Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea Magistrale in Fisica

Characterization of a Dual-energy X-ray Absorptiometry System for Soft Tissues Assessment and their correlations with Metabolic State

Relatore:
Prof. Gastone Castellani

Correlatore:
Prof. Enrico Giampieri

Presentata da:
Giulio Serra

Sessione I
Anno Accademico 2014/2015

Sommario

Il crescente utilizzo di sistemi di analisi *high-throughput* per lo studio dello stato fisiologico e metabolico del corpo, ha evidenziato che una corretta alimentazione e una buona forma fisica siano fattori chiave per la salute. Inoltre l'aumento dell'età media della popolazione evidenzia l'importanza delle strategie di contrasto e prevenzione delle patologie legate all'invecchiamento. Una dieta sana è il primo mezzo di prevenzione per molte patologie, pertanto capire come il cibo influisce sul corpo umano è di fondamentale importanza.

In questo lavoro di tesi abbiamo affrontato la caratterizzazione dei sistemi di imaging radiografico *Dual energy X-ray Absorptiometry* (DXA). I sistemi DXA sono principalmente utilizzati per ricavare la massa di un materiale, in presenza di un altro, attraverso la conoscenza dei rispettivi coefficienti di attenuazione dei raggi X a diverse energie.

Dopo aver stabilito una metodologia adatta per l'elaborazione di dati DXA su un gruppo di soggetti sani non obesi, la *Principal Components Analysis* (PCA) ha evidenziato alcune proprietà emergenti dall'interpretazione delle componenti principali in termini delle variabili di composizione corporea restituite dalla DXA. Le prime componenti sono chiaramente associabili a degli indici macroscopici di descrizione corporea (come BMI e WHR). Inoltre, queste componenti sono sorprendentemente stabili al variare dello status dei soggetti in età, sesso e nazionalità.

Dati di analisi metabolica, ottenuti tramite *Nuclear Magnetic Resonance Spectroscopy* (MRS) su campioni di urina, sono disponibili per circa mille persone anziane (provenienti da cinque paesi europei) di età compresa tra i 65 ed i 79 anni, non affetti da patologie gravi. I dati di composizione corporea sono altresì presenti per questi soggetti.

L'algoritmo di *Non-negative Matrix Factorization* (NMF) è stato utilizzato per esprimere gli spettri (ottenuti tramite MRS) come una combinazione di fattori di base, ognuno dei quali interpretabile come espressione di un singolo metabolita. Anche in questo caso (come per le componenti della PCA) i fattori sono stabili, il che significa che gli spettri metabolici dei soggetti provenienti da diversi paesi sono composti dallo stesso pattern di metaboliti.

Attraverso un'analisi a singolo cieco sono stati trovati alti valori di correlazione tra le variabili di composizione corporea e lo stato metabolico dei soggetti. Questi risultati suggeriscono la possibilità di derivare la composizione corporea dei soggetti a partire dal loro stato metabolico.

Abstract

The increasing use of *high-throughput* analyses for the study of body physiological and metabolical status, has highlighted that proper nutrition and good physical fitness are of key factors for the human health. Moreover the increase in the average age of the population raises a critical importance to identify strategies able to contrast the age-related diseases. A good diet is the first step for the prevention of several pathologies.

In this thesis work we have addressed the characterization of a *Dual energy X-ray Absorptiometry* (DXA) system. DXA is an X-ray imaging technique primarily used to derive the mass of one material in the presence of another through knowledge of their unique X-ray attenuation at different energies.

After establish the proper method and preprocessing operation over a group of healthy and normal-weight subjects, *Principal Components Analysis* (PCA) has shown emergent properties of body composition variables from DXA.

The interpretation of first few components is really clear and can be view as good descriptive indexes of the body composition. Moreover these components are surprisingly stable across different subject status, age, gender and nationality.

Cohort includes about thousand elderly people, ranging from 65 to 79 years, free of major overt diseases and came from five different European countries. Metabolic analyses obtained trough *Nuclear Magnetic Resonance Spectroscopy* (MRS) on urine samples are considered.

A *Non-negative Matrix Factorization* (NMF) algorithm is used to express the original spectra (from MRS) as a combination of basis factors that can be understood as single metabolite.

We have found "stable" factors by NMF and this suggest us that metabolic spectra of subjects coming from different countries are compound by the same pattern of metabolites.

A blind analysis design is chosen for the correlation analysis between body composition variables and metabolic state of subjects. The high values obtained suggest the possibility to derive the "body shape" from the metabolic state.

Contents

Chapter 1

Introduction

The population is projected to become older in European Region. Thus the median age of the total population is likely to increase in all countries without exception due to the combined effect of the existing structure of the population, persistently low fertility and continuously increasing number of survivors to higher ages [1].

This demographic change emphasizes the critical importance of identifying strategies able to counteract or slow down ageing and the onset of age-related diseases and disabilities, and so contribute to increase the number of elderly European citizens in good health, and reducing age-related medical and social costs.

Within this scenario, the European consortium NU-AGE: *New dietary strategies addressing the specific needs of elderly population for an healthy ageing in Europe* (nu-age.eu) targets nutrition as a major modulator of inflaming and other age-related functional outcomes.

This thesis investigate how composition of body soft tissues is correlated with the metabolic state of ageing subjects and the difference of these correlations varying on living countries of the subjects.

Dual-energy X-ray Absorptiometry (DXA) is an X-ray imaging technique to derive the mass of one material in the presence of another through knowledge of their unique X-ray attenuation at different energies.

Body composition measurements with DXA can look beyond weight and traditional body mass index (BMI) to determine distribution of body fat, lean and bone mass. DXA exams provide regional and total body information of body composition.

The DXA scanner uses small X-ray dose beam composed of two energy levels and is based on a model that forces all tissue types into three classes (based on

their X-ray attenuation properties): bone minerals, fat and lean (fat-free). These differences in absorption are used to determine bone mineral density (BMD) and body composition values, and can be used to predict total body fat, fat-free mass, and total body bone mineral.

Second chapter presents DXA technique for bone and soft tissue determination. We will clarify how the integrated dual-energy measurements from a single projection can quantitatively determine the mass of intervening materials, the advantages and the limitations of this technique.

In the later chapter we explain the statistical tools needed for the next chapter. Particular attention will be placed in the explanation of the methods for the simplification of the data, such as the Principal Component Analysis and the Non Negative Matrix Factorization, used to processing the body composition variables from DXA and the metabolic data respectively.

The fourth chapter can be seen as compound of two part: in the first part we analyse the body composition variables of healthy subjects to define method, preprocessing operations and factors to be taken into account for analysis of data from DXA and establish reference values for body composition on healthy people. Moreover we investigate how soft tissues distribution in the body are related to blood lipid concentrations.

In the second part we will apply results obtained from the healthy database on another database, composed by elderly people, to study the correlation between DXA variables and the metabolites.

Chapter 2

Dual-energy X-ray absorptiometry

Dual energy X-ray absorptiometry (DXA) is an X-ray imaging technique primarily used to derive the mass of one material in the presence of another through knowledge of their unique X-ray attenuation at different energies. Two images are made from the attenuation of low and high average X-ray energy. DXA is a special imaging modality that is not typically available with general use X-ray systems because of the need for special beam filtering and near perfect spatial registration of the two attenuations. Dedicated commercial DXA systems first became available in the late 1980s [2].

DXA is an extension of an earlier imaging technique called dual-energy photon absorptiometry (DPA). DPA has been used for about 15 years to measure bone and soft-tissue composition. In DPA a radionuclide source (usually gadolinium ^{153}Gd)) is used to generate the gamma rays. In DXA, the radionuclide source has been replaced by a low current X-ray tube, which allows a much higher photon flux to be generated. This results in higher resolution images and, hence, precision, and a much faster scan time. Due to these improvements, DPA has been superseded by DXA [3].

This chapter focuses on Dual-energy X-ray absorptiometry (DXA) theoretical background, the first section provides background, fundamentals interactions occurring between photon and matter at energy range used in DXA will be reported. In the later sections, DXA technique for bone and soft tissue determination will be described.

2.1 Interactions of X-ray with Matter

When traversing an absorbing medium, photons may experience various interactions with the atoms of the medium. These interactions involve either the nuclei of the absorbing medium or the orbital electrons of the absorbing medium:

1. The interactions with nuclei may be direct photon-nucleus interactions (photodisintegration) or interactions between the photon and the electrostatic field of the nucleus (pair production).

2. The photon-orbital electron interactions are characterized as interactions between the photon and either (1) a loosely bound electron (Thomson scattering, Compton effect, triplet production) or (2) a tightly bound electron (photoelectric effect, Rayleigh scattering).

A loosely bound electron is an electron whose binding energy (E_B) is small in comparison with photon energy ($h\nu$): $E_B \ll h\nu$, thus the electron is considered as a "free" electron. A tightly bound electron is an electron whose binding energy $E_B \sim h\nu$. For a photon interaction to occur with a tightly bound electron $E_B \lesssim h\nu$. An interaction between a photon and a tightly bound electron is considered an interaction between a photon and the atom as a whole.

Pair, and triplet, production can only occurs when the photon energy exceed 1.02 MeV. Since such high energies never occurs during DXA scanning, these interactions will not be discussed in this section.

The most important parameter used for characterization of X-ray penetration into absorbing media is the linear attenuation coefficient μ. This coefficient depends on energy $h\nu$ of the photon and atomic number Z of the absorber, and may be described as the probability per unit path length that a photon will have an interaction with the absorber.

The functional relationship between the thickness of an absorber and intensity of a photon beam attenuated by the absorber can be derived using differential calculus. Supposing narrow beam geometry tecnique that implies a narrowly collimated source and detector. The absorber decreases the intensity $I(0)$ measured without the absorber in place to $I(x)$ measured with absorber thickness x in the beam.

A layer of thickness $\mathrm{d}x'$ of the absorber reduces the beam intensity by $\mathrm{d}I$ and the fractional reduction in intensity $\mathrm{d}I/I$, is proportional to the linear attenuation coefficient μ, measured in [cm^{-1}], and to the layer thickness $\mathrm{d}x'$:

$$-\frac{\mathrm{d}I}{I} = \mu \mathrm{d}x', \tag{2.1}$$

where the negative sign indicates that the signal decreases as the absorber thickness increases, and μ represents the probability that a photon interact in a unit thickness of absorber layer dx'.

After integration over absorber thickness from 0 to x and over intensity from the initial intensity $I(0)$ to intensity $I(x)$ at absorber thickness x, we get

$$\int_{I(0)}^{I(x)} \frac{\mathrm{d}I}{I} = -\int_{0}^{x} \mu \, \mathrm{d}x' \qquad \text{or} \qquad I(x) = I(0)e^{-\int_0^x \mu \, \mathrm{d}x'} \tag{2.2}$$

For a homogeneous medium the attenuation coefficient μ is constant and Equation 2.2 reduces to the standard exponential relationship valid for monoenergetic photon beams:

$$I(x) = I(0)e^{-\int_0^x \mu \, \mathrm{d}x'} = I(0)e^{-\mu x} \tag{2.3}$$

In addition to linear attenuation coefficient μ, *mass attenuation coefficient* is defined as the linear attenuation coefficient divided by the mass per unit volume ρ of the absorber: $\mu_m = \mu/\rho$, measured in [cm^2/g]. When μ_m is used in Equation 2.2, the thickness is expressed in [g/cm^2].

X-rays in the energy range used for DXA interact with tissue using three processes (that will be discussed below): photoelectric absorption, Compton (inelastic) scattering and Rayleigh (coherent) scattering [2]. To determine the total attenuation from all three attenuation interactions, one simply sums the mass attenuation coefficients from each effect:

$$\frac{\mu}{\rho} = \frac{\tau}{\rho} + \frac{\sigma_{_C}}{\rho} + \frac{\sigma_{_R}}{\rho} \tag{2.4}$$

where τ/ρ, $\sigma_{_C}/\rho$ and $\sigma_{_R}/\rho$ are the photoelectric, Compton and Rayleigh mass attenuation coefficients respectively.

2.1.1 Thomson Scattering

The scattering of low energy photon ($h\nu_0 \ll m_e c^2$) by essential free electrons of an absorber, is described adequately by non relativistic classical theory of Joseph J. Thomson [4]. Thomson assumed that the incident photon beam set each quasi-

free electron of the absorbed atom into a forced resonant oscillation and then used classical theory to calculate the cross section for re-emission of the electromagnetic radiation as a result of the induced dipole oscillation of the electron (see Figure 2.1).

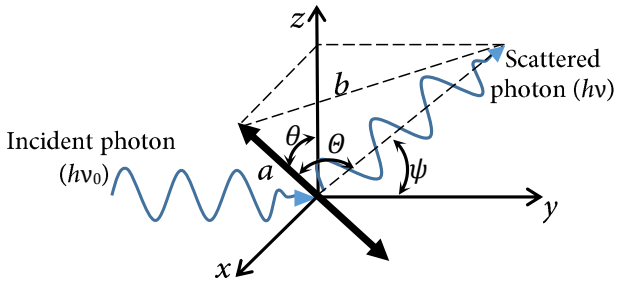

Figure 2.1: Schematic diagram of Thomson scattering, where the incident photon($h\nu$) is scattered and emitted with a scattering angle θ. Note that angles θ and Θ are not coplanar.

The differential electronic cross section per unit solid angle for Thomson scattering is:

$$\frac{\mathrm{d}\sigma^{e}_{Th}}{\mathrm{d}\Omega} = \frac{r_e^2}{2}(1+\cos^2\theta) \qquad \text{in } \left[\mathrm{cm}^2/(\text{electron}\cdot\text{steradian})\right]. \tag{2.5}$$

Figure 2.2 show the differential electronic cross section against the angle θ in polar coordinate system. The graph show that $\mathrm{d}\sigma^{e}_{Th}/\mathrm{d}\theta$range from 39.7mb/electron $\cdot$ steradian at $\theta = \pi/2$ to 79.4mb/electron $\cdot$ steradian for $\theta = 0$ and $\theta = \pi$.

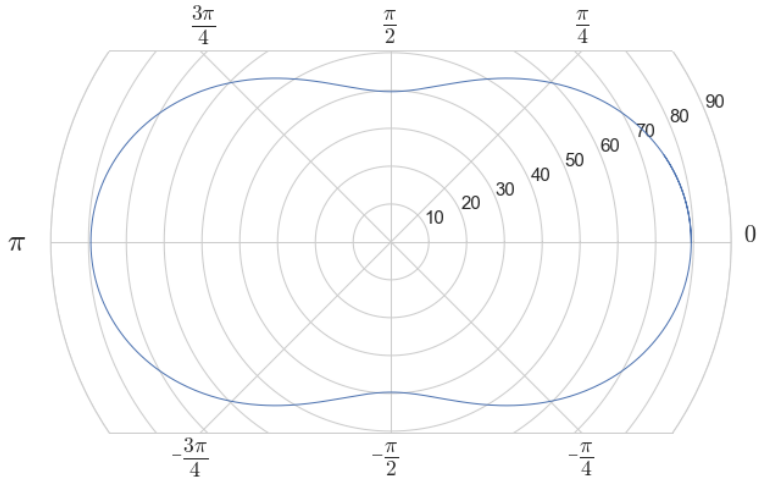

Figure 2.2: Differential electronic cross section $\mathrm{d}\sigma^{e}_{Th}/\mathrm{d}\Omega$ per unit solid angle against the scattering angle θ for Thomson scattering.

2.1.2 Compton Scattering

An inelastic collision of a photon of energy $h\nu_0$ with a loosely bound orbital electron of an absorber is called Compton effect (Compton scattering). The effect is also known as incoherent scattering.

In theoretical studies of the Compton effect an assumption is made that the photon interacts with a free and stationary electron. A photon, referred to as a scattered photon with energy $h\nu$ that is smaller than the incident photon energy $h\nu_0$, is produced in Compton effect and an electron, referred to as a Compton (recoil) electron, is ejected from the atom with kinetic energy E_{e^-}.

Compton scatter is the predominant interaction of X-ray in diagnostic energy range with soft tissue. This interaction is most likely to occur between photon and valence shell electron, see Figure 2.3.

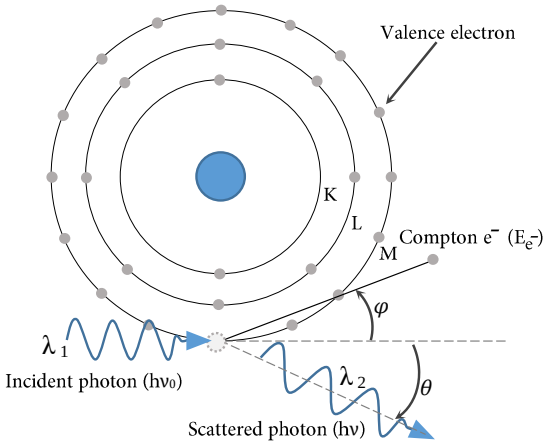

Figure 2.3: Compton scattering, the figure show the incident photon with energy $h\nu_0$, interacting with a valence electron that results in the ejection of the Compton electron E_{e^-} and the simultaneous emission of a Compton scattered photon of energy $h\nu$ emerging at an angle θ relative to the incident photon. K, L and M are electron shells.

Supposing the electron as free ($h\nu_0 \gg$ of binding energy) and applying the conservation of energy and momentum get:

$$\Delta\lambda = \lambda - \lambda_0 = \frac{h}{mc}(1 - \cos\theta) \tag{2.6}$$

where $\lambda_0 = 2\pi\hbar c/h\nu_0$ is the wavelength of the incident photon and $\lambda = 2\pi\hbar c/h\nu$ of the scattered photon. $\Delta\lambda$ depends only on the scattering angle θ and is independent of the energy of the incident photon $h\nu_0$.

The relationship for the energy of the scattered photon $h\nu$ as a function of the incident photon energy, $h\nu_0$, and the scattering angle, θ, is:

$$h\nu = h\nu_0 \frac{1}{1+\varepsilon(1-\cos\theta)}. \tag{2.7}$$

where $\varepsilon = h\nu_0/mc^2$ is defined as the incident photon energy normalized to electron rest mass energy ($mc^2 = 511\,\text{keV}$).

The Equation 2.7 is plotted in Figure 2.4 for various values oh θ. From this equation some conclusions can be made:

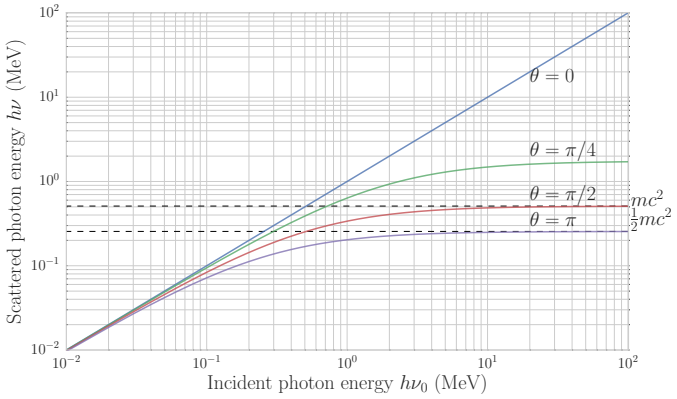

Figure 2.4: Scattered photon energy $h\nu$ against the incident photon energy $h\nu_0$ for $\theta = 0$, $\pi/4$, $\pi/2$ and π.

