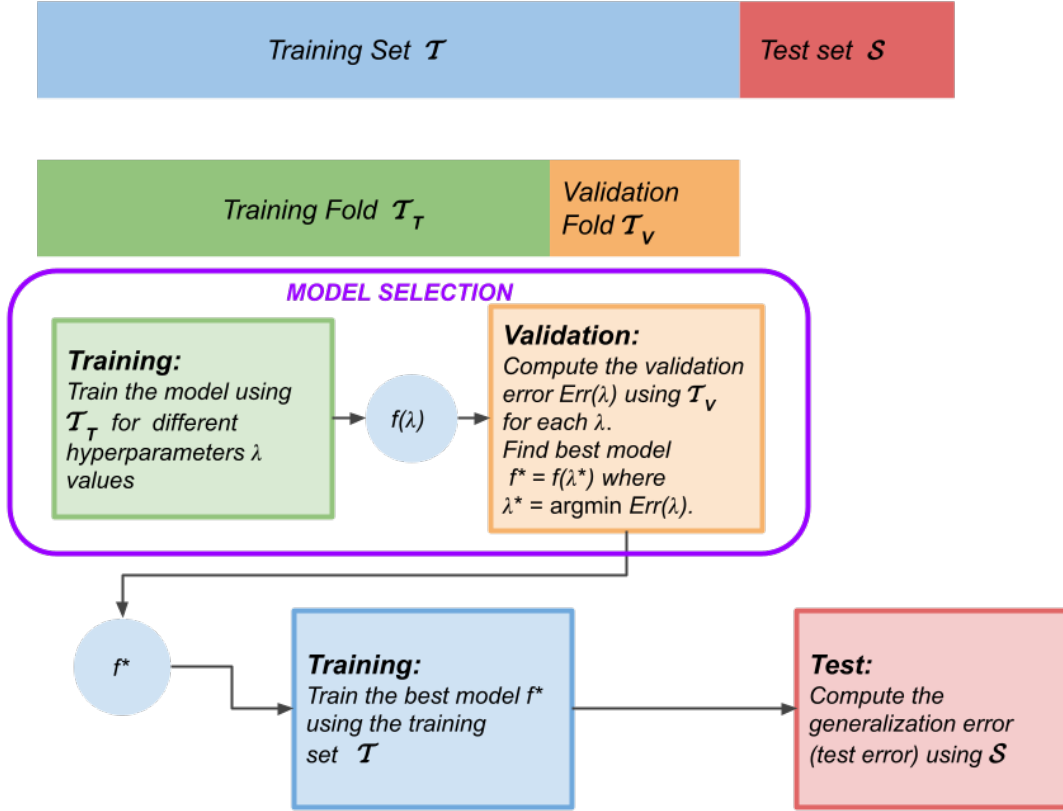


Figure 3.3: Workflow of a supervised ML algorithm. Common choices for data partitions are 50-60 % for training set and 20-30 % for validation and test set.

small. When λ is small the algorithm will prefer solutions that fit the data well (i.e. which minimize $\hat{\mathcal{E}}(f)$), when λ is large, on the contrary, solutions that minimize the regularizer will be preferred.

In general, a regularization method can be defined as any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error. Regularization is one of the central concerns of the field of ML, rivaled in its importance only by optimization[40].

Underfitting and overfitting can be considered in terms of bias and variance of an estimator. In general, the mean square error expected value of a given estimator can be decomposed into a bias term and a variance term [53].

Let consider the relation (3.2). The function f_D is an estimator of the target function f^* , thus its expected mean squared error can be decomposed in a bias term, a variance term, and an irreducible error term (i.e. the variance of the noise ϵ).

The relationship between bias and variance is tightly linked to the ML concepts of capacity, underfitting and overfitting. When generalization error is measured by the MSE increasing capacity tends to increase variance and decrease bias.

Thus, the overfitting underfitting trade-off is commonly called *bias-variance tradeoff*, where bias refers to underfitting and variance refers to overfitting, as shown in figure 3.2.

Hyperparameter and validation set An hyperparameter is a parameter that is set before the learning process begins. This means that hyperparameters are not involved in the minimization process as variable to minimize.

These parameters are tunable a-priori and can directly affect how well a model trains. Hyperparameters directly drive the underfitting/overfitting trade-off, for example, λ in equation (3.9) is an hyperparameter. As hyperparameters are fixed a priori, how can we find their best values, i.e. the values which optimize the underfitting/overfitting trade-off?

Using training set to tune hyperparameters is not possible, because we want to choose hyperparameters values that will best generalize and thus we need to test various hyperparameters choices using data that was not used to train the algorithm (exactly as for the test set). However, we cannot even tune hyperparameters using the test set, because that would be cheating. We are only allowed to use the test set once, to report the final performance. If we peek at the test data by using it to tune hyperparameters, it will no longer give a realistic estimate of generalization performance, as the generalization error will be typically underestimated.

The typical solution to this problem relies on the definition of a validation set. Let consider the training set \mathcal{T} . We divide the training set in two disjoint subsets, \mathcal{T}_T and \mathcal{T}_V , as illustrated in figure 3.3. The former is used to train the algorithm (i.e. optimize parameters for any given hyperparameters choice) and so it is used as an actual training set. We will call \mathcal{T}_T the training fold to distinguish it from the training set \mathcal{T}_T .

The subset \mathcal{T}_V is used to estimate the generalization error after training for any specific hyperparameters, and so it is used as a test set for hyperparameters optimization.

The \mathcal{T}_T subset is still typically called the training set, even though this may be confused with the larger pool of data used for the entire training process. The subset \mathcal{T}_V used to guide the selection of hyperparameters is called the validation set.

Typically, one uses about 70-80 percent of the training data for training and 20-30 percent for validation, but an absolute rule does not exist.

After hyperparameters optimization is complete, the generalization error may be estimated using the test set.

The general unsupervised algorithm is summarized in figure 3.3; however the approach discussed here requires a data-rich situation which is often not the case when dealing with real problems.

3.1.2 Supervised Learning and small size datasets

Assessment of generalization performance of a ML model is extremely important in practice, since it guides the choice of learning method or model, and gives us a measure of the prediction ability of the ultimately chosen model.

Choosing the best model and estimating its generalization error could be very complicated issues when datasets size are not large enough.

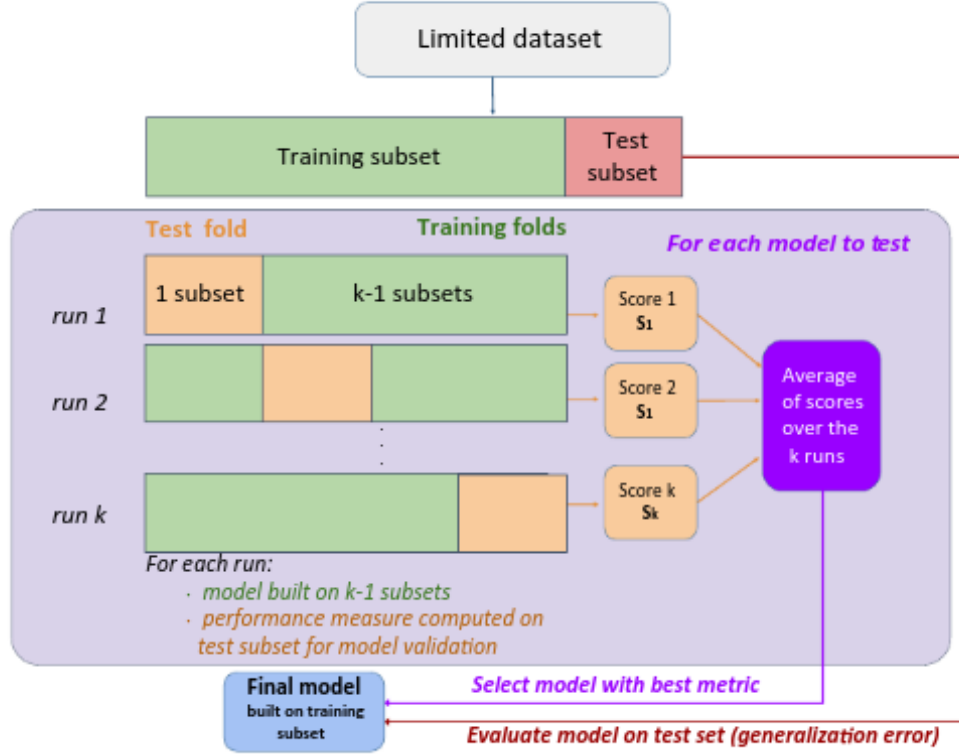


Figure 3.4: Supervised ML flowchart for limited datasets. The iterative process in gray area represents a k -fold validation procedure. Once the model have been optimized, the generalization error is evaluated on the test set. Courtesy of [18]

In this section we discuss both this issues in the framework of relatively small size datasets.

If we are in a data-rich situation, the best approach for both problems is the one discussed in the previous section and illustrated in figure 3.3.

Before continuing let fix some notation.

The main point is the same of the previous section: we want to estimate the function f^* which relate an input set X to an output set Y . As discussed in the previous section, typically models have hyperparameters which control the models complexity and which are needed to be tuned a-priori. Thus we denote the estimated f by $f_\lambda(\mathbf{x})$, where λ represents hyperparameters model dependencies. Having said this, for brevity we will often suppress the dependence of $f(\mathbf{x})$ on λ .

Model selection and cross-validation If we are not in a data-rich situation, using one validation set to optimize the algorithm could lead to a bad hyperparameter optimization.

Small size data increases the probability of sampling a validation set that is not representative of the population.

In a such case, the hyperparameters optimization will not holds in general, as it would be related to a not representative validation set.

One might be tempted to choose a large percentage validation partition. However

this choice leads to a small training set and thus to inadequately trained models: this is a bad solution too.

Cross validation allows to circumvent this problem. Even though many cross-validation techniques exist, all the cross-validation techniques are based on a common idea, which I will briefly explain below.

Cross validation methods are iterative processes based on many validation steps. described in previous section. and it is illustrated in figure 3.4.

Each iteration is actually equivalent to a validation step (see figure 3.3), but the validation and training set will change during each iteration. Cross-validation methods can be resumed as follows

- the training set is splitted in k disjoint set, i.e. $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_k\}$, called training folds
- the splitting leads to k validation steps: let consider the generic i -th step. The i -th step is described as follows
 - the i -th subset is treated as the validation set, while the model is trained using the $\mathcal{T} \setminus \mathcal{T}_i$ subsets, obtaining the i -th estimated function f_{λ}^i . As occurs in validation, the algorithm is trained for different value of λ by minimizing a cost function of the type (3.9).
 - the i -th generalization error is computed using \mathcal{T}_i . Thus we have

$$\hat{\mathcal{E}}_i(f^i, \lambda) = \frac{1}{n} \sum_{(\mathbf{x}, y) \in \mathcal{T}_i} L(f_{\lambda}^i(\mathbf{x}), y) \quad (3.10)$$

where L is a given loss function

- once all the $\hat{\mathcal{E}}_i(f_{\lambda}^i, \lambda)$ are computed, we get the final cross-validation error defined as the average of all the k steps validation errors, i.e.

$$\text{Err}_{CV}(\lambda) = \frac{1}{k} \sum_{j=1}^k \hat{\mathcal{E}}_j(f_{\lambda}^j, \lambda) \quad (3.11)$$

Minimizing the error (3.11) we obtain the best hyperparameter $\lambda^* = \arg \max_{\lambda} \text{Err}_{CV}(\lambda)$.

Once hyperameters have been setting via cross-validation, the algorithm is trained on the whole training set \mathcal{T} using the optimized hyperparameters, and the generalization error is computed on the set set to obtain an estimation of the algorithm performance. The generical supervised ML method in the framework of limited dataset is summarized in figure 3.4.

However, this is a point which presents a criticality: we have said that cross-validation is necessary because in order to circumvent . However, as shown in figure 3.3, we have to consider that the numerosity of the test set is typically similar to that of the validation set. Therefore, we conclude that the test set suffers of the same problems of the validation set: using only one test set may leads to a generalisation error which is not representative of the quality of the model, just as using only one validation set can lead to a bad optimisation of the hyperparameters.

Generalization error issues The final step of a supervised ML algorithm is to estimate the generalization error, as the generalization algorithm ability is what we are really interested in.

Given a limited dataset, two approaches are possible for estimating generalization error:

- we can use an a-priori fixed test set \mathcal{S} , as discussed in previous section and as illustrated in 3.4. However, using only one randomly chosen test set to compute generalization error could lead to the same criticality of using only one validation set to optimize hyperparameters.
- as we did for model optimization, we can compute the generalization error using the cross-validation approach (other techniques can be used, but cross-validation is the simplest and most widely used). In this paragraph we will discuss the drawbacks of this choice.

In this context it is important to distinguish between *expected generalization error* and *conditional generalization error*.

Let $f_{\mathcal{T}}$ a function trained on the given training set \mathcal{T} .

The conditional generalization error $\text{Err}_{\mathcal{T}}$ is the generalization error related to an algorithm trained on a given training set \mathcal{T} , i.e. it is the generalization error related to the function $f_{\mathcal{T}}$. It can be defined as

$$\text{Err}_{\mathcal{T}} = E_{X,Y}\{L(f_{\mathcal{T}}(X), Y)\} \quad (3.12)$$

We remark that here the training set is fixed, and generalization error refers to the error for this specific training set.

The expected generalization error is quite different: it is the expected value of the conditional generalization error for all possible training sets, i.e.

$$\text{Err} = E_{\mathcal{T}}\{\text{Err}_{\mathcal{T}}\} \quad (3.13)$$

The expected conditional prediction error error is what we are really interested in knowing, as it estimates the generalization error of our specific model which is trained on a specific \mathcal{T} .

At this point, it is interesting to wonder about what quantity the cross-validation estimates. Unfortunately, cross-validation is actually an estimator of Err rather than $\text{Err}_{\mathcal{T}}$ [53]; heuristically when we cross-validate the model, we train it on different training fold samples. The cross-validation error (3.11) is therefore constructed by averaging over models trained on different sets and not on the specific training set given.

Despite theoretical aspects, cross validation could be sometimes used to estimates generalization error, but caution must be exercised.

If stable ML models are considered, which are not particularly sensitive to the training set, cross-validation may be a legitimate choice.

On the contrary, if we are using methods like tree-based algorithms, cross-validation can underestimate the true error by 10%, because the search for best tree is strongly affected by the validation set. In these situations only a separate test set will provide an unbiased estimate of test error [53].

3.1.3 Unsupervised Learning

Unsupervised learning (also informally called "learning without a teacher") algorithms experience a dataset containing many features, then learn useful properties of the structure of this dataset by minimizing (or maximizing) an objective function.

A generic unsupervised learning problem is characterized by having a dataset $D \in \mathbb{R}^{N \times p}$ made of N p -dimensional observation. Thus

$$D = \{\mathbf{d}_1, \dots, \mathbf{d}_N \mid \mathbf{d}_i \in \mathbb{R}^p \quad \forall i \in [1, N]\} \quad (3.14)$$

The underlying assumption of unsupervised learning is that each observation $\mathbf{d} \in D$ is a realization of a multivariate stochastic variable $\mathbf{X} = [X_1, \dots, X_p]$ which has a joint probability density function (pdf) $P_{\mathbf{X}}(x_1, \dots, x_p)$.

According to the assumption $P_{\mathbf{X}}(x_1, \dots, x_p)$ is the pdf which generates all possible data. It is important to underline that this pdf is typically unknown.

The goals of unsupervised learning are [53]:

- highlighting relations and patterns (more or less hidden) which could subsist within the dataset D
- inferring the probability distribution $P_{\mathbf{X}}(x_1, \dots, x_p)$ that generated the dataset D itself

The dimension of D is sometimes much higher than in supervised learning, and the properties of interest are often more complicated than input/output predictions. These difficulties are somewhat mitigated by the fact that D represents all of the variables under consideration; one is not required to infer how the properties of $P_{\mathbf{X}}$ change, conditioned on the changing values of another set of variables.

Furthermore, a remarkable difference with respect to supervised is that of in unsupervised learning, there is no such direct measure of success. It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms. One must resort to heuristic arguments not only for motivating the algorithms, as is often the case in supervised learning as well, but also for judgments as to the quality of the results. Practically speaking, in unsupervised learning we have no test set to check the quality of our prediction.

Typical unsupervised learning techniques are dimensionality reduction and clustering.

Principal components analysis, multidimensional scaling, self-organizing maps, and principal curves, for example, attempt to identify low-dimensional manifolds within the D space that represent high data density. This provides information about the associations among the variables and whether or not they can be considered as functions of a smaller set of latent variables.

Cluster analysis attempts to find multiple convex regions of the D space that contain modes of $P_{\mathbf{X}}(x_1, \dots, x_p)$. This can tell whether or not $P_{\mathbf{X}}(x_1, \dots, x_p)$ can be represented by a mixture of simpler densities representing distinct types or classes of observation.

3.2 Resampling techniques

Data sampling refers to statistical methods for selecting observations with the objective of estimating a population parameter. In statistics, resampling is the method that consists of drawing repeated samples from the original data [50]. Resampling techniques are widely used both in statistic and in machine learning; resampling allows to improve the estimate of population parameters and help to quantify the uncertainty of estimates as well as it is a powerful tool to optimize and test machine learning algorithms.

In this section I will discuss the random bootstrap technique and oversampling/undersampling techniques and stratified sampling. Furthermore, in supervised learning framework, a very important and widely used resampling technique is the cross-validation. This technique is the solution to a specific problem that occurs in supervised learning in the presence of small datasets. Therefore, for the sake of a better narrative, cross-validation has been discussed in the machine learning dedicated section 3.1

Bootstrap Bootstrap is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, and correlation coefficient. It is often used as a robust alternative to procedures based on parametric assumptions, especially when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors [141].

Let consider a sample made of N elements, say $X = \{x_1, \dots, x_N\}$. From X we resample with replacement B other samples of numerosity equal to N . Thus we obtain a bootstrap set $\{X_1^*, \dots, X_B^*\}$, where $X_i^* = \{x_{(i)1}^*, \dots, x_{(i)N}^*\}$. In each bootstrap extraction, each element has probability $1/N$ of being extracted.

Suppose we are interested in studying a statistical parameter T . Let \hat{T} the estimator of T : given a sample, we estimate T by computing the estimator \hat{T} using the elements of the sample.

We notice that bootstrap provide us a set $\{X_1^*, \dots, X_B^*\}$ of samples: for each sample X_i^* we obtain an value \hat{T}_i^* for the estimator \hat{T} .

This approach allows us to obtain B estimates $\{\hat{T}_1^*, \dots, \hat{T}_B^*\}$ of T : from this bootstrap statistic the bootstrap mean, bootstrap variance, bootstrap percentiles etc can be computed.

A great advantage of bootstrap is its simplicity. It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of the distribution, such as percentile points, proportions, odds ratio, and correlation coefficients. Bootstrap is also an appropriate way to control and check the stability of the results.

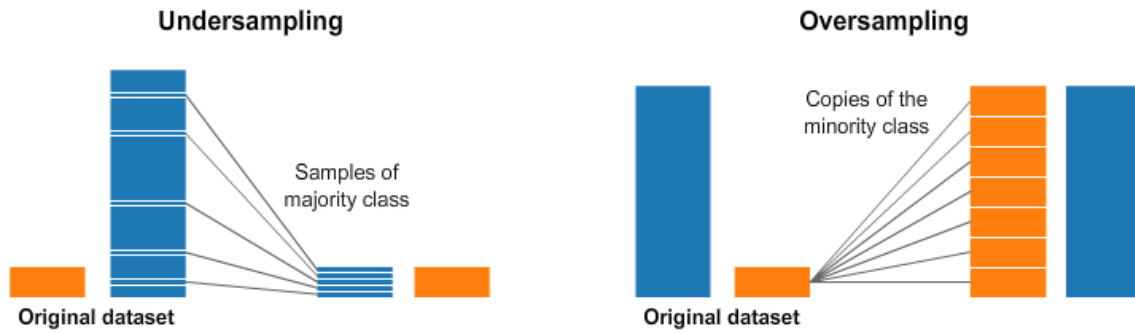


Figure 3.5: Schematic illustration of undersampling and oversampling

Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality[48].

Oversampling and undersampling techniques Let consider a two values classification problem where the output set $\{g_1, g_2\}$ is very unbalanced, say 1:100.

Training a classifier using such an imbalanced dataset usually could leads to very poor accuracy in predicting the minority class . This is because the algorithm minimizes the cost function as a whole, without taking into account the prediction accuracy of the individual classes. If the classes are very unbalanced the minimization algorithm makes it more convenient to minimize the error on the most represented class and this leads to a poor prediction of the least represented class.

Many methods have been proposed to circumvent this problem, and they can be roughly divided into cost-sensitive learning methods and resampling methods.

The former are based on modifying the cost function giving a class-dependent weight to the prediction error; basically they use cost-sensitive loss functions that give greater weight to the prediction error of the minority class (e.g. [94, 86, 166, 98, 8]).

The latter class of methods relies on resampling techniques to balance the dataset. Resampling consists of drawing repeated samples from original data; two kind of resampling are considered: the under-sampling which consists of removing samples from the majority class and the over-sampling, which consist of adding more examples from the minority class. The drawback of oversampling is overfitting, as it uses more examples from the minority class in order to balance the classes [43], while the drawback of under-sampling is that it may potentially discard useful or important samples [57].

The simplest resampling techniques are random oversampling and random undersampling.

Random oversampling involves supplementing the training data with multiple copies of some of the minority classes. Instead of duplicating every sample in the minority class, some of them may be randomly chosen with replacement.

Random undersampling consists is randomly remove samples from the majority class,



Figure 3.6: Schematic illustration of stratified sampling.

with or without replacement.

An illustration of random sampling is given in figure 3.5.

Random over/under sampling are widely used, as they are very simple to implement and all in all they are quite efficient [57].

Many more sophisticated oversampling and undersampling techniques can be considered, e.g. SMOTE (Synthetic Minority Over-sampling Technique)[29], ADASYN (Adaptive synthetic sampling approach)[77]. However, the discussion of these techniques goes beyond the scope of this thesis.

Stratified sampling In a stratified sample, the population is divided into homogeneous subpopulations called strata, based on specific characteristics (e.g. age, gender, location, etc.). Every member of the population should be in exactly one stratum.

Each stratum is then sampled using another probability sampling method, such as the simple random sampling, allowing to estimate statistical measures for each subpopulation, as is illustrated in figure 3.6.

Stratified sampling can be used when the population can be exhaustively partitioned into disjoint subgroups.

Stratified sampling is used to highlight differences between groups in a population, as opposed to simple random sampling, which treats all members of a population as equal, with an equal likelihood of being sampled.

3.3 Decision tree based ML algorithms

Many supervised ML algorithms are based on decision trees (DT). In this section I will first discuss decision trees, then I will introduce classification and regression trees, and finally I will describe the Random Forest method.

3.3.1 Decision Trees

First of all, we will introduce the notion of DT. In decision analysis, a DT can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

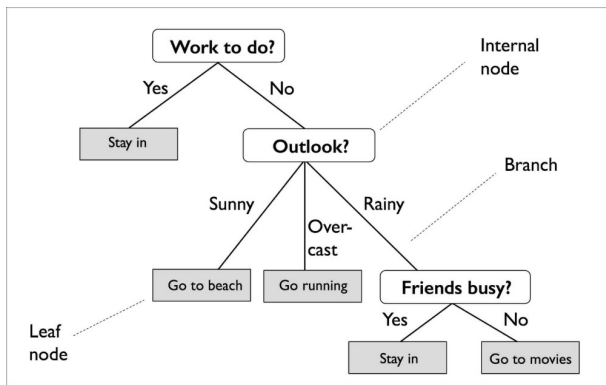


Figure 3.7: Flowchart of a simple decision tree

nodes or end-nodes. The decision nodes are the ones where the data gets fragmented according to the rules given in branches, and the leaves are one where we get the output. Furthermore, a subtree is any tree we can obtain by pruning a tree. A subtree of a tree T is a tree consisting of a node in T and all of its descendants in T . Pruning is reducing the size of decision trees by removing sections of the tree, where the size of a tree is the number of nodes in the tree.

DT is a structure for data classification, by recursive composition of elements to reach a logical decision [131], as illustrated in figure 3.7.

The main entities of a decision tree are the decision nodes and the leaves.

A decision tree is a flowchart-like structure, as illustrated in figure 3.7. Trees are made of nodes, leafs and branches. The tree elements are called nodes. The lines connecting elements are called branches. Nodes without children are called leaf

3.3.2 Classification and regresion trees (CART)

Tree-based ML methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful. Here we describe a popular method for tree-based regression and classification called CART.

The term CART (Classification And Regression Tree) is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. in 1984 [27]. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split [27].

Let us consider a regression problem with continuous response Y and inputs X_1 and X_2 , with $X_1, X_2 \in [0, 1]$. The top left panel of figure 3.8 shows a partition of the feature space by lines that are parallel to the coordinate axes. In each partition element we can model Y with a different constant. This is a fairly simple relation between predictors X_1 , X_2 and the response Y . However there is a problem: some of the resulting regions are complicated to describe.

To simplify matters, we restrict attention to recursive binary partitions like that in the top right panel of figure 3.8. We first split the space into two regions, and model the response by the mean of Y in each region.

We choose the variable and split-point to achieve the best fit. Then one or both of these regions are split into two more regions, and this process is continued, until some stopping rule is applied. For example, in the top right panel of 3.8, we first split at

$X_1 = t_1$. Then the region $X_1 \leq t_1$ is split at $X_2 = t_2$ and the region $X_1 > t_1$ is split at $X_1 = t_3$. Finally, the region $X_1 > t_3$ is split at $X_2 = t_4$. The result of this process is a partition into the five regions R_1, R_2, \dots, R_5 shown in the figure. The corresponding regression model predicts Y with a constant c_m in region R_m , that is

$$f(X) = \sum_{m=1}^5 c_m I((X_1, X_2) \in R_m) \quad (3.15)$$

where the function $I(\text{"condition"})$ is defined as follows

$$I(\text{"condition"}) = \begin{cases} 1 & \text{if "condition" is true} \\ 0 & \text{elsewhere} \end{cases} \quad (3.16)$$

This same model can be equivalently represented by the binary tree in the bottom left panel of figure 3.8. The full dataset sits at the top of the tree. Observations satisfying the condition at each junction are assigned to the left branch, and the others to the right branch. The terminal nodes or leaves of the tree correspond to the regions R_1, R_2, \dots, R_5 . The bottom right panel of figure 3.8 is a perspective plot of the regression surface from this model.

The central point here is that dividing the input space recursively creates partitions of the space that are equivalently represented with a single decision tree simply by placing conditions on the input variables.

A key advantage of recursive binary tree like the one we just described is its interpretability.

Regression Trees Let consider a regression problem. Suppose we have a dataset $\mathcal{D} = X \times Y$ made of N couples of input/output data described as follows

$$\mathcal{D} = \{(\mathbf{x}_1, y_1) \dots, (\mathbf{x}_N, y_N) \mid \mathbf{x}_i \in \mathbb{R}^p \text{ and } y_i \in \mathbb{R} \quad \forall i \in [1, N]\} \quad (3.17)$$

where X is the input set and Y the output set.

Let suppose there exist a target function f^* which relates X space and Y space such that $Y = f^*(X) + \epsilon$, where ϵ represents an irreducible noise (i.e. the relation is not deterministic) which follows a normal distribution with zero mean, namely $\epsilon \sim \mathcal{N}(0, \sigma)$. Our goal is to obtain an estimation f of f^* using a regression tree ML method, given the dataset \mathcal{D} .

As described before, the starting point is recursively input space partition. Let suppose recursively partitioning led us to R_1, \dots, R_M regions of the form illustrated in top right panel of figure 3.8.

According to equation (3.15) and (3.16), we model the response function as a sum of estimated values \hat{c}_m as follows

$$f(x) = \sum_{m=1}^M \hat{c}_m I(x \in R_m) \quad (3.18)$$

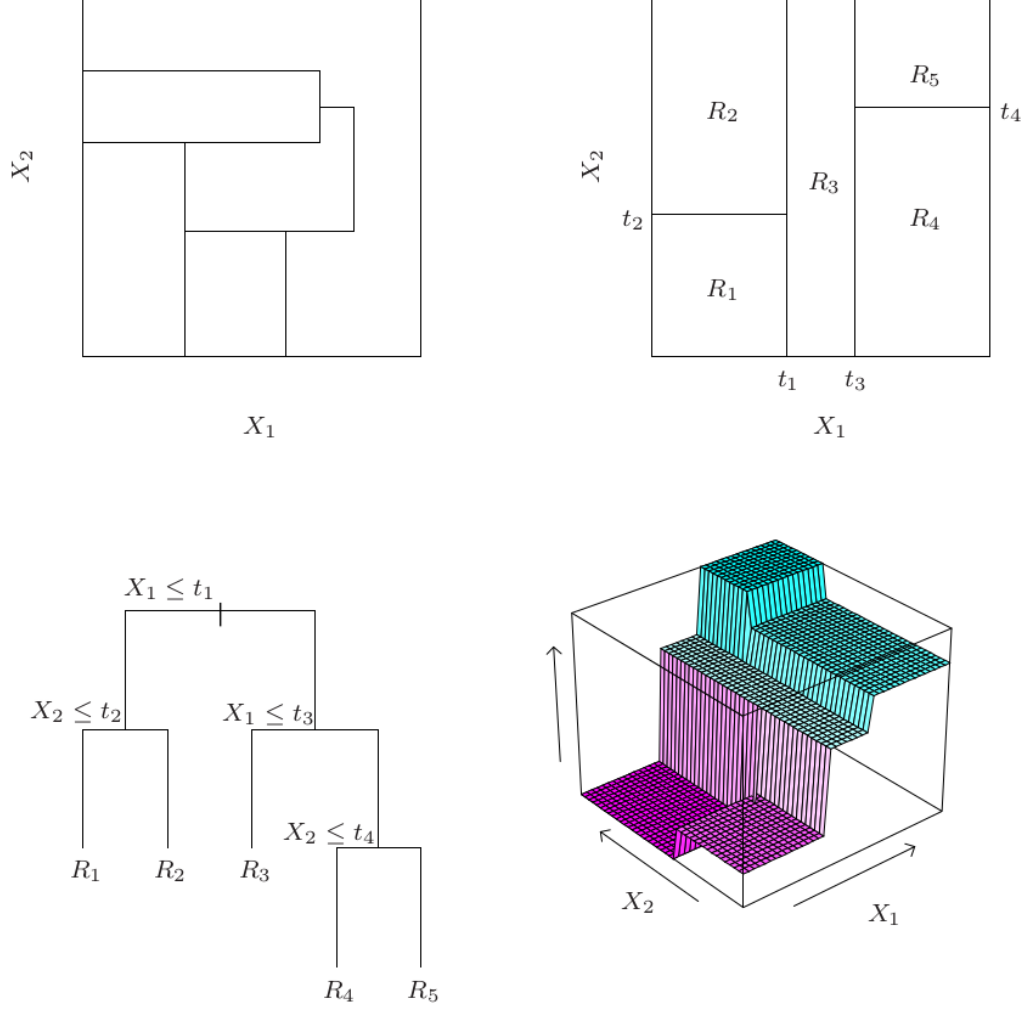


Figure 3.8: Top right panel shows a partition of a two-dimensional feature space by recursive binary splitting. Top left panel shows a general partition that cannot be obtained from recursive binary splitting. Bottom left panel shows the tree corresponding to the partition in the top right panel, and a perspective plot of the prediction surface appears in the bottom right panel. Courtesy of [53].

If we adopt a L_2 loss function, we would minimize the sum of squares $(y_i f(x_i))^2$. It is easy to see that the best \hat{c}_m for this choice is just the average of y_i in region R_m :

$$\hat{c}_m = \text{mean}(y_i \mid \mathbf{x}_i \in R_m) \quad (3.19)$$

Now finding the best binary partition in terms of minimum sum of squares is generally computationally infeasible. Hence we proceed with a smarter hierarchical algorithm.

Starting with all of the data, the algorithm's first step is dividing space in two "best" partition: considering a splitting variable j and split point s , space can be divided in two partition as follows

$$R_1(j, s) = \{X \mid X_j \leq s\} \text{ and } R_2(j, s) = \{X \mid X_j > s\} \quad (3.20)$$

We notice the two regions are uniquely defined by the couple (j, s) .

The "best" R_1, R_2 partition is the one which minimize sum of squares, thus it is defined

by the splitting variable j and split point s which solve

$$\min_{j,s} \left[\min_{c_1} \sum_{\mathbf{x}_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{\mathbf{x}_i \in R_2(j,s)} (y_i - c_2)^2 \right] \quad (3.21)$$

For any choice j and s , the inner minimization of equation (3.21) is the same problem of equation (3.19), thus is simply solved by

$$\hat{c}_1 = \text{mean}(y_i \mid \mathbf{x}_i \in R_1(j, s)) \quad (3.22)$$

$$\hat{c}_2 = \text{mean}(y_i \mid \mathbf{x}_i \in R_2(j, s)) \quad (3.23)$$

$$(3.24)$$

The outer minimization is the more interesting part: the determination of the (j, s) couple can be done very quickly by computationally scanning all the input space, computing the results for each variables couple, and then selecting the the one which mininmize the equation (3.21).

Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. Then this process is repeated on all of the resulting regions. This iterative hierarchical approach allows partitioning the space into smaller and smaller regions, and each partitioning corresponds to a new node in the DT. Iterating the procedure for a very high number of steps will create a very fine space partition that fits very well to our data (training set), but may have difficulty in generalizing to predict new data (i.e. overfitting). Remembering that each partition of the space corresponds to one and only one decision tree, we can map the related problem of finding an appropriate partition of the space to managing the size and properties of a decision tree. Thus, we will control the trade off between overfitting and underfitting by controlling the size of the decision tree itself. A decision tree that is too large (too small) with respect to the complexity of our problem will lead to overfitting (underfitting).

In DTs, the underfitting/overfitting trade-off is managed by the pruning operation. Let consider a subtree $T \subset T_0$. We denote leafs (terminal nodes) by m : it is important to notice that each leaf m represents a region R_m . Let $|T|$ the number of leafs in T and let $N_m = \#\{\mathbf{x}_i \in R_m\}$ the number of input data in R_m . Then, letting

$$\begin{aligned} \hat{c}_m &= \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} y_i \\ Q_m(T) &= \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 \end{aligned} \quad (3.25)$$

allows us to define the cost complexity criterion

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T| \quad (3.26)$$

The idea is to find, for each α , the subtree $T_\alpha \subseteq T_0$ to minimize $C_\alpha(T)$.

We notice two facts

- the number of leaf $|T|$ is actually the number of regions of the equivalent space partition

- the equation (3.26) is an application of the equation (3.9) where $|T|$ is the regularizer and α is an hyperparameter to tune
- α manage the tree complexity (i.e. underfitting/overfitting tradeoff): large values of α result in smaller trees T_α are preferred, while for smaller values of α , the goodness of fit have more weight respect to tree size. As the notation suggests, with $\alpha = 0$ the solution is the full tree T_0 .

It can be shown that for each α there is a unique smallest subtree T_α that minimizes $C_\alpha(T)$ [53].

Estimation of α is typically achieved by five or ten fold cross-validation[53]: we choose the value α to minimize the cross-validated sum of squares.

Classification Trees In a classification problem the outcome takes values in a finite label set $\bar{Y} = \{1, \dots, K\}$) which is the output space. The classification tree algorithm is equal to the regression one except for the criteria for assigning outputs, splitting nodes and pruning the tree.

First of all, in regression we estimated the values of f in various regions R_m using a sample mean which involved the response data y_i . This is not possible in classification. Furthermore in a classification problem, the goal is, given an input \mathbf{x}_i , to predict the output label $y_i \in \bar{Y}$. Thus, a modification is required: we define

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} I(y_i = k) \quad (3.27)$$

Concretely speaking, \hat{p}_{mk} is the proportion of the k -th label in the m -th region. Thus the predict class in the region R_m for classification problems became

$$\hat{c}_m = k_{max}(m) = \arg \max_k \hat{p}_{mk} \quad (3.28)$$

i.e. we assign to a given input $\mathbf{x} \in R_m$ the label $k_{max}(m)$, where $k_{max}(m)$ is the most represented class in R_m .

Furthermore we can not use the mean squared error for each leaf $Q_m(T)$ which have been used (see equation (3.25)) in regression tree. Most common error measures of $Q_m(T)$ used in classification tree are [53] :

$$Q_m(T) = \begin{cases} \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq \hat{c}_m) = 1 - \hat{p}_{m\hat{c}_m} & \text{Misclassification error} \\ \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}) & \text{Gini index} \\ - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk} & \text{Cross-entropy} \end{cases} \quad (3.29)$$

$\hat{p} \log \hat{p} + (1 - \hat{p}) \log (1 - \hat{p})$, respectively. They are shown in Figure 9.3. All three are similar, but the cross-entropy and the Gini index are differentiable, and hence are better for numerical optimization. In addition, cross-entropy and the Gini index are more sensitive to changes in the node probabilities than the misclassification rate.

Instability of Trees One major problem with trees is their high variance. Often a small change in the data can result in a very different series of splits.

The major reason for this instability is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it. One can alleviate this to some degree by trying to use a more stable split criterion, but the inherent instability is not removed. It is the price to be paid for estimating a simple, tree-based structure from the data.

In next section algorithms (bagging and Random Forest) used to reduce tree's variance will be discussed.

3.3.3 Bootstrap aggregating

Bootstrap Aggregation (usually called Bagging), is a simple and very powerful ensemble method.

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.

Bagging can be considered as an application of the bootstrap technique to a machine learning algorithm: bagging allows to dramatically reduce the variance of an algorithm and thus it is very used for those ML methods affected by high variance and low bias, such as decision trees.

Let f^* a target function between a given input and output space we would like to estimate. Bagging works as follows:

- a training set and a test set are fixed *a-priori*
- B bootstrap samples are collected from the fixed training set
- each bootstrap sample is considered as a training set, and thus it is used to train the algorithm: $\{f_1, \dots, f_B\}$ estimation of the target function are obtained
- we summarize the set $\{f_1, \dots, f_B\}$ in a single classifier f_{bag} which has a better generalization performance than the single classifier.

In particular, for regression, we simply define f_{bag} as an average of the set:

$$f_{bag}(x) = \frac{1}{B} \sum_{i_1}^B f_i(x) \quad (3.30)$$

For classification we replace the mean by the most "voted" class: given an input x , we will have B output labels, one for each classifier. The final label that the classifier f_{bag} assigns will be the one most voted by the individual classifiers.

Bagging works very well for high variance and low bias classifier because averaging reduces variance and leaves bias unchanged [53].

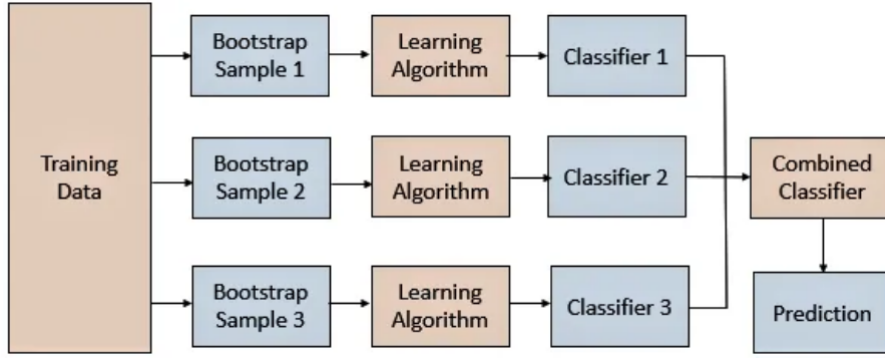


Figure 3.9: Sketch of a bagging workflow

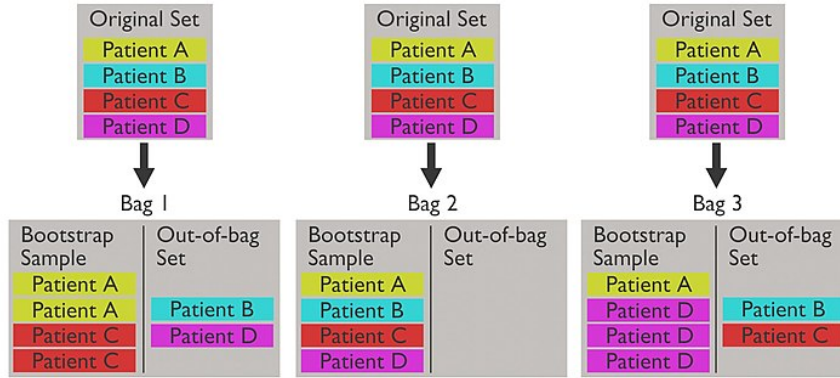


Figure 3.10: Schematic illustration of OOB samples due to a bagging algorithm.

Out of bag samples As bootstrap is a resampling with replacement technique, for each bootstrap sample taken from the training data, there could be samples left behind that were not included. These samples are called Out-Of-Bag samples (usually denoted by OOB).

The performance of each model on its left out samples when averaged can provide an estimated accuracy of the bagged models. This estimated performance is often called the OOB estimate of performance.

These performance measures are reliable to validate models parameters and can be used also for generalization error estimate.

Finally we notice that the OOB estimation of performance is very similar to a k fold cross-validation.

An illustration of OOB estimation is provided in figure 3.10.

3.3.4 Random Forest

Random forests [26] is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them.

As just mentioned in previous section, trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias. Since trees are notoriously noisy, they benefit greatly from the averaging.

An important issue will be highlighted now. Let consider B identically distributed

stochastic variables. Let suppose that each B has the same variance σ^2 and let denote the pairwise correlation by ρ . Then it can be shown that the variance of the average of the B variables is

$$\sigma_{Average}^2 = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \quad (3.31)$$

As B increases, the second term disappears, but the first remains, and hence the size of the correlation of pairs of bagged trees limits the benefits of averaging. The Random Forests key idea is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables.

More in detail, when growing a tree on a bootstrapped dataset, before each split, a number $m \leq p$ of the p input predictors are randomly selected as candidates for splitting. Intuitively, reducing m will reduce the correlation between any pair of trees in the ensemble, thus, as shown in (3.31) the variance of the average is reduced.

3.4 Principal Component Analysis

Principal components analysis (PCA) is a dimensionality reduction unsupervised ML technique which is widely used in high-dimensional data analysis.

The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set.

PCA is one of the simplest and most robust ways of doing such dimensionality reduction[137], .

Let us consider D -dimensional dataset made of n observations described by a data matrix $\mathbf{X} \in \mathbb{R}^{n \times D}$. For the sake of simplicity let us suppose the data have zero mean. The goal of PCA is to map \mathbf{X} in a lower dimensional matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times d}$, where $d < D$, such that the loss of information due to the use of $\tilde{\mathbf{X}}$ instead of \mathbf{X} is the least possible.

Thus, let us focus on the i -th observation \mathbf{x}_i . This vector is implicitly expressed in the canonical basis of \mathbb{R}^D , but it can be also expressed in any other basis of \mathbb{R}^D . Let us consider a generic orthonormal basis of \mathbb{R}^D , which we denote by $\{\mathbf{b}_1, \dots, \mathbf{b}_D\}$ and let us express \mathbf{x}_i in terms of the new basis $\mathbf{x}_i = \sum_{j=1}^D \langle \mathbf{b}_j, \mathbf{x}_i \rangle \mathbf{b}_j$.

Let B the subspace of \mathbb{R}^D spanned by the first d vector $\{\mathbf{b}_1, \dots, \mathbf{b}_d\}$, which we will call principal subspace, and let B_\perp its orthogonal complement spanned by $\{\mathbf{b}_{d+1}, \dots, \mathbf{b}_D\}$. We can rewrite \mathbf{x}_i as

$$\mathbf{x}_i = \tilde{\mathbf{x}}_i + \tilde{\mathbf{x}}_{\perp i}. \quad (3.32)$$

where

$$\tilde{\mathbf{x}}_i = \sum_{k=1}^d \langle \mathbf{b}_k, \mathbf{x}_i \rangle \mathbf{b}_k \quad (3.33)$$

$$\tilde{\mathbf{x}}_{\perp i} = \sum_{k=d+1}^D \langle \mathbf{b}_k, \mathbf{x}_i \rangle \mathbf{b}_k. \quad (3.34)$$

The first term is the projection of \mathbf{x}_i into B , while the second term is the projection of \mathbf{x}_i into B_\perp . We reduce the data dimensionality by using the B data projection instead of the data themselves, i.e. using $\tilde{\mathbf{x}}_i$ instead of \mathbf{x}_i .

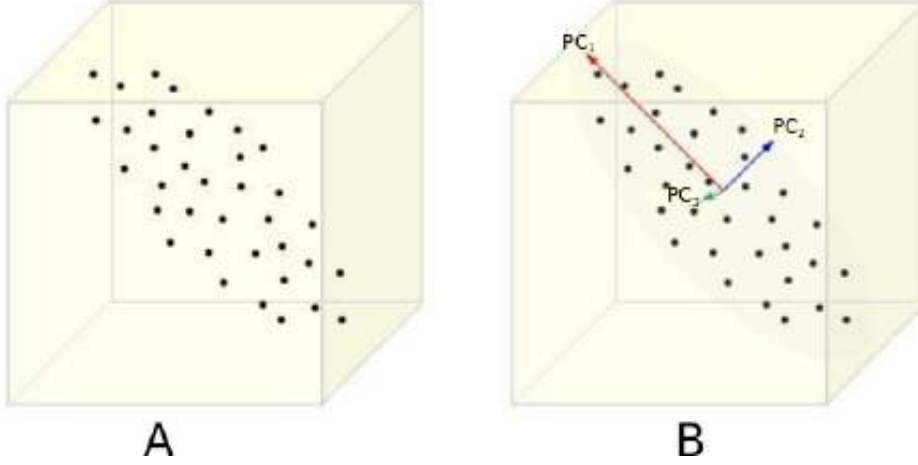


Figure 3.11: Original dataset A in space x, y, z is mapped onto a new space B built on principal components $PC1, PC2, PC3$ (linear combination of original coordinates). Axis $PC1$ is the one that better interprets dataset variability, while axis $PC3$ explains little variance. The best choice for reducing dimensionality in the example in figure could be to consider the projected data in $PC1$ and $PC2$ space (i.e. ignoring their component along $PC3$). In this way we reduce the complexity of the data and at the time we lose as little information as possible.

What is the error due to this choice? The error we make is a function of the subspace B we choose and it can be write as $\mathbf{L}_{i,m}(\mathbf{b}_1, \dots, \mathbf{b}_d) = \mathbf{x}_i - \tilde{\mathbf{x}}_i$. Considering all the data, the mean square error can be written as

$$\begin{aligned} \text{MSE}_m(\mathbf{b}_1, \dots, \mathbf{b}_d) &= \frac{1}{n} \sum_{i=1}^n \|\mathbf{L}_{i,m}(\mathbf{b}_1, \dots, \mathbf{b}_d)\|^2 = \\ &= \frac{1}{N} \sum_{i=1}^n \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \tilde{\mathbf{x}}_i)^T (\mathbf{x}_i - \tilde{\mathbf{x}}_i) \end{aligned} \quad (3.35)$$

Observing that $\tilde{\mathbf{x}}_i^T \mathbf{x}_i = \mathbf{x}_i^T \tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i$ because of the orthogonality between B and B_\perp , the equation 3.35 became

$$\text{MSE}_m(\mathbf{b}_1, \dots, \mathbf{b}_d) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i \quad (3.36)$$

Let us rewrite the second term of 3.36

$$\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_i^T \tilde{\mathbf{x}}_i = \frac{1}{n} \sum_{i=1}^n \left\langle \sum_{j=1}^d \langle \mathbf{b}_j, \mathbf{x}_i \rangle \mathbf{b}_j, \sum_{k=1}^d \langle \mathbf{b}_k, \mathbf{x}_i \rangle \mathbf{b}_k \right\rangle = \sum_{j=1}^d \frac{1}{n} \sum_{i=1}^n \langle \mathbf{b}_j, \mathbf{x}_i \rangle^2 \quad (3.37)$$

where in the last passage we used inner product bilinearity and basis orthonormality condition $\langle \mathbf{b}_j, \mathbf{b}_k \rangle = \delta_{jk}$.

The data have zero mean, so $\frac{1}{n} \sum_{i=1}^n \langle \mathbf{b}_j, \mathbf{x}_i \rangle^2$ is the by definition the variance of $\langle \mathbf{b}_j, \mathbf{x}_i \rangle$, i.e. the variance of the data along the \mathbf{b}_j direction, which we will call $\sigma_{\mathbf{b}_j}^2$. Thus equation 3.36 became

$$\text{MSE}_m(\mathbf{b}_1, \dots, \mathbf{b}_d) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i - \sum_{j=1}^d \sigma_{\mathbf{b}_j}^2 \quad (3.38)$$

The former MSE contribution is a constant term which is fixed once data are given, while the latter is function of the subspace $\mathbf{b}_1, \dots, \mathbf{b}_d$ we chose to project the data. Then,

minimize the MSE is equivalent to maximize the variance explained along the $\mathbf{b}_1, \dots, \mathbf{b}_d$ directions.

Before maximize the variance, it will be convenient to rewrite the problem using matrix notation. We observe that

$$\sum_{i=1}^n \langle \mathbf{b}_j, \mathbf{x}_i \rangle^2 = (\mathbf{X}\mathbf{b}_j)^T (\mathbf{X}\mathbf{b}_j) \quad (3.39)$$

$$\mathbf{C} = \mathbf{X}^T \mathbf{X} \quad (3.40)$$

where \mathbf{C} is the variance-covariance matrix. Then, we can write

$$\sum_{j=1}^d \frac{1}{n} \sum_{i=1}^n \langle \mathbf{b}_j, \mathbf{x}_i \rangle^2 = \frac{1}{n} \sum_{j=1}^d (\mathbf{X}\mathbf{b}_j)^T (\mathbf{X}\mathbf{b}_j) = \frac{1}{n} \sum_{j=1}^d \mathbf{b}_j^T \mathbf{X}^T \mathbf{X} \mathbf{b}_j = \frac{1}{n} \sum_{j=1}^d \mathbf{b}_j^T \mathbf{C} \mathbf{b}_j. \quad (3.41)$$

Thus, we want to maximize the RHS of 3.41. We also impose the constraint that $\mathbf{b}_j^T \mathbf{b}_j = 1$ for all $j = 1 \dots d$. This is a constrained optimization problem which can be solved using lagrangian multipliers method.

Let $\mathcal{L} = \sum_{j=1}^d [\mathbf{b}_j^T \mathbf{C} \mathbf{b}_j - \lambda_j (\mathbf{b}_j^T \mathbf{b}_j - 1)]$ the langrangian function, where the λ_j are the lagrangian multipliers. In order to find the $\mathbf{b}_1, \dots, \mathbf{b}_d$ which maximize 3.41, we need to put to zero at the same time all the d derivatives with respect to \mathbf{b}_k where $k = 1 \dots d$.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_k} = \mathbf{C} \mathbf{b}_k - \lambda_k \mathbf{b}_k = 0, \quad \forall k = 1 \dots d. \quad (3.42)$$

These d equations tell us that the \mathbf{b}_k which maximize the variance in the d -dimensional subspace within which we want project data are d eigenvectors of \mathbf{C} .

The last step is about finding these eigenvectors: \mathbf{C} is a D -by- D symmetric matrix, thus it has D eigenvectors. Which are the d whose maximize the variance? Let us consider the eigenvalue equation related to 3.42:

$$\mathbf{C} \mathbf{b} = \lambda \mathbf{b} \quad (3.43)$$

Solving this equation is equivalent to diagonalize \mathbf{C} . The eigenvectors of \mathbf{C} , which form an orthonormal basis, are called *Principal Components*. In eigenvectors basis we have that $\mathbf{C} = \text{diag}(\lambda_1, \dots, \lambda_D)$, then the eigenvalues λ_j represents the variance explained by the j -th eigenvector. Because we want to maximize the variance, the $\mathbf{b}_1, \dots, \mathbf{b}_d$ eigenvectors (i.e. Principal Components) which solve the problem are those are related to the greater d eigenvalues.

Now we can compute the lowest possible MSE given by 3.38.

Let us suppose that the grater d eigenvalues of variance-covariance matrix are $\lambda_1 \dots \lambda_d$, while $\lambda_{d+1}, \dots, \lambda_D$ are the lower. Let us rewrite the former contributions of 3.38 in the basis of the eigenvectors of \mathbf{C} . Following the same approach used in 3.37, we obtain

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i = \sum_{j=1}^D \sigma_{\mathbf{b}_j}^2 \quad (3.44)$$

Putting 3.44 in equation 3.38, and using the relation $\sigma_{\mathbf{b}_j}^2 = \lambda_j \quad \forall j = 1 \dots D$, we obtain

$$\text{MSE}_d = \sum_{j=1}^D \lambda_j^2 - \sum_{j=1}^d \lambda_j^2 = \sum_{j=d+1}^D \lambda_j^2 \quad (3.45)$$

The PCA algorithm is particularly useful when many variables are highly correlated (i.e. λ_j^2 are very different to each other): in this case most of the variance is explained by few Principal Components, then dimensionality reduction do not leads to lose much information. A pictorial PCA example is provided in figure 3.11.

Finally, it is important to observe that data expressed in the principal components basis are uncorrelated, because the variance-covariance matrix \mathbf{C} is diagonal: then, PCA can be used not only for dimensionality reduction, but also in order to uncorrelate variables which are correlated.

3.5 Statistical evaluators

3.5.1 Statistical test of parametric hypothesis

In statistics, hypothesis testing is used to test the validity of a hypothesis. In the framework of parametrical hypothesis test, an hypothesis can be defined as a statement about the parameters describing a population.

In this context, given an unknown parameter $\theta \in \Theta$, where Θ represents the domain of the stastical parameter (for example \mathbb{R}^n), an hypothesis H on θ is a statement of the type

$$H : \quad \theta \in \Theta_0 \subset \Theta \quad (3.46)$$

Hypothesis is divided in two broad class: simple and composite hypothesis. An hypothesis are called simple if Θ_0 contains only one element, otherwise are called composite.

Usually two hypothesis are considered in statistical test: The null and the alternative hypothesis.

The former, typically called H_0 , is the hypothesis to be verified. The latter, usually denoted with H_1 , is the hypothesis considered true when H_0 is false.

Typically H_1 is complementary with respect to H_0 , i.e. if H_0 is defined as $\theta \in \Theta_0 \subset \Theta$, then H_1 is defined as $\theta \in \Theta \setminus \Theta_0$.

A statistical test is a rule which allows us to decide whether, and to what extent, to accept or reject a null hypothesis, by examining the observations made on a sample statistic.

It is important to notice that testing hypothesis requires the sample statistic knowledge of the unknown parameter θ if the null hypothesis H_0 is true.

The decision to accept or reject the null hypothesis is affected by two types of error, which are defined as follows.

The *type I error* α is the rejection of a true null hypothesis

$$\alpha = \mathbb{P}(H_1|H_0) \quad (3.47)$$

The *type II error* β is the failure to reject a false null hypothesis.

$$\beta = \mathbb{P}(H_0|H_1) \quad (3.48)$$

The significance of the test is the maximum probability with which we are willing to risk making a first kind error. The significance is fixed a-priori by fixing α . Most common choice are $\alpha = 0.05$ and $\alpha = 0.01$.

The p-value[note 1] is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.[2][3] A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.

This means that, once α is fixed, if p-value $> \alpha$, then H_0 will be accepted, otherwise H_0 will be rejected.

The p-value helps to understand if the difference between the observed and the hypothesized result is due to the randomness introduced by the sampling, or if this difference is statistically significant, i.e. difficult to explain by the randomness due to the sampling.

Furthermore, another important quantity is the power of a test, which is defined as the quantity $W = 1 - \beta$ that measures the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true. However the determination of the power of a test is often difficult, because for its explicit calculation one also needs to know the sampling distribution of when the alternative hypothesis H_1 is true, which, moreover, is often not simple but composite.

Most common parametric tests are

- *Z-test*: A z-test is any statistical hypothesis test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. This hypothesis holds if the sample data come from a population with a normal distribution and known standard deviation σ . If the amount of data is large enough (typically > 30), both the assumption are not required and σ can be estimated using sample variance.
The z-test is typically used to determine whether a sample data set comes from a population with a particular mean μ . Moreover z-test can be used to determine if the sample mean of two different samples are compatible, i.e. if the two samples came from the same population.
- *T-test*: A t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.
The t-test is commonly used for the same purpose of z-test. The only difference is that t-test is used when population variance σ is unknown: in this case σ is estimated using sample variance. As sample data increase (usually > 30), t-test tends to be equivalent to z-test.
- *F-test*: An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. Let S_x^2 and S_y^2 two sample variances of two independent and identically distributed samples extracted from two populations which each has a normal distribution. Let suppose the two samples are made of n_x and n_y data respectively. The ratio $F = S_x^2/S_y^2$ follows a F-distribution F_{n_x-1, n_y-1} .

As the ratio between sample variances follows F distribution, the F-test is typically used to determine whether two sample variances coming from population with same variance or not.

- ANOVA (Analysis Of Variance): Let consider a sample made of n observation. We divide the samples in G groups according to multiple factors. We might be interested in knowing whether the sample means of these groups are compatible with each other, to see whether the effect of multiple factors is relevant or not with respect to given response variable y . Factors can be independent to each other or not, no independence assumptions are required.

As an example, we can intersted in studying the effect of weight, age and sex on the blood pressure (response variable y). We divide population in groups according to low/high weight, low/high age and M/F sex. The goal is checking whether the blood pressure means of considered groups are compatible or not.

ANOVA is an ensemble of statistical methods which offer a solution to this issue.

We focus on the simplest form of ANOVA only, sometimes called one way ANOVA. The one way ANOVA (just ANOVA for the sake of simplicity) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other, supposing the number of factors involved in group division are just one.

ANOVA can be seen as a generalization of t-test and z-test. When we have only two samples, t-test/z-test and ANOVA give the same results.

ANOVA is based on the assumption that all sample populations are normally distributed, but it is known to be robust to modest violations of this assumption. ANOVA tests the hypothesis that all group means are equal against the alternative hypothesis that at least one group is different from the others.

The central point of ANOVA is the following. Given n data divided into G groups, it is possible to decompose the variance into two components: the sample variance within groups (also called within sample variance σ_W) and sample variance between groups (also called sample between variance σ_B). Between and within variances are defined as follows

$$\sigma_B = \sum_{g=1}^G (\mu_g - \mu)^2 \frac{n_g}{n-1} \quad (3.49)$$

$$\sigma_W = \sum_{g=1}^G \sigma_g^2 \frac{n_g - 1}{n-1} \quad (3.50)$$

where μ is the sample mean of all n data, μ_g is the sample mean of g -th group, n_g is the number of data of the g -th group and σ_g is the sample variance of the g -th group.

As mentioned for F-test, the ratio between two sampling variances follows a F-statistic. Thus, ANOVA test whether σ_B and σ_W coming from two populations with same variance, through an F-test performed over the F-statistic $F = \frac{\sigma_B}{\sigma_W} \sim F_{G-1, n-G}$.

If F-test between σ_B and σ_W lead to reject the null hypothesis according to which the population came from the same variance, we reject also the ANOVA null hypothesis (i.e. all group means are equal), otherwise we accept the ANOVA null hypothesis.

3.5.2 Akaike Information Criterion and Bayesian Information Criterion

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data.

AIC is calculated from the number of independent variables used to build the model and the maximum likelihood estimate of the model (how well the model reproduces the data).

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.

Suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model. Let L be the maximum value of the likelihood function for the model. Then the AIC value of the model is given by the following formula [4]

$$\text{AIC} = 2k - \ln(L) \quad (3.51)$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

Another criterion for model selection is the Bayesian information criterion (BIC). It is very similar to AIC criterion, and as occurs in AIC criterion, the model with the lowest BIC is the model to prefer.

BIC is closely related to AIC, and as AIC is based on the number of model parameters and on the maximum likelihood value. BIC is defined as follows [136]

$$\text{BIC} = k \ln n - \ln(L) \quad (3.52)$$

where n is the sample size.

BIC is consistent in the sense that if the true model is among the candidates, the probability of selecting the true model approaches 1. On the other hand, AIC is minimax-rate optimal for both parametric and nonparametric cases for estimating the regression function [162]. Moreover, in the case of multivariate regression analysis, AIC is better than BIC in model selection [162]

Both AIC and BIC have no absolute meaning: what matters is the difference in AIC and BIC between models. Let $\Delta(\text{AIC})$ and $\Delta(\text{BIC})$ the difference between AIC and BIC of two models respectively. The rules of thumb are usually used. If the difference.

If $\Delta(\text{AIC})$ is [54]

- Less than 2, this indicates there is substantial evidence to support the candidate model (i.e., the candidate model is almost as good as the best model)
- Between 4 and 7, this indicates that the candidate model has considerably less support
- Greater than 10, this indicates that there is essentially no support for the candidate model (i.e., it is unlikely to be the best model)

If $\Delta(\text{BIC})$ is [54]

- Less than 2, it is not worth more than a bare mention
- Between 2 and 6, the evidence against the candidate model is positive
- Between 6 and 10, the evidence against the candidate model is strong
- Greater than 10, the evidence is very strong

3.5.3 Receiver Operating Characteristics

Receiver Operating Characteristics (ROC) is a technique for visualizing, organizing and selecting binary classifiers based on their performance. ROC charts have long been used in signal detection theory to depict the trade-off between hit rates and false alarm rates of classifiers [147, 75, 56].

Lets consider a two-class classification problem, where each instance I is mapped to one element of the set $\{p, n\}$ of positive and negative class labels, and denote with $\{Y, N\}$ the class predictions produced by the classifier model. Given a classifier and an instance, there are four possible outcomes:

- the True Positive (TP) outcome if a positive instance is classified as positive
- the False Positive (FP) outcome if a negative instance is classified as positive
- the True Negative (TN) outcome if a negative instance is classified as negative
- the False Negative (FN) outcome if a positive instance is classified as negative

Given a classifier and a set of instances (typically the test set), a 2-by-2 confusion matrix can be constructed representing the dispositions of the set of instances, as illustrated in figure 3.12.

		p	n		
<u>Hypothesized</u> <u>class</u>	Y	True Positives	False Positives	fp rate = $\frac{FP}{N}$	tp rate = $\frac{TP}{P}$
	N	False Negatives	True Negatives	precision = $\frac{TP}{TP+FP}$	recall = $\frac{TP}{P}$
				accuracy = $\frac{TP+TN}{P+N}$	
Column totals:		P	N	F-measure = $\frac{2}{1/\text{precision}+1/\text{recall}}$	

Figure 3.12: Confusion matrix for binary classification, and common performance metrics that stem from it.

The diagonal elements of the confusion matrix represent the correct decisions made, while the outer diagonal elements represents the misclassified instances.

Given a collection of TP and NP counts we can define the true positive rate tp and the false positive rate fp of a classifier as

$$tp = \frac{n(TP)}{n(TP) + n(FN)} \quad (3.53)$$

$$fp = \frac{n(FP)}{n(TN) + n(FP)} \quad (3.54)$$

$$(3.55)$$

where $n(X)$ represents the number of elements which belong to the class X .

Additional important terms associated with ROCs are *sensitivity* and *specificity*. Sensitivity measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition). We notice that sensitivity is, by definition, the same of tp and can be calculated using (3.53).

Specificity measures the proportion of negatives that are correctly identified (i.e. the proportion of those who do not have the condition (unaffected) who are correctly identified as not having the condition). Thus sensibility is by definition equal to the true negative ratio tn

$$tn = \frac{n(TN)}{n(TN) + n(FP)} \quad (3.56)$$

which is also equal to $1 - fp$.

ROC can be visualized as two-dimensional graphs in which tp rate is plotted on the y -axis and $1 - fp$ rate is plotted on the x -axis, as illustrated in figure 3.13. This graph depicts relative trade-offs between benefits (true positives) and costs (false positives). Several points in ROC space are important to note. The lower left point (0,0) represents a classifier of never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives. The opposite strategy, unconditionally issuing positive classifications, is represented by the upper right point (1,1).

To compare classifiers we may want to reduce ROC performance to a single scalar value representing the expected performance. A common method is to calculate the area under the ROC curve (AUC).

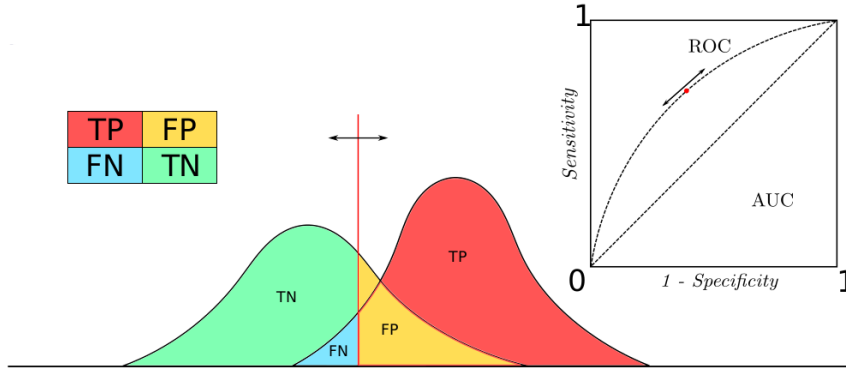


Figure 3.13: An exemplifying ROC curve. Step by step moving the threshold (red dot) leads to obtain couples of (tp, fp) values which compose the ROC curve

AUC ranges in $AUC \in [0, 1]$, with $AUC = 0.5$ corresponding to a random-guessing classifier. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

3.6 Image Registration

In chapter 2 all the steps to obtain a three-dimensional medical image from a PET examination were shown. In particular, in section 2.7 we have shown what is the final product of the whole procedure, i.e. a three-dimensional NIfTI image, known as the raw image, which is described in terms of a 3-D matrix $n - by - n - by - p$ whose elements are called voxels and whose dimension n and p are reconstruction dependent. The 3-D matrix can be considered as a discrete space in which the raw image is embedded, and this called the native space. Examples of native space images are give in figure 3.14.

In this section we point out that raw images are not directly usable in a data analysis, as they are not comparable with each other. We will then show how this problem can be solved by using so-called image registration.

Let us consider two or more medical images. They can be:

- multi-modality images of the same subject (for example one PET and one MRI)
- mono-modality images of different subjects (such as two PET od two different subjects)
- mono-modality images of the same subject acquired at different time

Whatever category the images you want to study belong to, the common point is that different images are not directly comparable, as they are embedded in spaces of different dimensionality and/or as they are not aligned with each other, as you can see in 3.14.

In other words, if we consider N images, the same anatomical point P will be described by N different sets of three coordinates: this represents a central problem in

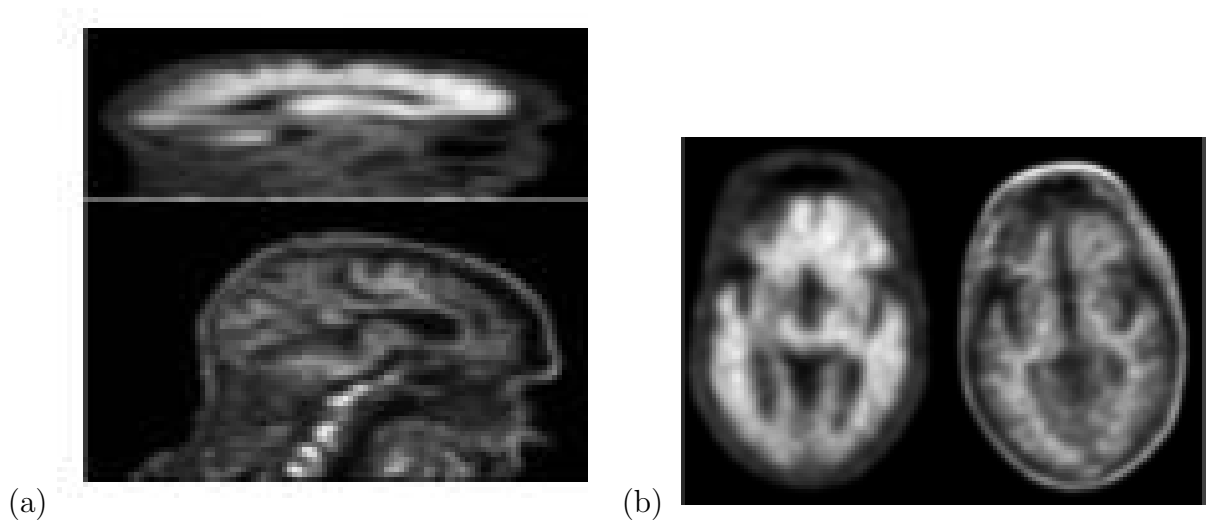


Figure 3.14: Examples of coronal (a) and axial (b) sections of two raw images.

medical imaging analysis.

Image registration is the process which permit to align different images in a common coordinates space through a transformation between the two spaces. The goal of image registration is to find a map which transform points from one image to homologous points on a second image. Registration brings one image to match another image, such that the same voxels refers roughly to the same structure in both brains.

The best transformation is the one which map into each other two points related to the same anatomical point. Let us consider the simplest registration problem, namely the one which involve two images.

A generic 3 – d image can be described by a function g defined on a M -by- N -by- P grid

$$\mathbf{x} \rightarrow g(\mathbf{x}) \in \mathbb{R} \quad (3.57)$$

where \mathbf{x} is a point a on the grid. Let $f(\mathbf{x})$ the *fixed image* and the let $m(\mathbf{y})$ the *moving image*, where \mathbf{x} and \mathbf{y} are points of the reference space and of the moving space respectively. Often the fixed image is a template, but you can register any two images with each other, there is nothing special in a template from the computation perspective. Registration is treated as an optimization problem with the goal of finding the best spatial mapping T^* which minimizes the distance d equation (3.58) (or maximize a similarity measure s (3.59) between the fixing and the moving image:

$$T^* = \arg \min_{T \in \mathcal{T}} d(f, I(T(m))) \quad (3.58)$$

$$T^* = \arg \max_{T \in \mathcal{T}} s(f, I(T(m))) \quad (3.59)$$

where \mathcal{T} is the space of the all possible transformations and where represents the action of a given interpolator on the transformed intensity $T(m)$. Indeed, it is important to observe that images are defined on a grid of integer values, so points mapped by T could not belong to a grid point (see figure 3.15 on the next page). When a transformation is applied to the input image, a new grid is obtained and an intensity interpolation algorithm

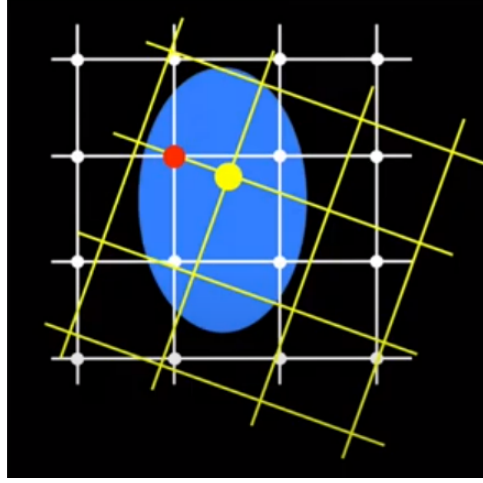


Figure 3.15: The red point, which belong to the moving space, is mapped by the transformation T in the yellow point. While the red point has a well-defined intensity value, since it corresponds to a grid point, the yellow point does not: the intensity value at this point is estimated by interpolation.

is necessary for the computation of new intensity values at every transformed grid point [157].

Image registration algorithms can broadly be classified in two categories according to the transformation models they use to relate the moving image space to the reference image space: affine transformations and elastic transformations.

The most general affine transformation is a composition of

- rotations
$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{pmatrix} \sin \theta & \cos \theta & 0 & 0 \\ -\cos \theta & \sin \theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$
- scaling
$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{pmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$
- translations
$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$
- shear
$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & a & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

and it is completely defined by 12 parameters.

Affine transformations action is global, as they are not functions of the application point. As a consequence, affine transformations cannot model local geometric differences between images.

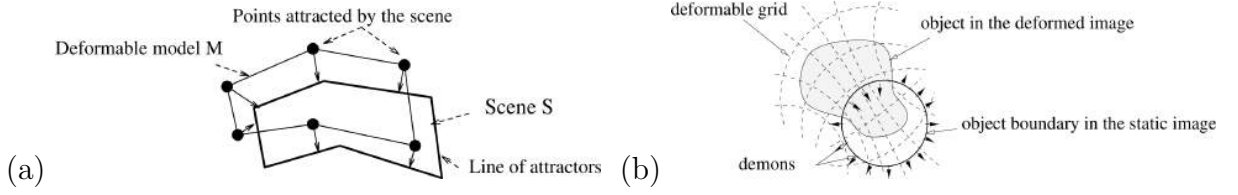


Figure 3.16: (a) Image registration as deformable model with attraction. (b) A deformed image, considered as a deformable grid, is diffusing through the contours of the objects in the static image, by the action of some effectors, called demons, situated on these interfaces.

On the contrary, elastic transformations are capable of locally warping the moving image in order to align it with the reference image.

As an example of elastic transformation, one of the most efficient methods is the Demons algorithm[149]. This algorithm consider non-rigid registration as a diffusion process where some entities (called demons) pushing voxels of the moving image according to local characteristics of the reference image, as illustrated in figure 3.16.

The forces which push voxels are inspired from the optical flow equations [19], and the registration procedure alternates several steps of computation of the forces and of regularization by means of a Gaussian smoothing. An improvement of the demons algorithm is the Diffeomorphic demons algorithm [155]. Diffeomorphisms are powerful assumptions in image registration, as they preserve the topology of objects. It is important to notice that the deformable registration process modifies the intensities of the moving image since it matches them on the target image. This means that in principle a perfect correspondence can be reached. However it must be taken into account that in this case the image information would be completely lost. Furthermore, diffeomorphisms are considered to be a good working hypothesis when no additional information about the spatial transformation is available. The image information can be recovered by analyzing the warp field through the Jacobian of the transformation.

The choice of the registration distance $d(f, T(m))$ (similarity measure $d(f, T(m))$) is a crucial issue in order to obtain a good image alignment. No distance is a priori better than another. The most widely used is cross-correlation equation (3.60), Mean Squared Error equation (3.61) and Mutual Information equation (3.62).

Let $f = I(T(m))$ the interpolated and transformed moved image. Let f_i and \tilde{f}_i the intensity related to i -th voxel ($i = 1 \dots N$), where N is the total number of voxels of the two images. The measures just mentioned are defined as follows

$$CC(\tilde{f}, f) = \frac{\sum_{i=1}^N (\tilde{f}_i - \bar{\tilde{f}})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^N (\tilde{f}_i - \bar{\tilde{f}})^2 \sum_{i=1}^N (f_i - \bar{f})^2}} \quad (3.60)$$

$$MSE(\tilde{f}, f) = \frac{1}{N} \sum_{i=1}^N (\tilde{f}_i - f_i)^2 \quad (3.61)$$

$$MI(\tilde{f}, f) = H(\tilde{f}) + H(f) - H(\tilde{f}, f) \quad (3.62)$$

where $H(f) = \int p_f(x) \log(p_f(x)) dx$ is the Shannon entropy and $H(\tilde{f}, f)$ is the joint entropy.

The Mutual Information is zero if the two sets f and \tilde{f} are independent, i.e. if $P(\tilde{f}, f) = P(\tilde{f})P(f)$.

By definition, CC metric has a maximum when images intensity are linearly related. If we consider two aligned images acquired with the same modality, we expect respective intensities are linearly related, as the two images convey the same information. Therefore, CC similarity measure may be a good choice in registering mono-modality images [69].

On the contrary, gray-level intensity values of medical images acquired in different modalities (such as MRI and PET) does not convey the same information. Then there is no reason to suppose a linear intensity relation between two different modality. This fact make CC not the best choice in registering multi-modality images. A better choice in this case could be the Mutual Information, because it can measures how much information is shared between images with respect to a probabilistic description of their respective intensities values.

The drawback of MI is its sensitive to noise. As noise in one or both images increases, the joint entropy of the two images increases, consequently the mutual information decreases [69].

Affine registration is generically used in aligning two images of the same patient (which can be acquired with two different modalities or not), while elastic registration is used in inter-subject registration in order to take into account the inter-subject anatomical variability.

Whichever registration algorithm and metric are chosen, a template or reference space is fundamental to standardize coordinates across many patients. One of the most commonly used space is the Montreal Neurological Institute (MNI) coordinate space¹.

3.6.1 ANTs software

Image registration is in practice performed using appropriate software which, through optimized algorithms, actually searches the best transformation which relates the moving and the fixed images, i.e. searches for the solution of the equation (3.58) or (3.59). Once the best transformation is found, the software applies it to the moving images, hence the registered image is finally obtained.

One of the most used medical image registration toolkit is ANTs (Advanced Normalization Tools) ², and during my PhD I used ANTs too in order register medical images, as discussed in section 4.2.

To achieve optimal registration, ANTs provides rigid, affine and diffeomorphic transformations which can be individually used as well as combined together. A rigid transformation is a particular type of affine transformation which not deform or scale the brain, it is just a roto-translation with no shear and no scaling.

¹Montreal Neurological Institute: <http://www.bmap.ucla.edu/portfolio/atlas>

²<http://stnava.github.io/ANTs/>

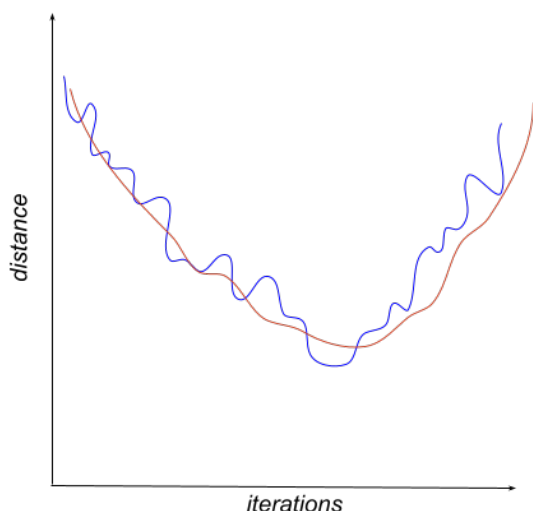


Figure 3.17: A naively illustration of a two-levels resolution approach for optimization: the blue curve represents the registration metric distance between a fixed and a moving non smoothed image, while the red curve represents the distance between a fixed and a moving smoothed image.

a sharper higher resolution version, and so on. At each level we use the transformation obtained from the previous level as starting position: as smoothing and sampling decreases, we get closer and closer to the global minimum with increasing precision.

The multi-level approach is superior to the single-scale version as it is much less likely to get trapped in a local minimum because of the smoothing effect of the pyramid, as schematically shown in figure 3.17. Furthermore, it is much faster because most iterations are performed at the coarsest resolution [151]

Each level is divided into a given number of iterations, which are the maximum number of iteration that the gradient descent can do in order to find the minimum.

ANTs allows a high level of customization of the registration algorithm through the choice of numerous parameters. The most important are

- the registration metric
- the transformation used (e.g. affine, diffeomorphic...)
- the number of levels and the maximum number of iterations per level
- the gradient step. In affine transformations gradient step state how big the linear shifts will be, while in diffeomorphic ones it tells the algorithm how much each point can move after each iteration.
- the shrink factor, used to control resolution
- the smoothing factor, used to manage smoothing

Focusing now on diffeomorphic transformation, after each iteration a gradient field which indicates how each voxel will shift in space is computed. This small deformation (or

The best transformation we are looking for is based on finding a global minimum of a distance/maximum of similarity measure (see eq. 3.58,3.59). However image registration is in general a non-convex optimization problem, thus a large number of local optima could be present. [112].

To try to overcome this optimization challenge, ANTS enables the use of a multi-level approach (also called resolution-pyramid-scheme) [112]: ANTs registration working principle is based on multiple resolution gradient descent framework [14]. ANTs improves the registration within each algorithm gradually at different resolutions, called *levels*. The ANTs multi-level approach working principle is described as follows: we start with "blurry" images (first level, low resolution, highly smoothed), register those to each other, then we go to the next level, with

updated gradient field) is combined with previous updates to form a total gradient deformation. Because each point can follow its own path, non-realistic deformations can occur, which may make images look like unrealistically stretched. To resolve this issue a penalty is usually added, such that shifts are not considered independently at each point. Thus, for non-rigid transformation only, we furthermore have two parameters which control the updated and the total gradient field:

- the update field variance, which is a parameter which serves as penalty on the updated field at each iteration. This parameter smooths the deformation computed on the updated gradient field, before this is added to previous deformations to form the total gradient field. Thus, for each point the deformation of neighboring points is taken into account as well, which avoids too much independent moving of points at each iteration (i.e. a point cannot move 2 voxels away in one direction if all its neighbors are moving 0.1 voxels away in the other direction).
- the total field variance, which is a parameter that serves as penalty on the total gradient field. It smooths the deformation computed on the total gradient field.

Once registration process has been computed, ANTs gives as output not only the registered image, but also the transformation used, and, in the case of diffeomorphic transformation, also the deformation field and its inverse.

Finally ANTs provides a tool called `AntsApplyTransform` which allows to apply a transformation previously computed and stored in PC memory.

3.7 Image semi-quantification

PET semi-quantification refers to all those algorithms which aim to measure the presence of a certain biomarker in some brain regions of interest (ROIs) using one (or more) static PET scan per subject.

Semi-quantification, which is also called relative quantification, differs from absolute quantification because it does not require kinetic models, dynamic scans and arterial blood samples.

Although semi-quantification is slightly less accurate than the absolute one, it has the great advantages of being both really simpler to achieve and less invasive for patients, so it is a widely used tool in medical imaging.

Semi-quantification allows us to estimate a biomarker both regionally and globally. Global semi-quantification refers to mean biomarker value across all the ROIs, while regional quantification refers to biomarker evaluation in each ROI.

ROIs are usually defined as regions of a brain atlas. A brain atlas is a spatial partition of a template in subsets which represent the ROIs.

Each ROI is uniquely defined by a gray-level intensity label.

As an example, the AAL atlas is shown in figure 3.18 on the next page

Regardless of the specific method, semi-quantification require two inputs (subject image and atlas) and provide semi-quantification values as output. More precisely, given a PET image P and an atlas $A = \{ROI_1, \dots, ROI_N\}$, a generic semi-quantification method

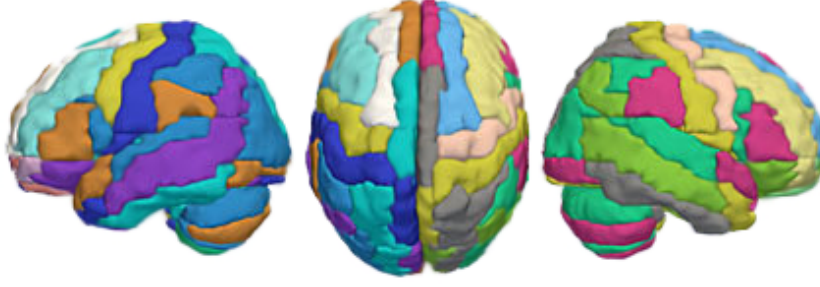


Figure 3.18: A pictorial 3 – D visualization of the AAL atlas

S is as a function

$$(P, A) \rightarrow S(P, A) = [s_{p1}, \dots, s_{pN}] \in \mathbb{R}^N \quad (3.63)$$

where s_{pi} is the semi-quantification value related to the i -th ROI of the p -th subject.

During my PhD I have dealt of semi-quantification: absolute quantification has been just mentioned above for completeness, but it will be never used or discussed anymore during this thesis.

Therefore, since there is no chance of misunderstanding and for the sake of simplicity, *I will from now on refer to semi-quantification just as quantification.*

In this thesis I focus and explain two quantification method: SUVR and ELBA.

3.7.1 Standardized Uptake Value ratio (SUVR)

The most used quantification method is the Standardized Uptake Value ratio (SUVR).

It is known that the radiotracer binds (aspecifically) to the biomarker. The amount of radiotracer at a given point is proportional to the image gray-level intensity at the same point. Thus it seems to be reasonable to define intensity-based measures. However the intensity range is not the same across different images, then a intensity normalization is required to compare intensities of different images. The intensity normalization problem is by fixed defining a reference region against which the intensity of any other ROI is assessed. Let R the ROI to quantify and let N a reference region. SUVR is an intensity based measure defined as

$$\text{SUVR} = \frac{f(I_{x \in R})}{f(I_{x \in N})} \quad (3.64)$$

where I_x is the intensity of the x -th voxel and where f is a function on the intensity which usually may be the mean, the median, generic n -th quantile, etcetera. In this thesis we set f as the mean, thus from now on SUVR will be defined as

$$\text{SUVR} = \frac{\frac{1}{n_R} \sum_{x \in R} I_x}{\frac{1}{n_N} \sum_{x \in N} I_x} \quad (3.65)$$

where n_R and n_N are the number of voxels which belong to R and to N respectively.

The reference ROI choice is quite arbitrary and constitutes an important issue. This ROI have to be as independent as possible from subjects clinical profile, thus it is

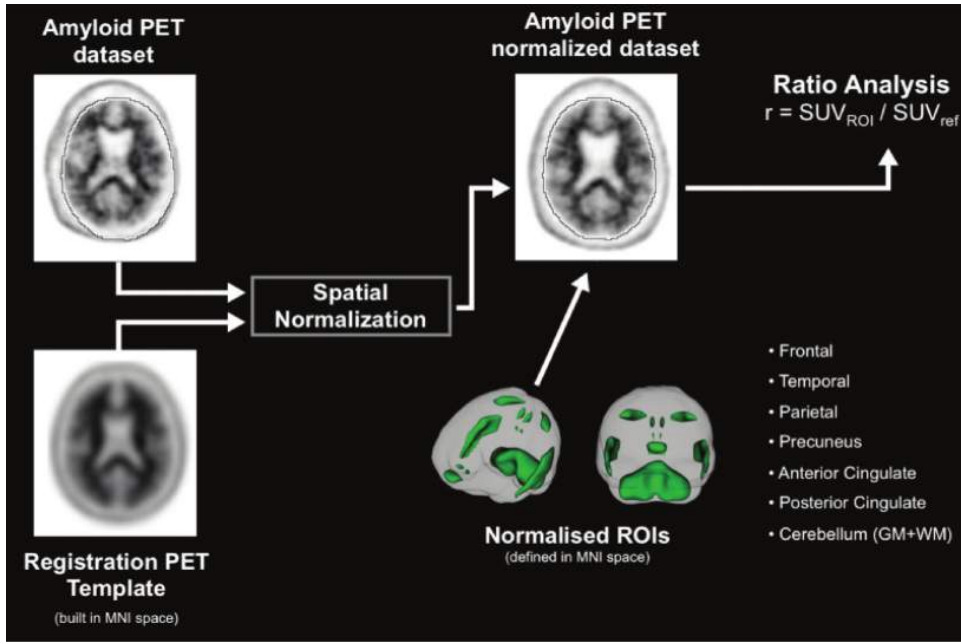


Figure 3.19: Schematic illustration of Standard Uptake Value ratio (SUVR) calculation: target PET is registered to a template image, and the uptake ratio is computed

usually chosen among those regions for which there is minimal specific binding of the tracer. Furthermore, the ability to correctly segment the chosen reference region must be taken into account, as a poor quality segmentation could lead to significant uncertainty in the count.

With regard to amyloid load quantification, there are several possible choices [134, 89] such as the encephalic trunk [89], the cerebellum GM [122, 89] or the entire cerebellum (both GM and WM) [35, 89, 104].

3.7.2 Evaluation of Brain Amyloidosis (ELBA)

Evaluation of Brain Amyloidosis (ELBA) is a fairly recent quantification measure [32]. ELBA is conceptually different from SUVR-like measures as it is not based on ROI intensity values, but rather on intensity distribution pattern.

The main advantage of quantifying the amyloid load using ELBA is that it does not require any reference region. The drawback of ELBA are basically two. While SUVR is a local measure of intensity, ELBA is an ensemble measure, so ELBA might lose precision in quantifying very small regions, as the number of voxels might not be sufficient to capture meaningful patterns. The second limitation is related to the appliWhile SUVR is a generic measure, ELBA is specific for amyloid quantification, as it is based on an empirical amyloid-specific observation.

The second limitation concerns the domain of applicability of SUVR and ELBA: while the former is a generic class of measures which can be used in many fields of medical imaging, the latter is specific for amyloid quantification, as it is based on an amyloid-specific working hypothesis.

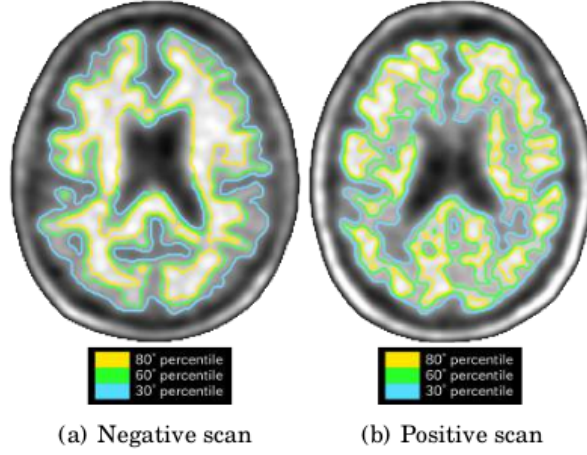


Figure 3.20: Three iso-intensity curves taken at different intensity values (quantile 30, 60 and 80 % of intensity distribution) for patients with low amyloid deposit (a) and for a patient with high and widespread amyloid deposit (b).

Here is the working hypothesis: *let us consider brain as a whole. Both geometrical appearances of iso-intensity surfaces and whole-brain intensity histograms are rather characteristic in typical negative and positive subjects. Positive subjects scans tend to show a sparser and more convoluted appearance of the iso-intensity surfaces with respect to negative scans*[32]. Figure 3.20 provides an example of what just said.

The aim of ELBA is to quantify these characteristics through two features: one that gauges the iso-intensity surface complexity and another that assess the histogram tendency toward higher/lower values for positive/negative scans.

Here I will describe the ELBA algorithm.

Geometric features Let B_i a brain region, let v the generical voxel $\in B_i$ and I_v the intensity of the voxel v .

First of all B_i is partitioned into n iso-intensity levels $0 < L_j < 1$ taken at equal quantile distances of the whole intensity distribution (n is typically set at 32 of 48).

Partitions consist of couples $\{s_j, V_j\}$, where s_j is surfaces and V_j is enveloped volumes defined as

$$V_j = \{v \in B_i \mid I_v \geq L_j\}$$

$$s_j = \sum_{v \in \partial V_j} 1 \quad (3.66)$$

$$(3.67)$$

where the ∂V_j denotes the boundary of V_j . As the sum have been computed on ∂V_j , s_j represents the number of voxels on the V_j boundary.

Let us consider the volume V_j of the j -th partition. The radius r_v^j of the equivalent sphere of volume V_j is

$$r_j^v = \left(\frac{3V_j}{4\pi} \right)^{\frac{1}{3}} \quad (3.68)$$

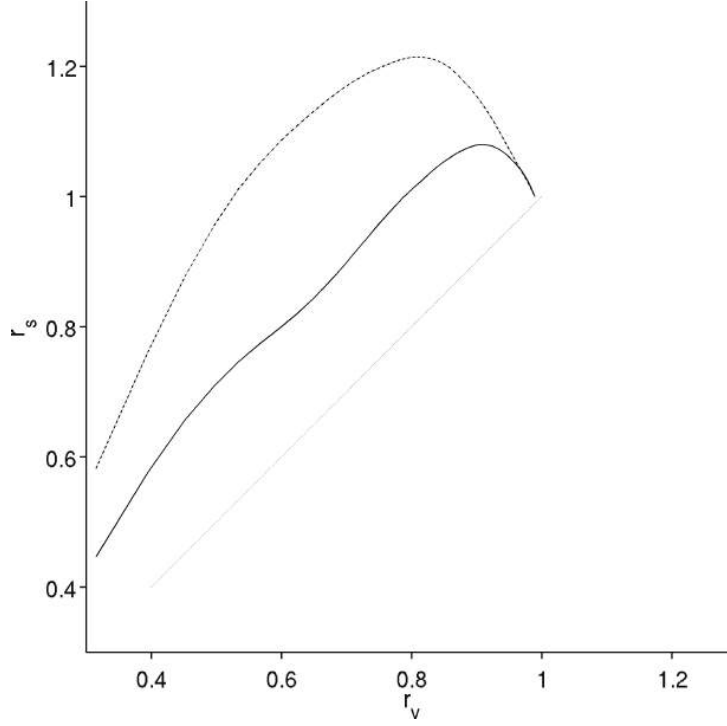


Figure 3.21: This plot illustrates the characteristic curve (r_v, r_s) for low/high amyloid burden scan (thick/dotted curve respectively). Values are normalized to the respective brain volume and boundary. The thin line is the bisector.

while the equivalent sphere having the same surface extent as s_j is

$$r_j^s = \left(\frac{s_j}{4\pi} \right)^{\frac{1}{2}} \quad (3.69)$$

Considering equations equation (3.69) and equation (3.68) on the previous page for all partitions, (i.e. for all $j = 1 \dots n$), we get a set of ordered pairs $\mathcal{R} = \{(r_1^v, r_1^s), \dots, (r_n^v, r_n^s)\}$. Plotting \mathcal{R} on a Cartesian plane lead us to obtain a characteristic curve inferiorly bounded by the bisector line $r_v = r_s$. The bisector line represents the limit for all s_j being actual spheres.

The area \mathcal{A} included between the bisector line $r_s = r_v$ and the characteristic curve $r^s(r^v)$ increases as s_j became rougher and notched.

When we subtracted the trivial bisector line, typically positive scans show a higher surface-to-volume ratio on the higher intensity levels (low r^v) with respect to the lower intensity levels (high r^v), and viceversa for negative scans.

The characteristic curve is integrated without the bisector area on the lower and higher half of its domain D (i.e. the range of r^v) to deliver the geometric feature G_i :

$$G_i = \frac{\int_{D1} (r^s(r) - r) dr}{\int_{D2} (r^s(r) - r) dr} \quad (3.70)$$

where $D1 = [\min(r^v), r^v/2]$ and $D2 = [r^v/2, \max(r^v)]$.

Intensity features The intensity feature is based on the intensity values in Bi . To characterize the different trends of intensity histograms in positive and negative subjects

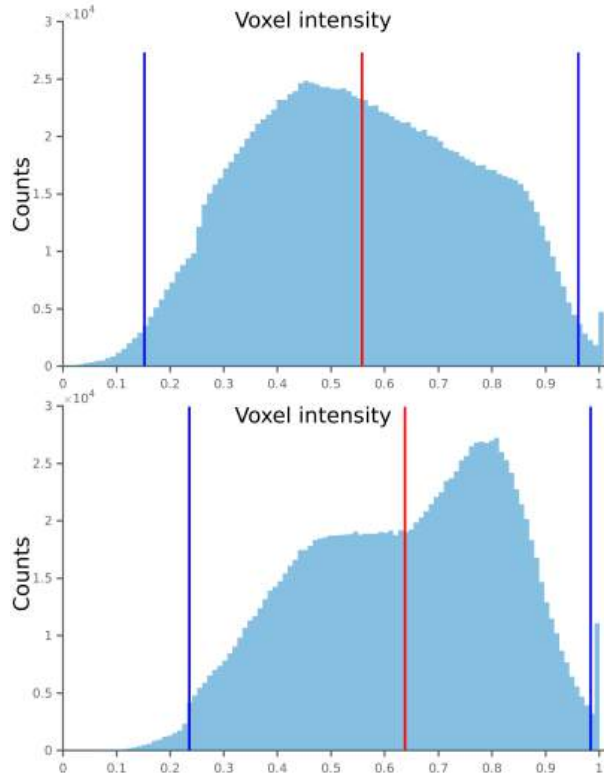


Figure 3.22: Comparison between the intensity histograms (cortical region) in the scan of a subject with limited amyloid burden (above) and one with large load (below). The tendency of the distribution toward high intensity values in the positive scan is evident

(see figure 3.22, we used the distance between the average intensity I_{mean} and the two intensity percentiles $q_1 = 1\%$ and $q_{99} = 99\%$ as described by the ratio

$$C_i = \frac{I_{mean} - q_1}{q_{99} - I_{mean}} \quad (3.71)$$

ELBA score The two adimensional image features G_i and C_i are combined with the geometric mean to provide the ELBA score for B_i which is defined as

$$E_i = \sqrt{G_i C_i} \quad (3.72)$$

3.8 Computing

Computational neuroimaging tools require substantial computational resources and the increasing availability of large image databases will further enhance this need [127].

In this context, a really effective and often indispensable solution when working with neuroimaging data is provided by parallel computing and outsourced analysis.

Parallel processing is the execution of program instructions by dividing them among multiple processing units (CPUs), this allows to remarkably reduce the running times, as jobs are performed concurrently instead of in sequence.

The typical parallel execution can be easily made nowadays on common multi-core desktop PCs, where different code fractions are taken on independently by each available

CPU.

Following the same principle, but adapting it on a large scale, it is possible to distribute jobs on high computing performance interconnected architectures: this allows to leverage the collective computational power of multiple machines, which are called *nodes*.

What was said represents the key-point of *clusters* and *grids*. A cluster is a group of nodes that are connected by a high bandwidth and low latency local area network (LAN). Clusters are often informally called *farm*.

Grids are larger distributed system solutions: a grid is typically a collection of interconnected and de-localized clusters owned by different institutes to leverage the collective computational power of these shared resources.

All the algorithms described in this thesis have been executed on a dedicated server farm within the INFN-GE ICT infrastructure. The ensemble counts 7 multi-CPU units (100 cores overall) with 2 Gb RAM per core.

A shared file system is available across the cluster, via fiber channel, with 10 Tb of disk space.

All servers are running Scientific Linux release 6.8 (Carbon) operating system, and have been equipped with several packages and libraries focused on neuroimaging such as ANTS and AntsRegistration³, Freesurfer⁴, ITK⁵ as well as a licensed calculus software Matlab⁶.

LONIpipeline⁷ The simple way with which we implemented parallel processing scheme on this network of servers is based on the LONI Pipeline Processing Environment [49].

The LONI Pipeline is a free distributed system for designing, executing, monitoring and sharing scientific workflows on grid computing architectures [128].

LONI Pipeline is a free distributed cross-platform java-based environment for designing, executing and monitoring scientific workflows on grid/cluster computing architectures.

Thanks to this environment it is possible to design a block programming paradigm where each block (or node) identifies a single executable. Streams can be designed where nodes are concatenated as directed graphs receiving input from former modules and sending output to following ones, as you can see in figure 3.23. Any command-line driven processing program or routine can be represented as a module, and submitted for execution to the different machines composing the calculus infrastructure.

³<http://picsl.upenn.edu/software/ants>

⁴<https://surfer.nmr.mgh.harvard.edu>

⁵<https://itk.org>

⁶<https://it.mathworks.com/products/matlab.html>

⁷<https://pipeline.loni.usc.edu/>

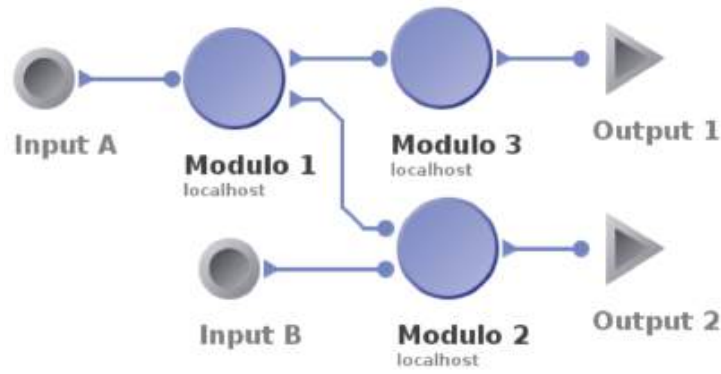


Figure 3.23: A basic LONI Pipeline workflow. Data streams flow from left to right: output of previous modules (Module 1) is fed as input to following ones (Module 2, Module 3). The final yields of the chain are comfortably collected in data-sink local folders Output 1 and Output 2.