1. For $\theta = 0$ then $h\nu = h\nu_0$, no energy is transferred to the recoil electron, we are dealing with Thomson scattering.

2. For $\theta > 0$, the energy of the scattered photon saturates at high values of $h\nu$; the larger is θ, the lower is the saturation value of $h\nu$ for $h\nu \longrightarrow \infty$.

3. For $\theta = \pi/2$, $h\nu = h\nu_0/(1+\varepsilon)$, and the saturation energy of the scattered photon is equal to the rest mass energy of the electron: $mc^2 = 511\,\text{keV}$.

4. For $\theta = \pi$, $h\nu = h\nu_0/(1+2\varepsilon)$ with saturation energy equal to : $mc^2/2 = 255\,\text{keV}$.

The results reported above show that photon scattered with angles θ larger than $\frac{\pi}{2}$ cannot exceed 511 keV in kinetic energy no matter how high is the incident photon energy $h\nu_0$. And for a given $h\nu_0$, $h\nu$ will be in the range between $h\nu_0/(1+2\varepsilon)$, for $\theta = \pi$, and $h\nu_0$ for $\theta = 0$.

Kinetic energy of the Compton (recoil) electron E_{e^-} depends on photon energy $h\nu_0$ and photon scattering angle θ. The relationship is determined using conservation of energy $h\nu_0 = h\nu + E_{e^-}$:

$$E_{e^-} = h\nu_0 - h\nu = h\nu_0 - h\nu_0 \frac{1}{1 + \varepsilon(1 - \cos\theta)} \tag{2.8}$$

For a given photon energy $h\nu_0$ the recoil electron kinetic energy ranges from a minimum value of $(E_{e^-})_{min} = 0$ for $\theta = 0$ (forward scattering), to a maximum value of

$$(E_{e^-})_{max} = h\nu_0 \frac{2\varepsilon}{1 + 2\varepsilon}$$

for $\theta = \pi$ (backscattering).

The ratio of the kinetic energy of the recoil electron $E_{e^-}(h\nu_0, \theta)$ to the energy of the incident photon $h\nu$ represents the fraction of the incident photon energy that is transferred to the electron in a Compton effect and is called the Compton energy transfer fraction $f_c(h\nu_0, \theta)$, expressed as follows:

$$f_c(h\nu_0, \theta) = \frac{E_{e^-}}{h\nu} = \frac{\varepsilon(1 - \cos\theta)}{1 + \varepsilon(1 - \cos\theta)}. \tag{2.9}$$

Figure 2.5 shows a plot of $f_c(h\nu_0, \theta)$ against θ for various incident photon energies in the range from 10 keV to 100 MeV.

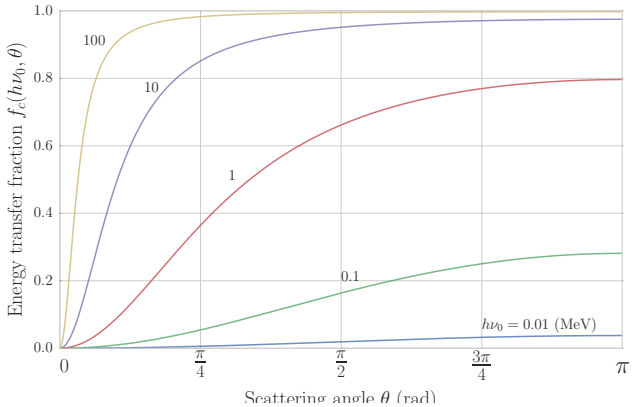

Figure 2.5: Compton energy transfer fraction $f_c(h\nu_0, \theta)$ against the scattering angle θ for various incident photon energies $h\nu_0$ in the range from 10 keV to 100 MeV.

The following features are notable:

1. For all $h\nu_0$ the Compton energy transfer fraction f_c is null: $(f_c(h\nu_0, \theta)|_{\theta=0} = 0)$.

2. For a given $h\nu_0$, $f_c(h\nu_0, \theta)$ increase with angle and saturate at $2\varepsilon/1 + 2\varepsilon$.

3. For a given θ, the larger is $h\nu_0$, the larger is $f_c(h\nu_0, \theta)$.

4. For fixed parameters ($h\nu_0$ and θ) the sum $h\nu + E_{e^-} = h\nu_0$.

The differential Klein-Nishina electronic cross section per unit solid angle for Compton effect $\mathrm{d}\sigma^e_{KN}/\mathrm{d}\Omega$ (in [(cm^2/electron)/steradian]) is:

$$\begin{aligned}\frac{\mathrm{d}\sigma^e_{KN}}{\mathrm{d}\Omega} &= \frac{r_e^2}{2}\left(\frac{\nu}{\nu_0}\right)^2\left\{\frac{\nu}{\nu_0}+\frac{\nu_0}{\nu}-\sin^2\theta\right\}\\ &= \frac{r_e^2}{2}(1+\cos^2\theta)F_{KN} = \frac{\mathrm{d}\sigma^e_{KN}}{\mathrm{d}\Omega}F_{KN}\end{aligned} \tag{2.10}$$

where ν_0 and ν are the frequencies of the incident and scattered photon respectively, θ is the scattering angle, r_e, as already mentioned, is the classical radius of electron ($2.82 \cdot 10^-15$m), $\mathrm{d}\sigma^e_{Th}/\mathrm{d}\Omega$ is the differential electronic cross section per unit solid angle for Thomson scattering, and $F_{KN}(h\nu_0, \theta)$ is the Klein-Nishina form factor, dependent on incident photon energy and photon scattering angle, that, for a free electron, is given as follows:

$$F_{KN}(h\nu_0,\theta) = \frac{1}{\left[1+\varepsilon(1-\cos\theta)\right]^2}\left\{1+\frac{\varepsilon^2(1-\cos\theta)^2}{\left[1+\varepsilon(1-\cos\theta)\right](1+\cos^2\theta)}\right\} \tag{2.11}$$

where, as said above, $\varepsilon = h\nu_0/mc^2$.

The Figure 2.6 show the differential electronic cross section for Compton effect $dsigma^e_{KN}/d\Omega$ against scattering angle θ for various values of ε.For $\varepsilon = 0$, the differential electronic cross section for Compton effect $dsigma^e_{KN}/d\Omega$ is equal to the differential electronic cross section for Thomson scattering (see Figure 2.2).

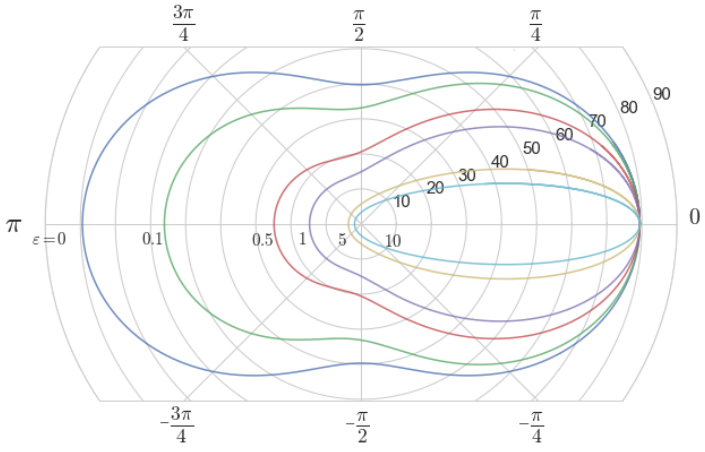

Figure 2.6: Differential electronic cross section $\mathrm{d}\sigma^e_{KN}/\mathrm{d}\Omega$ per unit solid angle against the scattering angle θ for Compton effect, for various values of ε.

After the photon-electron interaction, the ejected electron will lose its kinetics energy via excitation and ionization of atom in the surrounding material. The

photon may traverse the medium without interaction or may under go subsequent interactions such as Compton, Rayleigh or photoelectric interactions.

The total electronic Klein-Nishina cross section for the Compton scattering on a free electron σ^e_{KN} [in cm^2/electron] is calculated by integrating the differential electronic cross section per unit solid angle $\mathrm{d}\sigma^e_{KN}/\mathrm{d}\Omega$ over the whole solid angle, and the atomic cross section σ^a_{KN} is equal to:

$$\sigma^a_{KN} = Z(\sigma^e_{KN}) = Z\left(\int \frac{\mathrm{d}\sigma^e_{KN}}{\mathrm{d}\Omega}\,\mathrm{d}\Omega\right). \tag{2.12}$$

where Z is the atomic number of the absorber. The Klein-Nishina Compton electronic cross section σ^e_{KN} is given for free electrons and is thus independent of Z. This makes the atomic attenuation coefficient (cross section) σ^a_{KN} linearly dependent on Z.

The Compton mass attenuation coefficient σ_C/ρ, expressed in [cm^2/g], where ρ is the mass density of the absorber, is calculated from the Compton atomic cross section with the standard relationship:

$$\frac{\sigma_C}{\rho} = \frac{N_A}{A}\sigma^a_{KN} = \frac{ZN_A}{A}\sigma^e_{KN} \tag{2.13}$$

where N_A is the Avogadro number ($6.022 \cdot 10^{23}$ atom/mol) and A is the atomic mass number. Since $Z/A \approx 0.5$ for all elements with the exception of hydrogen for which $Z/A = 1$, σ_C/ρ is essentially independent of Z. In reality, $Z/A = 0.5$ for low atomic number absorbers but with increasing Z the ratio Z/A gradually falls to $Z/A \approx 0.4$ for very high atomic number absorbers, implying a small yet non-negligible Z dependence of σ_C/ρ.

In diagnostic radiology, Compton scattering is the most problematic interaction of photons with the body matter. First, the deflections in scattering events cause uncertainties in photon localization as it becomes difficult to keep the desired radiation transmission path, it reduces the contrasts in the image unless it is removed by collimation before the detector, and also lead to a lower signal-to-noise ratio. Second, it presents a radiation risk to the personnel using the equipment [5].

As we will see later, while the photoelectric effect dominates in materials with high atomic numbers, Compton scattering is more significant in materials with lower atomic numbers. Also, Compton scattering dominates with high-energy photons. Since the higher energy photons would cause a larger deflection angle, a

higher energy radiation is not desirable in radiological imaging.

2.1.3 Rayleigh Scattering

Rayleigh scattering is an interaction between a photon and absorber atom characterized by photon scattering on bound atomic electrons. The atom is neither excited nor ionized as a result of the interaction and after the interaction the bound electrons revert to their original state. The atom as a whole absorbs the transferred momentum but its recoil energy is very small and the incident photon scattered with scattering angle θ has essentially the same energy as the original photon. The scattering angles are relatively small because the recoil imparted to the atom produces no atomic excitation or ionization[4].

This interaction occurs mainly with very low energy diagnostic X-ray , as used in mammography (15 to 30 keV). During the Rayleigh scattering events, the electric field of the incident photon's electromagnetic wave expend energy, causing all of the electrons in the scattering atom to oscillate in phase. The atom's electron cloud immediately radiates this energy, emitting a photon of the same energy but in slightly different direction, see Figure 2.7.

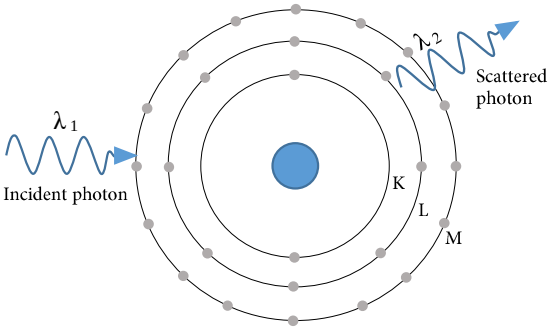

Figure 2.7: Rayleigh scattering, the figure show the incident photon(λ_1)interact with an atom and the scatter photon($\lambda_2 = \lambda_1$) is being emitted with the same energy. Rayleigh scattered photon are typically emitted in the forward direction fairly close to trajectory of the incident photon. K, L and M are electron shells.

The differential Rayleigh atomic cross section $\mathrm{d}\sigma^a_R/\mathrm{d}\Omega$ per unit solid angle is given as follows:

$$\frac{\mathrm{d}\sigma^a_R}{\mathrm{d}\Omega} = \frac{r_e^2}{2}(1+\cos^2\theta)\left\{F(x,Z)\right\}^2 = \frac{\mathrm{d}\sigma^e_{Th}}{\mathrm{d}\Omega}\left\{F(x,Z)\right\}^2 \tag{2.14}$$

where $\mathrm{d}\sigma^e_{Th}/\mathrm{d}\Omega$ is the differential Thomson cross section and $F(x,Z)$ is the atomic

form factor for Rayleigh scattering whit the momentum transfer variable $x = \sin(\theta/2)\lambda_0$ (λ_0 is the wavelength of the incident photon and Z is the atomic number of the absorber).

The differential Rayleigh atomic cross section $\mathrm{d}\sigma_{R}^{a}/\mathrm{d}\theta$ per unit scattering angle θ is

$$\begin{aligned}\frac{\mathrm{d}\sigma_{R}^{a}}{\mathrm{d}\theta} &= \frac{\mathrm{d}\sigma_{R}^{a}}{\mathrm{d}\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}\theta} = \frac{r_e^2}{2}(1+\cos^2\theta)\left\{F(x,Z)\right\}^2 2\pi\sin\theta \\ &= \pi r_e^2 \sin\theta(1+\cos^2\theta)\left\{F(x,Z)\right\}^2 .\end{aligned} \tag{2.15}$$

The Rayleigh atomic cross section σ_{R}^{a} can be calculated by integrating Equation 2.15 over all possible scattering angle θ from 0 to π:

$$\sigma_{R}^{a} = \int_0^{\pi} \frac{\mathrm{d}\sigma_{R}^{a}}{\mathrm{d}\theta}\,\mathrm{d}\theta, \tag{2.16}$$

and the Rayleigh mass attenuation coefficient σ_{R}/ρ [cm^2/g] is determined through the standard relationship:

$$\frac{\sigma_{R}}{\rho} = \frac{N_A}{A}\sigma_{R}^{a}. \tag{2.17}$$

In this interaction, electrons are not emitted end thus ionization does not occur. In general, the scattering angle increase as the X-ray energy decreases. In medical imaging, detection of the scattered X-ray will have a deleterious effect on image quality. However, this type of interaction has a low probability of occurrence in the diagnostic energy range. In soft tissue, Rayleigh scattering accounts for less than 5% of X-ray interactions above 70 keV and at most only 12% of interactions at approximately 30 keV [6]. Rayleigh interactions are also referred as to "coherent" scattering.

2.1.4 Photoelectric Absorption

An interaction between a photon and a tightly bound electron of an absorber atom is called photoelectric effect. In the interaction the photon is absorbed completely and the orbital electron is ejected with kinetic energy E_{e^-}. The ejected orbital electron is called photoelectron. The photoelectric interaction between a photon of energy $h\nu_0$ and a K-shell atomic electron is shown schematically in Figure 2.8. In contrast to Compton effect which occurs between photon and a loosely

bound electron ($E_B \ll h\nu_0$), the photoelectric effect occurs between a photon and a tightly bound electron ($E_B \lesssim h\nu_0$). The requirement for electron tight binding to the atom arise from consideration of total energy and momentum conservation [4].

The extra energy and momentum carried by the photon are transferred to the absorbing atom, however, because of the relatively large nuclear mass, the atomic recoil energy is exceedingly small and may be neglected. The kinetic energy E_{e^-} of the ejected photoelectron is assumed to be equal to the incident photon energy $h\nu_0$ less the binding energy E_B of the orbital electron:

$$E_{e^-} = h\nu_0 - E_B. \tag{2.18}$$

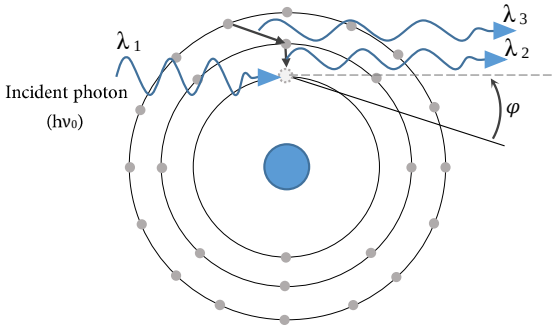

Figure 2.8: Photoelectric absorption, the figure show the incident photon with energy $h\nu_0$, interacting with an atom. In this case K-shell electron is ejected whit a kinetic energy $E_{e^-} = h\nu_0 - E_B$. The vacancy created in the k shell results in the transition of an electron from L shell to k-shell. The difference in their binding energies, results in a K_α characteristic X-ray. This electron cascade will continue resulting in the production of other characteristic X-rays of lower energies. Although not showed in this figure, Auger electron of various energies can be emitted in lieu of the characteristic X-ray emission.

The vacancy that results from the emission of the photoelectron from a given shell will be filled by a higher shell electron and the transition energy will be emitted either as a characteristic (fluorescence) photon or as an Auger electron, the probability for each governed by the fluorescence yield ω, as will be discussed later.

The angular distribution of photoelectrons depends on the incident photon energy $h\nu_0$. The photoelectron emission angle ϕ is defined as the angle between the incident photon direction and the direction of the emitted photoelectron, similarly to the definition of the recoil electron angle ϕ in Compton scattering (see Fig 2.8). At low $h\nu_0$ of the order of 10 keV photoelectrons tend to be emitted at angles close

to 90^o to the incident photon direction, hence in the direction of the electric vector of the incident photon. As $h\nu_0$ increases, however, the photoelectron emission peak moves progressively to more forward photoelectron emission angles. Figure 2.9 show the directional distribution of photoelectron emission for various incident photon energies $h\nu_0$ in the range from $h\nu_0 = 10\,\mathrm{keV}$ with maximum emission angle $\phi_{max} \approx 70^o$ to $h\nu_0 = 10\,\mathrm{MeV}$ with $\phi_{max} \approx 2^o$. The ordinate plots $\mathrm{d}n/\mathrm{d}\phi$, the relative number of photoelectrons ejected between two cones with half-angles of ϕ and $\phi + \mathrm{d}\phi$ for a given incident photon energy $h\nu_0$.

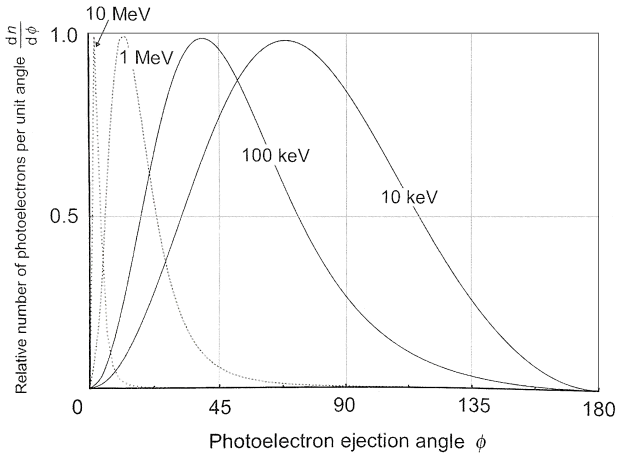

Figure 2.9: Angular distribution of photoelectrons ejected between two cones with half angles of ϕ and $\phi + \mathrm{d}\phi$ for given incident photon energy $h\nu_0$range from 10 keV to 10 MeV. Angle ϕ is the photoelectron emision angle definited as the angle between the incident photon direction and the direction of the emitted photoelectron. All peaks in angular distribution are normalized to 1.

The atomic cross section (attenuation coefficient) for the photoelectric effect τ^a as a function of the incident photon energy $h\nu_0$ exhibits a characteristic sawtooth structure in which the sharp discontinuities, referred to as absorption edges, arise whenever the photon energy coincides with the binding energy of a particular electron shell. Since all shells except the K shell exhibit a fine structure, the τ^a curve plotted against the incident photon energy $h\nu_0$ also exhibits a fine structure in the L, M, etc. absorption edges. Three distinct energy regions characterize the atomic cross section τ^a:

1. Region in the immediate vicinity of absorption edges.

2. Region at some distance from the absorption edge.

3. Region in the relativistic region far from the K absorption edge.

Theoretical predictions for τ^a in region (1) are difficult and uncertain. For region (2) the atomic attenuation coefficient for K-shell electrons a $\tau^a_{_K}$ is given as follows:

$$\tau^a_{_K} = \alpha^4(\sigma^e_{_{Th}})Z^n\sqrt{\frac{32}{\varepsilon^7}} \tag{2.19}$$

where $\varepsilon = h\nu_0/mc^2$, $\sigma^e_{_{Th}}$ and Z have the usual meaning, α is the fine structrure constant and n is the power for Z dependence of $\tau^a_{_K}$ ranging from $n = 4$ at relatively low photon energies to $n = 4.6$ at high photon energies.

In region (3) ($\varepsilon \gg 1$), $\tau^a_{_K}$ is:

$$\tau^a_{_K} = \frac{1.5}{\varepsilon}\alpha^4 Z^5(\sigma^e_{_{Th}}). \tag{2.20}$$

The following conclusions may be reached with regard to energy and atomic number dependence of $\tau^a_{_K}$:

1. The energy dependence of a $\tau^a_{_K}$ is assumed to go as $1/(h\nu_0)^3$ at low photon energies and gradually transforms into $1/(h\nu_0)$ at high $h\nu_0$.

2. The energy dependence for regions (2) and (3) can be identified from Figure 2.10 that displays the atomic cross section for the photoelectric effect τ^a against incident photon energy for various absorbers ranging from hydrogen ($Z = 1$) to lead (Z=82).

3. Absorption edges are clearly shown in Figure 2.10, the K absorption edges are identified for aluminum (1.56 keV), copper (8.98 keV) and lead (88 keV). The fine structures of the L and M absorption edges are also displayed.

4. The atomic number Z dependence ($\tau^a \propto Z^n$) of τ^a, where n ranges from 4 to ~ 5, is also evident from Figure 2.10.

The mass attenuation coefficient for the photoelectric effect τ/ρ is calculated from the atomic cross section τ^a with the standard relationship

$$\frac{\tau}{\rho} = \frac{N_A}{A}\tau^a \tag{2.21}$$

where A and ρ are the atomic number and density, respectively, of the absorber.

The benefit of photoelectric absorption in X-ray transmission imaging is that there are no additional non-primary photons to degrade the image. The photoelectric process predominates when lower energy photons interact with high Z materials. In fact, photoelectric absorption is the primary mode of interaction

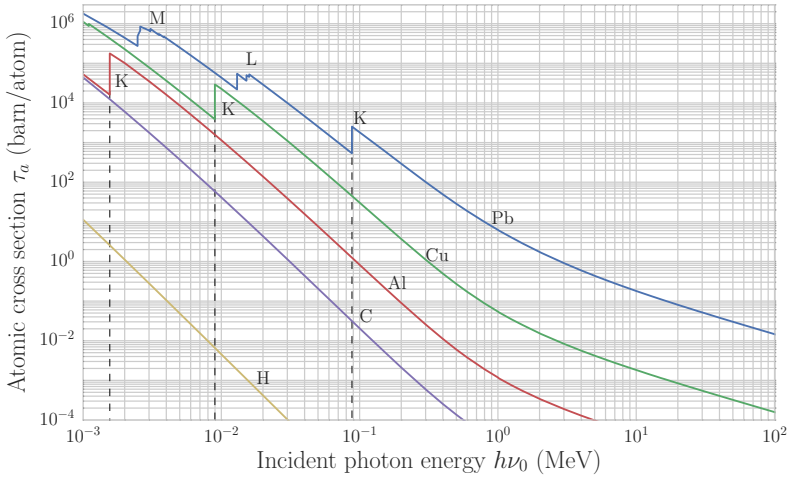

Figure 2.10: Photoelectric atomic cross section τ^a against photon energy $h\nu_0$ for various absorbers. Data are from the NIST.

of diagnostic X-rays with screen phosphors, radiographic contrast materials, and bone. Conversely, Compton scattering will predominate at most diagnostic photon energies in materials of lower atomic number such as tissue and air [6].

2.2 Physics of Absorptiometry

In this section we will clarify how the integrated dual-energy measurements from a single projection can quantitatively determine the mass of intervening materials.

As we know from the first part of this chapter, the linear attenuation coefficient is density (ρ) dependent, a convenient practice when working with tissues that differ in density is to calculate the mass attenuation coefficient μ/ρ. This removes the physical density dependence of the linear attenuation coefficient [7]. In the diagnostic energy range, the two principal means for attenuation of X-ray are photoelectric absorption and Compton scattering. In this energy range, the mass attenuation coefficient of a material can be approximated to:

$$\frac{\mu}{\rho} \simeq \frac{\tau}{\rho} + \frac{\sigma_C}{\rho} \tag{2.22}$$

where τ/ρ, σ_C/ρ are the photoelectric and Compton mass attenuation coefficients respectively and the mass attenuation coefficients due to Rayleigh scattering σ_R/ρ has been neglected [8].

Attenuation of monoenergetic photons (see Equation 2.3) using mass attenua-

tion coefficient can be rewrite as:

$$\mathrm{I} = \mathrm{I}_0 e^{-\mu x} = \mathrm{I}_0 e^{-\frac{\mu x \rho}{\rho}} = I_0 e^{-\frac{\mu}{\rho}\sigma} \tag{2.23}$$

or in term of log attenuation:

$$ln\left(\frac{\mathrm{I}}{\mathrm{I}_0}\right) = -\frac{\mu}{\rho}\sigma \tag{2.24}$$

were I and I_0 were used instead $\mathrm{I}(x)$ and $\mathrm{I}(0)$ respectively, $\sigma = x\rho$ is the areal density, expressed in [g/cm^2], and μ/ρ is the total mass attenuation coefficient as given in Equation 2.4. Since in DXA, like many other radiographic methods, pixel area is constant and known, σ represents total mass of the absorber system's volume element (voxel).

At any given photon energy, the mass attenuation coefficient (μ/ρ) of an element is constant and known from experimental studies, when photons at two different energies (H and L) are passed through an absorber:

$$\begin{aligned} ln\left(\frac{\mathrm{I}}{\mathrm{I}_0}\right)_H &= -\left[\frac{\mu}{\rho}\right]_H \sigma, \\ ln\left(\frac{\mathrm{I}}{\mathrm{I}_0}\right)_L &= -\left[\frac{\mu}{\rho}\right]_L \sigma. \end{aligned} \tag{2.25}$$

Attenuation at the lower energy can be expressed as a ratio (R) to attenuation observed at the higher energy, for a homogeneous absorber, R is simply the ratio of the component's mass attenuation coefficient at the two energies:

$$R = \frac{ln(\mathrm{I}/\mathrm{I}_0)_L}{ln(\mathrm{I}/\mathrm{I}_0)_H} = \frac{-\left[\frac{\mu}{\rho}\right]_L \sigma}{-\left[\frac{\mu}{\rho}\right]_H \sigma} = \frac{\left[\frac{\mu}{\rho}\right]_L}{\left[\frac{\mu}{\rho}\right]_H}, \tag{2.26}$$

Given the high and low energy each element has a characteristic R value. In Figure 2.11 are reported the values of R (40 keV and 70 keV were chosen as the high and low energy respectively) for elements that compose the main components presents in human body.

There is a strong dependence of R on the atomic number, as expected from the expressions of the mass attenuation coefficients obtained in previous sections. Elements with lower atomic number have lowest R values.

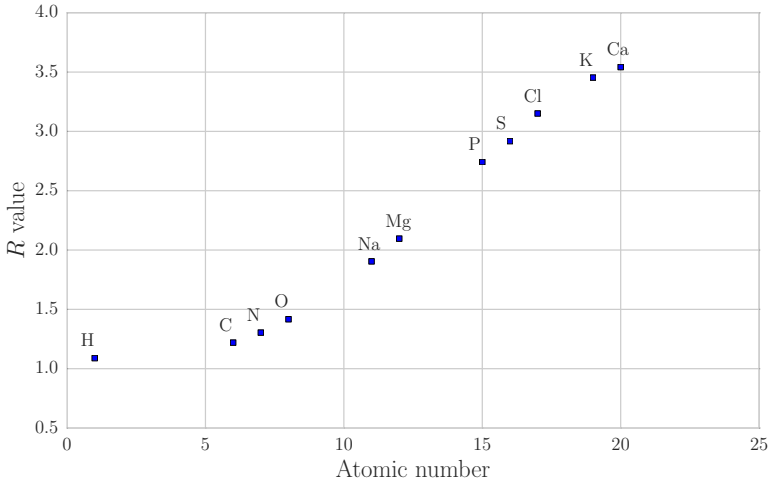

Figure 2.11: Relationship between R for 40 keV and 70 keV photons and the atomic number of elements. Organic compounds consist mainly of elements with low R values. Mineral elements (such as Na, K, P, Cl, and Ca) include high R comnonents. Data are from the NIST.

For heterogeneous absorbers Equation 2.24 becomes:

$$ln\left(\frac{\mathrm{I}}{\mathrm{I}_0}\right) = \sum_i \frac{\mu_i}{\rho_i}\sigma_i. \tag{2.27}$$

2.2.1 Biological Composition Standards

The biological standards of the body composition, since establish the exact atomic composition of tissues, are relevant to DXA measurements. In the molecular model, the body is represented as five compartments: water, protein, mineral, glycogen and lipid, summarized in Table 2.1.

Water, makes up over 60% of human body composition. The density of water is 1.0 g/cm^3 at 37 °C. Fat tissues (or lipids) are not water soluble and can be subdivided into categories based on their complexity. However DXA cannot distinguish between chemically extracted fat and the connective tissue and cellular membranes since all lipids have similar X-ray attenuation properties. This must be taken into account when using DXA in body composition models. For a reference man, approximately 90% of the body's lipid is fat. For the purposes of this thesis, protein is defined as almost all compounds that contain nitrogen and range in complexity from simple amino acid to nucleoproteins. The term mineral is used to describe the inorganic molecules in the body that contain metal elements such as calcium, sodium and potassium. Mineral is found in the body as either osseous or extra-

Table 2.1: Summary of fractional presence and density widely used to model each component. These are presented to be of assistance in understanding the modelling of DXA. (*) A reference man is defined as "being between $20-30$ years of age, weighing 70 kg, is 170 cm in height, and lives in a climate with an average temperature of 10 to 20 °C. He is Caucasian and is a Western European or North American in habitat and custom". Data from[2]

Component	Fraction in reference man*	Density g/cm^2
Water	0.6	1.000
Fat	0.19	0.900
Protein	0.15	1.34
Mineral	0.053	2.982
Glycogen	0.006	1.52

osseous, with the osseous component being by far the largest. Carbohydrates are principally stored as glycogen and are found in the cytoplasm of cells. The overall body content of glycogen is very small, less than 1%, (higher concentrations are found in muscle and liver tissue).

Given a compound, R can be calculated by Equation 2.26 knowing the mass fraction values (*component mass/compound mass*). In Table2.2 are show R values for some commonly body composition components. The theoretical R values for fatty acids and triglycerides were calculated by averaging the R values of various fatty acids and triglycerides respectively. Note that for soft tissue R value ranging from 1.21 for lipids to 1.30 for lean compound, while is twofold higher (2.86) for bone.

Table 2.2: Calculated μ/ρ values at 40 keV and 70 keV for commonly body composition components. (*)μ/ρ reported, are computed by averaging μ_m of various elements of the same category. Data from NIST, chemical structure of compound from [7].

	Mass Fraction									μ/ρ at [keV]		
Component	H	C	N	O	Na	Mg	P	S	Ca	40	70	R
Protein	0.070	0.532	0.161	0.227	-	-	-	0.010	-	0.236	0.183	1.291
Glycogen	0.062	0.444	-	0.494	-	-	-	-	-	0.238	0.183	1.301
Water	0.111	-	-	0.889	-	-	-	-	-	0.264	0.194	1.357
Fatty acids*	-	-	-	-	-	-	-	-	-	0.227	0.188	1.212
Triglycerides*	-	-	-	-	-	-	-	-	-	0.228	0.187	1.218
Bone	-	0.020	-	0.428	0.014	0.005	0.169	-	0.364	0.904	0.316	2.861

DXA was developed to determine the mass and composition of any two known materials when physical measurements of the materials, such as overall thickness, are either not available. The three component model used for DXA is a simplifi-

cation of the molecular model as shown in Figure 2.12.

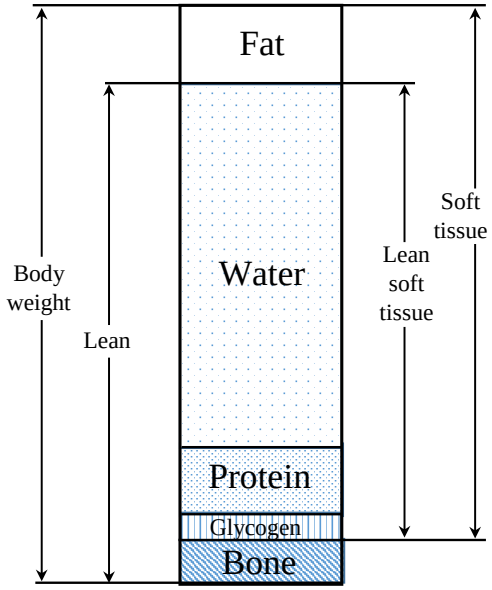

Figure 2.12: The five component molecular model of body composition compared to the three component model for DXA.

The components of the body are grouped into these three classes (based on their X-ray attenuation properties): bone minerals, fat and lean (non-fat). The model forces all tissue types into these three groups (the distinction between water, protein and glycogen is lost). More, distributions of tissue is lost, for example the distinction between subcutaneous adipose tissue (SAT) and visceral adipose tissue (VAT) is lost for trunk measurements when both are projected in the same pixels. This limitation is true for most of projective composition models.

The X-ray properties of these classes are dissimilar due to their differing proportions of high atomic number elements (as shown in Table2.2) Bone mineral contains a large percentage of calcium and phosphorus, whereas soft tissue is composed nearly completely of hydrogen, carbon and oxygen. However, there is a slight difference between the lean and fat components of soft tissue, since the lean compartment components contain traces of potassium, chlorine, sulphur and calcium, primarily as electrolytes, while fat contains none.

To measure bone, fat and lean content of human body, DXA gets around the limitation of two classes taking into account that bone mineral in the body is concentrated in dense local regions (bones). Thus it is possible to sort the pixel into those which contain bone and those which do not, ad to analyse the two types differently:

- The no-bone containing pixels are analysed for fat and lean as the two materials.

- The bone containing pixels are analysed for bone and soft tissue as the two materials. The specific mix of fat and lean is treated as "soft tissue" in the bone pixels and must be estimated, since it cannot be measured.

Since the fat/lean distribution vary significantly from subject to subject , to estimate the soft tissue composition, hidden by bone, DXA must be based on local measurements of fat and lean.

2.3 Principles of DXA

DXA was developed to solve the mass density of two unknown materials when physical measurements of the materials, such as overall thickness, are either not available or practical. Three fundamental assumptions are used to determine bone density using two energies:

1. Transmission of X-rays through the body within two energy windows can be accurately described by a monoexponential attenuation process (Equation 2.27).

2. Each image pixels of the human body can be described as a two component system, i.e. soft tissue and bone mineral, or when bone is not present, fat and lean mass.

3. The soft tissue overlaying the bone in the image has a composition and X-ray properties that can be predicted by the composition and X-ray properties of the tissue near but not overlaying the bone.

For simplicity, the equations will be derived for two monochromatic X-ray beams (DPA equation) at high and low energy. The log attenuation equation for each beam is:

$$\begin{aligned} J^H &= \mu_B^H \sigma_B + \mu_S^H \sigma_S, \\ J^L &= \mu_B^L \sigma_B + \mu_S^L \sigma_S \end{aligned} \tag{2.28}$$

where μ is equal to μ/ρ in Equation 2.27, J is equal to $ln\,(\mathrm{I}/\mathrm{I}_0)$, σ is the areal density expressed in [g/cm^2] and the H and L superscripts represent the high

and low energy X-ray beams respectively. B and S denote bone and soft tissue respectively.

Elimination of σ_S gives:

$$\sigma_B = \frac{J^L - R_S J^H}{\mu_B^L - \mu_B^H RS} \tag{2.29}$$

where R_S is the ratio value for the soft tissue:

$$R_S = \frac{\mu_S^L}{\mu_S^H}. \tag{2.30}$$

In the earlier technology of DPA the radionuclide source ^{153}Gd was used because its photon emissions at 44 and 103 keV were close to the ideal energies for in-vivo measurement of the lumbar spine. At photon energies above 100 keV there is little difference in the mass attenuation coefficients of bone and soft tissue and transmission measurements reflect essentially the total mass of tissue in the beam. Photon energies around 40 keV are ideal for the low energy beam because there is good contrast between bone and soft tissue without excessive attenuation to limit the signal reaching the detector.

When a DXA scan is analysed the basic data processed create a pixel-by-pixel map of Bone Mineral Density (BMD) over the entire scan field calculated from Equation 2.29. However, because of the effects of variable soft tissue composition and beam hardening the numerator in Equation 2.29 may take non-zero values in the soft tissue regions adjacent to bone [9]. Beam hardening effect is the tendency of low energy X-rays to be preferentially absorbed to high energy X-rays (due to dependency of absorption coefficient by the energy), which shifts the average beam energy to a higher value.

The soft tissue "hidden" is assumed to be the same composition as the nearby soft tissue (in no-bone containing pixels). This is a reasonable assumption for such a regional scan. This is called the weighted linear distribution model and is appropriate for areas such as the femur and long bones[10]. In regions such as the upper torso, more approximations are necessary that are usually proprietary and have loose physical interpretation to solve for R.

Then these surrounding regions provide a reference area of comparable thickness and soft tissue composition from which a line-by-line correction is applied to the BMD values in the bone region.

An edge detection algorithm is first used to find the bone edges (Figure 2.13).