LONI Pipeline responds to the scheme:

client job submission \longrightarrow remote analysis \longrightarrow client output retrieving

this approach saves client machines from all of the computational burden, and users from waiting for jobs completion connected.

LONI Pipeline grants jobs independent execution: the same steps are performed for each single image without any interaction.

One of the main advantages of LONI Pipeline is that it allows users to surgically access and interact with workflow execution collecting information, status and intermediate results at each level of execution as well as pausing/interrupting it at will.

LONI Pipeline parallel execution allows to process a large quantity of images ($\sim 10^3$) at the same time. This leads to a huge amount of saving time, as the computational time to process many images is roughly the same of a single-image execution on a regular desktop PC. For all these reasons, LONIpipeline is widely used in neuroimaging case studies, as you can see in [33, 34, 31, 30].

3.9 Database

3.9.1 Database building and managing

The practice of data sharing is growing in society, particularly in the scientific community, as vast amounts of data continue to be acquired [117]. With the rapid advances being made in neuroimaging technology, data acquisition, and computer networks the successful organization and management of neuroimaging data has become more important than ever before[152].

As neuroimaging databases have grown in size and complexity, an effort to quickly and efficiently store, manage and browsing data is required.

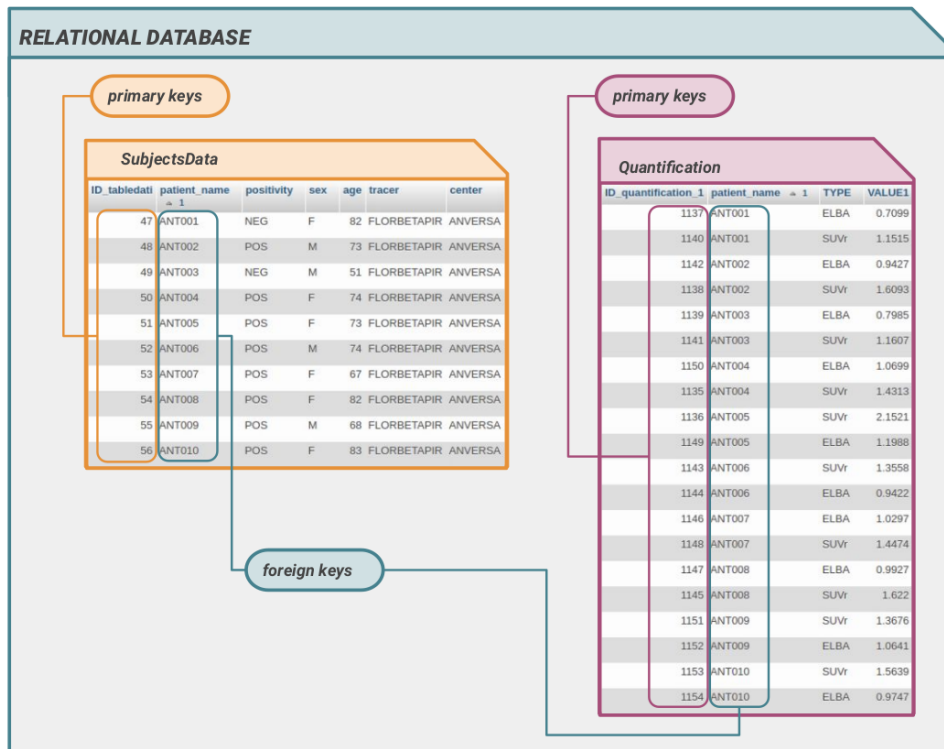


Figure 3.24: *SubjectsData* and *Quantification* are two tables of a relational database. The column *patient name* contain the foreign keys, which easily allow to relate the data stored in the two tables. Both the primary keys uniquely identify rows (i.e. data record) in each table.

A database can be defined as an organized collection of data, which is usually stored and accessed electronically from a computer system. The access to a database is usually provided by a database management system (DBMS), which is a software system that enables users to define, create, maintain and control access to the database [38]. There are different typologies of databases (hierarchical, graphical, relational etc...), but the relational one is probably the most popular. A relational database is managed by a relational database management system (RDBMS) and it can be defined as a database based on the relational model of data [36].

More in detail, a relational database is characterized by the presence of two elements: *table* and *keys*.

Tables are the "containers" where all the data is stored, and they are made up of rows and columns. Each column represents a specific datatype (attribute), while each row represent a data record.

Keys define the relational structure of a database and they are divided into *primary keys* and *foreign keys*.

Primary keys are defined independently for each table: for each rows in a table, there exist one and only one primary key value (i.e. if a table has n rows, then it has also n different primary key values).

Foreign keys allow to relate the information in one table to that in another table.

How a RDBMS works is graphically explained in figure 3.24.

MySQL MySQL is a widely used RDBMS based on SQL (Structured Query Language) programming language. MySQL has many advantages: it is easy to use, it is free and open source, it is reliable (has been around since 1995) and secure (you need to be logged

through a password to browsing the database, and the password itself is encrypted in MySQL), and it has a large community of developers who can help answer questions. Furthermore it can be accessed using phpMyAdmin, as I will better explain in the next paragraph.

Users with administrator privileges can import data and create, copy, drop, rename and modify databases tables, while other users are allowed to browse the database only after a username and password login.

phpMyAdmin phpMyAdmin is one of the most popular MySQL administration tools: it is a free software written in PHP which allows you to administer a MySQL database via any browser.

The software is intended for both database administrators and users, who can access the database after logging in through user name and password.

An intuitive web interface enables all logged users, according to their privileges, to browse databases and tables, to import/export data to various formats (CSV, SQL, XML, Open-Document Text and Spreadsheet, Word and LATEX) and to manage databases, tables, columns, users and permissions.

However, you still have the ability to directly execute any SQL statement through SQL queries via PhpMyAdmin: creating complex queries can be made easier through the use of Query-by-example (QBE).

Chapter 4

Materials

4.1 AmyDB Database

Using MySQL I built a relational database, called AmyDB, where I stored all the neuroimaging data made available to me. AmyDB can be also accessed via phpMyAdmin.

AmyDB is multicentric and multitracer amyloid-PET database made of 1001 subjects from a naturalistic population. Data were provided by 14 EADC (European Alzheimer's Disease Consortium) clinical centers and other 7 memory clinics and excellence centers for Alzheimer Disease.

Images have been acquired using 4 different tracers (PIB (146), Florbetaben (280), Florbetapir (369), and Flutemetamol (190)). Each subject has at least one amyloid-PET late scan.

DICOM files have been provided for 562 subjects. However DICOM files do not give us a complete knowledge about image acquisition and reconstruction for all subjects: acquisition scanners (manufacturer and model) are known for 543 subjects, while PET reconstruction methods as described in the DICOM files are known for 514 subjects.

Of these 514, detailed reconstruction information are given for 290 images only (see table 4.4), while for the other images only roughly reconstruction information are known: the type of reconstruction (iterative vs analytical), the inclusion of PSF in reconstruction method and the use of TOF in photons detection.

The global cortical $A\beta$ load of each late scan has been assessed by 4 independent expertise nuclear medicine physicians through a dichotomic evaluation (POS/NEG label). It is important to point out that even though database subjects are labelled by a POS/NEG amyloid assessment label, they can not be respectively considered as case and controls. This is because controls needs to be healthy subjects: as PET scanning is an invasive procedure, in clinical practice PET examination is usually performed only on subjects with symptoms and clinical pictures which justify such an exam.

Moreover basic demographics and neuropsychological assessments are provided for all subjects: age, sex, MMSE (Mini Mental State Examination), Education (years). Further neuropsychological details were not available for the whole dataset and have been discarded in this analysis.

All these demographic data are summarized in tables 4.1.

Inclusion-exclusion database criteria could be different center by center, as they depend both on country own's legislation and at least on physician's opinion. This fact leads to a between-centers heterogeneous demography, size samples and clinic (see table 4.1).

Furthermore, it is important to notice that AmyDB is also very heterogeneous with regard to images acquisition: images have been acquired with different scans and acquisition protocols, as illustrated in tables 4.1 and 4.2.

Images belonging to the same clinical center are usually acquired with the same scanner and with same reconstruction method, even though this is not always true, as illustrated in tables 4.3 and 4.4.

Finally, 198 images have been evaluated regarding their quality by an experienced nuclear medicine physician: three dichotomic labels "High Quality"(HQ)/"Low Quality"(LQ)for evaluating three different quality facets of images have been provided. Quality labels will be used in chapter 6.

Center	Tracer	Sample	Sex(F-M)	Amy(N-P)	Age	MMSE	Education
GNV	BEN	15	6 - 9	5 - 10	72.60 (7.80) [55.00 - 82.00]	27.67 (1.95) [24.00 - 30.00]	10.53 (5.17) [0.00 - 18.00]
HSR	BEN	49	28 - 21	26 - 23	69.80 (7.68) [51.00 - 87.00]	25.62 (3.36) [16.00 - 30.00]	10.91 (4.81) [0.00 - 17.10]
MAN	BEN	32	15 - 17	13 - 19	65.78 (9.78) [48.00 - 85.00]	24.39 (4.16) [13.00 - 30.00]	11.53 (4.30) [3.00 - 21.00]
PDV	BEN	86	40 - 46	54 - 32	69.31 (9.92) [43.00 - 86.00]	25.85 (3.17) [16.00 - 30.00]	12.90 (4.85) [0.00 - 30.00]
FBB	BEN	9	4 - 5	3 - 6	68.00 (6.73) [56.00 - 77.00]	24.36 (3.62) [18.26 - 28.00]	12.33 (5.15) [4.00 - 18.00]
UPG	BEN	38	22 - 16	21 - 17	68.21 (6.30) [53.00 - 82.00]	26.15 (4.64) [16.00 - 30.00]	13.11 (4.42) [4.00 - 21.00]
TVG	BEN	51	38 - 13	33 - 18	69.67 (6.69) [56.00 - 80.00]	25.56 (4.21) [13.00 - 30.00]	12.74 (4.13) [3.60 - 19.30]
ANT	PIR	74	35 - 39	31 - 43	71.55 (7.53) [51.00 - 85.00]	25.07 (3.68) [15.00 - 30.00]	15.69 (4.39) [5.00 - 30.00]
BRE	PIR	75	42 - 33	36 - 39	71.09 (6.63) [54.00 - 84.00]	25.32 (3.12) [18.00 - 30.00]	10.89 (4.98) [0.00 - 21.00]
GEN	PIR	57	27 - 30	19 - 38	72.93 (5.01) [57.00 - 83.00]	26.07 (3.62) [13.00 - 30.00]	9.86 (4.64) [0.00 - 21.00]
GNV	PIR	41	21 - 20	26 - 15	72.48 (7.81) [55.00 - 87.00]	25.27 (4.31) [13.00 - 30.00]	13.47 (4.06) [0.00 - 20.00]
MAR	PIR	26	14 - 12	17 - 9	77.77 (7.34) [68.00 - 90.00]	27.62 (4.16) [12.00 - 30.00]	11.54 (3.08) [7.00 - 17.00]
MON	PIR	14	8 - 6	3 - 11	71.93 (5.20) [64.00 - 82.00]	21.21 (3.83) [16.00 - 26.00]	9.29 (4.05) [5.00 - 18.00]
CUN	PIR	33	11 - 22	11 - 22	69.24 (7.10) [48.00 - 82.00]	27.00 (2.18) [18.00 - 30.00]	10.33 (4.41) [5.00 - 17.00]
PAV	PIR	19	10 - 9	5 - 14	75.89 (5.85) [65.00 - 87.00]	24.10 (5.72) [13.00 - 30.00]	7.43 (3.83) [3.60 - 17.00]
PER	PIR	10	7 - 3	5 - 5	76.70 (2.58) [72.00 - 80.00]	25.70 (4.08) [19.00 - 30.00]	7.50 (3.72) [4.00 - 13.00]
UBS	PIR	20	12 - 8	11 - 9	64.75 (8.88) [51.00 - 81.00]	25.76 (2.82) [19.00 - 29.00]	12.61 (3.90) [5.00 - 17.10]
GEN	MOL	15	8 - 7	5 - 10	69.67 (7.68) [54.00 - 79.00]	26.73 (3.17) [18.00 - 30.00]	10.60 (4.34) [5.00 - 17.00]
PAR	MOL	44	26 - 18	37 - 7	62.09 (8.21) [41.00 - 82.00]	27.86 (1.81) [23.00 - 30.00]	13.37 (3.46) [6.00 - 20.00]
PRT	MOL	97	49 - 48	29 - 68	69.57 (8.85) [48.55 - 87.98]	23.59 (4.74) [13.00 - 30.00]	12.48 (5.16) [3.60 - 30.00]
ROM	MOL	18	8 - 10	6 - 12	63.78 (7.86) [48.00 - 79.00]	23.08 (7.00) [4.00 - 30.00]	12.00 (3.36) [5.00 - 16.00]
FBB	MOL	14	11 - 3	6 - 8	79.14 (4.37) [68.00 - 84.00]	23.38 (4.98) [13.00 - 29.00]	11.50 (4.00) [4.00 - 19.30]
UBS	MOL	2	1 - 1	0 - 2	71.00 (5.66) [67.00 - 75.00]	20.50 (6.36) [16.00 - 25.00]	13.15 (8.70) [7.00 - 19.30]
COI	PIB	68	34 - 34	27 - 41	65.51 (7.23) [49.00 - 77.36]	23.03 (5.87) [4.00 - 30.00]	8.49 (4.75) [2.00 - 15.00]
LIS	PIB	78	52 - 26	21 - 57	65.72 (7.69) [43.00 - 81.00]	23.95 (3.89) [14.00 - 30.00]	12.69 (3.99) [4.00 - 17.00]

Tracer	Sample	Sex(F-M)	Amy(N-P)	Age	MMSE	Education
BEN	280	153 - 127	155 - 125	69.04 (8.40) [43.00 - 87.00]	25.68 (3.74) [13.00 - 30.00]	12.25 (4.67) [0.00 - 30.00]
PIR	369	187 - 182	164 - 205	72.02 (7.29) [48.00 - 90.00]	25.52 (3.84) [12.00 - 30.00]	11.74 (4.97) [0.00 - 30.00]
MOL	190	103 - 87	83 - 107	68.02 (9.33) [41.00 - 87.98]	24.76 (4.77) [4.00 - 30.00]	12.43 (4.55) [3.60 - 30.00]
PIB	146	86 - 60	48 - 98	65.62 (7.45) [43.00 - 81.00]	23.51 (4.93) [4.00 - 30.00]	10.73 (4.83) [2.00 - 17.00]

Table 4.1: Top table: demographic data per center and per tracer. Bottom table: demographic data per tracer. Amy(N-P) is the number of negative and positive subjects respect to $A\beta$ load. BEN, PIR, MOL stand for Florbetaben, Florbetapir and FLutemetamol respectively.

Rec-Scan	Tracer	Sample	Sex(F-M)	Amy(N-P)	Age	MMSE	Education
IPT -BGR40mCT	BEN	9	4 - 5	3 - 6	68.00 (6.73) [56.00 - 77.00]	24.36 (3.62) [18.26 - 28.00]	12.33 (5.15) [4.00 - 18.00]
I -1080	BEN	15	6 - 9	5 - 10	72.60 (7.80) [55.00 - 82.00]	27.67 (1.95) [24.00 - 30.00]	10.53 (5.17) [0.00 - 18.00]
UKN -DSC 690	BEN	21	12 - 9	10 - 11	69.52 (8.30) [51.00 - 87.00]	25.67 (2.43) [21.00 - 30.00]	12.06 (4.54) [4.00 - 17.10]
UKN -DSC STE	BEN	27	16 - 11	15 - 12	69.89 (7.44) [54.00 - 81.00]	25.61 (4.05) [16.00 - 30.00]	10.09 (5.00) [0.00 - 17.10]
I -DSC STE	BEN	45	34 - 11	31 - 14	69.64 (6.92) [56.00 - 80.00]	26.18 (3.62) [18.00 - 30.00]	13.20 (3.94) [3.60 - 19.30]
I -DSC ST	BEN	25	14 - 11	14 - 11	67.60 (7.30) [53.00 - 82.00]	25.95 (4.80) [16.00 - 30.00]	14.08 (4.58) [4.00 - 21.00]
IP -BGR40mCT	BEN	29	13 - 16	11 - 18	65.08 (10.64) [48.00 - 85.00]	24.33 (4.23) [13.00 - 30.00]	12.17 (3.97) [4.00 - 21.00]
I -BGRmMR	BEN	73	36 - 37	47 - 26	69.48 (9.80) [43.00 - 84.00]	25.83 (3.25) [16.00 - 30.00]	12.93 (4.94) [0.00 - 30.00]
IPT -BGR40mCT	PIR	74	41 - 33	36 - 38	71.14 (6.66) [54.00 - 84.00]	25.27 (3.13) [18.00 - 30.00]	10.85 (5.00) [0.00 - 21.00]
IPT -DSC 690	PIR	20	13 - 7	10 - 10	65.05 (8.89) [51.00 - 81.00]	25.76 (2.82) [19.00 - 29.00]	12.46 (3.78) [5.00 - 17.10]
IPT -BGR64mCT	PIR	5	1 - 4	0 - 5	67.60 (8.08) [60.00 - 80.00]	27.40 (1.52) [26.00 - 29.00]	13.40 (3.51) [8.00 - 17.00]
IPT -1080	PIR	27	9 - 18	11 - 16	69.67 (7.13) [48.00 - 82.00]	26.96 (2.33) [18.00 - 30.00]	9.93 (4.39) [5.00 - 16.00]
I -1080	PIR	57	27 - 30	19 - 38	72.93 (5.01) [57.00 - 83.00]	26.07 (3.62) [13.00 - 30.00]	9.86 (4.64) [0.00 - 21.00]
IPT -BGR128mCT	PIR	41	21 - 20	26 - 15	72.48 (7.81) [55.00 - 87.00]	25.27 (4.31) [13.00 - 30.00]	13.47 (4.06) [0.00 - 20.00]
I -DSC ST	PIR	10	7 - 3	5 - 5	76.70 (2.58) [72.00 - 80.00]	25.70 (4.08) [19.00 - 30.00]	7.50 (3.72) [4.00 - 13.00]
IP -DSC 600	PIR	12	8 - 4	2 - 10	71.67 (5.57) [64.00 - 82.00]	20.50 (3.66) [16.00 - 25.00]	9.50 (4.36) [5.00 - 18.00]
I -DSC 600	PIR	2	0 - 2	1 - 1	73.50 (2.12) [72.00 - 75.00]	25.50 (0.71) [25.00 - 26.00]	8.00 (0.00) [8.00 - 8.00]
IPT -UKN	PIR	19	10 - 9	5 - 14	75.89 (5.85) [65.00 - 87.00]	24.10 (5.72) [13.00 - 30.00]	7.43 (3.83) [3.60 - 17.00]
IPT -BGR40mCT	MOL	13	11 - 2	6 - 7	79.15 (4.54) [68.00 - 84.00]	23.25 (5.17) [13.00 - 29.00]	11.38 (4.14) [4.00 - 19.30]
IT -BGR64mCT	MOL	18	8 - 10	6 - 12	63.78 (7.86) [48.00 - 79.00]	23.08 (7.00) [4.00 - 30.00]	12.00 (3.36) [5.00 - 16.00]
I -1080	MOL	15	8 - 7	5 - 10	69.67 (7.68) [54.00 - 79.00]	26.73 (3.17) [18.00 - 30.00]	10.60 (4.34) [5.00 - 17.00]

Reconstr	Tracer	Sample	Sex(F-M)	Amy(N-P)	Age	MMSE	Education
IPT	BEN	9	4 - 5	3 - 6	68.00 (6.73) [56.00 - 77.00]	24.36 (3.62) [18.26 - 28.00]	12.33 (5.15) [4.00 - 18.00]
IP	BEN	29	13 - 16	11 - 18	65.17 (9.89) [48.00 - 85.00]	24.48 (4.08) [13.00 - 30.00]	11.93 (4.30) [3.00 - 21.00]
I	BEN	158	90 - 68	97 - 61	69.53 (8.52) [43.00 - 84.00]	26.14 (3.53) [16.00 - 30.00]	12.96 (4.68) [0.00 - 30.00]
IPT	PIR	186	95 - 91	88 - 98	70.96 (7.62) [48.00 - 87.00]	25.54 (3.66) [13.00 - 30.00]	11.18 (4.74) [0.00 - 21.00]
IP	PIR	12	8 - 4	2 - 10	71.67 (5.57) [64.00 - 82.00]	20.50 (3.66) [16.00 - 25.00]	9.50 (4.36) [5.00 - 18.00]
I	PIR	70	34 - 36	26 - 44	73.33 (4.99) [57.00 - 83.00]	26.03 (3.59) [13.00 - 30.00]	9.57 (4.56) [0.00 - 21.00]
IPT	MOL	14	11 - 3	6 - 8	78.29 (5.44) [67.00 - 84.00]	22.69 (5.34) [13.00 - 29.00]	11.95 (4.50) [4.00 - 19.30]
IT	MOL	18	8 - 10	6 - 12	63.78 (7.86) [48.00 - 79.00]	23.08 (7.00) [4.00 - 30.00]	12.00 (3.36) [5.00 - 16.00]
I	MOL	17	9 - 8	5 - 12	70.53 (7.62) [54.00 - 79.00]	26.53 (3.02) [18.00 - 30.00]	10.53 (4.20) [5.00 - 17.00]

Table 4.2: Top table: demographic data per roughly reconstruction information, scan model and tracer. Bottom table: demographic data per roughly reconstruction information and tracer. Regarding reconstruction information, I stands for iterative method, P denote the inclusion of PSF in reconstruction and T denote the using of TOF technology to detect photons. With regard to scan model, BGR stands for Biography, while DSC stand for Discovery. BEN, PIR, MOL stand for Florbetaben, Florbetapir and FLutemetamol respectively, while Amy(N-P) is the number of negative and positive subjects respect to $A\beta$ load.

Rec - Scan	BRE	GEN	GNV	HSR	MAN	MON	PDV	CUN	PAV	PER	ROM	FBB	UBS	UPG	TVG
IPT - BGR40mCT	74	0	0	0	0	0	0	0	0	0	0	22	0	0	0
IPT - DSC 690	1	0	0	0	0	0	0	0	0	0	0	0	20	0	0
IPT - BGR64mCT	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
IPT - 1080	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0
IT - BGR64mCT	0	0	0	0	0	0	0	1	0	0	18	0	0	0	0
I - BGR40mCT	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
I - 1080	0	87	0	0	0	0	0	0	0	0	0	0	0	0	0
IPT - BGR128mCT	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0
UKN - DSC 690	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0
UKN - DSC STE	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0
I - DSC STE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45
I - DSC ST	0	0	0	0	0	0	0	0	0	10	0	0	0	25	0
I - DSC 690	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
IP - BGR40mCT	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0
IP - DSC 600	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0
I - DSC 600	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
IPT - UKN	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0
I - BGRmMR	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0

Scanner	BRE	GEN	GNV	HSR	MAN	MON	PAD	CUN	PER	ROM	FBB	UBS	UPG	TVG
BGR40mCT	74	0	0	0	29	0	0	0	0	0	23	0	0	0
DSC 690	1	0	0	21	0	0	0	0	0	0	0	22	0	0
BGR64mCT	0	0	0	0	0	0	0	6	0	18	0	0	0	0
1080	0	87	0	0	0	0	0	27	0	0	0	0	0	0
BGR128mCT	0	0	41	0	0	0	0	0	0	0	0	0	0	0
DSC STE	0	0	0	27	0	0	0	0	0	0	0	0	0	45
DSC ST	0	0	0	0	0	0	0	0	10	0	0	0	25	0
DSC 600	0	0	0	0	0	14	0	0	0	0	0	0	0	0
BGRmMR	0	0	0	0	0	0	73	0	0	0	0	0	0	0

Reconstr	BRE	GEN	GNV	MAN	MON	PDV	CUN	PAV	PER	ROM	FBB	UBS	UPG	TVG
IPT	75	0	41	0	0	0	32	19	0	0	22	20	0	0
IP	0	0	0	29	12	0	0	0	0	0	0	0	0	0
IT	0	0	0	0	0	0	1	0	0	18	0	0	0	0
I	0	87	0	0	2	73	0	0	10	0	1	2	25	45

Table 4.3: Tables describing the numerosity of the couple (reconstruction information, scanner) versus clinical centers (top), of the scanner versus clinical centers (middle), of reconstruction information versus clinical centers(bottom).

It should be noted a strong relation between scanner-center and reconstruction information and center I,P,T,BGR and DSC have been defined in table 4.2.

Reconstruction Details	BRE	GEN	GNV	MAN	CUN	ROM	FBB
PSF+TOF3i21s	74	0	0	0	32	0	0
OSEM3D+TOF3i21s	0	0	0	0	1	0	0
PSF+TOF4i21s	0	0	0	0	0	0	13
OSEM3D5i24s	0	0	0	0	0	0	1
PSF+TOF5i21s	0	0	41	0	0	0	9
OSEM2D4i14s	0	6	0	0	0	0	0
OSEM2D6i16s	0	27	0	0	0	0	0
OSEM2D2i24s	0	5	0	0	0	0	0
OSEM2D4i16s	0	49	0	0	0	0	0
PSF3i24s	0	0	0	29	0	0	0
OSEM3D+TOF4i21s	0	0	0	0	0	18	0

Table 4.4: Detailed reconstruction information are known for 290 images. In this table you can find the numerosity of detailed reconstruction methods versus clinical centers. Here "i" and "s" stand for iteration and subset respectively.

In the final part of this section I will describe the database structure. AmyDB consists of 3 tables:

- **Subjects.** This table is basically used to summarize all the database subjects as well as to define the foreign keys once for all. It contains all the anonymized subject code-names which will be used as foreign keys
- **SubjectsData.** This table contains:
 - patients basic demographic and clinical information at the time of image acquisition
 - PET images file paths (raw paths, registered paths, DICOM file paths, and transformation matrices/warp fields related to each registration process has been done)
 - three dichotomic quality labels (HQ/LQ) for the assessment of three different facets of scans quality
 - a two-value label which serves as registration quality flag
 - data acquisition information and provenance, such as acquisition center, tracer, used scanner and image reconstruction protocol (when available)
- **Quantification.** All the quantification information are stored in this table: each data record contains the global and regional semi-quantification values computed using ELBA and SUVr methods (described in sections 3.7.1 and 3.7.2) as well as the atlas used to quantify

4.2 Image preprocessing

In this section I describe the image preprocessing which basically consists of two steps: image registration and image quantification. These two steps are necessary to organize the raw data (i.e. DICOM files or raw images (see section 2.7)) into quantification matrices that allow us to perform subsequent analyses.

4.2.1 Image preprocessing overview

Before describing in detail the registration and quantification algorithms I used, I will describe the relation between registration and quantification. This will allow a better understanding of the registration and quantification algorithms and the choice of related parameters that will be discussed in the next sections.

The goal of my registration and quantification algorithm is to register each subject's image in the MNI space using as fixed image the ICBM152 template¹(which is an MRI structural template) and then to quantify images using two brain atlases and two different quantification method, ELBA and SUVr.

¹In this text, the acronyms MNI and ICBM152 are used indiscriminately to identify the template-space and the template image itself.

Registration and quantification in case of patient's structural images are available Having at disposal, in addition to PET, also patient's structural image (e.g. MRI) allows to perform an optimal registration and quantification process, which is described here below

- Let consider a specific subject. The subject's PET raw image is co-registered on the subject's MRI raw image I_M through an affine transform. We denoted the co-registered PET image by I_P .
- Then I_M is mapped into the template using an affine transform denoted by A , obtaining the affine registered image $I_{MA} = A(I_M)$
- The affine image I_{MA} is mapped into the template using a diffeomorphic transform W , leading us to obtain the diffeomorphic registered image $I_{MW} = W(I_{MA})$. We notice that both A and W are invertible maps which relate the patient's anatomical structures and those one of the template.
- Then we apply the affine transform A on the PET image I_P . We obtain the PET affine registered image $I_{PA} = A(I_P)$. This is the image we will quantify.
- Let R a given atlas which we would like to use for quantification. Typically R is a parcellation of the template, hence it is embedded in MNI space and it is by definition aligned with the template.
- Let consider the W previously computed. Applying the inverse transform W^{-1} we are able to map the atlas R on I_{PA} , obtaining $R_A = W^{-1}(R)$. We emphasize that R_A is the patient-like representation of the atlas R .
- Finally quantification can be performed using the couple (I_{PA}, R_A) as generically described in function 3.63

This algorithm may seem cumbersome, but it is necessary to get a good quantification result. Quantifying using the couple (R, I_A) instead of using (R_A, I_A) is not a good choice at all. Each patient has its own individual local variability which remains present in I_A , as an affine registration can not take into account local variability. The regions of the R atlas is not patient-dependent and hence using R could lead to wrong quantification, especially if small regions are considered.

Registration and quantification in case of patient's structural images are *not* available Sometimes the structural images of patients are not available. In particular, the AmyDB database I built does not have structural images. This means that the above process cannot be performed. It is therefore necessary to use an alternative procedures.

Actually the algorithm I have used is similar to the previous one but it is not driven by the structural images; in other words the above mentioned A and W transformations will not be computed using the structural images of the patients, but they will be computed using only the PET. Some tricks will be required to avoid problems associated with this different approach. Here I describe the algorithm I used

- Let consider a raw PET image I_P . We map I_P to the template using an affine transform denoted by A , obtaining the affine registered image $I_{PA} = A(I_P)$. As in the previously described algorithm, this is the image we will quantify.

- I_{PA} is mapped to the template using a diffeomorphic transform W
- Let consider the W previously computed. Applying the inverse transform W^{-1} to the atlas R , we obtain $R_A = W^{-1}(R)$.
- Exactly as for the other algorithm, quantification can be performed using the couple (I_{PA}, R_A)

In this framework, the diffeomorphic transformation plays an important and delicate role.

We notice that, in principle, a diffeomorphic registration is able to find a W such that the transformed image is mapped exactly to the template. Such a W is what we wish to have in the case where the structural image is available, so that the anatomical structures of the patient and the template are perfectly mapped. However, in the case we are considering now, such a transformation is not appropriate at all.

We are mapping a radiotracer distribution onto the structural MRI template. The radiotracer being lipophilic, binds nonspecifically to certain anatomical structures in which there is no β -amyloid, typically the skull and white matter. For this reason it is grossly possible to distinguish some anatomic macro-regions by observing a PET scan. However, it must be emphasized that a map between a PET and template is *not* a map between anatomical structures, but at least it can be considered as a map between a distribution of radiotracer carrying some, gross anatomical information and the anatomical structures of the template, which is very detailed.

So, at the end of the story, a diffeomorphic transformation W without any constraint on the deformation field, which acts locally on a small spatial scale (the typical length scale of the gray matter gyri, where the clinical signal, i.e. $A\beta$, accumulates) will lead to exchange clinical with anatomical information. Therefore the atlas obtained by $W^{-1}(R)$ will be absolutely improper to identify the patient-like atlas ROIs.

The above can be clarified with an example. Let us consider 2 subjects, with exactly the same brain anatomy (identical MRI), but with two different amyloid loads, one very positive and one very negative (different PET). Registering the two PET on a template without any deformation field small scale constraints could give two transformations W_1 and W_2 potentially very different from each other, as the registration algorithm starts from two different PET images. Since W_1 differs from W_2 , then $W_1(R) \neq W_2(R)$. However, by hypothesis, the two patients have the same anatomy, thus they must share the same R_A atlas. This example shows why such an approach is wrong.

To overcome, or at least mitigate, this problem, I put constraints on deformation fields so that deformations on the small spatial scale were strongly suppressed, leaving the possibility of slight deformations at a large spatial scale, which is typically that of the coarse structures a-specifically evidenced by the radiotracer distribution.

Structures that are typically modified by the constrained deformation field are the shape of the brain, ventricles and cerebellum.

To do that I used ANTs, that offers the possibility to control the computed and total gradient, as showed in 3.6.1. The computational details, such as the specific choice of parameters, of the registration algorithm are described in the next section.

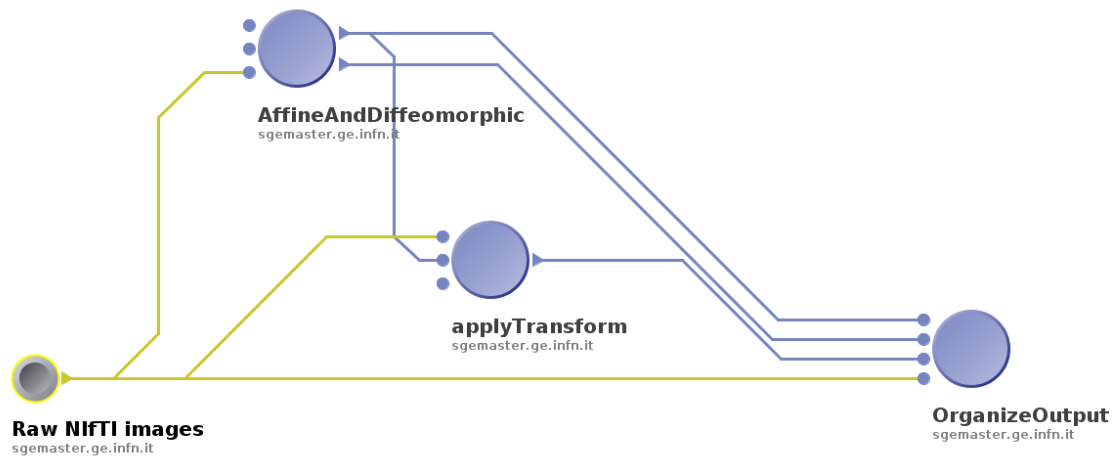


Figure 4.1: The pipeline for registration

4.2.2 Registration pipeline

The first preprocessing step is image registration. First of all I converted DICOM files using the *dcm2nii* software.