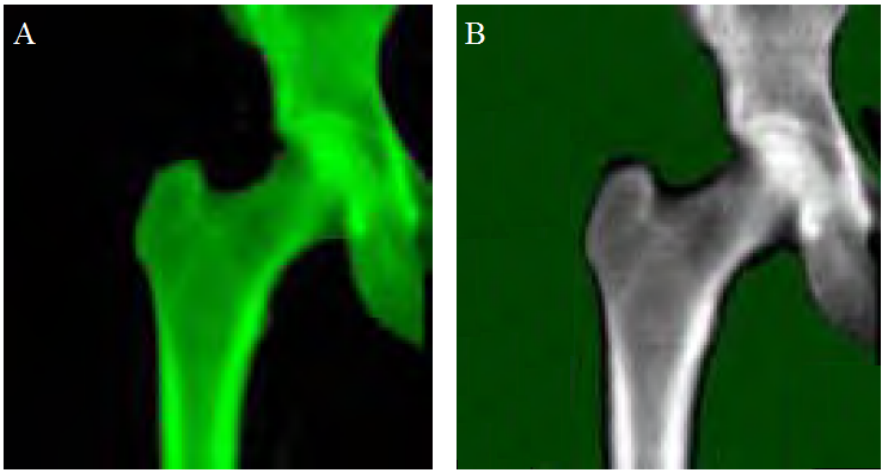

Figure 2.13: A femur projection by Lunar iDXA ME scanner: (A) results automatic bone point and (B) soft tissue point typing. From [11].

The total projected area of bone is then derived by summing the pixels within the bone edges and the reported value of BMD calculated as the mean BMD over all the pixels identified as bone. Finally, Bone Mineral Content (BMC) is derived by multiplying mean BMD by projected area:

$$BMC = BMD \cdot Area \tag{2.31}$$

Analogous of Equations 2.28 can be written for fat and lean tissue in pixels that not containing bone, and also in this case the equation is complicated by the beam hardening effect. Thus, it is common practice to describe the R as a function of high energy attenuation (HE), a surrogate for total mass [12]. In figure 2.14 step phantom calibration curves are shown, data are acquired on a Hologic system but the same is true for any DXA system. The R-value, which is the ratio of the low to high energy attenuation, is plotted on the vertical axis. Higher R-values correspond to increasing fat-free (lean) content. High energy attenuation, which is proportional to mass, is plotted on the horizontal axis.

In practice, DXA algorithms are more complex than those presented, and models used are proprietary and not available to the research community. Modern DXA scanner provide measures of total body and regional mass of different components.

By extending basic principles, triple photon absorptiometry might allow for the measurement of three different types of tissue. However, because there are only two attenuation processes (Compton scattering and the photoelectric effect) in the passage of diagnostic X-rays through tissue, the equations of triple photon absorptiometry have built in redundancy. Then within the diagnostic energy range

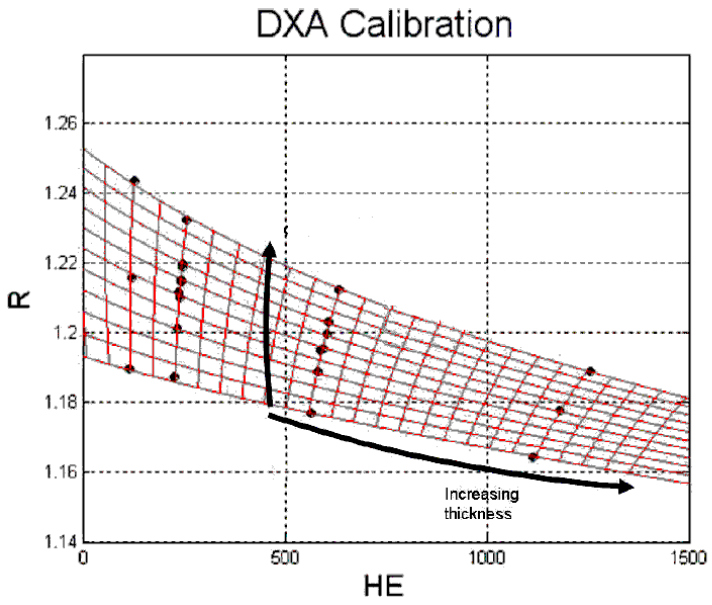

Figure 2.14: Step phantom calibration curves on a Hologic Delphi. This curve represented the R values as a composite of fat and lean tissue. Higher red lines correspond to higher presence of lean tissue. The black dots are phantom measurements at different thickness and composition. The red lines are the calibration function that was a best fit to the phantom data. Horizontal lines show iso-composition and vertical lines show iso-volume. Source [2].

with, measurements at more than two energies add no information, in practice it is possible to discriminate only two types of tissue [8].

2.4 DXA System

DXA systems have much in common with other medical X-ray imaging systems, with many of the same components, as you can see in Figure 2.15.

The scan speed and image quality are dictated by the X-ray beam geometry.

In commercially available DXA systems, the method by which low and high energy images are acquired varies according to manufacturer. For example, the exact X-ray tube voltage settings are unique to each manufacturer. The need for excellent spatial registration between low and high energy images is critical, since this affects the R values. Mis-registration can lead to substantial errors.

First generation bone densitometers use pencil beam geometry Figure 2.16 (A). The photon beam is tightly collimated with one photon source and one detector. The source and detector are rigidly coupled and moved together in a rectilinear manner to build an image of the bone being examined line by line. The disad-

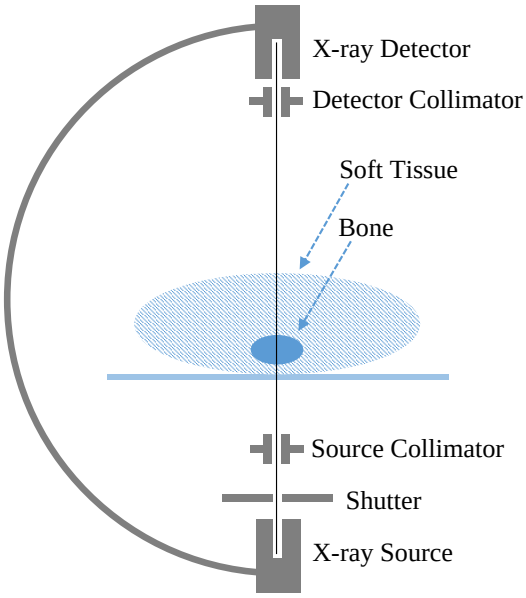

Figure 2.15: Typical DXA system including the X-ray tube, filtration, pre-patient aperture, examination table or surface, pre-detector aperture and detector. The components have a fixed geometry.

vantage of this technology is the relatively slow scan speed. However, the direct relationship between source and detector means that calculated bone and tissue masses are less likely to be artefactual.

The second generation of X-ray bone densitometer has a fan beam geometry, with a source which fans out in the short axis plane of the patient and is measured by an array of detectors in the same plane, Figure 2.16 (B). The bones are imaged in one pass along the long axis of the body providing an immediate advantage in scan speed (time for a whole body scan take 20 min for pencil beam, and less than 3 min for fan beam). The disadvantage of fan beam DXA is that the photon flux at the edges is lower than the middle of the image (due to the inverse square law). As a result, mass calculations may have some systematic error, although bone mineral density values have been shown to be unaffected.

It is important to note that fan and pencil beam systems project the three dimensional human body onto the two dimensional image in different ways, as is illustrated in Figure 2.17.

The pencil beam image is projected perpendicular to the plane of the table, the fan beam image may be projected under a certain angle in the direction parallel to the fan width. Thus, even if identical ROIs are outlined on the resulting images, these ROIs are projections of different physical volumes of interest. This difference between pencil and fan beam ROIs is one of definition. DXA ROI definitions are

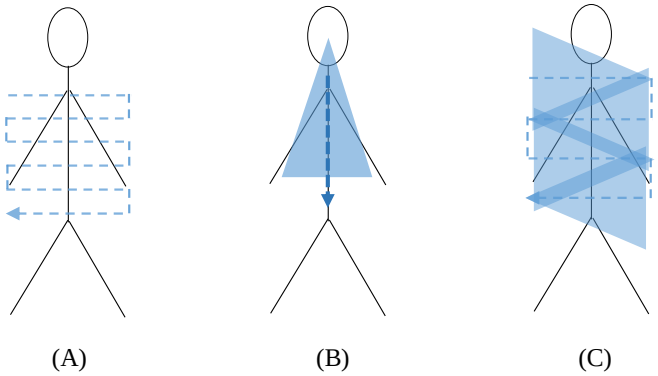

Figure 2.16: Illustration of the path of the X-ray beam (arrow) in the successive dual-energy X-ray absorptiometry (DXA) systems: (A) pencil beam, (B) fan beam, and (C) narrow fan beam (modified from [13]).

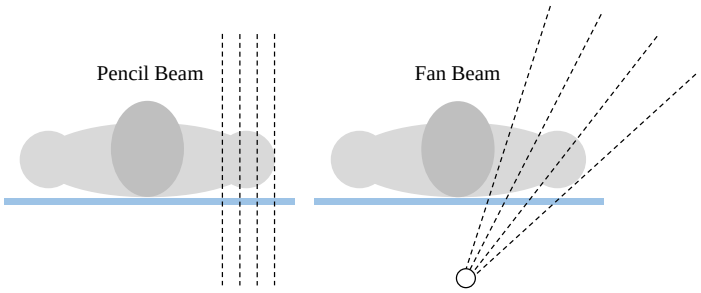

Figure 2.17: Pencil and fan beam geometries project the same ROI differently. The pencil beam image is projected perpendicular to the plane of the table, whereas the fan beam projection depends on the position of the object within the beam (modified from [2]).

arbitrary; both projections and measurements are equally valid [14].

An additional difference between pencil beam versus fan beam systems is the so called fan beam magnification. The size of the projected area of a volume of interest depends on the position of the object between the X-ray tube and the detector.

Thus, in fan beam systems, bone area appear to decrease the further the subject is from the source. Since the precise location of the bone along the X-ray path is generally not known, magnification errors in bone area can be challenging to correct. In addition, magnification and its associated error only occur in the dimension along the fan length. The image dimension in the direction of the scanning motion does not have any magnification.

The most recent advance has been the introduction of the narrow fan beam. Narrow fan beam is designed to overcome some of the limitations of the fan beam geometry. A small fan beam (about 4cm wide at the detector) in the long axis is measured by an array of detectors. The beam scans the bones in the short patient axis on each individual sweep along the long axis of the patient with some beam overlap (Figure 2.16 (C)). Although slightly slower than a fan beam scanner, the mass results should be more accurate as the photon flux has little variability in the area being measured and the magnification is really low (due to the beam overlap).

2.4.1 DXA radiation source

The replacement of the radionuclide source used in DPA with a X-ray tube improved the performance of dual photon absorptiometry by combining higher photon flux with a smaller diameter source.

The availability of an intense, narrow beam of radiation shortened scan times (from 20 minutes to 2 minutes), enhanced image definition (image resolution from 2 mm to 1 mm), and improved precision (BMD measurements from 2% to 1%) [9].

In all DXA systems on the market, the X-ray tubes used are standard tungsten anode tubes with focal spot sizes on the order of 0.5 to 1 mm.

The use of a X-ray tube requires the solution of several significant technical problems. A highly stable source is essential. Image noise must be limited by photon statistics and not by instabilities in the X-ray generator. Because X-ray tubes produce poly-energetic spectra rather than the discrete line emissions of a radionuclide, the effects of beam hardening are a potential source of error.

In beam hardening, lower energy photons are preferentially removed from the

radiation beam compared to higher energy photons, leading to a progressive shift in spectral distribution to higher effective photon energies with increasing body thickness. As a result the attenuation coefficients for bone and soft tissue in Equation 2.29 change with body thickness, and so vary from patient to patient and from site to site within the body.

However, there are differences in how the dual energy images are created. The two methods in use are K-edge filtering systems made by Norland (Norland, Cooper Surgical, Madison, WI, United States of America) and GE Lunar (GE Healthcare, Madison, WI, United States of America) and voltage switching systems made by Hologic [2].

In a K edge filter system, the X-ray tube is operated in a steady direct current mode and a K absorption edge filter splits a single X-ray spectrum into low and high energy components that mimic the emissions from $^{1}53Gd$. Because the two components have inherently narrow spectral distributions the problems associated with beam hardening are minimized (Figure 2.18). The Lunar DXA systems have a cerium (Z = 58) filter and use pulse height analysis at the detector to discriminate between high- and low-energy photons. Norland systems use a samarium (Z = 62) filter and separate detectors for high and low energy X-rays. Dynamic range is extended by a system for switching filters with different thicknesses of samarium into the beam. In these systems, since both high and low energy X-rays are intermixed, the energy separation is done at the detector using pulse height measurements.

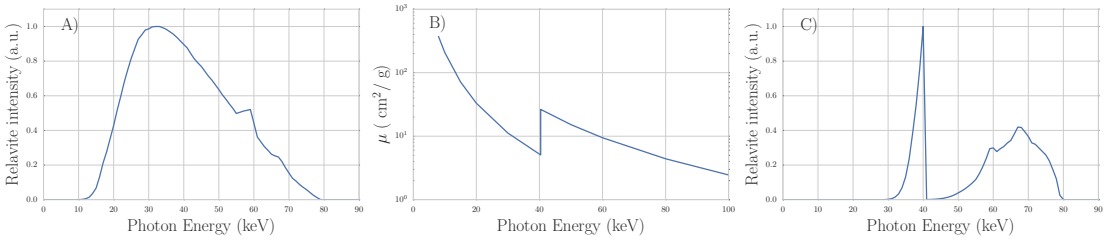

Figure 2.18: Pirnciples of K-edge filtration. (A) Unfiltered 80 kV spectrum (tungsten anode), attuation coefficent as function of photon energy of Cerium (Z = 58); (C) 80 kV spectrum firltered by 400 mg/cm^2 cerium. Spectrum have been normalized to same peak intensities. (A) and (C) data generated using [15] through online tool for the simulation of X-ray spectra by Siemens tool, (B) data from NIST.

In a voltage switching system, two X-ray tube voltage settings are used to create low and high energy images. The X-ray tube power supply switches between a low (70 kVp) and high (140 kVp) voltage setting during alternate half cycles of the power supply. The resulting pulses are very short, 8.33 ms for 60 Hz and 10 ms for

50 Hz systems. The spectral distribution (Figure 2.19) is wider than with the K-edge filter method and the consequent effects of beam hardening are corrected by a rotating calibration wheel (Figure 2.20) containing bone and soft tissue equivalent filters that measure the attenuation coefficients in the DPA equation and calibrate the scan image pixel by pixel. The filter, voltage switching and detectors are all electronically and mechanically synchronized to sequentially collect low and high energy information for each position of the X-ray gantry.

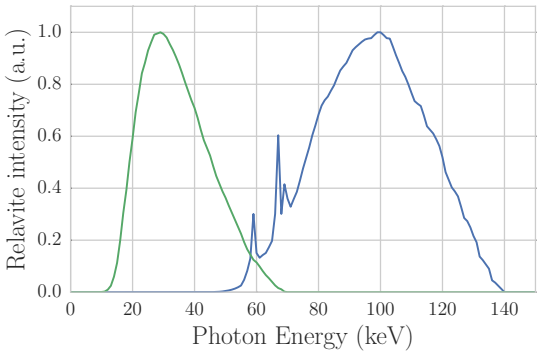

Figure 2.19: Dual kV spectrum with 4-mm Al filtration and 3.1 mm Cu filter in 140 kV beam. Data generated using [15] through online tool for the simulation of X-ray spectra by Siemens tool

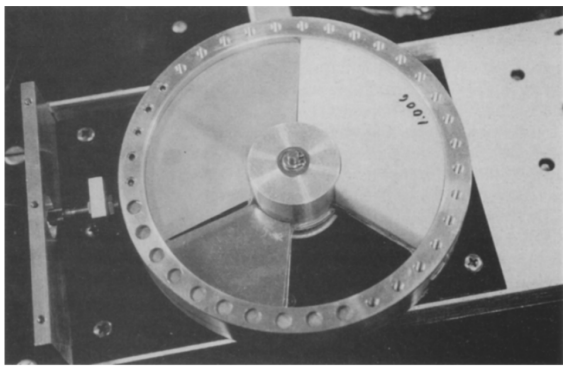

Figure 2.20: The calibration wheel used as the internal reference standard in Hologic scanners. The segments in the wheel include bone and soft tissue equivalent filters together with an empty air sector. Each of these 3 segments has separate high and low energy X-ray sectors with and without an additional brass filter. Image from [9]

2.4.2 Radiation dose

DXA systems generate ionizing radiation. Thus, subjects being scanned and equipment operators, consequently, receive a (small) radiation dose as a result of

any procedure. The absorbed dose to tissue is quantified as the amount of energy absorbed in a kilogram of tissue. The unit of measure is the Gray (Gy), where 1 Gy is equivalent to 1 J/kg. Another useful quantity of dose is *effective dose*, measured in sievert (Sv) where 1 Sv = 1 J/kg.

In summary: 1 Gy (physical quantity) is the deposit of a joule of radiation energy in a kg of matter or tissue. 1 Sv (biological effect) represents the equivalent biological effect of the deposit of 1 J of radiation energy in 1 Kg of human tissue.

Effective dose takes account not only of the amount of energy absorbed, but also the type of radiation and the susceptibility of the tissue to radiation damage. The effective dose is calculated as the sum of the absorbed doses to radio-sensitive organs multiplied by their associated weighting factors, w_T and w_R. The tissue weighting factors and radiation weighting factors are defined by National Council on Radiation Protection (NCRP). In other words, the effective dose, E, is the tissue weighted sum of the equivalent doses in all specific tissues and organs of the body, given by the expression:

$$E = \sum_T w_T \sum_R w_R D_{R,T} \tag{2.32}$$

where $D_{T,R}$ is the absorbed dose, w_R is the radiation weighting factor equal to one for diagnostic X-rays, and w_T is the tissue weighting factor for different tissues (Table 2.3).

Table 2.3: Tissue weighting factors, source NCRP [2]. (*) Remainder tissues: adrenals, extrathoracic region, gall bladder, heart, kidneys, lymphatic notes, muscle, oral mucosa, pancreas, prostate, small intestine, spleen, thymus, uterus/cervix.

Tissue	w_T	$\sum w_T$
Bone marrow, colon, lung, stomach, breast, remainder tissues*	0.12	0.72
Gonads	0.08	0.08
Bladder, oesophagus, liver, thyroid	0.04	0.16
Bone surface, brain, salivary glands, skin	0.01	0.04
Total		1

Basic Safety Standards (BSS) [16] require that all medical radiation exposures are appropriately justified.

The diagnostic benefit from DXA must outweigh the radiation detriment that might ensue. The ICRP recommends that both generic justification and individual justification are applied. For generic justification, the national professional bodies, in conjunction with national health authorities and the radiation protection

regulatory body, will have decided which DXA procedures generally improve the diagnosis or treatment, or provide necessary information about the exposed individuals. Individual justification considers whether the application of the particular DXA procedure to a particular individual is justified or not.

DXA procedures may also be used as part of a biomedical research project, such as in the role of a metric where the measurement of bone density or body composition is part of assessing the efficacy of the treatment under investigation.

In this situation, the benefit from the use of radiation is expected to be accrued by society, such as through improved health care options. The use of DXA procedures in this role must normally also be justified by an ethics committee. If a given DXA procedure is justified, then the BSS require that its performance is optimized. For DXA, this means ensuring that the patient dose is the minimum necessary (*ALARA* principles: As Low As Reasonably Achievable) to determine bone density or body composition to an appropriate level of certainty.

Patient effective doses in DXA depend on the type of beam (pencil beam, fan beam, narrow fan beam), the protocol or mode used for the scan (scan area, tube current, scan speed) and the body region being scanned.