Then I built a pipeline for images registration using LONIpipe and ANTs. ANTs has been employed with notable success in recent public, unbiased, international evaluation studies [15] and allows the use of various types of concatenated transformations, both affine and non-affine; each transformation can be controlled by the user through appropriate parameters in order to obtain the best result. ANTs is described in the section 3.6.1, and LONIpipe has been discussed in section 3.8.

The registration pipeline is illustrated in figure 4.1; here below I will report the description of the single modules of the pipeline.

Raw NiftI Images : here I listed the paths of the raw images to register (i.e. the moving images).

Template : here I listed the path of the fixed image, which is the template ICBM152.

AffineAndDiffeomorphic : this is the module used to register images through three consequential transformations: a rigid, an affine and a SyN transformation. SyN is one of the diffeomorphic transformation provided by ANTs, I used this one because it is considered as one of the top performing algorithms [88]. This module requires two input NiftI images, the moving raw image and the fixed image (ICBM152 template) and gives as output the diffeomorphic registered image, the global affine transformation file (as a composition of the two consequential rigid and affine transformations) as well as the total gradient field (and its inverse) which drives the SyN transformation. Here below is the bash pseudo-code I used in this module

```
antsRegistration \
```

```
...
```

```
-n Linear \
```

I used a linear interpolator. There are many interpolation schemes proposed in the

literature, but linear interpolation seems to be the best for higher accuracy in this framework[112]

```
...  
-t Rigid [0.1] \
```

I did a rigid transformation with 0.1 gradient step

```
-m MI \
```

MI stands for Mutual Information that is the metric I used for registration. I used MI because it is very for multi-modal image registration.

```
-c 1000x500x250x100 \
```

This parameter controls levels and iterations: I chose a 4 levels registration, 1000,500,250,100 are the maximum number of iteration steps for each level. Note that the number of iterations is reduced at each level because it is assumed that the algorithm approaches at each level more and more to the global optimization; in particular the start of level n is the global optimum find at level $n - 1$. For this reason, as we approach the optimization, the convergence steps can be fewer and fewer. It is not necessary to reduce the number of iterations, but it is strongly recommended to have an algorithm considerably faster.

```
-s 3x2x1x0 \
```

The sigma-smoothing values for each step: $\sigma = 3,2,1,0$. To convert the sigma in amount of mm you can use roughly a factor of 2.36. The above correspond roughly to 7mm, 5mm, 2mm and 0mm (no smoothing).Note, smoothing is applied before shrinking the image to lower resolution.

```
-f 8x4x2x1 \
```

The 4 level steps will have resolutions divided by 8,4,2,1

```
-t Affine[0.1] \
```

I did an affine transformation with 0.1 gradient step. The parameters of the affine transformation listed below are the same as those of the rigid one, so I report them without comments

```
-c 1000x500x250x100 \
```

```
-m MI \
```

```
-s 3x2x1x0 \
```

```
-f 8x4x2x1 \
```

```
-t SyN[0.1,7,1] \
```

I performed a diffeomorphic SyN transformation, with a gradient step of 0.1. The values 7 and 1 are the update field variance and total field variance penalties respectively (see section 3.6.1). As detailed discussed in section 4.2.1, we need to put hard constraints to deformation fields, just accepting small deformations on large spatial scales. As update field variance is very related to small scale local image stretching, I chose a big penalty for it, while I choose a smaller one for total field variance.

```
-m MI \
```

```
-c 100x75x50x25 \
```

I chose a 4 level registration again, but the number of convergence iterations is considerably smaller than in the rigid and affine transformations. This is because we are assuming that the rigid and affine registrations has already been sufficiently precise. Furthermore, we want small local deformations, so few iterations are sufficient for our purposes.

```
-s 3x2x1x0 \
```

```
-f 8x4x2x1 \
```

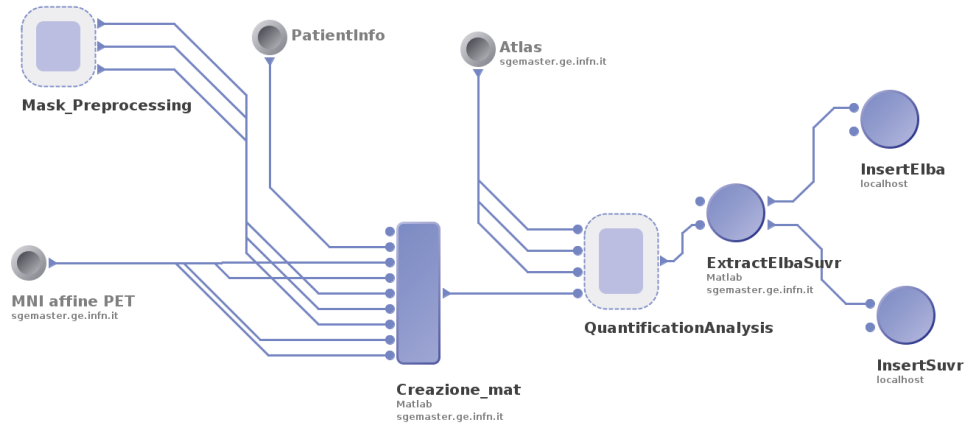


Figure 4.2: The quantification pipeline

ApplyTransform The composition of rigid and affine registration matrix is given in input together with raw image to the "ApplyTransform" module. To do that I used the `antsApplyTransform` command which allows to apply a given transformation to an input image. The output of this module is the affine registered image.

OrganizeOutput This module allows to organize all the files previously obtained in the database AmyDB. In particular I stored in AmyDB the diffeomorphic and affine registered images as well as the global affine transformations, the deformation fields and their inverse.

4.2.3 Quantification pipeline

Once the images were recorded I quantified them following the algorithm presented in section 4.2.1. To quantify images I used the pipeline in figure 4.2.

Images have been quantified using two methods, ELBA and SUVr, and they were quantified both globally and regionally using two different atlases, one made of 14 ROIs and one made of 50 ROIs.

The SUVr score has been computed using the whole cerebellum as reference region, which is illustrated in figure 4.3. I chose this reference ROI because is less prone to segmentation errors than the selection of the cerebellum gray matter alone or the brain stem, as reported in some works in literature [134].

Now I will briefly describe the most important quantification pipeline (4.2) modules:

MNI affine PET Here I listed all MNI affine registered images to quantify.

Mask Preprocessing As discuss in the algorithm described in section 4.2.1, here the patient-like atlas is computed: the diffeomorphic inverse transform is applied to all atlas ROIs by creating ROIs masks adapted to patients.

matFileCreation Here the ROIs mask and the affine registered images are stored in a MATLAB file which will be used for quantification

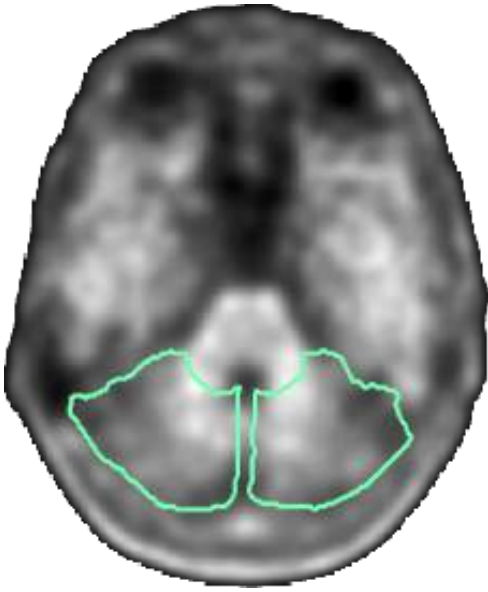


Figure 4.3: Axial section of a florbetapir PET image. Green lines enclose the cerebellum, often indicated as reference region; a good contrast between WM/GM (higher/lower part of the of highlighted areas) can be noticed

QuantificationAnalysis This module contains the computational implementation (MATLAB based) of ELBA and SUVr quantification methods as described in sections 3.7.2 and 3.7.1.

ExtractElbaSuvr, InsertElba and InsertSuvr these modules simply serves to save the quantification values obtained to text files and hence to insert them in the AmyDB database.

Chapter 5

Data Harmonization in PET imaging

In section 3.7 we explained what quantification is in the context of PET imaging and we give the operation description of two quantification methods, ELBA and SUVR.

In chapter 4 I described the AmyDB database highlighting its heterogeneity both clinically and related to image acquisition and reconstruction methodologies.

In chapter 2 I described the acquisition and reconstruction methods in PET imaging, showing all the factors that can affect both the two steps as well as highlighting the huge variability that is present in such steps, as many different PET scanners and reconstruction methods can be used, each of them with different properties, strengths and weaknesses.

Furthermore, in chapter 2, I emphasized how the perceived quality of PET images is strongly influenced by the *data provenance*, where data provenance is a shortcut to denote everything that involves the "history" of a given image, i.e. all the steps required to obtain a PET image as a 3D matrix of voxels (NIfTI image) starting from the radiotracer injection. Moreover, I point out that image quality can influence the quantification values.

We can think of quantification as a measurement process aimed at synthesizing the distribution of a signal of interest, i.e. the distribution of the radiotracer, through a vector of quantification. However this signal is in general affected by the data provenance, in particular the most important data provenance factors are

- scan detector properties(e.g. detector efficiency, dead time, energy and time window, TOF availability etc...)
- scatter and random coincidences
- type of acquisition (2D vs 3D)
- inclusion of a resolution model (PSF)
- type of reconstruction used, e.g. analytical or iterative
- the choice of a specific parameters set for a given reconstruction method, e.g. subset and iteration in OSEM reconstruction

I will refer to all these factors as *noise*. This definition is improper because this is not noise in the proper sense of the term: noise is typically a stochastic process which can be described in probabilistic terms.

Although some signal degrading factors in PET imaging are actual noise (e.g. Poissonian noise affecting photon counting), many other degrading factors are not properly noise: consider for example the artifacts introduced in the texture due to image reconstruction in iterative reconstruction methods.

However, here the main point is to distinguish between true signal (the relative distribution of radiotracer) and anything that degrades signal information. In this sense I operationally define noise as anything that degrades the signal we want to measure.

To summarize, we are dealing with a database, AmyDB, consisting of images with different provenance, from which we extract a measure of quantification that will be affected by the provenance itself, and we would like to perform a single study involving all these quantification measures.

The above can be framed within a problem of data harmonization. Indeed data harmonization refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies [70].

This issue will be deepened discussed below.

5.1 Multicentric and monocentric studies

Monocentric (or single-center) studies are studies conducted on a statistical sample from a single clinical center, whereas multicenter studies are based on the analysis of data from many centers.

Single center studies are logistically easier and cheaper with respect to multicentric ones. Furthermore they do not require prolonged negotiations on the study protocol and they typically deal with a less heterogeneous population, thereby diminishing confounding[21]. The primary shortcoming of single-center studies is their potentially limited external validity [42]. Interventions tested in a single clinical environment are not necessarily generalizable to a broader population [21].

Benefits of multicenter studies include a larger number of participants from different geographic locations, the possibility of inclusion of a wider range of population groups, and the ability to compare results among centers, all of which increase the generalizability of the studies. Multicenter trials are difficult to conduct, and when underpowered or poorly conducted may be even less useful than a single-center trial.[21]

Even though both monocentric and multicentric studies have pro and contra, the scientific community is certainly directed towards multicenter studies [41, 103, 2, 90]; indeed monocentric studies investigate samples whose size are typically small and this may lead to incomplete, or even misleading, results. [103].

However, multicentric studies, in order to achieve valid and generalizable results, must take into account the so-called *data harmonization* problem. Data from different clinical centers have a different provenance, and the data provenance could have a very big impact on studies results, this is especially true in PET multicentric studies,

where data provenance is highly varied and can have a significant impact on image quality and therefore on quantification and at the the end of the story on the results of the studies.

For this reason, together with the growing need for multicentric studies, there is increasing attention to the issue of data provenance and the development of harmonization techniques, e.g. [113, 156, 90, 2, 41, 42, 80, 5, 108].

IN 2017 Aide et. al [2] have pointed out that in addition to the issue of reconstruction method and scanner, other effects such as errors in the administration of the radiotracer, patient movement, blood glucose level and uptake period can have a significant impact on the quality of the image.

In 2015 Sunderland et. al. [146] published an oncological PET survey study in which they reported that differences in site-specific reconstruction parameters increased the quantitative variability among similar scanners, with postreconstruction smoothing filters being the most influential parameter: in their survey involving 237 PET/CT systems in 170 international imaging centers, with technology advancements spanning more than a decade and covering the three major PET manufacturers (GE Healthcare, Siemens and Phillips Healthcare made up approximately 56%, 34% and 10%), more than 100 reconstruction parameters were reported.

Moreover, in an international PET oncological survey, in 2011 Beyer et al. [23] reported that 52% of sites used alternative protocols with adapted reconstruction parameters, instead of the acquisition and reconstruction guidelines.

In 2019 a neuroimaging survey of Jovicic et al. [85] completed by 459 participants (MRI 53.6% of participants, EEG 30.3%, and PET-SPECT 16.1%) revealed a substantial lack of harmonization for analysis tools and the necessity of harmonize multivendor image reconstruction parameters.

In 2021 Verwer et. al [156] state that rigorous quality control and assurance are required in order to prevent that variability and differences between PET systems with regard to image quality can affect research conclusions or patient diagnostics; this is especially true when the effects studied are small (e.g. annual change in amyloid signal in Alzheimers disease), so that data from multiple centres need to be combined to form the large datasets needed to obtain statistically significant conclusions.

In 2016 Akamatsu et al. [5] pointed out that the SUVr quantification values of PET amyloid images are not directly comparable if the image reconstruction parameters have not been prior-calibrated by individual centers using phantoms, since quality features such as noise or resolution typically affects SUVr values.

At the end of the story, multicentric studies could be considered as double-edged sword: they are very important to improve results statistical significance, but the significance improvement can be considered as a real improvement only if the data provenance is taken into account and mitigated through harmonization.

5.2 Prospective and retrospective harmonization in PET imaging

PET harmonization methods can be divided into two categories: pre-processing and retrospective harmonization.

Prospective harmonization addresses the issue considering standardization of acquisition protocols and reconstruction settings. A Prospective strategy is by definition associated with less noise and improved statistical power, potentially allowing for more valid interpretations[42].

Typical prospective harmonization is based on phantom studies in which different PET scanner and/or reconstruction method are compared in order to obtain the best reconstruction parameter to achieve harmonized images. Phantom were used as a ground truth to test the radiotracer distribution reconstruction ability. Examples of these studies are [150, 102, 138, 97]. Prospective harmonization is related to standardization of acquisition protocols: in previously section we reported many literature which highlight the lack of standardization thus a retrospective harmonization is often required. I will not discuss further the prospective harmonization since it is not a topic I have addressed during my PhD.

Retrospective harmonization addresses the harmonization issue in the feature domain by either selecting features prior to the statistical analysis based on their robustness in order to rely only on features insensitive to multicenter variability, or by keeping all features and harmonizing their statistical properties so they can be pooled during the modeling step [42].

Here below I will briefly discuss the most used retrospective harmonization techniques.

Harmonization only based on imaging sites A naively approach for harmonizing across imaging sites is a simple z-score normalization. Consider an image-derived measurement y_{ijf} for imaging site i , subject j , and feature type f (e.g. SUVr). Once the mean μ_{if} and he standard deviation σ_{if} has been computed for each imaging site and feature (if more than one is present), the features harmonized value y_{ijf}^H is given by

$$y_{ijf}^H = \frac{y_{ijf} - \mu_{if}}{\sigma_{if}} \quad (5.1)$$

where subject j was scanned at site i .

a general linear model approach that includes site or scanner as a fixed effect covariate

Another approach can be used to take into account the site-specific effects is the linear regression. To this end, site becomes a regressor in a linear regression

$$y_{ijf} = \alpha_f + \gamma_{if} + \epsilon_{ijf} \quad (5.2)$$

where α_f is the mean of the feature of a reference site (the intercept), γ_{if} is an additive imaging site effect and ϵ_{ijf} is the residual for the subject j of the site i .

Thus the features harmonized value is the residual

$$y_{ijf}^H = y_{ijf} - \hat{\gamma}_{if} \quad (5.3)$$

where $\hat{\gamma}_{if}$ is the estimator of γ_{if} computed solving the corresponding ordinary least squares problem for the considered dataset.

Harmonization preserving demographic variability: linear model Let us consider a multicentric study in which N centers are involved. In general we are dealing with a multicentric sample S made of single-center samples S_1, \dots, S_N . If the study is not designed a priori, samples could be unbalanced with respect to demographic/clinical variability: in general there is no reason to expect that samples are equivalent with respect some demographic or clinical factors which may influence the output feature we are studying. For example, it could happen that the sample S_l has an average age of 65 years, while for sample S_m the average age is 75 years. Subject's age is obviously a confounding factor for amyloid quantification, as the amyloid burden typically increases as age increases. Fitting a model as described in equation (5.2) will leads to ignore this fact; indeed such a model is equivalent to consider the between-samples age demographic difference like if it were a center difference (where we point out that "center difference" refers to "data provenance difference" only). In this context we would like to take into account subject-specific information due to demographic/clinical variability. To fix this problem we add a vector variable \mathbf{k}_j to the model (5.2) to indicate a set of demographic/clinical confounding variables for each j -th patient (e.g. age, years of education...), obtaining

$$y_{ijf} = \alpha_f + \gamma_{if} + \mathbf{k}_j \beta_f + \epsilon_{ijf} \quad (5.4)$$

where \mathbf{k}_j is the set of confounding variables related to subject j and where β_f is the regression coefficient for the feature f . We notice that given a j -th subject, the confounding vector \mathbf{k}_j can be written as $\mathbf{k}_j = (k_{j1}, \dots, k_{jR})$, where R is the number of covariates considered. Thus \mathbf{k}_j can be considered as a matrix K_{jr} where j goes from 1 to the total number of subjects and r goes from 1 to R .

Then one can harmonize considering the partial residual

$$y_{ijf}^H = y_{ijf} - \hat{\gamma}_{if} \quad (5.5)$$

where $\hat{\gamma}_{if}$ is the estimator of γ_{if} computed solving the ordinary least squares problem given in equation (5.4).

Harmonization preserving demographic variability: ComBat Harmonization ComBat is an harmonization method introduced for genomic data harmonization by Johnson et al. [84] to correct the batch effect (BE). In genomic BE occurs when non-biological factors in an experiment cause changes in the data produced by the experiment; in particular BE refers to technical variation or non-biological differences between measurements of different groups of samples. If this systematic bias is not removed, its effect can mask important biological differences, at worst resulting in misleading inferences and conclusions [109].

In 2017 and 2018 Fortin et al proposed the ComBat method to harmonize diffusion tensor imaging ¹ data [63] and for harmonization of cortical thickness measurements obtained from different MRI scanners [62] respectively.

¹Diffusion tensor imaging is an MRI imaging technique

In 2018 Orlhac et al. [113] proposed ComBat to harmonize radiomic features extracted from FDG(Fluorodeoxyglucose)-PET oncological images.

The idea that underlies the use of ComBat for medical images is the analogy between batch effect and data provenance effect (e.g. scanner, reconstruction protocol...): both are effects due only to technical factors afferent to the acquisition of the data that introduce a systematic error that affects the output variable that we are interested to study.

ComBat can be considered as a generalization of a simple linear model given by equation (5.4): ComBat adds a site-specific scaling factor δ , yielding a model that adjusts for additive and multiplicative effects. In addition, ComBat uses empirical Bayes for inferring model parameters, which assumes that model parameters across features are drawn from the same distribution.

ComBat harmonization model assumes that the value y_{ijf} of the feature f for the subject j and center i can be written as follows [84]:

$$y_{ijf} = \alpha_f + K_{jr}\beta_f + \gamma_{if} + \delta_{if}\epsilon_{ijf} \quad (5.6)$$

where $\hat{\alpha}_f$ is the average value across i and j indices for feature y_{ijf} , K_{jr} is the matrix of the covariates of interest, β_f is the vector of regression coefficients corresponding to each covariate for the feature f , γ_{if} is the additive effect of center i on feature f , δ_{if} describes the multiplicative scanner effect, and ϵ_{ijf} is an te residual variance unexplained by the linear model, which is supposed to be normally distributed with a zero mean.

ComBat harmonization uses Bayes Empirical estimation for estimating γ_{if} and δ_{if} , supposing that γ_{if} follows an inverse gamma distribution and δ_{if} follows a normal distribution. We denoted the empirical Bayes estimation of γ_{if} and δ_{if} by $\hat{\gamma}_{if}$ and $\hat{\delta}_{if}$ respectively. The harmonized value is then obtained as [84]:

$$y_{ijf}^{ComBat} = \frac{y_{ijf} - \hat{\alpha}_f - K_{jr}\hat{\beta}_f - \hat{\gamma}_{if}}{\hat{\delta}_{if}} + \hat{\alpha}_f + K_{jr}\hat{\beta}_f \quad (5.7)$$

where $\hat{\alpha}_f$ and $\hat{\beta}_f$ are estimators of parameters α_f and β_f respectively. It is important to observe that K_{jr} is a matrix which summarizes the covariates of interest. This covariates of interest are preserved by ComBat algorithm.

5.3 Fixing the notation

At this point it is convenient to fix some notation once and for all to avoid misunderstandings.

Confounding variables Let consider a model which relates an input X to an output Y and let suppose that X and Y are correlated. We might tempt to say that X causes Y , but this is not true in general. Let consider a variable Z that is affects both X and Y ; Z could be responsible for the correlation between X and Y . In this framework Z is called *confounding variable* and the correlation between X and Y is called *spurious correlation*; we remark that spurious correlation never imply a causal relation between X and Y . For example, consider you want to study causes of AD. Analyses of the data show a high correlation between gray hair (input X) and AD (output Y), which may naively lead to

the conclusion that gray hair causes AD. However, the observed correlation between gray hair and AD is only due to a person's age (confounding Z). Therefore, the association between gray hair and AD is confounded by the common cause age. This form of bias is known as *confounding bias*.

Selection bias In neuroimaging studies, various types of bias can be present that can alter the conclusions one deduces from this study. In the first step, individuals have to be enrolled into the study. If subjects do not faithfully represent the overall population one wants to study, i.e., one obtains a non-random sample of a population, conclusions will be biased. This is referred to as *selection bias*. The selection bias could occur in the form of confounding variables. If samples from different clinical centers have a selection bias we say that the samples are *unbalanced*.

For example, in amyloid imaging, age, sex, MMSE and years of education can be considered as confounding variables. Samples from different clinical centers may not have an equivalent distribution with respect to clinical and/or demographic confounding variables. Take into account confounding variables are very important to avoid misleading results.

Batch variable versus center variable It is important to remark that the variability due to the provenance of the data is exclusively technical. The effect of data provenance on output measures will be denoted by *batch effect*. This term has been borrowed by biology but it is already used in medical imaging literature, e.g. [62, 113].

Now I will clarify a point that can lead to misunderstandings. It may happen that the same center acquires and reconstructs images with different protocols. Similarly, it can happen that different research centers share the same protocol. There is therefore *no one-to-one correspondence between the center variable and the data provenance*. The center variable *per se* does not play any direct role on image quality, as quality is related to the technical factors involved in creating images. As these factors are summarized in the data provenance, I will define a data provenance specific variable, called *batch variable*. It often happens that batch and center variables are equivalent as they carry the same information, but this is not always true. This fact can be observed in the table 4.3.

Batch effect and sample bias effect Thus, statistically significant differences in the distribution of a given output variable of interest between samples from different centers may be due to a combination of the following effects

- **batch effect (BE)**: all the factors (exclusively technical) related to data provenance which is supposed to affect the output variable (quantification, for our purpose).
- **sample bias effect (SE)**: it emerges if samples are not representative of the whole population, i.e. samples are unbalanced. It could be due to different enrollment criteria among centers or it could be due to other issues, such as geographical differences: the demographic/clinical variability of the world population may not be homogeneously distributed, but may depend on geographical location. Since multicenter studies bring together data from even very distant geographical locations, this variability cannot be excluded a priori.

Harmonization Here I emphasize a fundamental point: the ideal harmonization algorithm is the one that completely eliminates the BE while preserving the clinical and/or demographic variability related to the SE.

5.4 PET retrospective harmonization methods discussion

Let start this section with an observation: BE can be of marginal or major importance depending on how pronounced this effect is and depending on what we are interested in measuring. A measure can be more or less stable with respect to the quality of the image and therefore to the BE. For example, if we study the global quantification of amyloid load on the whole brain we expect that the BE is rather limited, while if we study the regional quantification on small regions it is safe to assume that the BE will have a quite important role.

Many multicentric amyloid PET studies take into account the BE using a simple linear model, at most preserving the SE by inserting known confounding variables into the model. This approach is not absolutely right or wrong, but in agreement with what has been said above, it can be correct if we believe that the BE has a marginal role with respect to what we want to measure.

ComBat provides an approach that is certainly more performative than a simple linear model, however although it has been used in FDG PET in oncology studies, as far as I know it has never been used in multicenter studies of amyloid PET imaging.

What are the limits of linear models and ComBat in taking into account the BE? Apart from the limitations related to the assumptions on which linear models and ComBat are intrinsically built, the most important even though rarely considered limitation common to both methods is the following: we know that both ComBat and linear models are able to take into account the batch effect by introducing a more or less sophisticated batch variable (BV) (i.e. γ_{if} in (5.4) and γ_{if} and δ_{if} in (5.6)). Furthermore those methods are able to preserve a set of confounding clinical/demographic variables (i.e. they can preserve the SE) introducing an appropriate term $K_{jr}\beta_f$ (see (5.4) and (5.6)).

However, what is rarely emphasized even in the literature is that the confounding variables preserved by these approaches are only the *known confounding variables* (KCV), while many other (more or less relevant) unknown confounding variables (also called latent) could be present. We will denote latent confounding variables by LCV.

Obviously the set of all confounding variables CVs is the disjoint union of KCVs and LCVs sets, thus we will improperly write $CV=KCV+LCV$ for simplicity.

In particular when we harmonize data using a method that specifies which is the batch (introducing a suitable set of BV) and which are the variables to be preserved KCV, we run into the possible trap of latent confounders; in practice LCVs are not preserved by definition: they are considered by the model as part of the batch effect. LCV's together with BE are therefore modeled within the BVs and eliminated by the harmonization

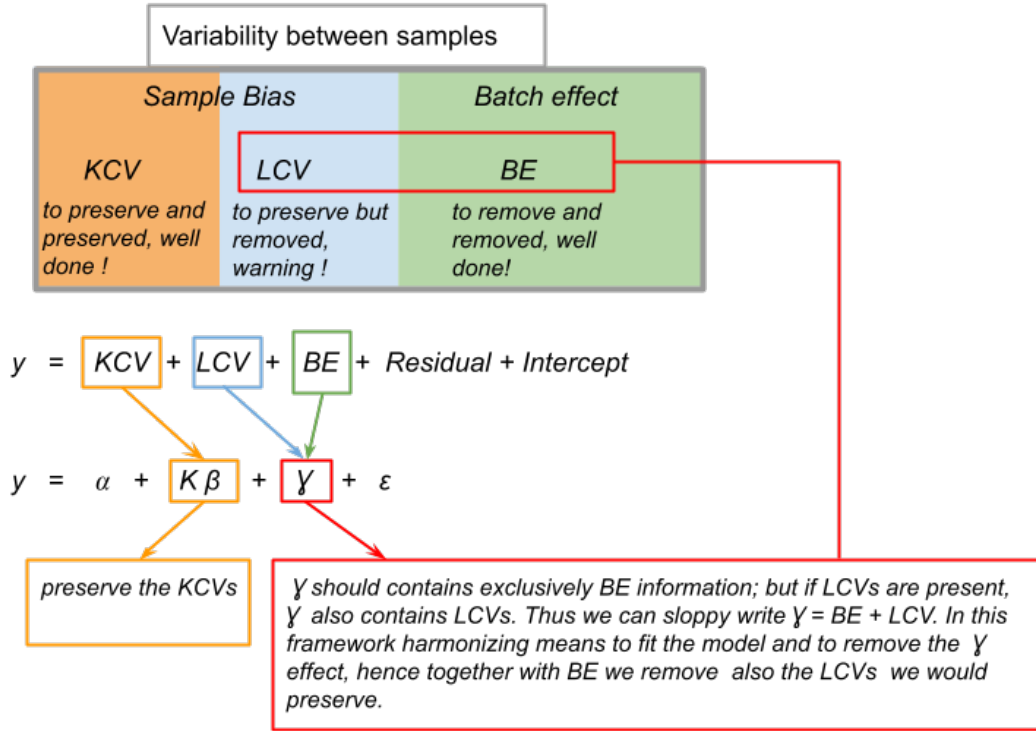


Figure 5.1: Variability between samples arises from sample bias and batch effect. We wish to remove the latter and preserving the former. However sample bias is splitted into a set of known (KCV) and unknown (LCV) confounding variables. The linear model in figure are the same of (5.4), but written with no indices for simplicity. Using such a linear model we are able to preserve only the former, while the latter are removed together with BE. it is important to point out that the same would happen using ComBat or any other model based on the scheme: "give me the variables to preserve, everything else is to eliminate".

algorithm. This delicate point is explained in figure 5.1.

Why do this actually happen? The point is that *we are not modeling the batch effect*. Rather we model just the KCVs we would preserve, and we assume that everything is not KCVs is due to the batch effect, therefore to be eliminated. However looking the set in figure 5.1 we notice that the complementary set of KCV is not BE, but is BE + LCV, that is what we are really removing. This problem is common to all the harmonization methods based on the schemes: "give me a set of variables to preserve, any other statistical difference between samples is to eliminate".

Approaches based on this scheme work ideally only when the KCVs coincide with all possible CVs. However, this scenario is essentially impossible to implement. If LCVs are present, but they have a small impact, then the error committed using such methods is reasonably negligible.

The difficulties for a realization of a scenario where LCVs would have a small impact are both practical and theoretical. At a theoretical level, since biological phenomena are very complex it is very hard to know a priori all the important confounders. Moreover, practically speaking, sometimes it happens that some important confounders are unknown.

It is important to remark that these studies are not conducted on statistical units where variability is controlled and limited to certain aspects, but are conducted on subjects who live their lives each in a potentially very different way. We do not know whether factors such as income or lifestyle (diet, sedentary lifestyle, night rest, etc...), can have a strong impact as confounding variables. However, what is reasonable to assume is that these variables are dependent on the geographical location of the enrolled subjects: this will generate unbalanced samples with respect to these variables. In addition, genetic variability and comorbidities may be also unknown confounding variables. We do not have an a priori model to identify all the confounding variables, and this might limit the use of the harmonization strategies I have presented.

Some scientific articles warn about the problem of confounding variables in harmonization [142, 158, 91, 110]; in particular Wachinger et al [158] notice that in practice the knowledge of all potential confounders is almost impossible and this could lead to easily lose relevant subject-specific information. However, harmonization also requires caution as it can easily remove relevant subject-specific information, therefore Wachinger suggest to use harmonization with caution.

In particular, in 2017 Lewinn et al. [93] examined whether sample composition influences age-related variation in global measurements of gray matter volume, thickness, and surface area in MRI imaging. They proved that an uneven sample composition across a number of basic socio-demographic (socioeconomic status, ethnicity and sex) characteristics had introduced significant bias on results.

In context of the biology, Nygaard et al [110] investigated various harmonization models including ComBat, and warns us that using not evenly distributed samples can lead to errors in data harmonization, inducing apparent batch differences, highlighting how this important point is often not taken seriously into account by the scientific community.

Furthermore, with regard to ComBat harmonization, the minimum number of patients required per imaging protocol to successfully apply ComBat remains to be comprehensively investigated [114] even though the method seems to work well using small size samples (~ 50 per "batch").

Chapter 6

Image quality estimation and data harmonization: proposal for a novel approach

6.1 Batch effect estimation using quality measures directly extracted from the PET images

In the previous section we evidenced the limits of the harmonization approaches discussed in section 5.2. A common point that limits these approaches is that the batch effect which explains some of the variability between samples is not explicitly modeled. Indeed, the methods I previously described are based on identifying a set of known confounding variables KCVs that one wants to preserve and on the implicit assumption that any other statistically significant difference between samples not explained by KCVs is attributable to the BE and therefore should be eliminated, as illustrated in figure 5.1.

In this chapter I will propose a method to explicitly estimate the BE. This will be useful to overcome the limitation just described.

Before going into detail, I will illustrate the general lines of the approach that I followed in order to introduce my idea.

The final goal of my work is to define a set of appropriate quality measures that allow to evaluate certain aspects of PET image quality. These aspects should be related to the data provenance, hence I am looking for measures which are able to assess features related to texture, noise and smoothness of the images themselves.