Many DXA units offer different acquisition modes, typically, the tube current and/or scanning speed is changed. The patient dose may change widely based on mode of examination [11] [17] [18]. Then the appropriate choices of parameters must be made for the particular individual undergoing the procedure.

To put these DXA patient doses into perspective, it is helpful to consider exposure from other sources. Human beings are constantly exposed to ionizing radiation from natural sources, including cosmic rays and naturally occurring radioactive material in foods, soil, water and air. This is collectively referred to as natural background radiation. The average annual natural background radiation dose to humans worldwide is about 2400 μSv (vary from 1000 to 10 000 μSv). Thus, in comparison, effective patient doses from DXA are small and are similar to those received on average from one or two days of exposure to natural background radiation. Adult effective doses, represented in μSv, for various radiological procedures and conditions are shown in Figure 2.21.

Operators that perform the DXA procedure may also receive a radiation dose due to scattered radiation from the patient. This scattered dose is much less than the dose in the primary beam. While the occupational dose limit prescribed in the BSS is 20 000 μSv/a averaged over five consecutive years with a limit of 50 000 μSv in any single year, the application of the principle of optimization of protection

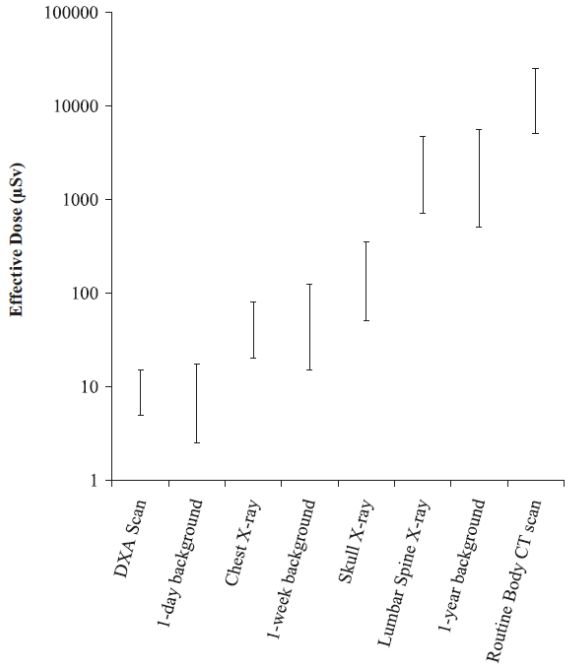

Figure 2.21: Adult effective doses for various procedures and exposures. Source [2]

means that occupational doses must be as low as reasonably achievable.

With DXA, occupational doses are determined primarily by the workload of the DXA unit (number of patients per day), the distance the technologist or other personnel are from the patient during the scan, and the type of scanner and the protocols/mode being used (Table 2.4).

Table 2.4: Radiation dose as function of distance from beam. The beam was attenuated with a water target having dimensions of 25x25x15 cm. Each measurement consisted of a static exposure at the maximum X-ray tube current and voltage of 2.5mA and 100kV. Source [11]

Dose (μSv/h)	Distance (cm)
44	37.5
13.2	75
5.3	112.5
2.6	150
1.3	200

In practical terms, the operator's desk should be positioned at more then 2 m from the scanner and the use of protective screens or shields may be not necessary. With these precautions, it is most likely that the operator dose will be in the lower range of acceptable occupational exposures.

2.4.3 DXA regions of interest

There are several regions of interest (ROIs) that can be defined, with each having information to offer. The optimal site depends on the intent of the scan. Typical ROIs, defined by the proprietary software, include five main corporeal districts: trunk, upper limbs, lower limbs, android region (a portion of the abdomen included between the line joining the two superior iliac crests and extended toward the head up to the 20% of the distance between this line and the chin) and gynoid region (a portion of legs leaving from the femoral great trochanter, directed caudally up to a distance double of the android region). For bone density, most ROIs currently defined (spine, femur) are useful for the diagnosis and prevention of diseases related to osteoporosis. Only the whole body scan mode can measure fat, lean and bone total mass. Figure 2.22 shows an acquisition of a whole body scan, highlighting the five main districts in which are calculate the amount of fat and lean tissues.

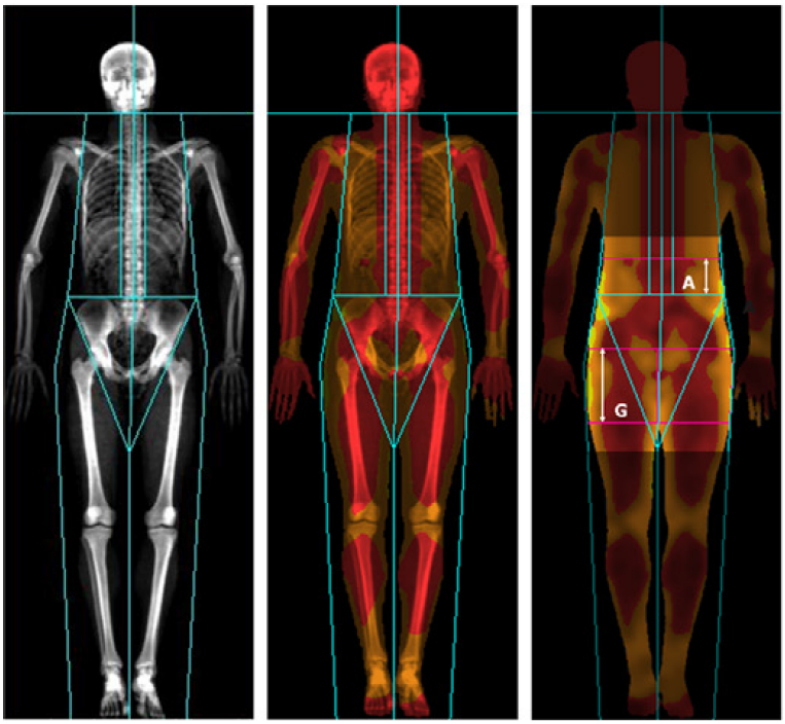

Figure 2.22: DXA examination of body composition (yellow for high fat percentage tissue, red for low fat percentage e lean tissue). ROIs are automatically drawn by the software according to anatomical landmarks (A and G are android and gynoid regions respectively).

2.4.4 DXA limitation

While the DXA technique is extremely accurate, with very low doses, it has some limitations. DXA is still a projective technique, the measurements of bone density is in units of grams per unit area since DXA does not have the ability to measure tissue thickness. Thus, DXA systems cannot tell the difference between thick low density bone and thin high density bone.

Degenerative changes, for example aortic calcifications and fractured vertebrae, are difficult to visualize and can cause significant bias to the BMD results. These biases are systematic and typically found in older populations beyond the age of 65 years.

Moreover as we have already said DXA can only solve for two materials simultaneously. Thus, soft tissue composition can only be solved in areas exclusive of bone, and bone mass can only be determined with an assumption of the soft tissue composition overlaying the bone. Since bone is typically contained in 40% or more of the body image pixels, the soft tissue composition has to be estimated from surrounding tissue. In some cases, accurate estimates cannot be made, such as the head, hands, feet and upper torso because there is not adequate soft tissue outside the bone projection, and manufacturers turn to proprietary methods to reference the soft tissue.

Generally, BMD values across manufacturers cannot be directly compared and are not interchangeable for several reasons. The comparison of patient data among different dual X-ray absorptiometry (DXA) scanners is complicated because no universally accepted cross-calibration procedure or standard currently exists.

Although operating on the same basic principles, the instruments show differences in scanner design, bone mineral calibration, and analysis algorithms. Lunar scanners rely on daily scanning of standards to provide a bone tissue equivalent calibration. Hologic uses an internal calibration system, which corrects for short-term instabilities. Also, the software used for analysis of the scans is manufacturer specific and unique, particularly with regard to the edge detection algorithms used for separating bone and soft tissue regions. This implementation results in differences in the defined bone area and BMC evaluated by different systems (then also BMD measurements is different because this is calculated as BMC divided by the measured area). For example, the differences between Hologic and GE Lunar are approximately 8% in BMD and 20% in BMC. There is also a lack of standardization

on the placement of ROIs [19].

2.4.5 Quality Control

Longitudinal Scanner Quality Control (QC) procedures consist of procedures used to monitor the performance of a single scanner over time. QC is carried out daily, before the measurements on patients. This procedure calibrates and verifies functionality as well as the accuracy and precision of the densitometer.

The QC procedure is completely automatized and use a the calibration block phantom (Figure 2.23) that consists of tissue-equivalent material with three bone-simulating chambers of known bone mineral content.

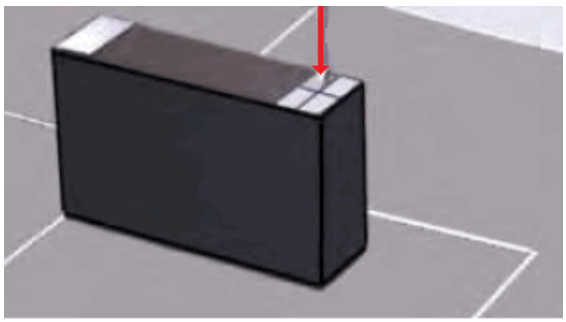

Figure 2.23: Black box phantom of tissue-equivalent material with three bone-simulating chambers. The red arrow represent the laser alignment system. Source [11]

These scans are analysed automatically and added to the QC database.

If the measurements falls outside acceptable limits, the phantom should be re-scanned. If the measurement from the second scan also falls outside the limits, the service provider for the system should be called.

Daily QC report is show at the end of each procedure, in Figure 2.24 is show the QC reports of last year.

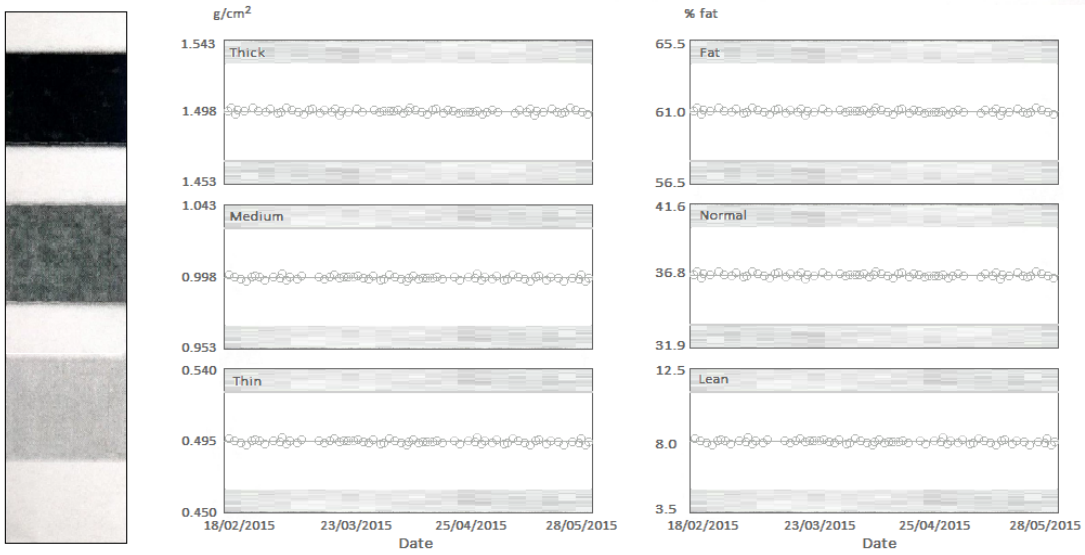

Figure 2.24: Report Print page: on the right the phantom scan is shoe, on the center and on the left result of calibration, for the BMD and soft tissue composition respectively, are show. Data from "Lunar iDXA" enCORE software v.16.

Chapter 3

Statistical Methods

In this chapter statistical tools, that we will use in the last part of this work, are explained.

3.1 Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test (KS test) is a non parametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare two samples (two-sample KS test). The two-sample KS test is one of the most simple and useful non-parametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

The KS statistic quantifies a distance between the empirical continuous distribution functions of two samples, is defined as:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|, \tag{3.1}$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively, and $\sup_x$ is the supreme function of the set of distances (see Figure3.1).

The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution. The null hypothesis is rejected if

$$D_{n,n'} > c(\alpha, n, n'). \tag{3.2}$$

The value of $c(\alpha, n, n')$ is given in tables that have been published.

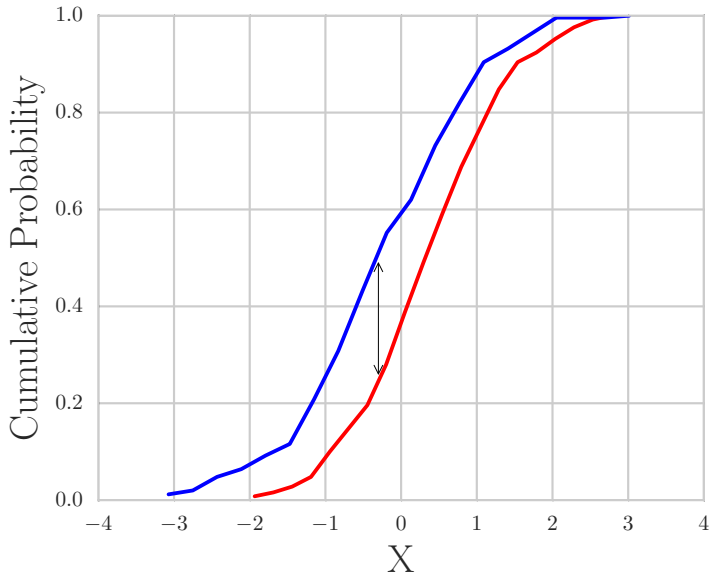

Figure 3.1: Illustration of the two-sample KS statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

Note that the two-sample KS test checks whether the two data samples come from the same distribution. This does not specify what that common distribution is (e.g. whether it's normal or not normal).

We will use this test to asses if different populations or subjects groups have similar enough distribution of observables to be treated as a single population or corrections need to be made. These corrections could range from dividing the two datasets and analyzing them separately, to explicitly model the difference between the groups.

3.2 Principal component analysis

The original purpose of principal component analysis (PCA) was to reduce a large number (p) of variables to a much smaller number (m) of principal components (PCs) whilst retaining as much as possible of the variation in the p original variables.

Although there are many other ways of applying PCA, this original usage is probably still the most prevalent single application [20].

In this work the application of PCA is performing not only to reduces the dimensionality of the problem, but to obtain m PCs ($m \ll p$) which can be readily

interpreted.

Since PCA projects original data onto directions which maximize the its variance, if some variables have a large variance and some small, PCA will load on the large variances.

To have a PCA independent of features variance, first pre-processing data to normalize its mean and variance as follow:

1. Let $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ where i denotes the ith subjects
2. Replace each $x^{(i)}$ with $x^{(i)} - \mu$
3. Let $\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)}\right)^2$ where j denotes the jth measurement
4. Replace each $x_j^{(i)}$ with $x_j^{(i)} \Big/ \sigma_j$

Steps (1-2) zero out the mean of data, steps (3-4) rescale each coordinates to have unit variance, which ensures that different attributes are all treated on the same scale.

Suppose that $\boldsymbol{x}$ is a vector of p random variables. Consider the case where the vector of random variables $\boldsymbol{x}$ has a known covariance matrix $\boldsymbol{\Sigma}$. This is the matrix whose (i,j)th element is the covariance between the ith and jth elements of x when $i \neq j$, and the variance of the jth element of $\boldsymbol{x}$ when $i = j$.

It turns out that for $k = 1, 2, \cdots, p$, the kth PC is given by $\boldsymbol{z}_k = \boldsymbol{\alpha}_k^\top \boldsymbol{x}$ where $\boldsymbol{\alpha}_k$ is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to its kth largest eigenvalue λ_k.

Furthermore, if $\boldsymbol{\alpha}_k$ is chosen to have unit length ($\boldsymbol{\alpha}_k^\top \boldsymbol{\alpha}_k = 1$), then $\text{var}(z_k) = \lambda_k$, where $\text{var}(z_k)$ denotes the variance of z_k.

To derive the form of the PCs, consider first $\boldsymbol{\alpha}_1^\top \boldsymbol{x}$; the vector α_1 maximizes $\text{var}\left[\boldsymbol{\alpha}_1^\top \boldsymbol{x}\right] = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1$. It is clear that, as it stands, the maximum will not be achieved for finite α_1 so a normalization constraint ($\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$) must be imposed.

To maximize $\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$:

$$\underset{\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1}{\arg\max} \left\{ \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 \right\} = \arg\max \left\{ \frac{\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1} \right\} \tag{3.3}$$

the standard approach is to use the technique of Lagrange multipliers formulation of the Rayleigh quotient.

The quantity to be maximised can be recognised as a Rayleigh quotient $R(\boldsymbol{\Sigma}, \boldsymbol{\alpha}_1)$, defined as:

$$R(\boldsymbol{\Sigma}, \boldsymbol{\alpha}_1) = \frac{\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1} \tag{3.4}$$

with the constraint $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$.

The problem is to find the critical points of $R(\boldsymbol{\Sigma}, \boldsymbol{\alpha}_1)$ and this is equivalent to find the critical points of

$$\mathcal{L}(\boldsymbol{\alpha}_1, \lambda) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda \left(\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 - 1 \right), \tag{3.5}$$

where λ is a Lagrange multiplier. The stationary points of $\mathcal{L}(\boldsymbol{\alpha})$ occur at

$$\frac{d\mathcal{L}(\boldsymbol{\alpha}_1)}{d\boldsymbol{\alpha}_1} = 0 \tag{3.6}$$

therefore

$$2\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} - 2\lambda \boldsymbol{\alpha}_1^\top = 0 \Rightarrow \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda \boldsymbol{\alpha}_1. \tag{3.7}$$

Therefore, the eigenvector $\boldsymbol{\alpha}_1$ of Σ is the critical points of the Rayleigh Quotient and their corresponding eigenvalue λ_1, is the stationary value of R.

To decide which of the p eigenvectors gives $\boldsymbol{\alpha}_1^\top \boldsymbol{x}$ with maximum variance, note that the quantity to be maximized is:

$$R(\boldsymbol{\Sigma} \boldsymbol{\alpha}_1) = \frac{\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1} = \lambda \frac{\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1} = \lambda \tag{3.8}$$

so λ must be as large as possible. Thus, $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$, and $\text{var}(\boldsymbol{\alpha}_1^\top \boldsymbol{x}) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda_1$, the largest eigenvalue.

In general, the kth PC of x is $\boldsymbol{\alpha}_k^\top x$ and $\text{var}(\boldsymbol{\alpha}_k^\top \boldsymbol{x}) = \lambda_k$, where λ_k is the kth largest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_k$ is the corresponding eigenvector. This will now be proved for $k = 2$; the proof for $k \geq 3$ is slightly more complicated, but very similar. The second PC, $\boldsymbol{\alpha}_2^\top \boldsymbol{x}$, maximizes $\boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2$ subject to being uncorrelated with $\boldsymbol{\alpha}_1^\top \boldsymbol{x}$, or equivalently subject to $\text{cov}[\boldsymbol{\alpha}_1^\top \boldsymbol{x}, \boldsymbol{\alpha}_2^\top \boldsymbol{x}] = 0$, where $\text{cov}(x, y)$ denotes the covariance between the random variables x and y.