The main difficulty in following this approach can be summarized in two factors. First, we cannot use quality measures that require knowledge of a reference image to use as a ground truth.

Second, we are focusing on quality variability between images, but we must take into account that inter-images variability are not only due to quality.

For example, let consider two images with different provenance: they will differ in quality because of different provenance. However they will also have a clinical and an anatomical variability: the radiotracer binds specifically to amyloid and provides clinical

information with respect to the amyloid burden itself. At the same time it binds a-specifically to white matter and bone structure, and these latter regions have an inter-individual variability too.

In literature image quality is commonly assessed a-priori using phantom studies. This allows to eliminate both the problem previously introduced: we have a ground truth (the reference image is given, as the actual radiotracer distribution is known) and we have no clinical and anatomical variability, as we are using phantom to characterize PET scanner quality performance.

However this approach is typically related on a prospective harmonization, in which acquisition parameters and image reconstruction protocols are tuned a-priori. This approach is not often used: in section 5.1 we reported a bunch of literature which highlights the problem of lack of PET protocols standardization and harmonization which often affect PET studies. Furthermore, acquisition parameters, image reconstruction methods and type of scanner used can be combined in really many ways, so it is very complicated to characterize them all. In addition, data provenance is not only reduced to factors measurable on phantoms, there are also factors related to patient movement, non-standardized injection dosage and uptake periods.

Therefore, we want to find a number of no-reference quality measures (i.e. measures that do not need a reference image) that are able to quantitatively describe only and exclusively quality, ignoring the clinical and anatomical variability of patients. In other words, the key point is decoupling quality from clinical/anatomical information.

How is it possible to obtain measures with such properties? At this point it is important to emphasize that we have one more degree of freedom we have ignored so far: the regions where these measures will be evaluated. Indeed it is not necessary to use the whole image to assess quality, but it will be enough to use an appropriate part of it.

My working hypothesis is that there exist some image's regions where the radiotracer uptake is a-specific bounded (no clinical variability) and in which the inter-subject anatomical and variability is so low that the only differences we can observe between different PET scans are due to data provenance. Therefore the request for a set of measures exclusively dependent on quality can be reformulated into a request for regions exclusively dependent on quality. Regions with such property are what I used to characterize image quality.

Thus, the points I will discuss in next sections are

- I will define appropriate ROIs where I will compute quality measures
- I will define 3 no-reference quality measures
- I will validate measures through visual analysis as well as by verifying the link with data provenance
- I will harmonize data using these measures and I will explore the consequence of my approach

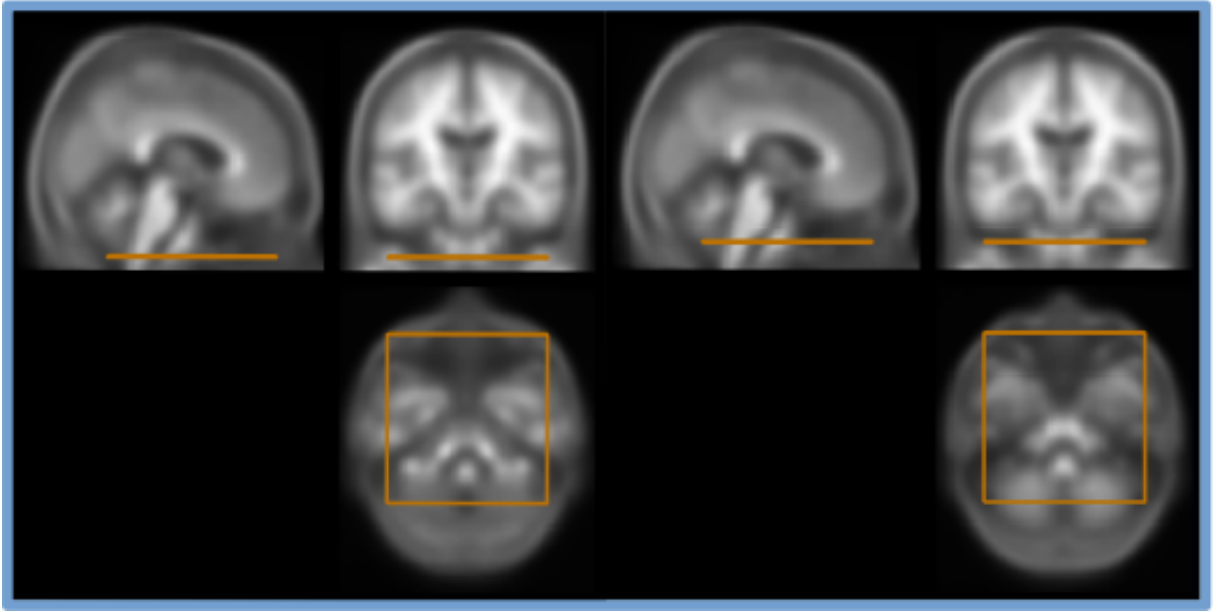


Figure 6.1: The lower (on the left, $z = 16$) and the higher (on the right, $z = 26$) ROI out of the 6 ROIs I have selected. z is the axial MNI coordinate.

6.2 ROIs selection

In this section I discuss the ROIs selection where I will perform quality measures. To evaluate just and only the PET images quality, I focus my attention only on ROIs with the following properties:

- very low inter-subject clinical and anatomical variability
- an anatomical structure with simple and clearly recognizable shapes
- large enough to be representative of the quality of the whole image

I selected for each image of the dataset six transaxial ROIs, namely $\{ROI_1, \dots, ROI_6\}$, defined by the following MNI coordinates:

$$ROI_i = \{x, y, z \mid x \in [40, 153], y \in [70, 189], z = 16 + 2(i - 1)\} \quad (6.1)$$

where $i \in [1, 6]$.

A figurative example of ROIs I chose can be found in figure 6.1.

6.3 Quality measures

In this section I will introduce and describe the set of quality measures I defined and used.

6.3.1 Watershed

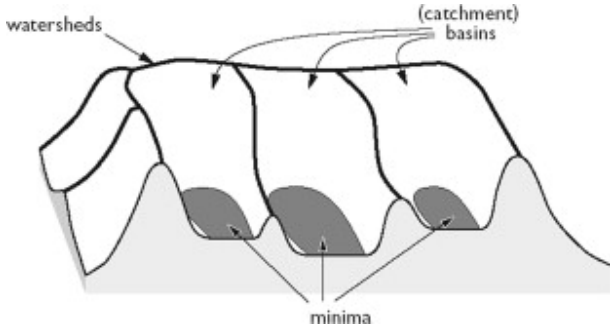


Figure 6.2: Pictorial representation of watershed transform working principle: all voxels in the same basin are denoted by a unique label. Obviously the three-dimensionality of the figure is a representation of the voxels intensity: the higher the intensity the higher the height. So valleys and ridges are defined by the intensity levels of the image.

representing its height. This transform identifies the ridges of the map: points surrounded by ridges are indexed with the same label. We will refer to these regions as the valleys of the image. An example of watershed transform is provided in figure 6.3.

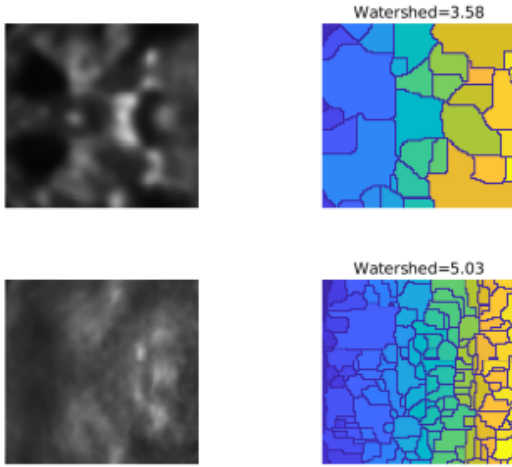


Figure 6.3: Watershed Measure. On the left two ROIs with low granularity (top image) and high granularity (bottom image) texture. On the right the respective watershed transforms. In the top image arises less watershed regions than in the bottom one, as expected.

The first quality measure I introduce aims to estimate the granularity of PET images texture.

Granularity can be considered as the size of the elementary particles that make up the fundamental structure of a texture[67].

The central idea is that texture granularity could be captured by applying a watershed transform to the ROIs considered. Watershed transform treats the image it operates upon like a topographic map, with the intensity level of each point

Watershed is used for image segmentation in various fields, including medical imaging [72]. A drawback of this segmentation method, which especially occurs in noisy medical image data, is that a large number of small regions arises. This is known as the *over-segmentation problem* [121].

My idea is to take advantage of over-segmentation in order to capture and quantify the image granularity. Since ROIs describe the same anatomical structure for each patient, and because we are assuming ROIs individual variability is negligible, we can conclude that the greater is the number of watershed regions, the greater is the texture granularity.

Let \mathcal{W} the watershed transform and let R a given ROI. I define the watershed measure W as follows

$$W = \log(N[\mathcal{W}(R)]) \quad (6.2)$$

where N is the operator which counts the number of watershed regions (i.e. the number of $\mathcal{W}(R)$ label). We notice that W is an a-dimensional quantity.

6.3.2 Delta Contrast

Watershed measure allows to quantify textures granularity, but it does not provide any information about the noise amplitude. Let us consider images in figure 6.4: although textures of the two ROIs considered are both grainy, they appear very different with regard to noise amplitude. I therefore defined a measure which allows us to quantify textures noise amplitude. I called this measure Delta Contrast (ΔC). Before explaining the procedure I followed to obtain ΔC I will give a mathematical definition of image contrast.

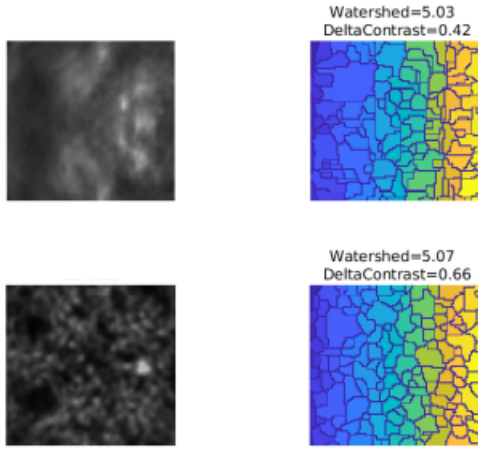


Figure 6.4: Top and bottom ROIs have almost the same grainy texture, but they appear very different about noise amplitude. ΔC allows to differentiate these two images capturing the noise amplitude.

Image contrast and co-occurrence matrix

Image contrast can be defined using co-occurrence texture analysis, which is a statistical method of examining texture that considers the spatial relationship between voxels of a given binary image (e.g. black and white image). Co-occurrence texture analysis is based on the so-called gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix.

First of all, the image is discretized in a given number of gray intensity level, namely N . Then the GLCM is obtained calculating how often a pixel with gray-level value i occurs adjacent to a pixel with the value j . Each element (i, j) in GLCM specifies the number of times that the pixel with intensity i occurred adjacent to a pixel with intensity j .

Once GLCM is computed some image properties such as contrast can be evaluated. Specifically, let g_{ij} the GLCM. The image contrast is defined as

$$C = \frac{\sum_{i,j} (i - j)^2 g_{ij}}{\sum_{i,j} g_{ij}} \quad (6.3)$$

The formula 6.3 tells us that image contrast is positively related with intensity variations of neighbor voxels. We notice that voxels intensity variations are determined by two contributions:

- the gray-level intensity variations due to noise profile, which worsens the quality of the image
- the gray-level intensity variations which highlight the anatomical structures, which improves image quality

Below I will propose a method to estimate just and only the factor due to noise by decoupling it from the contribution due to anatomical structures.

Contrast and noise amplitude relation Before going any further, I give a proof of what just said, namely that contrast increases as noise increases. Thus I considered all the images of AmyDB database: for each subject I added a zero mean gaussian noise of variable variance σ to the 6 ROIs defined by 6.1 and then I measured the respective contrast C as a function of the noise variance using the formula 6.3). I obtained $6 \text{ ROIs} \times 1001$ subject curves $C_i(\sigma)$ (where $i \in [1, \dots, 6 \times 1001]$). All the C_i showed an increasing trend, even sometimes slight fluctuations were present. These fluctuations, in addition to being small in amplitude, had a typical length that was very small relative to the increasing σ . Therefore, they are negligible fluctuations, while the increasing trend was very evident. However, the rate of contrast versus σ increment was image dependent. To visualize what has been said, I calculated the mean value and standard deviation of the contrast versus σ curves and plotted them in figure 6.5.

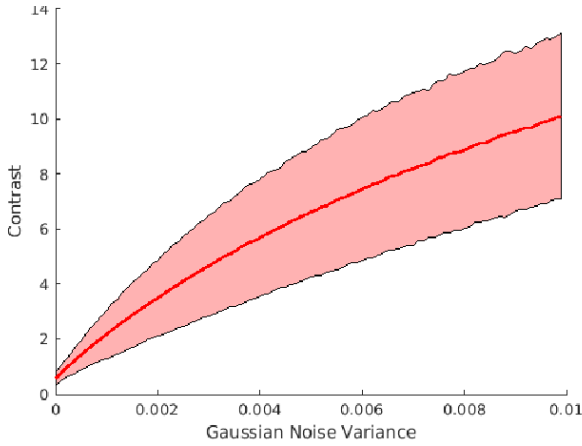


Figure 6.5: Gaussian noise (zero mean, variable variance) has been added to ROIs images, and respective contrast has been evaluated. Red curve is the mean across all contrast versus variance curves, while the light red area is the region between mean ± 3 standard deviation of contrast curves.'

Decoupling contrast due to noise from contrast due to anatomical structures Even though C is a monotone increasing function of the noise variance σ , it can not be used by itself to estimate noise amplitude, as contrast depends also on how well the anatomical structures are highlighted.

In order to decoupling the two contributions to contrast, it is important to notice that the local gray-level intensity variations due to noise have a smaller spatial length scale than the variations due to anatomical structures highlighting.

Thus, I applied to ROIs a gaussian filter of an appropriate radius (I chose 7 voxels radius), so that the filter affects almost uniquely the noise. This allows to dump noise while keeping anatomical structures highlighted at the same time.

So, to summarize, the contrast of the original image C_R is due to both noise and anatomical structures, the contrast of the filtered image C_{RF} is almost all due to the presence of anatomical structures; therefore, the greater the contrast variation between the filtered and the original image, the greater is the noise contribution to the total image contrast. Thus, I defined the measure Delta Contrast ΔC as follows

$$\Delta C = \frac{C_R - C_{RF}}{C_R} \quad (6.4)$$

So the greater is ΔC the greater is the noise amplitude. I divided by C_R in order to keep the measure range between 0 and 1.

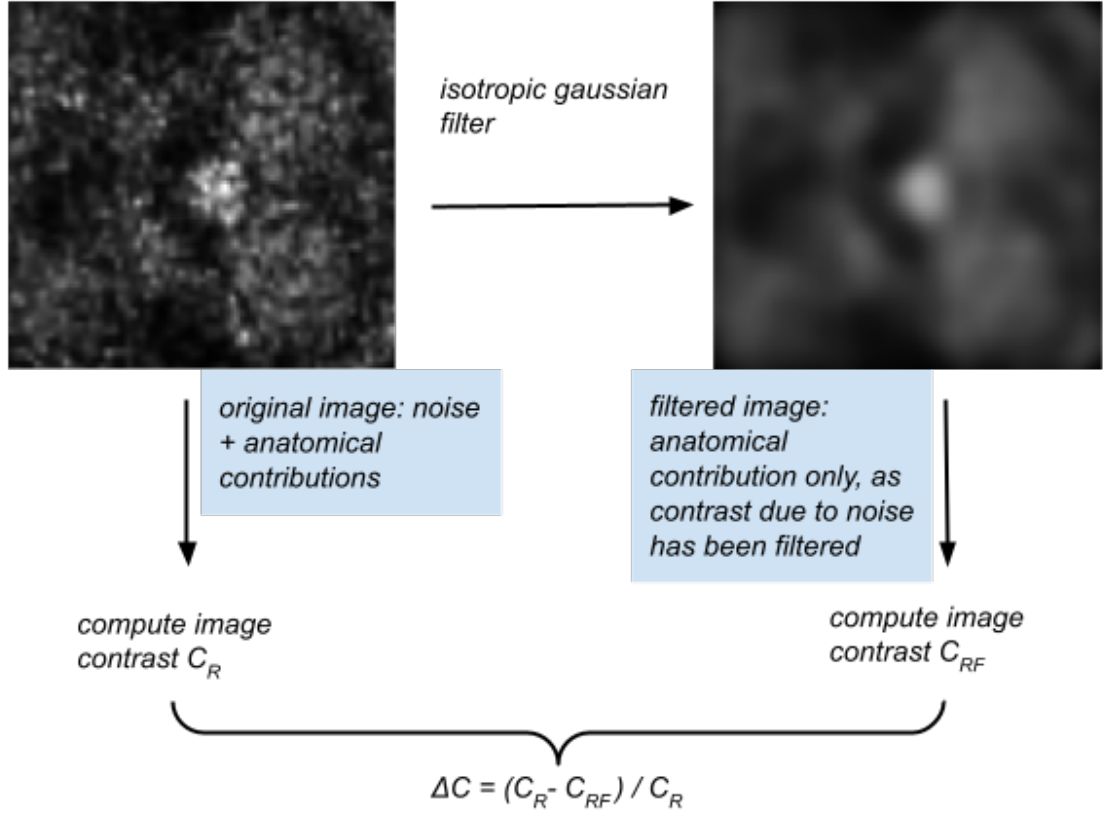


Figure 6.6: Illustration of the Delta Contrast algorithm

6.3.3 Acutance

Acutance is one of the two features, together with resolution, which determines the sharpness of an image [119]. Acutance is best described as how well a photographic medium handles edge contrast. While high acutance gives crisp, clean edges, low acutance gives fuzzy edges that are less distinct [119]. Acutance is related to intensity spatial gradient measured along images edges[125, 51].

Therefore I will propose an algorithm based on spatial gradient to measure acutance. The algorithm can be divided in two parts: the ROIs edges detection and the gradient evaluation on the edges previously detected.

Edges detection In this first step an automated edges detection method is provided. A gaussian filter of 7 voxels amplitude is applied to a given ROI in order to dump noise while keeping the anatomical structures highlighted. Filtering is a necessary step to identify the edges of mean features of images (i.e. anatomical structures) ; otherwise noise could dominate the edges detection, as you can see in figure 6.7.

Once the image has been filtered, I applied a 3 voxel range-filter to highlight the edges of the filtered image. The edges region will be denoted by E . This procedure is illustrated in figure 6.8.

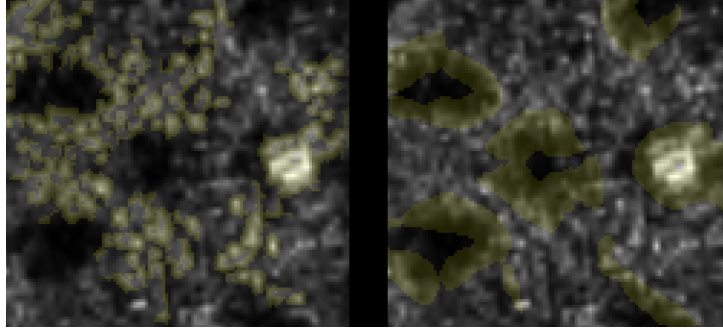


Figure 6.7: Edge detections in a highly noisy ROI: the yellow regions represent the edges identified by the algorithm. On the left edges detection without gaussian filter: edges are not detected. On the right edges detection using filtering: edges are correctly detected.

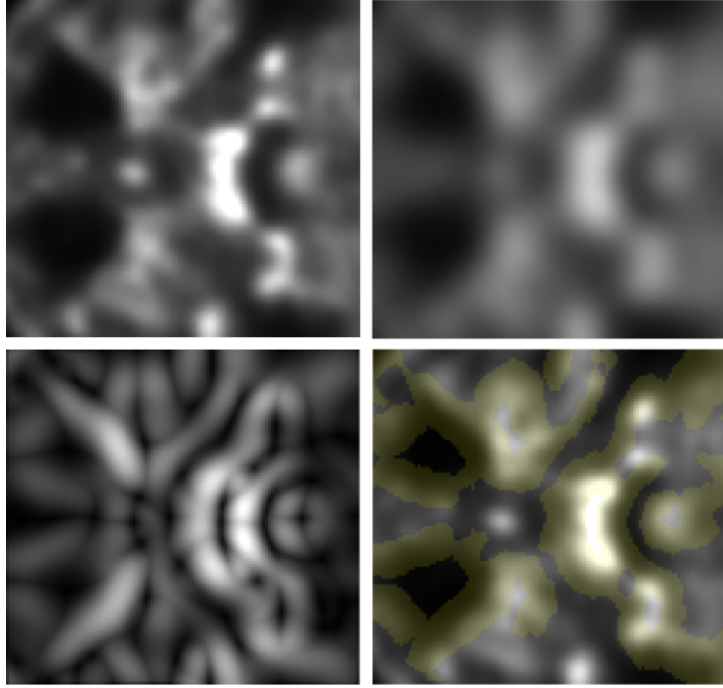


Figure 6.8: The 4 steps of edge detections. Top left: The ROI considered. Top right: the ROI filtered by gaussian filter. Bottom left: range filter has been applied to the top right image. Bottom right: ROI's main edges (shown in yellow) detection by application of range filter mask to the initial ROI.

Gradient computation The second step is related to the evaluation of the edges gradient intensity. The simple way to do that is to consider the mean value of the absolute intensity gradient computed over the edges previously detected. Such a straightforward approach tends to overestimate acutance, especially in noisy images: the edge's gradient intensity could be increased by the local intensity variations of the noise profile. A good sharpness measure should have the ability to distinguish between sharpness due to original high frequency detail of an image and sharpness due to high frequency noise detail [118].

Therefore I propose a method to avoid acutance overestimation caused by high level noise profile.

Let $R(x, y)$ a function representing the gray-level intensity map a given ROI, where x, y are the spatial coordinates, and let $R_F(x, y)$ the ROI filtered by an isotropic gaussian filter. For this purpose I chose a 7 voxels radius.

Let us now consider the gradient of both functions, namely ∇R and ∇R_F , then let compute the following projection

$$G = \nabla R \cdot \frac{\nabla R_F}{|\nabla R_F|} \quad (6.5)$$

I evaluate the acutance using the mean across the edges E of absolute value of the projection G by the following formula

$$A = \frac{1}{N} \sum_{x,y \in E} |G(x,y)| \quad (6.6)$$

where N is the number of voxels which belong to the edges region E . Estimating acutance A using formula 6.6 allows to reduce the gradients noise contribution, as it is mitigated by the projection 6.5.

Indeed the filtered ROI $R_F(x,y)$ is related to anatomical structures, because noise has been filtered, thus the gradient vector $\frac{\nabla R_F}{|\nabla R_F|}$ locally gives the maximum increase direction of the anatomical structures, ignoring noise profile. Noise intensity variations are typically isotropic, using this method the gradient components not parallel to the direction of interest (the one defined by the maximum intensity variation of the anatomical structures $\frac{\nabla R_F}{|\nabla R_F|}$) are more or less strongly suppressed (in particular the orthogonal components are set to zero).

Therefore acutance computed by 6.6 gives a more accurate estimation then a simple evaluation of gradient across the edges.

6.3.4 Natural Image Quality Evaluator (NIQE)

NIQE[105] is no-reference Natural Scene Statistic (NSS) based quality measure. It is implemented in Matlab and it is characterized by having a very good correlation (about 90 percent) between quality predicted scores and human judgments of visual quality [105].

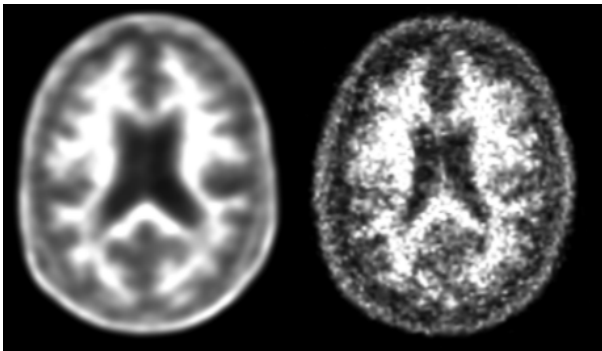


Figure 6.9: An example of a low NIQE (better quality) and high NIQE (worst quality) transaxial section of two different subjects.

NSS model assumes that natural images possess certain regular statistical properties[140]. The presence of distortions will change the statistics property of natural images. Therefore, the key idea of NSS-based metrics is to quantify the image quality degradations by measuring the losses of naturalness. NSS-based algorithms are achieved by measuring the variation of image statistics, which are characterized by the fitting parameters of NSS model, across different distortions[165].

NIQE is based on the construction of a "quality aware" collection of statistical features based on a simple space domain NSS model. These features are derived from a corpus of

natural, undistorted images. NIQE algorithm is based on fitting "quality aware" features to a multivariate Gaussian model.

In particular, the quality of a given image is expressed as the distance between the multivariate Gaussian fit of the NSS features extracted from the test image, and a multivariate Gaussian model of the "quality aware" features extracted from the corpus of natural images.

6.3.5 Matrix Dimension

Finally, the last quality measure I used is a discrete numerical quantity related to the resolution of the image: the matrix dimension, which has been briefly discussed in section 2.7. The matrix dimension basically is the dimension of the square grid used in image reconstruction: the higher is the matrix dimension, the higher is the number of raw image voxels.

Resolution can be defined as the number of distinguishable elements per measurement unit [101]. Images with higher resolution allows us to perceive the fine details contained in an image. As the theoretical best resolution is directly proportional to the numbers of image voxels, an higher matrix dimension is related to a better resolution [101].

Matrix dimension can be easily obtained from raw PET images embedded in native space.

6.4 Quality measures extraction and post-processing

I have implemented Watershed, Delta Contrast and Acutance measures in Matlab (NIQE is already available as a function of Matlab). So I have developed an automated procedure (Matlab based) which extracts ROIs (defined as in equation 6.1), evaluates the quality measures on each of the 6 ROIs and averages them. Therefore, for a given patient, a 4-dimensional vector consisting of the average across ROIs of quality measures is provided.

The size of the matrix was instead extracted in advance from the raw image in the native space, and it has been treated as a categorical variable. This is because the values of matrix dimension, even though they are actually number, take values in set of a few natural numbers, namely 128, 256, 336 and 512.

At the end of the story I therefore obtained a 1001-by-5 matrix, where the generic element i, j corresponds to the j -th measure of the i -th patient.

The quality measures I defined are not homogeneous about their range values, then a z -score transformation has been applied to make them comparable. Some measures are quite correlated, as you can see in figure 6.10 (a).

In order to have a set of orthogonal measures with respect to correlation, I applied a PCA transform (see section 3.4). The PCA measures I obtained, denoted by $\{q_1, q_2, q_3, q_4\}$, are linear combinations of quality measures I previously introduced. It is worth emphasizing that no information has been lost because of PCA transformation: $\{q_1, \dots, q_4\}$ are just a different basis representation of quality measures. Here I report the PCA transform

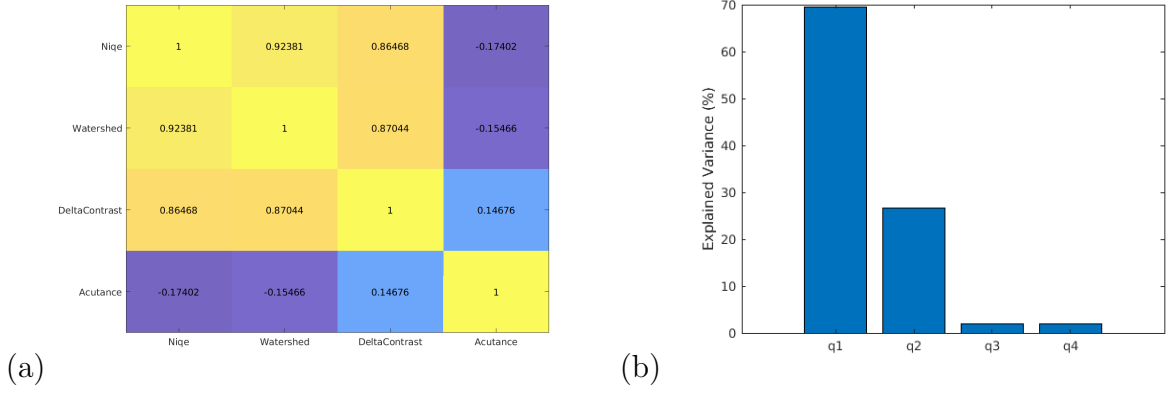


Figure 6.10: (a)Quality measures correlation matrix. (b)Principal components explained variance.

$$\begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} = \begin{bmatrix} 0.58 & 0.58 & 0.56 & -0.06 \\ -0.08 & -0.06 & 0.25 & 0.96 \\ -0.33 & -0.45 & 0.78 & -0.26 \\ 0.74 & -0.67 & -0.06 & 0.04 \end{bmatrix} \begin{bmatrix} N_z \\ W_z \\ \Delta C_z \\ A_z \end{bmatrix} \quad (6.7)$$

where $N_z, W_z, \Delta C_z, A_z$ are NIQE, Watershed, DeltaContrast, Acutance z-scored measures.

Finally, the cumulative PCA explained variance is reported in figure 6.10 (b).

6.5 Quality Measures Validation

In this section a validation of quality measures previously described will be provided.

6.5.1 Independence between quality measures and clinical profiles

First of all, we test the independence between quality measures and patients amyloid positivity visual assessment (P-N). This is an important step, as quality measures should be related to date provenance only and not to clinical variability of patients.

I considered 1000 bootstrap samples and I performed a two tailed t-test to compare the distributions of each quality measure in positive and negative populations. The null hypothesis I tested was that $\text{mean}_N[m_i] = \text{mean}_P[m_i]$ where m_i is the i -th quality measures and where mean_N and mean_P denotes the mean of q_i across negative and positive subjects respectively. Acceptance of the null hypothesis (therefore large p-values) are a necessary condition for measures validation, as rejection of the null hypothesis means that quality measures distinguish between the patient's clinic.

Because positivity and negativity numbers are not balanced for each center, (see table 4.1), to avoid possible spurious correlation between center and positivity, a balanced bootstrap has been used.

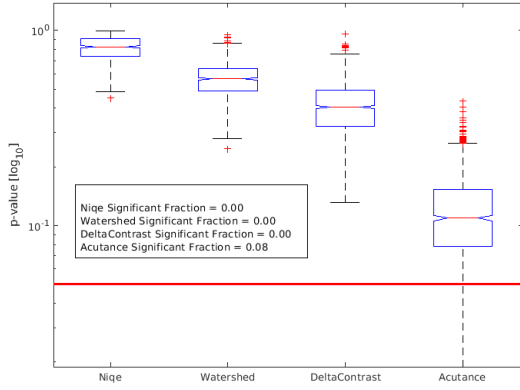


Figure 6.11: P-value distribution per quality measures of 1000 bootstrap samples. For each quality measure, p-values were obtained from a t-test which assumed as null hypothesis that the measure's averages on positive and negative subjects was equal. The red line represents the 5% significance level. The fractions of significant p-values (< 0.05) for each measure are shown in the box.

and image edge sharpness, giving to each of these aspects a two-values label of quality (HQ,LQ). Results are summarized in figure 6.12.

Results suggest that NIQE, Watershed and Delta Contrast measures are independent from P - N assessment, while Acutance p-values distribution it is very often independent although in a small fraction of bootstrap samples the null hypothesis that positive and negative samples come from the same population was rejected, having set the test significance to 5%

6.5.2 Visual validation

Quality measures which I have defined (Watershed, DeltaContrast, Acutance) have been visually validated by an experienced nuclear medicine physician who viewed and assessed 150 images. In particular, for each measure the physician focused his attention on the feature evaluated by the measure itself, giving a dichotomous evaluation. Thus, the physician evaluated texture graininess, noise amplitude,

6.5.3 Testing the ability of data provenance reconstruction

During my thesis I extensively discussed the relation between image quality and data provenance. In particular, data provenance are often homogeneous within centers, hence PET scans coming from the same clinical center are often comparable about quality. Therefore it is reasonable to expect that, given the images, quality measures have the ability to trace back to respective clinical centers and to certain aspects of reconstruction methods.

In this section I test whether quality measures have this ability.

I explored relation between quality measures and data provenance using two random forest classifiers, one predicting acquisition centers and one predicting a main aspects of data provenance, namely the possible use of TOF and/or PSF during acquisition and reconstruction steps. in order to investigate the quality measures prediction ability with regard to the two output classes just mentioned. The input of random forest were the 5 quality measures which I introduced in section 6.3.

Training set and test set have been randomly chosen with the following proportion: 70% for the former, 30% for the latter. The database is imbalanced about output classes (see tables 4.1 and 4.3): as discussed in section 3.2, training a classifier on an imbalanced dataset can lead to loss in accuracy in minority classes predictions. To avoid this problem, we randomly oversampled the training set in order to balance the numerosity of the output classes.