Noting that:

$$\text{cov}[\boldsymbol{\alpha}_1^\top \boldsymbol{x}, \boldsymbol{\alpha}_2^\top \boldsymbol{x}] = \boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1, \tag{3.9}$$

the quantity to be maximized is

$$\mathcal{L}(\boldsymbol{\alpha}_1, \lambda, \phi) = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda \left(\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 - 1 \right) - \phi \boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1, \tag{3.10}$$

where λ, ϕ are Lagrange multipliers.

Differentiation with respect to $\boldsymbol{\alpha}_2$ and multiplication on the left by $\boldsymbol{\alpha}_1^\top$ gives:

$$\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 - \phi \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 0, \tag{3.11}$$

which, since the first two terms are zero and $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$, reduces to $\phi = 0$. Therefore,$\boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda \boldsymbol{\alpha}_2 = 0$, so λ is once more an eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_2$ the corresponding eigenvector.

Again,$\lambda = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2$ so λ is to be as large as possible. Assuming thatΣ does not have repeated eigenvalues, λ cannot equal λ_1. If it did, it follows tha$\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1$, violating the constraint $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 = 0$. Hence λ is the second largest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_2$ is the corresponding eigenvector.

It can be shown that for the kth PC, the vector of coefficients $\boldsymbol{\alpha}_k$ is the eigenvector of $\boldsymbol{\Sigma}$ corresponding to λ_k, the kth largest eigenvalue.

It should be noted that the sign of any PC is completely arbitrary. If every coefficient in a PC,$z_k = \boldsymbol{\alpha}_k^\top \boldsymbol{x}$, has its sign reversed, the variance of z_k is unchanged, and so is the orthogonality of $\boldsymbol{\alpha}_k$ with all other eigenvectors.

3.3 Non-Negative Matrix Factorization

A fundamental problem in this data-analysis is to find a suitable representation of the data. A useful representation typically makes latent structure in the data explicit, and often reduces the dimensionality of the data to facilitate the application of statistics and computational methods.

Non-Negative Matrix Factorization (NMF) is a useful decomposition for multivariate data, oriented to find a positive part-based linear representation of non negative data [21].

We formally consider algorithms for solving the following problem, given a non-negative matrix $\boldsymbol{V}$, find non-negative matrix factors $\boldsymbol{W}$ and $\boldsymbol{H}$ such that:

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}. \tag{3.12}$$

NMF can be applied to the statistical analysis of multivariate data in the following manner. Given a set of of multivariate n-dimensional data vectors, the vectors are placed in the columns of an $n \times m$ matrix $\boldsymbol{V}$ where m is the number of examples in the data set. This matrix is then approximately factorized into an $n \times r$ matrix $\boldsymbol{W}$ and an $r \times m$ matrix $\boldsymbol{H}$.

Usually r is chosen to be smaller than nor m, so that $\boldsymbol{W}$ and $\boldsymbol{H}$ are smaller than the original matrix $\boldsymbol{V}$ (see Figure3.2). This results in a compressed version of the original data matrix.

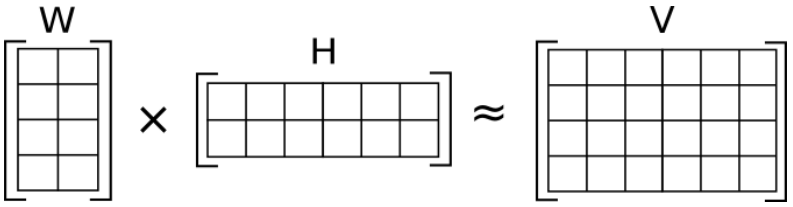

Figure 3.2: Illustration of approximate non-negative matrix factorization: the matrix $\boldsymbol{V}$ is represented by the two smaller matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, which, when multiplied, approximately reconstruct $\boldsymbol{V}$.

It can be rewritten column by column as $\boldsymbol{v} \approx \boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{v}$ and $\boldsymbol{h}$ are the corresponding columns of $\boldsymbol{V}$ and $\boldsymbol{H}$. In other words, each data vector $\boldsymbol{v}$ is approximated by a linear combination of the columns of $\boldsymbol{W}$, weighted by the components of $\boldsymbol{h}$. Therefore $\boldsymbol{W}$ can be regarded as containing a basis that is optimized for the linear approximation of the data in $\boldsymbol{V}$. Since relatively few basis vectors are used to represent many data vectors, good approximation can only be achieved if the basis vectors discover structure that is latent in the data [22].

To find an approximate factorization $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$, we first need to define cost functions that quantify the quality of the approximation. Such a cost function can be constructed using some measure of distance between two non-negative matrices A and B . One useful measure is simply the square of the Euclidean distance between A and B:

$$\|\boldsymbol{A} - \boldsymbol{B}\|^2 = \sum_{i,j} (\boldsymbol{A}_{i,j} - \boldsymbol{B}_{i,j})^2 \tag{3.13}$$

This is lower bounded by zero, and clearly vanishes if and only if $\boldsymbol{A} = \boldsymbol{B}$.

As shown in [23] both PCA and NMF represent a signal as a linear combination of basis, but with qualitatively different results.

In NMF approach only additive combinations are allowed, because the non-zero elements of W and H are all positive. In contrast to PCA, no subtractions can occur. For these reasons, the non-negativity constraints are compatible with the

intuitive notion of combining parts to form a whole, which is how NMF learns a factor-based representation of the signal.

3.4 Linear Regression Model

A statistical model is a simplistic and necessary representation of the reality derived from experimental observations and logical deductions.

A general linear model that is used to determine Y from a knowledge of x is usually written in the form:

$$Y = \mu(X) + \varepsilon \tag{3.14}$$

where Y, the response variable, and ε are random variables, $\mu(X)$ is a function of variables X defined in domain D and named predictor variable. The function μ is definite to be the deterministic portion of Y and ε the random (stochastic) portion. If μ is of first order we talk about linear model.

Linear regression is an approach for modelling the relationship between Y, the response variable, and X, the predictors variables. The case of one explanatory variable is called *simple linear regression*.

More formally, linear regression represent a method to estimate the expected value with condition to an dependent variable Y, given the values of independent variables $X_1, \ldots X_k$:

$$\mathbb{E}[Y|X_1, \ldots, X_k]. \tag{3.15}$$

Linear regression has two main practical uses:

- If the goal is the prediction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.

- Given a variable y and a number of variables $X_1, \ldots, X_p$ that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j, to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y.

Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n samples, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of predictors variables x_i is linear. This relationship is modelled through an unobserved random variable (error variable) ε_i that adds noise to the linear relationship. Thus the model takes the form:

$$Y = X\beta + \varepsilon, \tag{3.16}$$

where $Y = [y_1, \ldots, y_n]^\top$ is an $n \times 1$ observable random vector, X is a $n \times p$ matrix where the rows $x_i = [x_{i,1}, \ldots, x_{i,p}]$ is the observable, $\beta = [\beta_1, \ldots, \beta_p]^\top$ is a $p \times 1$ vector of unknown parameters and $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^\top$ is a $n \times 1$ random vector such that the expected value $\mathbb{E}(\varepsilon = 0$.

Ordinary Least Squares

Ordinary least squares (OLS) method is used for estimating the unknown parameters in the regression models.

The goal of OLS is to minimizing the differences between the observed responses and the responses predicted by the linear approximation of the data (this is the sum of the squared vertical distances between each data point in the set and the corresponding point on the regression line).

The OLS estimator provides minimum-variance mean-unbiased estimation in absence of multicollinearity, the errors have the same finite variance and are uncorrelated. Under the additional assumption that the errors be normally distributed, OLS is the maximum likelihood estimator.

The quantity $y_i - x_i^\top \beta$, called the residual for the ith observation, measures the vertical distance between the data point (x_i, y_i) and the hyperplane $y = x^\top \beta$, and thus assesses the degree of fit between the actual data and the model. The sum of squared residuals is a measure of the overall model fit:

$$S(\beta) = \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 = (y - X\beta)^\top (y - X\beta). \tag{3.17}$$

The value of β which minimizes this sum is called the OLS estimator for the parameter. The function $S(\beta)$ is quadratic in β with positive-definite Hessian, and

therefore this function possesses a unique global minimum at $\beta = \hat{\beta}$:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} S(\beta) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i \tag{3.18}$$

or equivalently in matrix form,

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{3.19}$$

After we have estimated β, the fitted values (or predicted values) from the regression will be:

$$\hat{y} = X\hat{\beta}. \tag{3.20}$$

It is common to assess the goodness-of-fit of the OLS regression by comparing how much the initial variation in the sample can be reduced by regressing onto X. The *coefficient of determination* R^2 is defined as a ratio of explained variance to the total variance of the dependent variable y:

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}. \tag{3.21}$$

where $\overline{y}$ is the mean value of y.

In order for R^2 to be meaningful, the matrix X of data on regressors must contain a column vector of ones to represent the constant whose coefficient is the regression intercept.

Often is reported an adjustment of the coefficient R^2 to take account of the increasing of the R^2 when extra explanatory variables are added to the model.

R^2_{adj} include the information about the number of explanatory terms in a model relative and the sample size.

The R^2_{adj} can be negative, and its value will always be less than or equal to that of R^2. Unlike R^2, the R^2_{adj} increases when a new explanator is included only if the new explanator improves the R^2 more than would be expected by chance.

If a set of explanatory variables with a predetermined hierarchy of importance are introduced into a regression one at a time, with the R^2_{adj} computed each time, the level at which R^2_{adj} reaches a maximum, and decreases afterward, would be the regression with the ideal combination of having the best fit without excess or unnecessary terms.

The R^2_{adj} is often defined as

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{3.22}$$

where p is the total number of regressors in the model (not counting the constant term), and n is the sample size.

Despite the meaning of the parameter is still the same, many equivalent definitions have been made depending on the field of application.

In our analysis the used definition is

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{p}. \tag{3.23}$$

3.5 Support Vector Machine and Regression

In machine learning, support vector machine (SVM) is a supervised learning model, with associated learning algorithms, that analyses data and recognizes patterns, used for classification and regression analysis.

In practice SVM constructs a hyperplane in a high-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

Given some training data $\mathcal{D}$, a set of n points of the form

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\, y_i \in \mathbb{R}\}_{i=1}^{n}. \tag{3.24}$$

In SVM regression, our goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible [24]. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this.

We begin by describing the case of linear functions f, taking the form

$$f(\mathbf{x}) = \langle w, \mathbf{x} \rangle + b \tag{3.25}$$

where $\mathbf{x} \in \mathbb{R}^p$, $b \in \mathbb{R}$ and $\langle \cdot, \cdot \rangle$ denotes the dot product in $\mathbb{R}^p$.

Flatness in this case means that one seeks a small w, then imposing to solve the problem as a convex optimization problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|w\|^2 \\ &\text{subject to} && \begin{cases} y_i - \langle w, \mathbf{x}_i \rangle - b \leq \varepsilon \\ \langle w, \mathbf{x}_i \rangle + b - y_i \leq \varepsilon \end{cases} \end{aligned} \tag{3.26}$$

where $\|w\|^2 = \langle w, w \rangle$.

How ever an additional slack variable is add to to cope with otherwise infeasible constraints of the optimization problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) \\ &\text{subject to} && \begin{cases} y_i - \langle w, \mathbf{x}_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, \mathbf{x}_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \end{cases} \end{aligned} \tag{3.27}$$

where ξ_i and ξ_i^* are slack variables that specify the upper and the lower training errors subject to an error tolerance ε, and C is a positive constant that determines the degree of penalized loss when a training error occurs.

Figure 3.3 show the situation graphically. Only the points outside the shaded region contribute to the cost with linear penalization.

The next step is to make the SV algorithm non-linear. This could be achieved by simply preprocessing the training patterns $\mathbf{x}_i$ by a map $\Phi : \mathbb{R}^p \to \mathcal{F}$ into some feature space $\mathcal{F}$:

$$f(\mathbf{x}) = \langle w, \Phi(\mathbf{x}) \rangle + b \tag{3.28}$$

and replacing this new expression of f in 3.27.

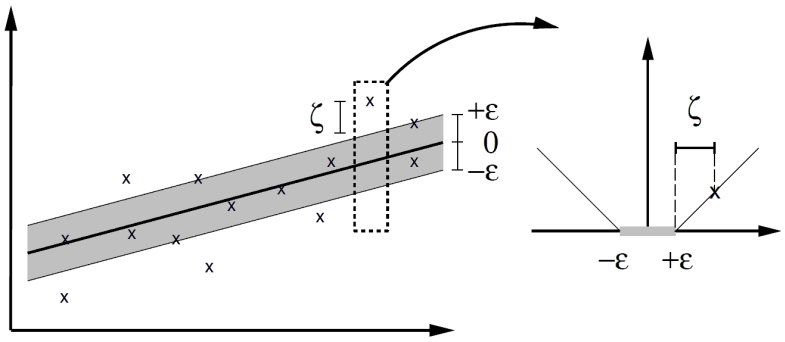

Figure 3.3: Illustration of the two-sample KS statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

3.6 Nearest Neighbours Classification and Regression

K-Nearest Neighbours algorithm (k-NN) is a non-parametric method used for classification and regression[25], where the input consists of the k closest (the nearest neighbours) training data point in the feature space. In k-NN classification, usually, an object is classified by a majority vote of its neighbours. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbours.

Neighbours-based regression can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based the mean of the labels of its nearest neighbours, see Fiure 3.4.

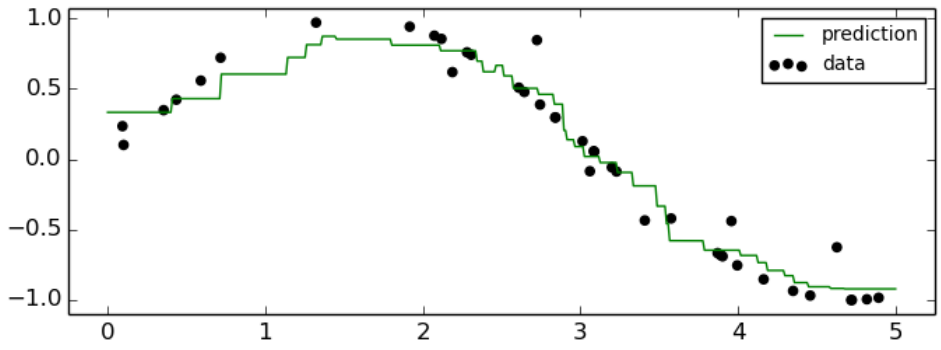

Figure 3.4: An example of k-NNR with $k = 5$ and uniform weight. Source [26].

The basic nearest neighbours regression uses uniform weights: that is, each point in the local neighbourhood contributes uniformly to the classification of a query point. Alternatively it can be advantageous to weight points such that

nearby points contribute more or with another function of the distance.

3.7 Cross validation method

In machine learning process finding out the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called over-fitting.

Cross Validation (CV) is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).

The goal of CV is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like over-fitting, give an insight on how the model will generalize to an independent dataset.

Two types of CV can be distinguished, exhaustive and non-exhaustive. In exhaustive CV methods all possible ways to divide the original sample into a training and a validation set are evaluated (e.g. leave one-out CV or leave p-out CV). Non-exhaustive CV methods do not compute all ways of splitting the original sample (e.g. k-fold CV).

Chapter 4

Data analysis

In the first part of this chapter we will analyse the body composition variables of healthy and normal weight subjects, recruited to achieve a wide age range, to investigate how fat tissue, and its distribution in the body, are related to blood lipid concentrations and to understand how a careful examination of body composition, such as DXA, affects these correlations.

The purpose of this analysis is to define method, preprocessing operations and factors to be taken into account for analysis of data from DXA and establish reference values for body composition on healthy people.

In the second part we will apply results obtained from the healthy database on another database, composed by elderly people. Metabonomic data of the patients will also be used to study the correlation between DXA variables and the metabolites.

4.1 BC assessment of healthy people

Cohort include 177 subject, 92 men and 85 women, from northern Italy. For each subject requirements of normal weight and good health are satisfied. Table 4.1 show mean and standard deviation of body mass index (BMI) defined as the body mass in kilograms divided by the square of the body height in meters, separately for males and females, in age range on ten years.

BMI values between 18.5 and 24.9 are commonly recognized as indicators of normal weight in both males and females.

For each subject DXA measurements of the whole body are acquired, more than 30 variables are available to assets the fat and lean content, Bone Mineral Content (BMC) and Density (BMD) in several standard Regions Of Interest (ROIs).

Table 4.1: Mean value of BMI in $[\mathrm{Kg/m^2}]$. n denote the number of subjects in the correspondent age range. Standard deviation is reported as uncertain.

	Male		Female	
Age	n	BMI	n	BMI
20-30	19	23 ± 2	20	22 ± 3
30-40	17	25 ± 2	16	24 ± 4
40-50	20	25 ± 3	21	24 ± 3
50-60	16	25 ± 3	20	24 ± 3
60-70	20	25 ± 3	8	23 ± 1

In Table 4.2 mean values of fat mass, lean (non bone) tissues mass, BMC and BMD of main ROIs are show.

Table 4.2: Fat mass, lean (non bone) tissues mass, BMC and BMD for main ROIs. Standard deviation is reported as uncertain.

Region	Fat [Kg]	Lean [Kg]	BMC [Kg]	BMD $[\mathrm{g/cm^2}]$
Whole body	20 ± 6	47 ± 9	2.7 ± 0.5	1.1 ± 0.1
Upper Limbs	2.3 ± 0.8	5.8 ± 1.8	0.4 ± 0.1	0.75 ± 0.10
Lower Limbs	7 ± 2	16 ± 4	1.0 ± 0.2	1.2 ± 0.2
Trunk	10 ± 4	22 ± 5	0.8 ± 0.2	0.96 ± 0.13
Android	1.7 ± 0.9	3.4 ± 0.7	0.04 ± 0.01	/
Gynoid	4 ± 1	6.7 ± 1.8	0.3 ± 0.1	/

In our analysis we want to describe fat and lean tissues distribution of subject, then we focus only on the soft tissues composition, and discard information about bone mineral content and density.

4.1.1 Preprocessing

Individuals can differ remarkably in body fat distribution, in particular these differences are consistent between men and women, both lean and obese. Women, compared to men, have higher percent body fat and deposit it in a different pattern, with relatively more adipose tissue in the hips and thighs, independently of total body fat [27].

In our analysis, features from each subject are normalized on his weight, in this way all measurements are intra- and inter-patients comparable.

For the same body mass index (BMI), women typically present with $\sim 10\%$ higher body fat compared to men. Aging increases adiposity in both sexes, but again, women are characterized by higher percent body fat throughout the entire life span [27].

In our database these difference are visible, in Table 4.3 are show the mean (μ) and variance (σ^2) of the fractional fat and lean content of upper (trunk, upper limbs and android regions) and lower (lower limbs and gynoid regions) part of body.

Table 4.3: Difference between males and females composition of upper and lower part of body. Upper region fat (lean) fraction is computed by sum of trunk, upper limbs and android fat (lean) mass (each term normalized on total weight of subject). Lower by sum of lower limbs and gynoid term. Standard deviation is reported as uncertain.