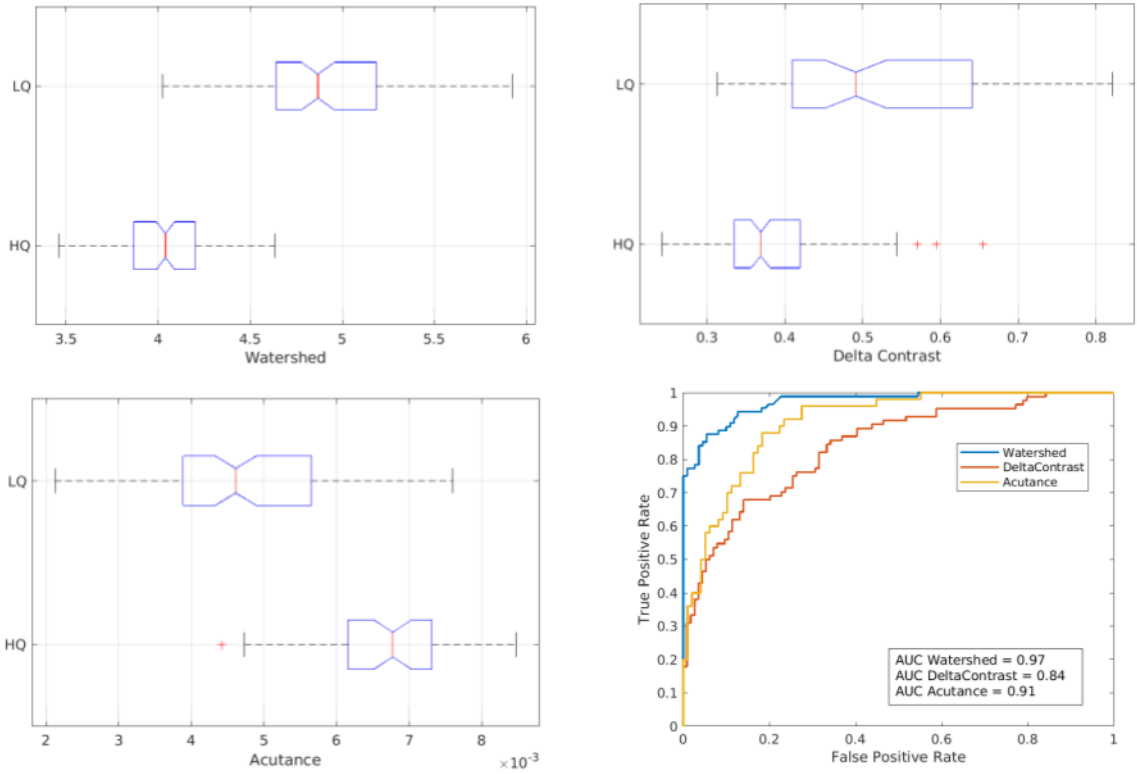


Figure 6.12: Visual validation results of quality measures. Boxplots of each measure versus visual label (HQ/LQ) are reported. Furthermore in the bottom right figure the ROC curves of each measures and the respective AUC are reported.

Total accuracy [92.5%]					
True Class	iterative	96.9%	2.1%	1.0%	
	iterative+PSF		83.3%	16.7%	
	iterative+PSF+TOF	7.1%	4.0%	88.9%	
	iterative+TOF		4.2%		95.8%
		iterative	iterative+PSF	iterative+PSF+TOF	iterative+TOF
Predicted Class					

Figure 6.13: Reconstruction methods confusion chart.

I chose for both the classifiers 300 classification trees. The minimum number of leaf and number of randomly selected predictors have been considered as hyperparameters optimized by minimizing the out of bag error. Once hyperparameter has been optimized, I trained the algorithm on the whole training set, then I tested the generalization ability on the test set. The prediction results for the test set are summarized in confusion charts 6.14 and 6.13.

Centers-quality measures relation

Let focus the attention on the wrong prediction of center confusion charts illustrated in 6.14. As we are considering this

chart as a validation check for measures, it is important to notice that it is not reasonable to expect a close to zero prediction error.

Indeed it may happen that images acquired in different centers may have a similar data provenance, (see *batch variable versus center variable* in section 5.3) therefore such images may be rather homogeneous with respect to image quality. When such a situation occurs, it is reasonable to expect that the quality measures are not able to distinguish between

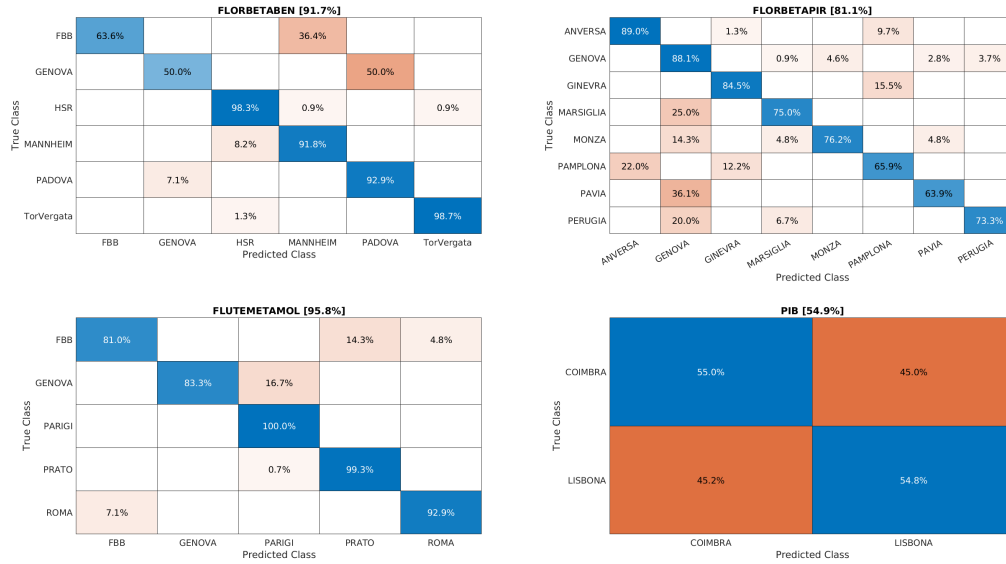


Figure 6.14: Confusion chart of the random forest prediction ability tracer by tracer. The total accuracy is provided for each tracer.

centers. An extreme example is given by the PIB tracer confusion chart: PIB images formally comes from two different centers, Coimbra and Lisbona. However, even though these two clinical centers enroll populations independently (thus are considered distinct), they conduct PET examinations using the same PET scanner, acquisition and registration protocols, hence obtaining images of equal quality on average. The fact that the prediction error is close to 50 percent for PIB images demonstrates that the quality measures I chose are not able to distinguish between the two centers. This represents a further validation of measures, as it shows that they are sensitive only to image quality (batch effect) and not to clinical and demographic variability that might be present between different independently enrolled samples.

Reconstruction methods-quality measure relation The use of TOF and PSF, as discussed in chapter 2, is a very important factors in determining image quality, though obviously is not the only one. The pretty good accuracy ($\sim 90\%$) shown in confusion chart 6.13 shows that quality measures have a good ability to discern such factors: this is a good validation check for measures themselves.

It might be interesting to test whether quality measures are also able to distinguish between images that have been reconstructed with different number of subsets and iterations as well as with different scanners. However, the paucity of data with regard to such information, coupled with the complex interactions between acquisition/reconstruction variables would make this test not feasible.

For example, if we would know the relation between quality measures and the number of iterations, it would be good to have a large number of images in which all other sources of variability are fixed (e.g., scanner, PSF/TOF inclusion, number of subsets). Indeed if other variables involved also vary, it becomes very difficult ,if not impossible, to isolate the effect of the parameter of interest (the number of iterations) on quality measures. The confusion chart in figure 6.13 is partially exempt from this criticality because the

use of PSF and TOF is a very important factor, which is considered dominant over other parameters variability [6, 11, 39, 7, 22]: this makes the confusion chart 6.13 significant even though other degrees of freedom are not fixed.

6.6 Harmonization of quantification values

In this section I will use the quality measures previously defined to harmonize the quantification values using a linear model, then I will explore the consequences of this harmonization.

Here I will consider 4 different linear models in order to explain the quantification values, denoted by S , in terms of combinations of the following covariates: the center variable C (categorical), the tracer variable T (categorical), the vector of PCA scored quality measures and the matrix dimension M_{dim} (M_{dim} is considered as categorical), denoted by $Q = \{q_1, q_2, q_3, q_4, M_{dim}\}$ and the vector K (both categorical and numerical) which contains demographic/clinical variables (namely age, sex, MMSE, Education).

Thus I considered the following linear models

$$S = \alpha + \beta C + \gamma T + \epsilon \quad (6.8)$$

$$S = \alpha + \beta Q + \gamma T + \epsilon \quad (6.9)$$

$$S = \alpha + \beta C + \gamma K + \delta T + \epsilon \quad (6.10)$$

$$S = \alpha + \beta Q + \gamma K + \delta T + \epsilon \quad (6.11)$$

where ϵ is the residual term.

Then I fitted these models and from each one I computed the residual variance ϵ .

To get more robust results I made this step with using 1000 balanced bootstrap samples. The results reported in figure 6.15 show that models fitted using the center as a covariate variable lead to a residual variance distribution with lower mean than models fitted using quality measures. This seems to suggest that the center variable explains more variance than the quality variable. It is interesting to note that the gap between residual variance of center and quality models could be due to the fact that the center variable contains, in addition to information on data provenance (batch effect) also "hidden" demographic/clinical information not considered in the model. In other words, in agreement with what was said in the section 5.4, the center variable might contain LCVs, while the quality measures which are specific to the batch effect, do not. This could account for the difference between such distributions of residuals.

Furthermore, as already happened in the section related to the validation of the measurements (section 6.5.3), also in this section the PIB plays a particular role. We know that the data provenance of the PIB images is homogeneous: for this reason we should expect that the quality measures do not reduce the residual variance of the PIB. The fact that this actually happens, as shown in figure 6.15, demonstrates the consistency of the results obtained.

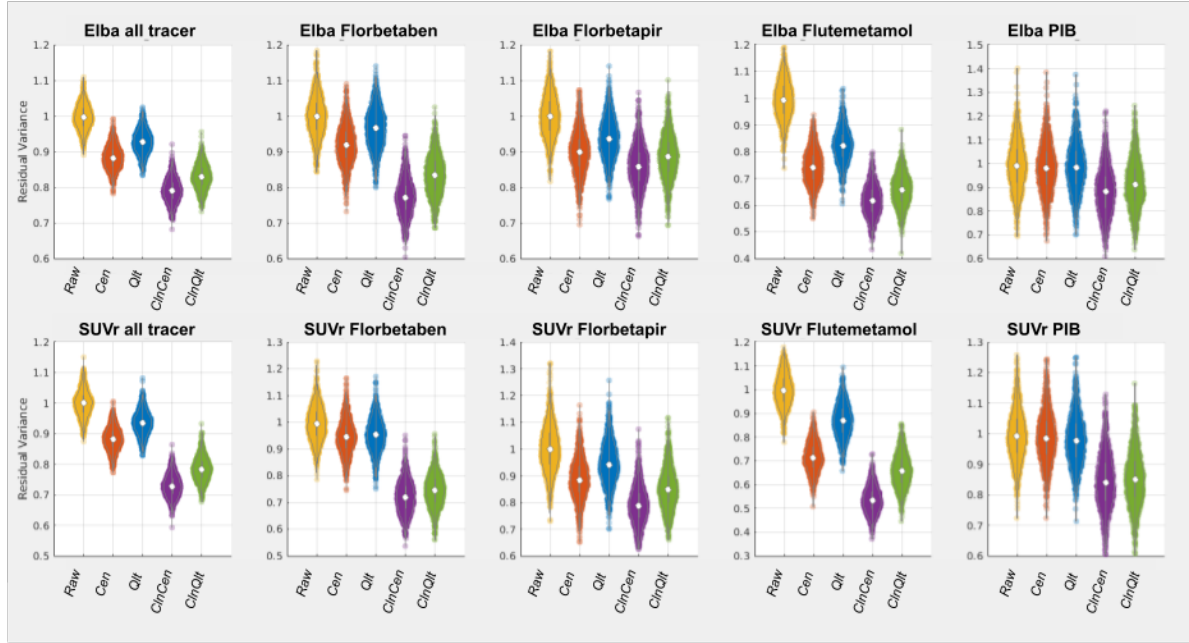


Figure 6.15: The residual variance distribution for each model. *Raw* denotes the quantification variaces, i.e. the not harmonized output. *Cen*, *Qlt*, *ClnCen* and *ClnQlt* denotes the residual variance of model fitted using as covariates respectively center label 6.8, quality measures 6.9, both clinical vector K and center label 6.10, both clinical vector K and quality measure 6.11.

6.7 Estimation of AIC and BIC for different models

One of the main problems of the linear fit that I have considered is that the 4 models considered have a number of degrees of freedom different from each other. The use of a different number of degrees of freedom (i.e. covariates) could make the results of the fit not comparable. In this section I will use the AIC and BIC methods introduced in section 3.5.2 in order to test the goodness of fit using methods that account for the difference in degrees of freedom.

I performed a bootstrap sampling repeated 1000 times on the dataset, resulting in 1000 bootstrap samples. For each bootstrap sample I fitted the 4 models considered above and I calculated their AIC and BIC. I considered the quantification data all together, without separating by individual tracer, in order to have fits that were more robust. Results for AIC and BIC are shown in the figure 6.16 and 6.17 respectively.

It is interesting to note that the models 6.8 and 6.10 (those containing the center labels) always better fit the data with respect to the models 6.9 and 6.11 (those containing the quality measures), both using the AIC and the BIC criteria. This shows that models fitted using the center label explain better the quantification data than models fitted using the quality measures.

Moreover in the figures 6.16 and 6.17 I have reported the boxplots of the differences of AIC and BIC between models fitted using the quality measures and models fitted using center labels. Observing these graphs is interesting to note that the difference of AIC and BIC is always extremely significant, as it is much greater than 10 (see section 3.5.2).

I notice that this confirm results illustrated in 6.15.

This suggests that the center label could describe not only batch variability, (data

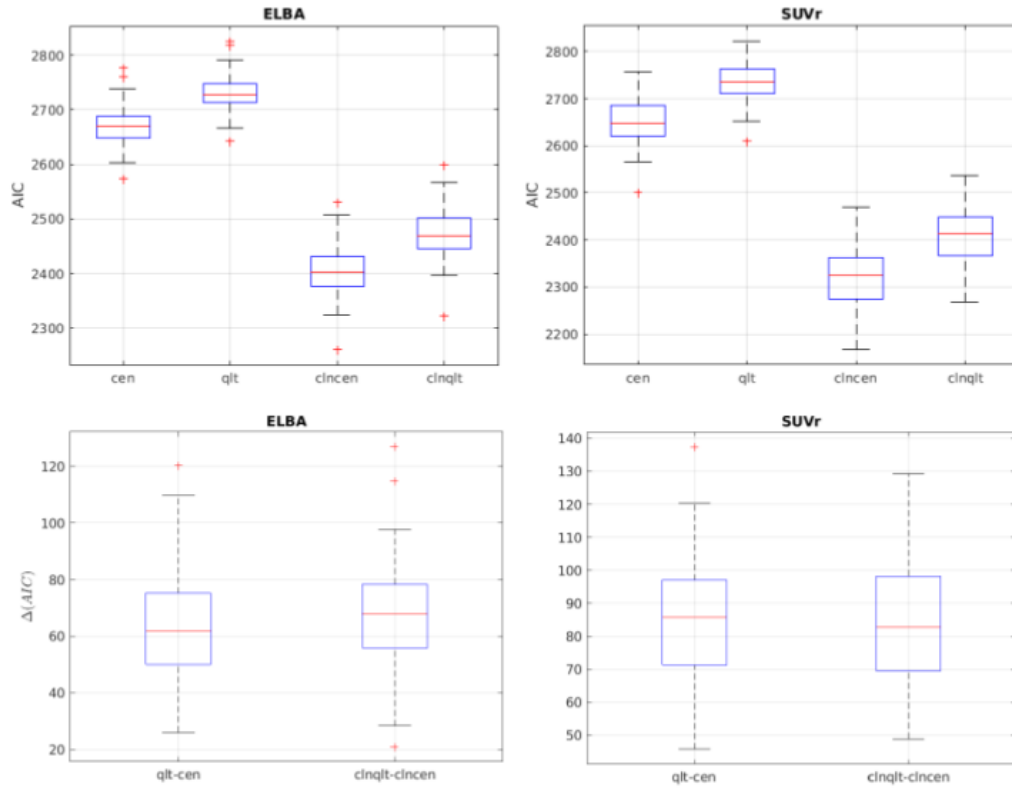


Figure 6.16: *cen* and *qlt* denotes the models fitted using center label (6.8) and quality measures (6.9) respectively. In addition, models denoted by *cln* were fitted using patient clinical and demographic information as well (see equations 6.10 and 6.11). In top boxplots you can find the AIC values for models fitted using the two quantification methods ELBA and SUVR as output. In bottom boxplots the differences in AIC (related to top boxplots) for the models with (*clncen-clnqlt*) and without (*cen-qlt*) clinical/demographic information are shown

provenance), but also additional demographic and/or clinical information about the sample. In other words, the center label could contain latent confounding information (the LCV's illustrated in figure 5.1). This greater amount of demographic/clinical information contained in the center label fitted models than in the quality measures fitted models could justify a significantly better fit.

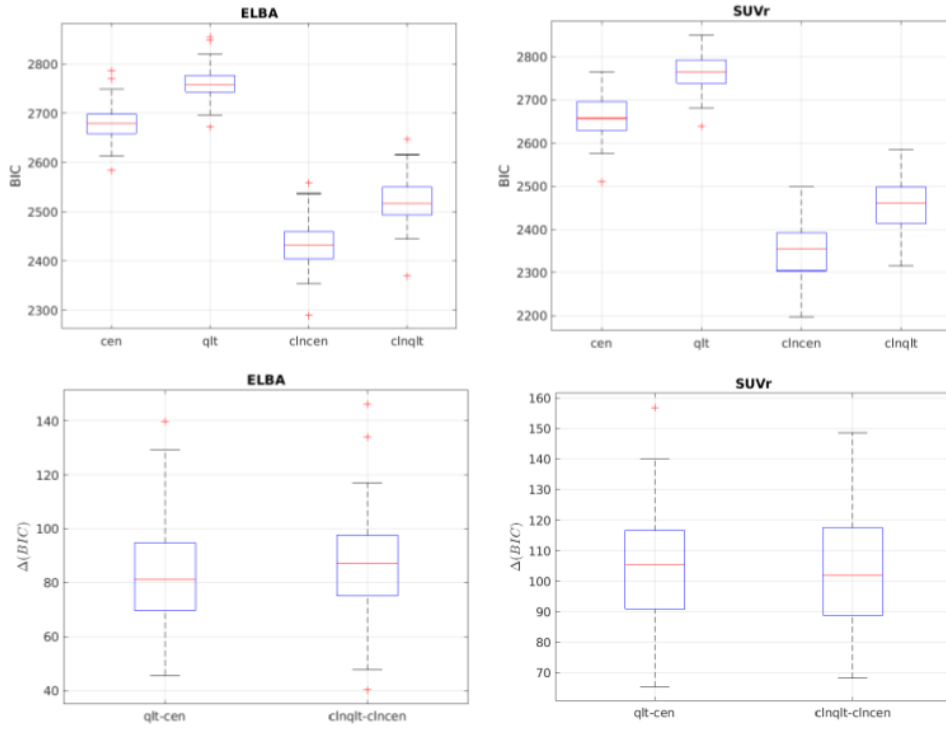


Figure 6.17: *cen*, *qlt*, *clncen* and *clnqlt* have been define in figure 6.16. In top boxplots you can find the BIC values for models fitted using the two quantification methods ELBA and SUVR as output. In bottom boxplots the differences in BIC (related to top boxplots) for the models with (*clncen-clnqlt*) and without (*cen-qlt*) clinical/demographic information are shown.

6.8 Discussion

In this section I will discuss the limitations and the prospects of this work.

The final goal of my work was to model the batch effect, in order to lay the foundations for an a-posteriori harmonization that is not affected by the problem of latent confounding variables. To achieve this aim, I defined and used a set of quality measures which I evaluated on specific ROIs. Then I performed a simple harmonization based on a linear regression model and I explored some of the consequences of this harmonization by comparing corrections to quantification using my approach with corrections made using a linear model in which data provenance has been taken into account through a categorical center variable. This latter approach is commonly used in the literature.

In chapter 4.2 I described the registration and quantification algorithm I used. Specifically, registration was done in the absence of MRI structural image guidance. This may have introduced errors into the registration and consequently quantification process. Repeating the analysis using a database where the MRI structural image is also provided could be interesting in the future.

The quality measures I described in section 6.3 were validated using different approaches. However, there is no guarantees that quality measures cover the whole spectrum of image quality: they may capture some aspects of quality, but not describe it completely. If this were true, the quality measures would not fully describe the batch effect. The part of the batch effect not explained by quality would end up being considered

as part of the latent confounding variables and hence it will be wrongly preserved.

A possible criticism may be related to the choice of ROIs. I only used transaxial regions, and I chose regions relatively close to the edge of the image. In my opinion these ROIs selection were the best possible choice compatible with the properties defined in section 6.2. However, regions too close to the image's edge could be affected by artifacts in the acquisition / reconstruction phase. Even though it is true that the anatomical ROIs I chose are relatively close to the edge of the registered image, it is also true that the same anatomical ROIs considered in the raw images embedded in native space are not so close to image's edge. Indeed, registration step remove some clinically not relevant pieces of native images actually acquired which are located outside the domain of the ICBM template used for registration. So, in my opinion, we can consider this ROIs selection acceptable. Furthermore, from a visual analysis of the image, I did not find any particular problems related to this topic.

Another criticism is related to the determination of the relation between quantification and quality carried out using linear regression models. Data are not homogeneously distributed in the quality space, nor are they numerically homogeneous with respect to the acquisition methods and clinical amyloid load dichotomic assessment. For this reason the fits of the considered linear models could enhance the inhomogeneity of the data rather than capture the actual relation between quantification and quality. To avoid this problem, I considered a balanced bootstrap approach.

Moreover it would be interesting to consider the relations between my quality measures and those that are currently used in literature; however to do this comparison, phantom images are required. As discussed in section 2.8, all the quality measures used in literature (as far as I know) require a ground truth reference, typically the knowledge of the actual radiotracer distribution, which is always known in phantom studies, but unknown in clinical images. For this reason such a comparison using clinical images is not feasible. However the measures I introduced could be used in the context of phantom studies and be compared with those used in literature.

An interesting consideration can be made regarding the generalizability to other type of imaging methods of the approach I proposed. I built and validated quality measures using amyloid PET images. This does not imply that they cannot be used in other imaging fields, but their validity outside the amyloid PET context would have to be demonstrated. I believe they could be used directly (or with little modification), in other types of PET images (such as tau PET or FDG PET images) thanks to the texture and visual perception similarity between these imaging types. For example, in the context of FDG PET, they could be used to harmonize radiomic features extracted from the images themselves. This is a major problem in FDG PET imaging.

However it must be borne in mind that the approach I have proposed requires the definition of a given set of ROIs in which perform quality measures. Such regions must show the effect of data provenance at best, and at the same time they should have a very low anatomical inter-subject variability and do not contains any clinical information. Indeed ROIs selection is an imaging-specific issue: ROIs I have chosen for amyloid neuroimaging PET may not be appropriate in other types of PET. In addition, the FDG-PET are also

carried out to other anatomical district (such as liver, lung, whole body), making the ROIs defined in the equation 6.1 unusable.

Furthermore I believe that quality measures I defined are not appropriate for MRI features harmonization (an equally important issue), since MRI texture are too far from PET images typical texture. However it is important to point out that the general aspects of the approach that I followed could be used also in the field of MRI images. It would be a matter of finding a set of quality measures and ROIs appropriately defined ad-hoc for MRI.

A further test for the quality measures could be to consider a given sinogram, thus reconstruct images from sinogram using different techniques, for example using an OSEM reconstruction and varying the number of subsets and iterations, and therefore investigate how the image quality varies accordingly.

Furthermore, it could be interesting to verify the consequences of harmonization approach I proposed on regional quantification values, in order to better appreciate its effects.

Finally, I mainly dealt with the issues of quality measures, while the harmonization of the variables in itself was done with a linear regression model. Although harmonization through linear models is often used, more sophisticated techniques are present in literature, for exaple ComBat. It would be interesting to combine them with the quality measures I defined and more in genereal with the basics of my work.

Conclusion

My PhD work is placed in the context of PET imaging postreconstruction harmonization.

When PET images from different research centers are gathered in a single multicentric study, quantities of interest extracted from images themselves could be affected by data provenance, i.e. any technical effects due to the use of different protocols and scanners in data acquisition and images reconstruction. Furthermore, samples coming from different centers could not be statistically equivalent with regard to the distribution of demographic variables. A good postreconstruction harmonization method should preserve the demographic inter-individual variability while removing effects introduced by data provenance.

Postreconstruction harmonization methods currently used in literature are able to achieve this aim only if all the demographic confounding variables are known. The complexity that characterizes biological mechanisms makes in practice impossible to identify all the confounding variables to be preserved.

For this reason, I proposed an harmonization method based on modeling the data provenance effects directly on clinical images through the use of 5 quality measures evaluated in appropriate regions of interest (ROIs). Using this harmonization scheme do not requires to provide a matrix of confounding to preserve, since the effect of data provenance can be estimated on images themselves and therefore removed.

Specifically, I defined 3 quality measures based on images texture properties, moreover I used a quality metric already implemented in Matlab, and I considered the reconstruction matrix dimension to take into account image resolution.

ROIs selection was also an important issue, as ROIs are required to be as independent as possible from individual variability in order to highlight the effect of data provenance on images texture only.

Quality measures have been validated both visually and using a random forest algorithm to test their ability to predict some important image reconstruction properties, which are known to be closely related to perceived quality. Furthermore, as quality measures are required to be related to image quality only, I tested the independence between quality and cortical amyloid load, which was visually assessed by expert physicians through a dichotomic positive/negative label.

Therefore I used a linear regression model to harmonize quantification values using quality measures as covariates and quantification values as output.

I explored the consequences of this harmonization by comparing corrections to quantification using my approach with corrections made using a linear model in which data provenance has been taken into account through a categorical center variable. This latter approach is commonly used in the literature.

In particular I considered the residual variances of models and I also computed their Aikake Information Criterion and Bayesian Information Criterion values. These three results are consistent with each other: harmonization performed using categorical center variables are significantly different with respect to harmonization based on quality measures. Moreover, the center-based harmonization has a lower residual variance than the quality-based harmonization. It is plausible that differences in residual variance are related to information carried by the center variable: it might contains "hidden" relevant unknown demographic variables (e.g. comorbidities, genetics, lifestyle) which could be related with the disease (i.e. cortical amyloid load) and therefore reduces the residual variance.

Quality measures have been built and validated using a multicentric amyloid PET database. It is important to remark that although the main principles of my proposal may be generalizable to other typologies of medical imaging, quality measures and in particular ROIs selection are image's type dependent. However I believe quality measures could be successfully used in other PET imaging modalities (such as in FDG PET and tau PET), even though its validity outside the context of amyloid PET imaging must be demonstrated.

I am currently exploring in detail the consequences of my quality based harmonization and I am considering the possibility of integrating the basics of my proposal with more sophisticated harmonization algorithms present in literature, such as ComBat. To conclude, I am presently using my harmonization method in a multicentric study (in collaboration with San Martino Hospital of Genoa¹) which aims to find a relation between cerebral microbleeds and regional amyloid burden.

¹<https://www.ospedalesanmartino.it/>

Bibliography

- [1] Richard G. Abramson, Kirsteen R. Burton, John Paul J. Yu, Ernest M. Scalzetti, Thomas E. Yankeeelov, Andrew B. Rosenkrantz, Mishal Mendiratta-Lala, Brian J. Bartholmai, Dhakshinamoorthy Ganeshan, Leon Lenchik, and Rathan M. Subramaniam. Methods and Challenges in Quantitative Imaging Biomarker Development, jan 2015.
- [2] Nicolas Aide, Charline Lasnon, Patrick Veit-Haibach, Terez Sera, Bernhard Sattler, and Ronald Boellaard. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *European Journal of Nuclear Medicine and Molecular Imaging*, 44:17–31, 2017.
- [3] Marco Aiello, Carlo Cavaliere, Antonio D’Albore, and Marco Salvatore. The Challenges of Diagnostic Imaging in the Era of Big Data. *Journal of Clinical Medicine*, 8(3):316, mar 2019.
- [4] Hirotugu Akaike. A New Look at the Statistical Model Identification. pages 215–222. 1974.
- [5] Go Akamatsu, Yasuhiko Ikari, Tomoyuki Nishio, Hiroyuki Nishida, Akihito Ohnishi, Kazuki Aita, Masahiro Sasaki, Masayuki Sasaki, and Michio Senda. Optimization of image reconstruction conditions with phantoms for brain FDG and amyloid PET imaging. *Annals of Nuclear Medicine*, 30(1):18–28, jan 2016.
- [6] Go Akamatsu, Kaori Ishikawa, Katsuhiko Mitsumoto, Takafumi Taniguchi, Nobuyoshi Ohya, Shingo Baba, Koichiro Abe, and Masayuki Sasaki. Improvement in PET/CT image quality with a combination of point-spread function and time-of-flight in relation to reconstruction parameters. *Journal of Nuclear Medicine*, 53(11):1716–1722, nov 2012.
- [7] Go Akamatsu, Katsuhiko Mitsumoto, Kaori Ishikawa, Takafumi Taniguchi, Nobuyoshi Ohya, Shingo Baba, Koichiro Abe, and Masayuki Sasaki. Benefits of point-spread function and time of flight for PET/CT image quality in relation to the body mass index and injected dose. *Clinical Nuclear Medicine*, 38(6):407–412, jun 2013.
- [8] Roberto Alejo, Jose M. Sotoca, R. M. Valdovinos, and Gustavo A. Casañ. The multi-class imbalance problem: Cost functions with modular and non-modular neural networks. In *Advances in Intelligent and Soft Computing*, volume 56, pages 421–431. Springer Verlag, 2009.

- [9] Adam M. Alessio, Paul E. Kinahan, and Thomas K. Lewellen. Modeling and incorporation of system response functions in 3D whole body PET. In *IEEE Nuclear Science Symposium Conference Record*, volume 6, pages 3992–3996, 2004.
- [10] B. Ali, A. Afshan, and M.B. Kakakhel. The Impact of Varying Number of OSEM Iterations on Standardized Uptake Value and Image Quality of Discovery STE PET/CT Scanner. *Journal of Global Oncology*, 4(Supplement 2):68s–68s, oct 2018.
- [11] P Alongi, M Picchio, V Bettinardi *Journal of Nuclear ...*, and Undefined 2012. Impact of time-of-flight (TOF) and point-spread-function (PSF) PET on whole-body oncologic studies. *Soc Nuclear Med*, 53(supplement 1), 2012.
- [12] Georgios I. Angelis. Novel Spatiotemporal Image Reconstruction for High Resolution PET Imaging in Neuroscience. Technical report, University of Manchester, 2011.
- [13] Anonymous. Biomarkers In Risk Assessment: Validity And Validation. *Environmental Health*, page 144, 2001.
- [14] Bb Avants, Nick Tustison, and Gang Song. Advanced Normalization Tools (ANTs). Technical report, 2009.
- [15] Brian B. Avants, Nicholas J. Tustison, Michael Stauffer, Gang Song, Baohua Wu, and James C. Gee. The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, 8(APR):44, apr 2014.
- [16] Ramsey Badawi. Introduction to PET Physics - 2D mode and 3D mode, 1999.
- [17] Ramsey D Badawi. Introduction to PET Physics: Image Reconstruction.
- [18] Solveig Badillo, Balazs Banfai, Fabian Birzele, Iakov I. Davydov, Lucy Hutchinson, Tony Kam-Thong, Juliane Siebourg-Polster, Bernhard Steiert, and Jitao David Zhang. An Introduction to Machine Learning. *Clinical Pharmacology and Therapeutics*, 107(4):871–885, apr 2020.
- [19] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, feb 1994.
- [20] Jen Beatty. The Radon Transform and Medical Imaging. *CBMS-NSF Regional Conference Series in Applied Mathematics*, (85), 2014.
- [21] Rinaldo Bellomo, Stephen J. Warrillow, and Michael C. Reade. Why we should be wary of single-center trials. *Critical Care Medicine*, 37(12):3114–3119, dec 2009.
- [22] V. Bettinardi, L. Presotto, E. Rapisarda, M. Picchio, L. Gianolli, and M. C. Gilardi. Physical Performance of the new hybrid PETCT Discovery-690. *Medical Physics*, 38(10):5394–5411, 2011.
- [23] Thomas Beyer, Johannes Czernin, and Lutz S. Freudenberg. Variations in clinical PET/CT operations: Results of an international survey of active PET/CT users. *Journal of Nuclear Medicine*, 52(2):303–310, feb 2011.

- [24] W. Dean Bidgood, Steven C. Horii, Fred W. Prior, and Donald E. Van Syckle. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging, 1997.
- [25] Kaj Blennow, Niklas Mattsson, Michael Schöll, Oskar Hansson, and Henrik Zetterberg. Amyloid biomarkers in Alzheimer’s disease. *Trends in Pharmacological Sciences*, 36(5):297–309, may 2015.
- [26] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.
- [27] Leo Breiman, Jerome Friedman, and Charles Stone. *Classification and Regression Trees - 1st Edition*. Chapman and Hall/CRC, 1984.
- [28] Philippe P. Bruyant. Analytic and iterative reconstruction algorithms in SPECT. *Journal of Nuclear Medicine*, 43(10):1343–1358, 2002.
- [29] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Technical report, 2002.
- [30] Andrea Chincarini, Paolo Bosco, Piero Calvini, Gianluca Gemme, Mario Esposito, Chiara Olivieri, Luca Rei, Sandro Squarcia, Guido Rodriguez, Roberto Bellotti, Piergiorgio Cerello, Ivan De Mitri, Alessandra Retico, and Flavio Nobili. Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer’s disease. *NeuroImage*, 58(2):469–480, sep 2011.
- [31] Andrea Chincarini, Paolo Bosco, Gianluca Gemme, Mario Esposito, Luca Rei, Sandro Squarcia, Roberto Bellotti, Lennart Minthon, Giovanni Frisoni, Philip Scheltens, Lutz Frölich, Hilka Soininen, Pieter Jelle Visser, and Flavio Nobili. Automatic temporal lobe atrophy assessment in prodromal AD: Data from the DESCRIPA study. *Alzheimer’s and Dementia*, 10(4):456–467, jul 2014.
- [32] Andrea Chincarini, Francesco Sensi, Luca Rei, Irene Bossert, Silvia Morbelli, Ugo Paolo Guerra, Giovanni Frisoni, Alessandro Padovani, and Flavio Nobili. Standardized Uptake Value Ratio-Independent Evaluation of Brain Amyloidosis. *Journal of Alzheimer’s Disease*, 54(4):1437–1457, oct 2016.
- [33] Andrea Chincarini, Francesco Sensi, Luca Rei, Irene Bossert, Silvia Morbelli, Ugo Paolo Guerra, Giovanni Frisoni, Alessandro Padovani, and Flavio Nobili. Standardized Uptake Value Ratio-Independent Evaluation of Brain Amyloidosis. *Journal of Alzheimer’s Disease*, 54(4):1437–1457, oct 2016.
- [34] Andrea Chincarini, Francesco Sensi, Luca Rei, Gianluca Gemme, Sandro Squarcia, Renata Longo, Francesco Brun, Sabina Tangaro, Roberto Bellotti, Nicola Amoroso, Martina Bocchetta, Alberto Redolfi, Paolo Bosco, Marina Boccardi, Giovanni B. Frisoni, and Flavio Nobili. Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer’s disease. *NeuroImage*, 125:834–847, jan 2016.
- [35] Christopher M. Clark, Julie A. Schneider, Barry J. Bedell, Thomas G. Beach, Warren B. Bilker, Mark A. Mintun, Michael J. Pontecorvo, Franz Hefti, Alan P. Carpenter, Matthew L. Flitter, Michael J. Krautkramer, Hank F. Kung, R. Edward Coleman, P. Murali Doraiswamy, Adam S. Fleisher, Marwan N. Sabbagh, Carl H.

- Sadowsky, P. Eric M. Reiman, Simone P. Zehntner, and Daniel M. Skovronsky. Use of florbetapir-PET for imaging β -amyloid pathology. *JAMA - Journal of the American Medical Association*, 305(3):275–283, jan 2011.
- [36] E. F. Codd. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6):377–387, jun 1970.
 - [37] Ann D. Cohen, Susan M. Landau, Beth E. Snitz, William E. Klunk, Kaj Blennow, and Henrik Zetterberg. Fluid and PET biomarkers for amyloid pathology in Alzheimer’s disease, jun 2019.
 - [38] T M Connolly and C E Begg. *Database Systems: A practical approach to design, implementation and management*. Pearson Education Limited, 1998.
 - [39] Maurizio Conti and Bernard Bendriem. The new opportunities for high time resolution clinical TOF PET, feb 2019.
 - [40] Ian Goodfellow Courville, Yoshua Bengio, and Aaron. *Deep learning* *Deep Learning*, volume 29. MIT Press, 2016.
 - [41] R. Da-ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, R. De Crevoisier, J. F. Rameé, O. Pradier, U. Schick, D. Visvikis, and M. Hatt. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports*, 10(1):10, dec 2020.
 - [42] R. Da-Ano, D. Visvikis, and M. Hatt. Harmonization strategies for multicenter radiomics investigations, dec 2020.
 - [43] Samrat Jayanta Dattagupta. A performance comparison of oversampling methods for data generation in imbalanced learning tasks. Technical report, 2018.
 - [44] H. Olaya Dávila, S. A. Martínez Ovalle, H. Pérez, and H. Castro. Determination of Spatial Resolution of Positron Emission Tomograph of Clear PET-XPAD3/CT System. *Universal Journal of Physics and Application*, 11(4):97–101, 2017.
 - [45] Michel Defrise, Paul E Kinahan, and Christian J Michel. Image Reconstruction Algorithms in PET. Technical report, 2006.
 - [46] Mustafa Demir, Türkey Toklu, Mohammad Abuqbeitah, Hüseyin Çetin, H. Sezer Sezgin, Nami Yeyin, and Kerim Sönmezolu. Evaluation of pet scanner performance in pet/mr and pet/ct systems: Nema tests. *Molecular Imaging and Radionuclide Therapy*, 27(1):10–18, feb 2018.
 - [47] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm . Technical Report 1, 1977.
 - [48] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212, sep 1996.
 - [49] Ivo D. Dinov, John D. Van Horn, Kamen M. Lozev, Rico Magsipoc, Petros Petrosyan, Zhizhong Liu, Allan MacKenzie-Graham, Paul Eggert, Douglas S. Parker, and Arthur W. Toga. Efficient, distributed and interactive neuroimaging data analysis using the LONI Pipeline. *Frontiers in Neuroinformatics*, 3(JUL):22, jul 2009.

- [50] Hisham El-Amir, Mahmoud Hamdy, Hisham El-Amir, and Mahmoud Hamdy. Data Resampling. In *Deep Learning Pipeline*, pages 207–231. Apress, 2020.
- [51] Salaheddin G. Elkadiki and Rangaraj M. Rangayyan. <title>Objective characterization of image acutance</title>. *Medical Imaging 1994: Image Perception*, 2166:210–218, apr 1994.
- [52] Jonathan Engle and Dan Kadrmas. Modeling the spatially-variant point spread function in a fast projector for improved fully-3D PET reconstruction. *Journal of Nuclear Medicine*, 48(supplement 2):417P–417P, 2007.
- [53] Trevor et. all. Hastie. Springer Series in Statistics The Elements of Statistical Learning. Technical Report 2, 2009.
- [54] Frank J. Fabozzi, Sergio M. Focardi, Svetlozar T. Rachev, and Bala G. Arshanapalli. Appendix E: Model Selection Criterion: AIC and BIC. In *The Basics of Financial Econometrics*, pages 399–403. John Wiley & Sons, Inc., Hoboken, NJ, USA, mar 2014.
- [55] Aziz Fajar, Riyanarto Sarno, Chastine Fatichah, and Achmad Fahmi. Reconstructing and resizing 3D images from DICOM files. *Journal of King Saud University - Computer and Information Sciences*, dec 2021.
- [56] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, jun 2006.
- [57] Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V. Chawla. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary, apr 2018.
- [58] Bruce Fischl and Anders M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055, sep 2000.
- [59] Isabel Fortier, Paul R. Burton, Paula J. Robson, Vincent Ferretti, Julian Little, Francois L’Heureux, Mylène Deschênes, Barthia M. Knoppers, Dany Doiron, Joost C. Keers, Pamela Linksted, Jennifer R. Harris, Geneviève Lachance, Catherine Boileau, Nancy L. Pedersen, Carol M. Hamilton, Kristian Hveem, Marilyn J. Borugian, Richard P. Gallagher, John McLaughlin, Louise Parker, John D. Potter, John Gallacher, Rudolf Kaaks, Bette Liu, Tim Sprosen, Anne Vilain, Susan A. Atkinson, Andrea Rengifo, Robin Morton, Andres Metspalu, H. Erich Wichmann, Mark Tremblay, Rex L. Chisholm, Andrés Garcia-Montero, Hans Hillege, Jan Eric Litton, Lyle J. Palmer, Markus Perola, Bruce H.R. Wolffenbuttel, Leena Peltonen, and Thomas J. Hudson. Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. *International Journal of Epidemiology*, 39(5):1383–1393, oct 2010.
- [60] Isabel Fortier, Dany Doiron, Paul Burton, and Parminder Raina. Invited commentary: Consolidating data harmonization - How to obtain quality and applicability? *American Journal of Epidemiology*, 174(3):261–264, aug 2011.

- [61] Isabel Fortier, Parminder Raina, Edwin R. Van den Heuvel, Lauren E. Griffith, Camille Craig, Matilda Saliba, Dany Doiron, Ronald P. Stolk, Bartha M. Knoppers, Vincent Ferretti, Peter Granda, and Paul Burton. Maelstrom Research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, 46(1):103–115, jun 2017.
- [62] Jean Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, feb 2018.
- [63] Jean Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A. Elliott, Kosha Ruparel, David R. Roalf, Theodore D. Satterthwaite, Ruben C. Gur, Raquel E. Gur, Robert T. Schultz, Ragini Verma, and Russell T. Shinohara. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170, nov 2017.
- [64] M. Furudate, K. Ito, G. Irie, J. Ando, and A. Miyamoto. Nuclear medicine, 1983.
- [65] Gene Gindi, Mindy Lee, Anand Rangarajan, and I. George Zubal. Bayesian Reconstruction of Functional Images Using Anatomical Information as Priors. *IEEE Transactions on Medical Imaging*, 12(4):670–680, 1993.
- [66] Boris Glavic. Big data provenance: Challenges and implications for benchmarking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8163 LNCS:72–80, 2014.
- [67] S. Alireza Golestaneh, Mahesh M. Subedar, and Lina J. Karam. The effect of texture granularity on texture synthesis quality. In Andrew G. Tescher, editor, *Applications of Digital Image Processing XXXVIII*, volume 9599, page 959912. SPIE, sep 2015.
- [68] Kuang Gong, Simon R. Cherry, and Jinyi Qi. On the assessment of spatial resolution of PET systems with iterative image reconstruction. *Physics in Medicine and Biology*, 61(5):N193–N202, feb 2016.
- [69] Arthur Ardeshir Goshtasby. Theory and Applications of Image Registration, 2017.
- [70] Peter Granda, Christof Wolf, and Reto Hadorn. Harmonizing Survey Data. In *Survey Methods in Multicultural, Multinational, and Multiregional Contexts*, pages 315–332. wiley, may 2010.
- [71] Alexander M. Grant, Timothy W. Deller, Mohammad Mehdi Khalighi, Sri Harsha Maramraju, Gaspar Delso, and Craig S. Levin. NEMA NU 2-2012 performance studies for the SiPM-based ToF-PET component of the GE SIGNA PET/MR system. *Medical Physics*, 43(5):2334–2343, may 2016.
- [72] V. Grau, A. U.J. Mewes, M. Alcañiz, R. Kikinis, and S. K. Warfield. Improved watershed transform for medical image segmentation using prior information. *IEEE Transactions on Medical Imaging*, 23(4):447–458, apr 2004.
- [73] Ida Häggström. Quantitative Methods for Tumor Imaging with Dynamic PET. Technical report, 2014.

- [74] Akifumi Hagiwara, Shohei Fujita, Yoshiharu Ohno, and Shigeki Aoki. Variability and Standardization of Quantitative Imaging: Monoparametric to Multiparametric Quantification, Radiomics, and Artificial Intelligence. *Investigative Radiology*, 55(9):601–616, sep 2020.
- [75] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.
- [76] Jennifer R. Harris, Paul Burton, Bartha Maria Knoppers, Klaus Lindpaintner, Marianna Bledsoe, Anthony J. Brookes, Isabelle Budin-Ljosne, Rex Chisholm, David Cox, Mylène Deschênes, Isabel Fortier, Pierre Hainaut, Robert Hewitt, Jane Kaye, Jan Eric Litton, Andres Metspalu, Bill Ollier, Lyle J. Palmer, Aarno Palotie, Markus Pasterk, Markus Perola, Peter H.J. Riegman, Gert Jan Van Ommen, Martin Yuille, and Kurt Zatloukal. Toward a roadmap in global biobanking for health. *European Journal of Human Genetics*, 20(11):1105–1111, nov 2012.
- [77] Haibo He, Yang Bai, Eduardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1322–1328, 2008.
- [78] G T Herman. *Fundamentals of Computerized Tomography*, volume 224 of *Advances in Pattern Recognition*. Springer London, London, 2009.
- [79] Gabriella V Hirsch, Corinna M Bauer, and Lotfi B Merabet. Using structural and functional brain imaging to uncover how the brain adapts to blindness. *Journal of Psychiatry and Brain Functions*, 2(1):7, 2015.
- [80] Clément Hognon, Florent Tixier, Olivier Gallinato, Thierry Colin, Dimitris Visvikis, Vincent Jaouen, and al Stan. Standardization of Multicentric Image Datasets with Generative Adversarial Networks. Technical report, 2019.
- [81] Ching Han Hsu. A study of lesion contrast recovery for iterative PET image reconstructions versus filtered backprojection using an anthropomorphic thoracic phantom. *Computerized Medical Imaging and Graphics*, 26(2):119–127, mar 2002.
- [82] H. Malcolm Hudson and Richard S. Larkin. Accelerated Image Reconstruction Using Ordered Subsets of Projection Data. *IEEE Transactions on Medical Imaging*, 13(4):601–609, 1994.
- [83] L. Jodal, C. Le Loirec, and C. Champion. Positron range in PET imaging: An alternative approach for assessing and correcting the blurring. *Physics in Medicine and Biology*, 57(12):3931–3943, 2012.
- [84] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, jan 2007.
- [85] Jorge Jovicich, Frederik Barkhof, Claudio Babiloni, Karl Herholz, Christoph Mulert, Bart N.M. van Berckel, and Giovanni B. Frisoni. Harmonization of neuroimaging biomarkers for neurodegenerative diseases: A survey in the imaging community of perceived barriers and suggested actions. *Alzheimer’s and Dementia: Diagnosis, Assessment and Disease Monitoring*, 11:69–73, 2019.

- [86] Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. Technical Report 8, 2018.
- [87] Saki Kimoto, NAOKI HASHIMOTO, Ayano Shoji, Yuji Tsutsui, Kazuhiko Himuro, Shingo Baba, Akihiko Takahashi, and Masayuki Sasaki. The evaluation of the spatial resolution of ^{11}C -, ^{18}F - and ^{64}Cu -PET images on a clinical PET/CT scanner using Monte Carlo Simulation and phantom examination. *Journal of Nuclear Medicine*, 59(supplement 1):2102, 2018.
- [88] Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson, Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, jul 2009.
- [89] William E. Klunk, Robert A. Koeppe, Julie C. Price, Tammie L. Benzinger, Michael D. Devous, William J. Jagust, Keith A. Johnson, Chester A. Mathis, Davneet Minhas, Michael J. Pontecorvo, Christopher C. Rowe, Daniel M. Skovronsky, and Mark A. Mintun. The Centiloid project: Standardizing quantitative amyloid plaque estimation by PET. *Alzheimer’s and Dementia*, 11(1):1–15.e4, 2015.
- [90] Gitte M. Knudsen, Melanie Ganz, Stefan Appelhoff, Ronald Boellaard, Guy Bormans, Richard E. Carson, Ciprian Catana, Doris Doudet, Antony D. Gee, Douglas N. Greve, Roger N. Gunn, Christer Halldin, Peter Herscovitch, Henry Huang, Sune H. Keller, Adriaan A. Lammertsma, Rupert Lanzenberger, Jeih San Liow, Talakad G. Lohith, Mark Lubberink, Chul H. Lyoo, J. John Mann, Granville J. Matheson, Thomas E. Nichols, Martin Nørgaard, Todd Ogden, Ramin Parsey, Victor W. Pike, Julie Price, Gaia Rizzo, Pedro Rosa-Neto, Martin Schain, Peter J.H. Scott, Graham Searle, Mark Slifstein, Tetsuya Suhara, Peter S. Talbot, Adam Thomas, Mattia Veronese, Dean F. Wong, Maqsood Yaqub, Francesca Zanderigo, Sami Zoghbi, and Robert B. Innis. Guidelines for the content and format of PET brain data in publications and archives: A consensus paper. *Journal of Cerebral Blood Flow and Metabolism*, 40(8):1576–1585, aug 2020.
- [91] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, mar 2012.
- [92] William R. Leo. *Techniques for Nuclear and Particle Physics Experiments*. Springer Berlin Heidelberg, 1994.
- [93] Kaja Z. Lewinn, Margaret A. Sheridan, Katherine M. Keyes, Ava Hamilton, and Katie A. McLaughlin. Sample composition alters associations between age and brain structure. *Nature Communications*, 8(1), dec 2017.
- [94] Charles X. Ling and Victor S. Sheng. Cost-Sensitive Learning and the Class Imbalance Problem. Technical report, 2008.

- [95] B. Lipinski, H. Herzog, E. Rota Kops, W. Oberschelp, and H. W. Müller-Gärtner. Expectation maximization reconstruction of positron emission tomography images using anatomical magnetic resonance information. *IEEE Transactions on Medical Imaging*, 16(2):129–136, 1997.
- [96] Markus Nowak Lonsdale and Thomas Beyer. Dual-modality PET/CT instrumentation-Today and tomorrow. *European Journal of Radiology*, 73(3):452–460, mar 2010.
- [97] Gerry Lowe, Bruce Spottiswoode, Jerome Declerck, Keith Sullivan, Mhd Saeed Sharif, Wai-Lup Wong, and Bal Sanghera. Positron emission tomography PET/CT harmonisation study of different clinical PET/CT scanners using commercially available software. *BJR—Open*, 2(1):20190035, nov 2020.
- [98] Huijuan Lu, Yige Xu, Minchao Ye, Ke Yan, Zhigang Gao, and Qun Jin. Learning misclassification costs for imbalanced classification on gene expression data. *BMC Bioinformatics*, 20(S25):681, dec 2019.
- [99] Maria Lyra, Agapi Ploussi, Maritina Rouchota, and Stella Synefia. Filters in 2D and 3D cardiac SPECT image processing, 2014.
- [100] Roger Magoulas, Ben Lorica, Brady Forrest, Jerry Michalski, Sarah Milstein, Peter Morville, Nathan Torkington, and David Weinberger. Technologies and Techniques for Large-Scale Data 18: Single Server and Distributed Data/Parallel Processing Clusters 19: A Data Architecture for Fast Platforms 23: Data Partitioning. 2009.
- [101] Henri Maître. *From Photon to Pixel Revised and Updated 2nd Edition - ISTE*. Wiley, 2017.
- [102] Nikolaos E. Makris, Marc C. Huisman, Paul E. Kinahan, Adriaan A. Lammertsma, and Ronald Boellaard. Evaluation of strategies towards harmonization of FDG PET/CT studies in multicentre trials: Comparison of scanner validation phantoms and data analysis procedures. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(10):1507–1515, oct 2013.
- [103] Granville James Matheson, Pontus Plavén-Sigraý, Jouni Tuisku, Juha Rinne, David Matuskey, and Simon Cervenka. Clinical brain PET research must embrace multi-centre collaboration and data sharing or risk its demise, feb 2020.
- [104] Niklas Mattsson, Philip S. Insel, Susan Landau, William Jagust, Michael Donohue, Leslie M. Shaw, John Q. Trojanowski, Henrik Zetterberg, Kaj Blennow, and Michael Weiner. Diagnostic accuracy of CSF Ab42 and florbetapir PET for Alzheimer’s disease. *Annals of Clinical and Translational Neurology*, 1(8):534–543, aug 2014.
- [105] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a ‘completely blind’ image quality analyzer. Technical Report 3, 2013.
- [106] A. Michael Morey and Dan J. Kadrmas. Effect of varying number of OSEM subsets on PET lesion detectability. *Journal of Nuclear Medicine Technology*, 41(4):268–273, dec 2013.

- [107] William W. Moses. Fundamental limits of spatial resolution in PET. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 648(SUPPL. 1):S236, aug 2011.
- [108] Agneta Nordberg, Stephen F. Carter, Juha Rinne, Alexander Drzezga, David J. Brooks, Rik Vandenberghe, Daniela Perani, Anton Forsberg, Bengt Långström, Noora Scheinin, Mira Karrasch, Kjell Nägren, Timo Grimmer, Isabelle Miederer, Paul Edison, Aren Okello, Koen Van Laere, Natalie Nelissen, Mathieu Vandembulcke, Valentina Garibotto, Ove Almkvist, Elke Kalbe, Rainer Hinz, and Karl Herholz. A European multicentre PET study of fibrillar amyloid in Alzheimer’s disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(1):104–114, jan 2013.
- [109] Gift Nyamundanda, Pawan Poudel, Yatish Patil, and Anguraj Sadanandam. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. *Scientific Reports*, 7(1):1–10, dec 2017.
- [110] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, jan 2016.
- [111] Silje Kjærnes Øen, Lars Birger Aasheim, Live Eikenes, and Anna Maria Karlberg. Image quality and detectability in Siemens Biograph PET/MRI and PET/CT systems phantom study. *EJNMMI Physics*, 6(1):16, dec 2019.
- [112] Johan Ofverstedt, Joakim Lindblad, and Natasa Sladoje. Fast and Robust Symmetric Image Registration Based on Distances Combining Intensity and Spatial Information. Technical Report 7, 2019.
- [113] Fanny Orlhac, Sarah Boughdad, Cathy Philippe, Hugo Stalla-Bourdillon, Christophe Nioche, Laurence Champion, Michael Soussan, Frederique Frouin, Vincent Frouin, and Irene Buvat. A postreconstruction harmonization method for multicenter radiomic studies in PET. *Journal of Nuclear Medicine*, 59(8):1321–1328, aug 2018.
- [114] Fanny Orlhac, Frédérique Frouin, Christophe Nioche, Nicholas Ayache, and Irène Buvat. Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology*, 291(1):53–59, apr 2019.
- [115] Vladimir Y. Panin, Frank Kehren, Christian Michel, and Michael Casey. Fully 3-D PET reconstruction with system matrix derived from point source measurements. *IEEE Transactions on Medical Imaging*, 25(7):907–921, jul 2006.
- [116] Larry A. Pierce, Brian F. Elston, David A. Clunie, Dennis Nelson, and Paul E. Kinahan. A digital reference object to analyze calculation accuracy of PET standardized uptake value. *Radiology*, 277(2):538–545, nov 2015.
- [117] Jean Baptiste Poline, Janis L. Breeze, Satrajit Ghosh, Krzysztof F. Gorgolewski, Yaroslav O. Halchenko, Michael Hanke, Karl G. Helmer, Daniel S. Marcus, Russell A. Poldrack, Yannick Schwartz, John Ashburner, and David N. Kennedy. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, 6(MARCH), 2012.

- [118] Payal Prajapati, Zunnun Narmawala, Nikunj P. Darji, S. Manthira Moorthi, and R. Ramakrishnan. Evaluation of Perceptual Contrast and Sharpness Measures for Meteorological Satellite Images. *Procedia Computer Science*, 57:17–24, 2015.
- [119] David Präkel. *The visual dictionary of photography*, volume 47. Routledge, 2009.
- [120] Anchalee Prasansuklab and Tewin Tencomnao. Amyloidosis in Alzheimer’s disease: The toxicity of amyloid beta ($A\beta$), mechanisms of its accumulation and implications of medicinal plants for therapy, 2013.
- [121] Bernhard Preim and Charl Botha. Image Analysis for Medical Visualization. In *Visual Computing for Medicine*, pages 111–175. Elsevier, jan 2014.
- [122] Julie C. Price, William E. Klunk, Brian J. Lopresti, Xueling Lu, Jessica A. Hoge, Scott K. Ziolko, Daniel P. Holt, Carolyn C. Meltzer, Steven T. DeKosky, and Chester A. Mathis. Kinetic modeling of amyloid binding in humans using PET imaging and Pittsburgh Compound-B. *Journal of Cerebral Blood Flow and Metabolism*, 25(11):1528–1547, nov 2005.
- [123] A. Rahmim, J. Tang, M. A. Lodge, S. Lashkari, M. R. Ay, R. Lautamäki, B. M.W. Tsui, and F. M. Bengel. Analytic system matrix resolution modeling in PET: An application to Rb-82 cardiac imaging. *Physics in Medicine and Biology*, 53(21):5947–5965, nov 2008.
- [124] Arman Rahmim, Jinyi Qi, and Vesna Sossi. Resolution modeling in PET imaging: Theory, practice, benefits, and pitfalls. *Medical Physics*, 40(6):1–35, 2013.
- [125] Rangaraj M. Rangayyan, Nema M. El-Faramawy, J. E.Leo Desautels, and Onsy A. Alim. Measures of acutance and shape for classification of breast tumors. *IEEE Transactions on Medical Imaging*, 16(6):799–810, 1997.
- [126] E. Rapisarda, V. Bettinardi, K. Thielemans, and M. C. Gilardi. Image-based point spread function implementation in a fully 3D OSEM reconstruction algorithm for PET. *Physics in Medicine and Biology*, 55(14):4131–4151, 2010.
- [127] Alberto Redolfi, Richard McClatchey, Ashiq Anjum, Alex Zijdenbos, David Manset, Frederik Barkhof, Christian Spenger, Yannik Legré, Lars Olof Wahlund, Chiara Barattieri Di San Pietro, and Giovanni B. Frisoni. Grid infrastructures for computational neuroscience: The neuGRID example. *Future Neurology*, 4(6):703–722, 2009.
- [128] David E. Rex, Jeffrey Q. Ma, and Arthur W. Toga. The LONI Pipeline Processing Environment. *NeuroImage*, 19(3):1033–1048, jul 2003.
- [129] Gabriel Reynés-Llompart, Aida Sabaté-Llobera, Elena Llinares-Tello, Josep M. Martí-Climent, and Cristina Gámez-Cenzano. Image quality evaluation in a modern PET system: impact of new reconstructions methods and a radiomics approach. *Scientific Reports*, 9(1), 2019.
- [130] Steve Ross. Q.Clear. *GE Healthcare, White Paper*, pages 1–9, 2014.

- [131] J. A. S. Sá, A. C. Almeida, B. R. P. Rocha, M. A. S. Mota, J. R. S. Souza, and L. M. Dentel. Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy. pages 1–5. Associacao Brasileira de Inteligencia Computacional - ABRICOM, jun 2016.
- [132] Gopal B. Saha. *Basics of PET Imaging: Physics, Chemistry, and Regulations*. 2010.
- [133] Gopal B. Saha. *Basics of PET Imaging*. Springer International Publishing, 2016.
- [134] Mark E. Schmidt, Ping Chiao, Gregory Klein, Dawn Matthews, Lennart Thurfjell, Patricia E. Cole, Richard Margolin, Susan Landau, Norman L. Foster, N. Scott Mason, Susan De Santi, Joyce Suhy, Robert A. Koeppe, and William Jagust. The influence of biological and technical factors on quantitative analysis of amyloid PET: Points to consider and recommendations for controlling variability in longitudinal data. *Alzheimer's and Dementia*, 11(9):1050–1068, sep 2015.
- [135] R E Schmitz, A M Alessio, and P E Kinahan. The Physics of PET / CT scanners [Online]. pages 1–16, 2013.
- [136] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, mar 2007.
- [137] Cosma Rohilla Shalizi. Advanced data analysis from an elementary point of view. *Book Manuscript*, page 801, 2013.
- [138] Naoki Shimada, Go Akamatsu, Keiichi Matsumoto, Hiromitsu Daisaki, Kazufumi Suzuki, Keiichi Oda, Michio Senda, Ukihide Tateishi, and Takashi Terauchi. A multi-center phantom study towards harmonization of FDG-PET: variability in maximum and peak SUV in relation to image noise. *Journal of Nuclear Medicine*, 61(supplement 1):1396, 2020.
- [139] Ayano Shoji, Keishin Morita, Saki Kimoto, Naoki Hashimoto, Yuji Tsutsui, Kazuhiko Himuro, Shingo Baba, and Masayuki Sasaki. The influence of the subset number on the quality of OSEM-reconstructed PET images. *Journal of Nuclear Medicine*, 58(supplement 1):1109, 2017.
- [140] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, mar 2001.
- [141] S. Sinharay. An overview of statistics in education. In *International Encyclopedia of Education*, pages 1–11. Elsevier Ltd, 2010.
- [142] Stephen M. Smith and Thomas E. Nichols. Statistical Challenges in Big Data Human Neuroimaging. *Neuron*, 97(2):263–268, 2018.
- [143] Robert Splinter. *Positron emission tomography*. Springer-Verlag, London, 2010.
- [144] Steven Staelens, Yves D’Asseler, Stefaan Vandenberghe, Michel Koole, Ignace Lemahieu, and Rik Van de Walle. A three-dimensional theoretical model incorporating spatial detection uncertainty in continuous detector PET. *Physics in Medicine and Biology*, 49(11):2337–2350, jun 2004.
- [145] Kyle Strimbu and Jorge A. Tavel. What are biomarkers?, nov 2010.

- [146] John J. Sunderland and Paul E. Christian. Quantitative PET/CT scanner performance characterization based upon the society of nuclear medicine and molecular imaging clinical trials network oncology clinical simulator phantom. *Journal of Nuclear Medicine*, 56(1):145–152, jan 2015.
- [147] John A. Swets. Measuring the accuracy of diagnostic systems. *Science Science*, 240(4857):1285–1293, 1988.
- [148] Hidemasa Takao, Osamu Abe, and Kuni Ohtomo. Computational analysis of cerebral cortex, aug 2010.
- [149] J. P. Thirion. Image matching as a diffusion process: An analogy with Maxwell’s demons. *Medical Image Analysis*, 2(3):243–260, sep 1998.
- [150] Yuji Tsutsui, Hiromitsu Daisaki, Go Akamatsu, Takuro Umeda, Matsuyoshi Ogawa, Hironori Kajiware, Shigeto Kawase, Minoru Sakurai, Hiroyuki Nishida, Keiichi Magota, Kazuaki Mori, and Masayuki Sasaki. Multicentre analysis of PET SUV using vendor-neutral software: the Japanese Harmonization Technology (J-Hart) study. *EJNMMI Research*, 8(1):83, dec 2018.
- [151] M Unser, A Aldroubi, and C R Gerfen. A multiresolution registration procedure using spline pyramids. Technical report, 1993.
- [152] John Darrell Van Horn and Arthur W. Toga. Is it time to re-prioritize neuroimaging databases and digital repositories? *NeuroImage*, 47(4):1720–1734, 2009.
- [153] Rik Vandenberghe, Katarzyna Adamczuk, Patrick Dupont, Koen Van Laere, and Gaël Chételat. Amyloid PET in clinical practice: Its place in the multidimensional space of Alzheimer’s disease, jan 2013.
- [154] S. Vandenberghe, E. Mikhaylova, E. D’Hoe, P. Mollet, and J. S. Karp. Recent developments in time-of-flight PET, dec 2016.
- [155] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, Nicholas Ayache, Inria Sophia Antipolis, and Sophia Antipolis. Diffeomorphic Demons Using ITK ’s Finite Difference Solver Hierarchy. Technical Report 154, 2007.
- [156] E. E. Verwer, S. S.V. Golla, A. Kaalep, M. Lubberink, F. H.P. van Velden, V. Bettinardi, M. Yaqub, T. Sera, S. Rijnsdorp, A. A. Lammertsma, and R. Boellaard. Harmonisation of PET/CT contrast recovery performance for brain studies. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–15, jan 2021.
- [157] Luc Vinet and Alexei Zhedanov. *A ’missing’ family of classical orthogonal polynomials*, volume 44. 2011.
- [158] Christian Wachinger, Anna Rieckmann, and Sebastian Pölsterl. Detect and correct bias in multi-site neuroimaging datasets. Technical report, 2021.
- [159] Yuchuan Wei, Ge Wang, and Jiang Hsieh. An intuitive discussion on the ideal ramp filter in computed tomography (I). *Computers and Mathematics with Applications*, 49(5-6):731–740, apr 2005.

- [160] Miles N. Wernick and John N. Aarsvold. Emission Tomography: The Fundamentals of PET and SPECT, 2004.
- [161] Frank Wübbeling. PET image reconstruction. Technical report, 2012.
- [162] Yuhong Yang. Can the strengths of AIC and BIC be shared. *Biometrika*, 92(4):937–950, 2005.
- [163] Gengsheng L. Zeng. Revisit of the ramp filter. *IEEE Transactions on Nuclear Science*, 62(1):131–136, feb 2015.
- [164] Henrik Zetterberg and Barbara B. Bendlin. Biomarkers for Alzheimer’s diseasepreparing for a new era of disease-modifying therapies. *Molecular Psychiatry*, 26(1):296–308, jan 2021.
- [165] Yazhong Zhang, Jinjian Wu, Xuemei Xie, Leida Li, and Guangming Shi. Blind image quality assessment with improved natural scene statistics model. *Digital Signal Processing: A Review Journal*, 57:56–65, oct 2016.
- [166] Yong Zhang and Dapeng Wang. A cost-sensitive ensemble method for class-imbalanced datasets. *Abstract and Applied Analysis*, 2013, 2013.
- [167] Yang-Ming Zhu. Ordered subset expectation maximization algorithm for positron emission tomographic image reconstruction using belief kernels. *Journal of Medical Imaging*, 5(04):1, nov 2018.
- [168] Susanne Ziegler, Bjoern W. Jakoby, Harald Braun, Daniel H. Paulus, and Harald H. Quick. NEMA image quality phantom measurements and attenuation correction in integrated PET/MR hybrid imaging. *EJNMMI Physics*, 2(1):1–14, 2015.