		Male		Female	
		μ	σ	μ	σ
Upper Body	Fat	0.2	0.2	0.2	0.2
	Lean	0.5	0.2	0.4	0.2
Lower Body	Fat	0.1	0.2	0.2	0.2
	Lean	0.4	0.2	0.3	0.1

As we expect, gender is a strong discrimination factor, in fact males and females have important difference in distribution of fat and lean mass. Lean soft-tissue mass is greater in males than in females irrespective of all segments, on the other hand fat tissue mass is greater in females.

Age-dependent changes in body composition, namely a decrease in lean mass and an increase in fat mass, are often observed in normal populations, we expect that a subject will lose muscular tone getting older, and this effect corresponds to an increase of the fat to lean mass ratio.

The Pearson's r, as measure of linear correlation between fat, lean and their ratio respect to age, are reported in Table 4.4. Pearson's r varies between -1 and +1 with 0 implying no correlation. Positive correlations imply that as one variable increases, so does the other variable. Negative correlations, in contrast, imply that as one variable increases, the other decreases.

Interestingly, whole body lean tissue mass decreasing and the ratio of fat to lean mass (see Figure 4.1) increasing with age in males whereas in females these phenomenons are weaker.

Despite these correlations between variables DXA and age be very interesting to determine how the body composition changes with ageing, if you want to analyse correlation with other age or gender dependent factors, the collinearity of this factors and body composition variables with age and gender, can affect the

Table 4.4: Correlation coefficients between fat, lean and fat-lean ratio and age for males and females. The p-value indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation as the one observed.

	Male		Female	
	r	p	r	p
Fat Tissue	0.40	$\ll 0.01$	0.22	0.04
Lean Tissue	-0.40	$\ll 0.01$	-0.20	0.06
Fat Lean Ratio	0.40	$\ll 0.01$	0.20	0.06

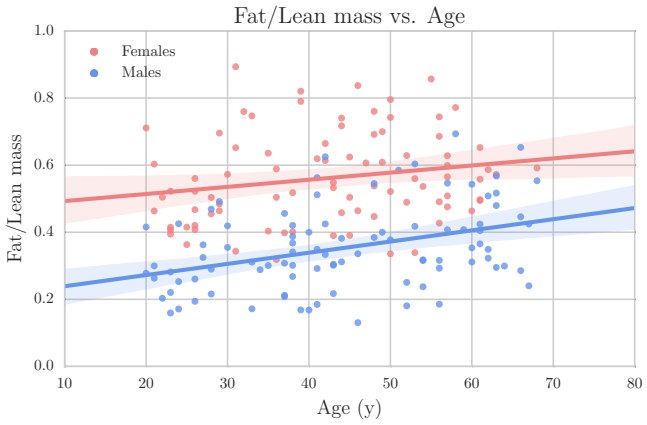

Figure 4.1: Age and ratio of fat to lean soft-tissue mass in males and females. Fat to lean soft-tissue mass increases with age more in males than in females. The shaded indicates the 95% of confidence interval for the regression line.

results. Moreover, due to the strong link between age/gender and BC variables, the behaviour of the data respect to other variables can be difficult to see.

Age and gender of the subject are confounding factors which must be taken into account in order to get correct results. To correct the effect of these confounding factors we need to stratify the database accordingly. Stratification, often used to control background characteristics, is the process of dividing members of the population into homogeneous, mutually exclusive and exhaustive subgroups.

Another method to overcome this problem is to perform linear simple regression between variables and to take the residues. This method allows us to not split the database, keeping a larger sample size.

For each variables the mean for males and females is subtracted and Kolmogorov–Smirnov test is performed to test whether this two samples are drawn from the same distribution. The equivalence of the distributions after the correction is a requirement for this procedure.

Table 4.5 show the K-S statistic and the p-value. The high values of p-value suggest that the distributions of variables for males and females are the same, less than a constant factor. This feature can be clearly seen also in FigureA.1 in Appendix A.

Table 4.5: Kolmogorov–Smirnov staisctics (K-S stat) and p-values for the two-sided K-S test for H_0 that the variables for males and females are drawn from distributions similar enough for this dataset size.

	Fat mass		Lean mass	
	KS-stat	p-value	KS-stat	p-value
Whole body	0.104	0.700	0.104	0.700
Upper limbs	0.069	0.980	0.179	0.103
Lower limbs	0.073	0.967	0.066	0.987
Trunk	0.085	0.891	0.093	0.817
Android	0.084	0.902	0.118	0.543
Gynoid	0.152	0.236	0.179	0.106

We performed a linear regression between DXA variables with age and gender, without interactions between this two regressors. For the following discussion we will use the resulting residuals of this regression.

Residual distribution are rather normally distributed (see Figure A.1 in Appendix A.), this allows us to ignore the differences in age and gender in the database.

In Table 4.6 are show the the third central moment, called skewness (that

represent the lopsidedness of the distribution) and the fourth central moment, called Kurtosis (that is a measure of whether the distribution is tall and skinny or short and squat, compared to the normal distribution).

Table 4.6: Skewness and Kurtosis of resides of considered variables.

	Fat mass		Lean mass	
	Skewness	Kurtosis	Skewness	Kurtosis
Whole body	0.11	−0.48	−0.10	−0.53
Upper limbs	−0.01	0.25	−0.05	0.50
Lower limbs	0.02	−0.19	0.01	−0.20
Trunk	0.23	−0.48	−0.02	−0.46
Android	0.48	−0.14	0.22	1.04
Gynoid	0.10	0.22	0.19	0.00

Negative values of skewness indicate that the "mass of the distribution" is concentrated on lower values, the distribution is said to be left-tailed, positives values indicate a right-tailed distribution.

The exact interpretation of kurtosis is is not so clear. For distributions with skewness close to zero, the classical interpretation is that kurtosis measures both the "peakedness" of the distribution and the heaviness of its tail. Positive kurtosis denotes a distribution with more acute peak around the mean and fatter tails than Normal distribution (e.g. Student's t-distribution, logistic distribution, etc). A distribution with negative kurtosis has a lower, wider peak around the mean and thinner tails. As reference the kurtosis value for logistic distribution (with unit variance) is 1.2, and for uniform distribution is -1.2.

Values in Table 4.6 suggest us that the resulting distributions, for considered variables, after preprocessing operations are rather normally distributed.

4.1.2 Principal Components Analysis

Figure 4.3 show the coefficients of first three PCs, that explain more than 80% (see Figure 4.2) of the total variance of data. The coefficients were ordered for easier viewing and interpretation.

It can be seen that the first PC take in account masses of lean versus fat soft tissues, the second seem to consider the upper (trunk and android regions) in contrast to lower (gynoid and lower limbs regions) fat content and the third component is related to composition of the appendicular regions(upper and lower limbs) versus composition of the central body regions (trunk, android and gynoid).

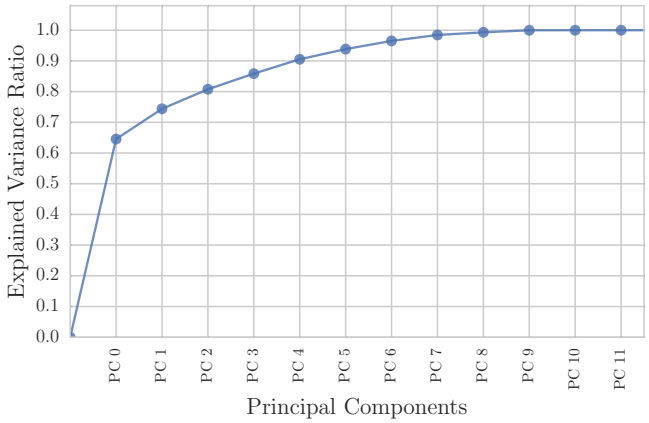

Figure 4.2: The plot shows that most of the variance (64.5% of the variance) can be explained by the first principal component. The second and the third principal components still bears some information (9.9% and 6.4% respectively). Together, the first three principal components contain 80.8% of the information.

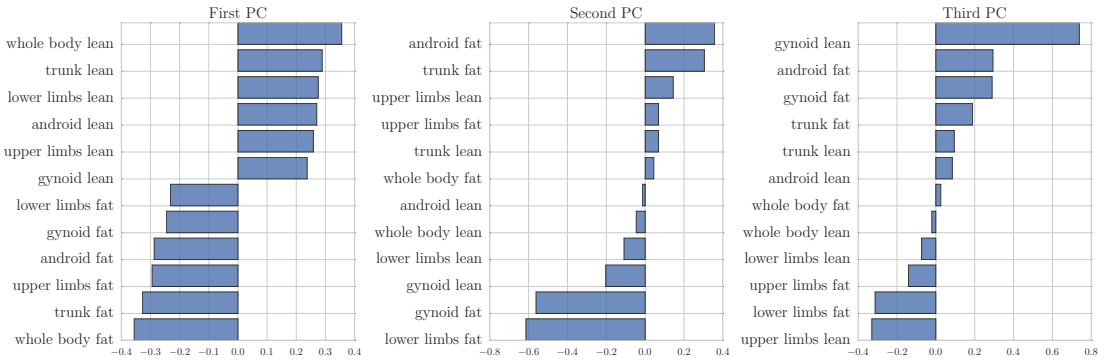

Figure 4.3: First three PCs of healthy people body composition database.

Interpretation of other PCs are not clear, and are not reported here. It should be emphasized that the interpretation of the PC is an approximation of the real meaning, which is often more subtle than is realized. These components will be included in the following analysis, even if their interpretation is not as direct as for the other ones.

4.1.3 Body composition and cholesterol

We will now study the relationship between the body composition described with the DXA PCs and the blood lipid concentration. The value of total cholesterol, triglycerides (TG), and HDL cholesterol (high-density lipoprotein cholesterol, also called "good" cholesterol) are available in most subject, see Table 4.7.

Table 4.7: Aviable data for lipid concentrations.

	Total	Males	Females
Tot CH	172	89	83
TG CH	170	87	83
HDL CH	98	46	52

In figure4.4 are show the distributions of these variable.

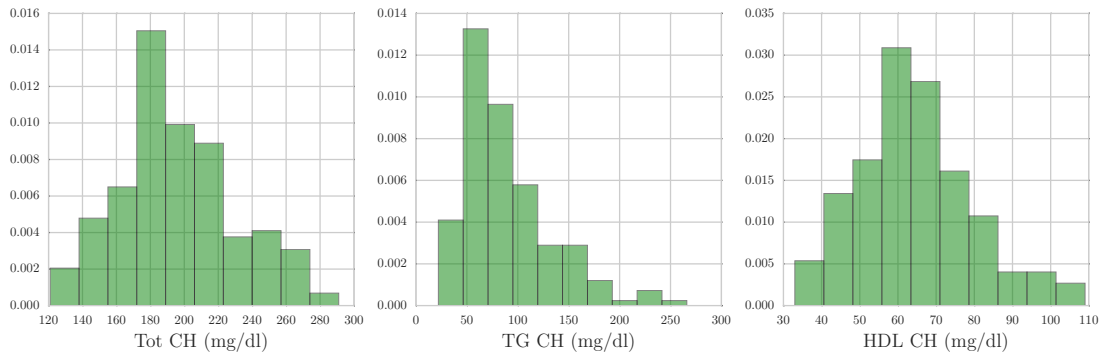

Figure 4.4: Distribution of of total cholesterol (Tot CH), triglycerides (TG CH), and HDL cholesterol (HDL CH)for healthy database. As you can see

Multiple linear regression models are fitted to investigate the relationship between body composition and blood lipid concentration variables. The predictors variables are the PC, computed by PCA. Also, more sophisticate methods are performed: support vector regression to evaluate the correlation using non linear transformation of data and K-nearest neighbours to evaluate the local behaviour of data, in particular KNN strongly depends on the number of neighbours considered,

then we have applied this method twice, one considering 5 nearest neighbours, the other considering 10. Table 4.8 report the results.

Table 4.8: Coefficient of determination R^2 of linear regression model (LRM), Support Vector Regression (SVR), and K-Nearest Neighbours Regression with 5 (5NNR) and 10 (10NNR) neighbours between the principals components and the blood lipid concentrations.

Variables	LRM	SVR	5NNR	10NNR
Tot CH	0.121	0.092	0.253	0.127
TG CH	0.151	0.091	0.288	0.100
HDL CH	0.196	0.118	0.251	0.189

As you can see linear model gives the best results, despite their coefficients of determination does not exceed 0.2. Nonlinear methods (SVM) returns worse results than linear method and this suggests us that there aren't other kind of relationship between the body composition variables and lipids blood concentration. Moreover increasing the parameter k of the NNR method the performance declines, then the high coefficients of determination obtained considering 5 nearest neighbours are due only to the local scale of this method.

Tables 4.9 and 4.10 reports the results of two type of cross validation. Cross validation technique tell us the predictive power of fitted models. The negative values show in Table 4.9 are due only to the software algorithms and not have a physical meaning. These results suggest as that lipids blood concentration are not correlated with the body composition variables (as reported by some articles in the literature [28]).

Table 4.9: Result of k-fold (k=10) cross validation of linear regression model (LRM), Support Vector Regression (SVR), and K-Nearest Neighbours Regression with 5 (5NNR) and 10 (10NNR) neighbours between the principals components and the blood lipid concentrations.

Variables	LRM	SVR	5NNR	10NNR
Tot CH	−0.38	−0.03	−0.24	−0.16
TG CH	−0.17	−0.04	−0.15	−0.14
HDL CH	−0.84	−0.17	−0.19	−0.04

4.2 NUAGE database

To study the effects of the diet on health and ageing factors, seniors across Europe has been recruited, body composition and metabonomics data has been

Table 4.10: Result of "leave-one-out" cross validation of linear regression model (LRM), Support Vector Regression (SVR), and K-Nearest Neighbours Regression with 5 (5NNR) and 10 (10NNR) neighbours between the principals components and the blood lipid concentrations.

Variables	LRM	SVR	5NNR	10NNR
Tot CH	0.00	0.00	0.01	0.00
TG CH	0.00	0.00	0.01	0.01
HDL CH	0.00	0.00	0.00	0.00

collected.

Metabonomics, means the quantitative measurement of the dynamic metabolic response of living systems to stimuli, has become an increasingly shared practices in research. The two most common used techniques are mass spectrometry and *Nuclear Magnetic Resonance Spectroscopy* (MRS) either in isolation or in conjunction [29].

The body composition of subjects is assessed by whole body DXA scan, fat and lean tissues distribution of main corporeal districts are available.

Cohort includes 1115 elderly people, ranging from 65 to 79 years (including an appropriate number of men and women, see Table 4.11), free of major overt diseases, will be recruited by 5 European centres with a great experience in conducting dietary intervention studies and research on the elderly, located in UK (Norwich), the Netherlands (Wageningen), Poland (Warsaw), France (Clermont-Ferrand) and Italy (Bologna). These centres have been strategically selected to represent different geographical areas covering as a whole a NEWS approach (Northern, Eastern, Western and Southern Europe).

Table 4.11: Database composition.

Centres	Males	Females
Italy	127	133
United Kingdom	68	118
Netherlands	103	131
Poland	103	140
France	95	97

4.2.1 Body composition variables

As we already saw in Section 4.1, before using body composition variables some operations are necessary to take into account age and gender of subjects,

that can introduce errors when correlations models are performed. Therefore the variables are normalized on subject's total weight, detrendized respect to age and standardized.

All operations are made separately for each centre. Age range of NU-AGE database is rather small (less then 15 years), and consequently the estimated coefficients can be unstable.

In general data from different scanners are not directly comparable and cross-calibration operations are necessary to compare data among different DXA scanners (as we said in first chapters). Then we compare the coefficients of linear regression between DXA data and age of subjects, performed on the Italian subset of NU-AGE database with with coefficients obtained in previous sections, because these data are made by the same scanner. The values of coefficients are reported in Table A.1 and plotted in Figure A.2 in Appendix A.

As you can see all values lie within a band around the mean with a width of two standard deviations.

Then we suppose that the bias error, due to use of different scanner and to the lack of cross-calibration procedure, is systematic and constant and apply the de-trending coefficients computed on Healthy database to the entire NU-AGE database, conscious that this is a rough approximation.

After these operations the PCA was done on the covariance matrix and first three PC resulting from each centre are compared.

The first three PC of PCA computed on healthy database (see Section 4.1) are reported as reference values.

Before we attempt to interpret the PCs, some explanation is necessary. When we interpret PCs it is usually only the general pattern of the coefficients that is really of interest, not values of coefficients, which may give a false impression of precision.

As you can see in Table 4.12, the PCs are rather interpretable. Comparing coefficients of first PC it is seen that the coefficients do not vary significantly, and represent fat versus lean soft tissues mass.

Second PC seem to consider the lower (gynoid and lower limbs regions) in contrast to upper (trunk and android regions) fat content.

The third component is more variable depending on the considered subset but still to be related to composition of the appendicular regions (upper and lower limbs) versus composition of the central body regions (trunk, android and gynoid).

The other components, although they are not so easily interpretable, are still

Table 4.12: Simplified version of the coefficients of first three PCs, + and − indicate the sign of most relevant coefficients of PC. H indicates values of coefficients found for healthy database, reported as reference values.

First PC	H	IT	NL	PL	FR	Second PC	H	IT	NL	PL	FR	Third PC	H	IT	NL	PL	FR
T. body fat	−	−	−	−	−	Legs fat	−	−	−	−	−	Arms lean	−	−	−	−	−
Trunk fat	−	−	−	−	−	Gynoid fat	−	−	−	−	−	Legs fat	−	0	0	0	+
Arms fat	−	−	−	−	−	Gynoid lean	−	0	0	0	+	Arms fat	−	−	−	−	−
Android fat	−	−	−	−	−	Legs lean	0	−	−	−	−	Legs lean	0	0	0	0	0
Gynoid fat	−	−	−	−	−	T. body lean	0	0	0	0	0	T. body lean	0	0	0	0	0
Legs fat	−	−	−	−	−	Android lean	0	−	0	0	0	T. body fat	0	0	0	0	0
Gynoid lean	+	+	+	+	+	T. body fat	0	0	0	0	0	Android lean	0	+	+	+	+
Arms lean	+	+	+	+	+	Trunk lean	0	0	0	0	+	Trunk lean	0	+	+	0	+
Android lean	+	+	+	+	+	Arms fat	0	0	0	0	−	Trunk fat	+	+	+	+	0
Legs lean	+	+	+	+	+	Arms lean	+	+	0	+	−	Gynoid fat	+	0	0	+	+
Trunk lean	+	+	+	+	+	Trunk fat	+	+	+	+	+	Android fat	+	+	+	+	0
T. body lean	+	+	+	+	+	Android fat	+	+	+	+	+	Gynoid lean	+	0	+	0	+

quite similar between centres. We can say that if we consider all patients as coming from a single center, we make a negligible error, but on the other hand allows us to obtain not split the database, keeping a larger sample size, thus increasing power and robustness of our analysis.

Since android and gynoid ROIs are missing for data of patients came from UK centre, their PC are not reported in Table 4.12. In later discussion data from UK centre will be consider, and information of android and gynoid ROIS of all other centre will be ignored. This approach is chosen because, as can be see in Figure , gynoid and android region overlap the lower limbs and trunk region respectively, then the information derived from these regions are rather redundant and can be safety ignored, moreover the interpretation of new PC, obtained from PCA performed, is not affected by this operation.

4.2.2 Metabolites

A blind analysis design is chosen for the metabolites analysis. A blind analysis is an analysis in which the final result, and the individual data on which it is based, are kept hidden from the analyst until the analysis is essentially complete. The principal motivation is to avoid experimenter's (concious or subconscious) bias such as looking for bugs, additional sources of uncertainty or to drop an event when a result does not conform to expectation [30].

Un-targeted metabolic profiles of patients, performed by Nestlé Institute of Health Sciences, are obtained by MRS on urine samples, using NOESY pulse sequence to get measurements of metabolites relative concentration, Figure 4.5.

In un-targeted analyses, the chemical identification of metabolites, that usually follows data acquisition and requires libraries for labels annotation [31], is not

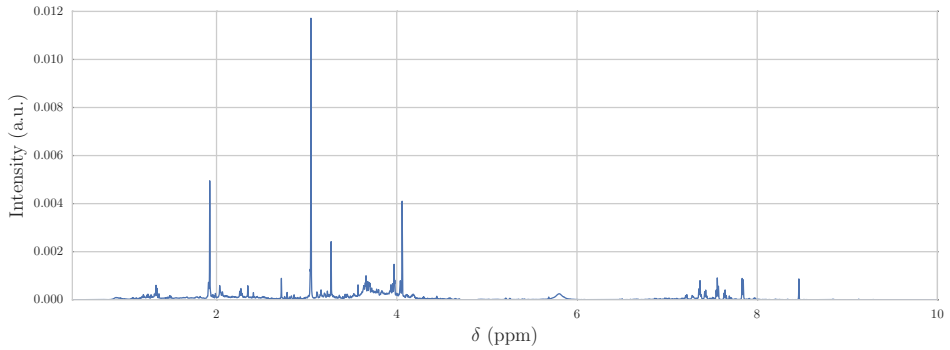

Figure 4.5: Spectrum obtained by MRS. To each peak (pattern of peaks), correspond a metabolite (small ensemble of metabolites).

available.

The NOESY is the most utilized NMR (*Nuclear Magnetic Resonance*) pulse sequence for the collection of metabolomics MRS data from biological samples, such as blood plasma, serum, urine or homogenized tissue extracts, due to the capacity to suppresses solvent signals [32].

We can consider that, in a simplistic way, to each peak, or at least a pattern of peaks, correspond a metabolite or a small ensemble of metabolites. Ideally peak have delta shape, but this condition is never satisfied and the width of peaks can be estimated through autocorrelation. Autocorrelation is the cross-correlation of a signal with itself at different points in time. More generally, it is the similarity between observations as a function of the position *lag* between them. If the peaks are delta function the autocorrelation is zero.

Visualizing the autocorrelation of a spectrum (Figure 4.6) and taking the full width at half maximum ($FWHM$) of the central peak of autocorrelation plot we have an approximatively measure of the mean width of peaks.

Autocorrelation is computed for all spectrum and mean value of $FWHM$, named $\overline{FWHM}$, is taken as approximate width of peaks.

Smoothing each spectra with a Gaussian filter with $\sigma = \overline{FWHM} = 10$ we ensure that: high frequencies noise is reduced and peaks are mostly preserved. These spectra will be analyzed using the NMF decomposition, and this smoothing will also improve the quality of the resulting components as it increase the correlation between neighbours values. NMF would otherwise use each one of the values in the spectra as a separate variable, without considering the spatial correlation between them.

Before performinf the final NMF an adequate number of factors must be found.

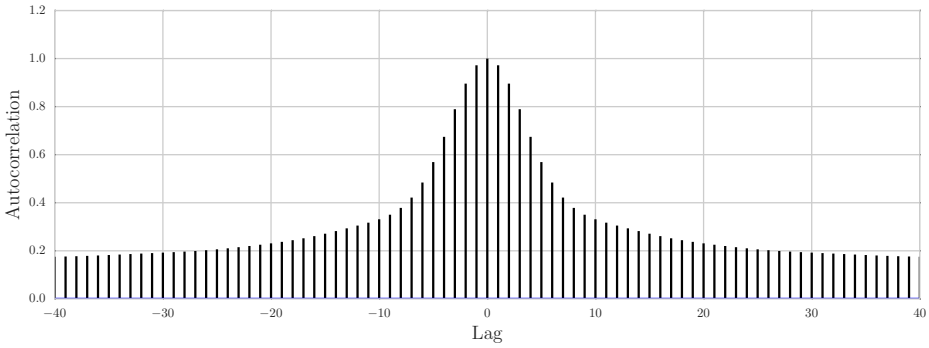

Figure 4.6: Autocorrelation plot of one spectrum. The $FWHM$ of indicates mean width of peaks. Autocorrelations should be near-zero for delta peaked spectrum.

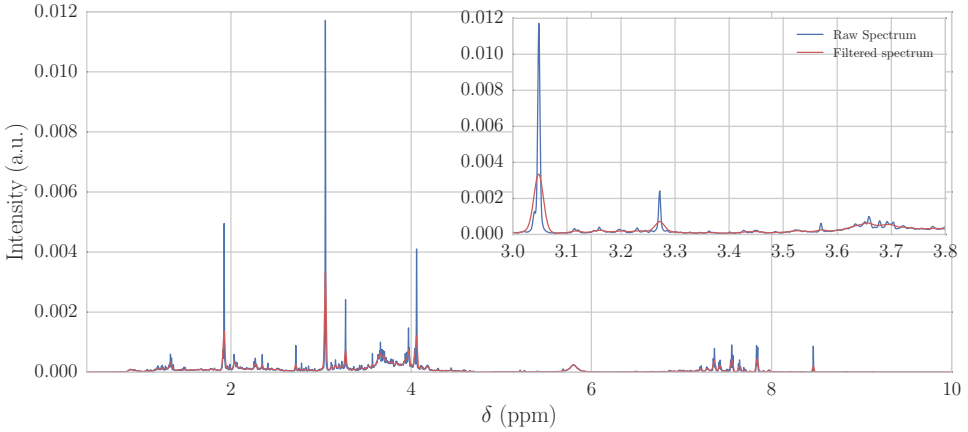

Figure 4.7: Comparison of original and filtered spectrum. As you can see apply a Gaussian filter involves in a reduction of high frequencies noise.

We have run the NMF algorithm several times varying the number of factors and measuring the reconstruction error as goodness criterion.

The reconstruction error is computed as the Frobenius norm of the matrix difference between the training data and the reconstructed data from the fit produced by the model. In Figure4.8 we show the the reconstruction error, normalized on total value of training data, as function of number of components, for each country.

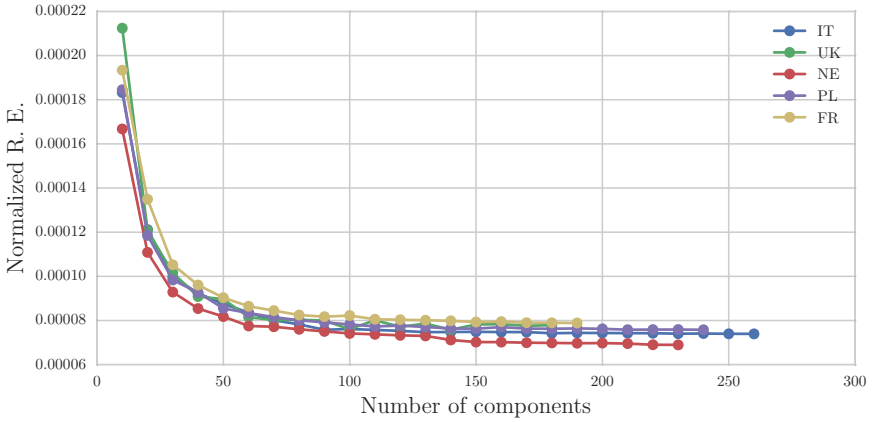

Figure 4.8: Normalized reconstruction error as function of number of factors, for each subset of database.

A number of components is set to 100, as a good compromise between reconstruction error and computational time.

NMF is performed separately for each centre, in Figure 4.9 are show the three first factors for each centre. As we can see the first factors is substantially the same for all centre, while the second factor of french people is the third for the polish people, and vice-versa. Then although we do not know the correspondence between the pattern of peaks and metabolites, we can assume that the second factor found for french patients, has the same meaning of the third factor for the polish patients.

Cross correlation between first few factor computed for different centre is performed (Table A.2) in Appendix A. As we can see, due to the high value of correlation, we can say that the first few factors are the same for all centres, with just a change in ranking.

Only the few first factors are clearly interpretable as linear combinations of patterns of peaks, then despite the behaviour show in TableA.2 in Appendix A is kept from later factors, the values of correlation are lower.

Then we can assume that considering all patients as coming from a single country, as we have do for body composition variables, we make a negligible error.

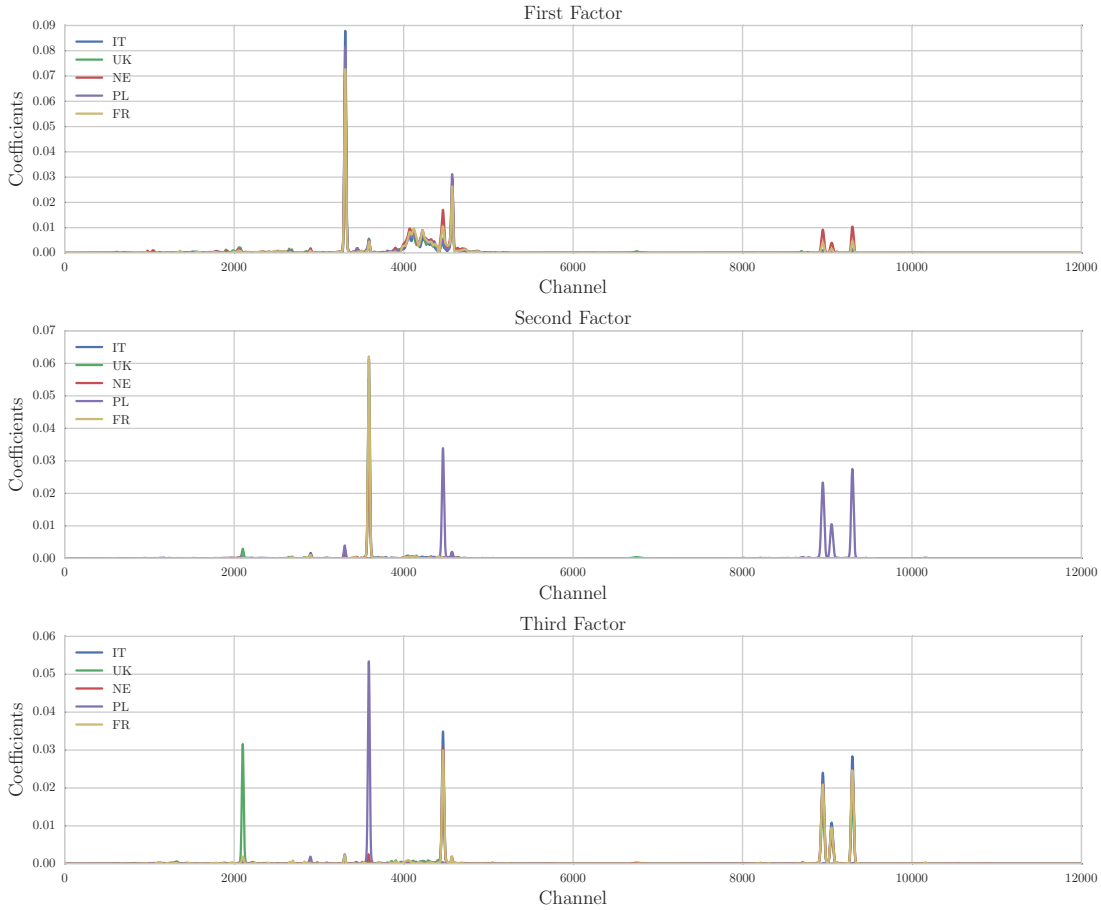

Figure 4.9: Comparison of first three factors for different countries. First factor is very similar for all subset of patient. second and third factors of french and polish subset seem the same but with inverted position.

A new NMF is performed neglecting the patients provenance, in this way the the factors have the same meaning for all centres, and the later result are easier interpretable.

4.2.3 Body composition and metabonomics

Multiple linear regression models are fitted to identify the relationship between metabonomic data and body composition variables. The predictors variables are the PC, computed by PCA, and the response variables are the factors founded by NMF.

When one considers a set of statistical inferences simultaneously, (we fit one hundred models for each centre), multiple testing problem may can occurs, also known as the *look-elsewhere effect*.

The p-value of a statistical test indicate the probability that a given result could be obtained, assuming random coincidence. If this p-value is less than some predetermined statistical significance threshold α, one considers the result "significant".

However, if one performs multiple tests ("looking elsewhere" if the first test fails) then obviously a p-value of $1/n$ is likely to occur after n tests [33].

For example, an event with $p < 0.05$ will probably be seen after 20 tests, even if there is no effect. In order to compensate for this, several correction methods of p-values exist. In our analysis the Benjamini-Hochberg adjustments is used.

Suppose that we have $H_1 \dots H_m$ null hypotheses tested and their corresponding p-values $P_1 \dots P_m$. We order these p-values in increasing order and denote them by $P_{(1)} \dots P_{(m)}$ (called a step-up procedure). Then for a given α, the Benjamini-Hochberg procedure returns $\frac{k}{m}\alpha$.

Table 4.13 show the classical and the adjusted coefficient of determination(R^2 and R^2_{adj} respectively) and the adjusted p-value q for the most common significant model of all countries.

In these regressions the principal components with *individual* $p-value$ greater than 0.05 are sequentially removed and only the significant PC are kept [34].

Individuals $p-value$ of a predictor is the probability of obtaining a result equal to or more extreme than that observed under the supposed true null hypothesis is $H_0 : \beta = 0$, where β is the coefficient of the predictor.

Detail of these models are show in Appendix A.

As you can see, despite R^2_{adj} values are rather small (not exceed 0.25), the results are very significant (the adjusted p-value q are very small) and then not

Table 4.13: Classical (R^2) and adjusted coefficient of determination (R^2_{adj}) and the adjusted p-value q of multiple regression between factors of NMF of metabonomics data and PC of PCA of body composition variables.

	IT			UK			NE			PL			FR		
Factors	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q
FAC9	0.13	0.11	< 0.01	0.16	0.13	< 0.01	0.16	0.14	< 0.01	0.23	0.22	< 0.01	0.23	0.20	< 0.01
FAC14	0.14	0.12	< 0.01	0.05	0.02	0.09	0.12	0.09	< 0.01	0.12	0.11	< 0.01	0.16	0.12	< 0.01
FAC16	0.08	0.06	< 0.01	0.08	0.06	0.01	0.09	0.06	< 0.01	0.17	0.15	< 0.01	0.13	0.10	< 0.01
FAC23	0.08	0.05	< 0.01	0.13	0.10	< 0.01	0.07	0.04	0.02	0.08	0.05	< 0.01	0.05	0.02	0.12
FAC27	0.13	0.10	< 0.01	0.18	0.15	< 0.01	0.12	0.10	< 0.01	0.06	0.04	< 0.01	0.06	0.03	0.10
FAC31	0.09	0.06	< 0.01	0.13	0.10	< 0.01	0.08	0.06	< 0.01	0.12	0.10	< 0.01	0.09	0.07	< 0.01
FAC50	0.11	0.09	< 0.01	0.14	0.11	< 0.01	0.15	0.13	< 0.01	0.24	0.21	< 0.01	0.12	0.08	< 0.01
FAC66	0.11	0.09	< 0.01	0.10	0.07	< 0.01	0.14	0.12	< 0.01	0.17	0.15	< 0.01	0.14	0.10	< 0.01

negligible. Moreover the figure from A.3 to A.10 in Appendix A show that the factors are easily interpretable as single metabolites or at least pattern of a small ensemble of metabolites.

Then we have perform new multiple linear regression models, this time to identify the relationship between body composition and metabonomic data, where the response variables are the PC from the PCA and the predictors variables are the factors founded by NMF.

The purpose of this new analysis is to highlight the possibility to link the fat distribution of a subject with his metabolic pattern. We evaluated the performance of two different approaches: in first approach only the factors that we have previously selected are considered, then all factors are considered. The result of these approach are show in Table 4.14 and Table 4.15 respectively.

Table 4.14: Classical (R^2), adjusted coefficient of determination (R^2_{adj}) and the p of multiple regression betweenPC of PCA of body composition variables and factors of NMF selected previously.

	IT			UK			NE			PL			FR		
	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p
PC0	0.08	0.05	0.01	0.10	0.06	0.02	0.10	0.06	0.00	0.12	0.09	0.00	0.06	0.02	0.17
PC1	0.05	0.02	0.11	0.12	0.08	0.00	0.10	0.07	0.00	0.03	0.00	0.47	0.12	0.08	0.00
PC2	0.08	0.05	0.01	0.05	0.01	0.25	0.03	0.00	0.53	0.04	0.00	0.37	0.04	0.00	0.44
PC3	0.13	0.11	0.00	0.06	0.02	0.20	0.05	0.02	0.17	0.10	0.07	0.00	0.06	0.02	0.19
PC4	0.06	0.03	0.04	0.05	0.00	0.38	0.10	0.07	0.00	0.15	0.12	0.00	0.05	0.01	0.27
PC5	0.11	0.08	0.00	0.08	0.04	0.07	0.06	0.03	0.08	0.12	0.09	0.00	0.07	0.02	0.12
PC6	0.13	0.11	0.00	0.09	0.05	0.03	0.03	0.00	0.44	0.20	0.17	0.00	0.07	0.03	0.08
PC7	0.01	−0.02	0.89	0.07	0.03	0.09	0.19	0.16	0.00	0.21	0.19	0.00	0.07	0.03	0.10

As you can see immediately from the comparison of Tables 4.14 and 4.15, considering all factors we obtain values of coefficient of determination R^2 very high and are still rather high even considering the corrected value R^2_{adj}.

The result of linear regression between the PC, as response variables, and the small sub-set of factors (as show in Tables 4.14) are very low due to raw cut of

Table 4.15: Classical (R^2), adjusted coefficient of determination (R^2_{adj}) and the p of multiple regression betweenPC of PCA of body composition variables and all factors of NMF.

	IT			UK			NE			PL			FR		
	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p
PC0	0.43	0.07	0.16	0.68	0.32	0.00	0.52	0.17	0.02	0.54	0.23	0.00	0.62	0.21	0.02
PC1	0.48	0.15	0.01	0.57	0.06	0.28	0.57	0.24	0.00	0.40	−0.02	0.61	0.64	0.26	0.01
PC2	0.39	0.01	0.42	0.55	0.03	0.41	0.45	0.04	0.30	0.48	0.11	0.08	0.51	−0.02	0.57
PC3	0.48	0.15	0.02	0.60	0.12	0.13	0.47	0.08	0.16	0.51	0.18	0.01	0.46	−0.13	0.88
PC4	0.46	0.12	0.05	0.60	0.14	0.11	0.58	0.27	0.00	0.47	0.11	0.08	0.56	0.09	0.20
PC5	0.53	0.24	0.00	0.57	0.07	0.27	0.46	0.05	0.25	0.49	0.13	0.05	0.54	0.03	0.39
PC6	0.52	0.22	0.00	0.56	0.04	0.36	0.40	−0.05	0.75	0.52	0.19	0.01	0.55	0.06	0.28
PC7	0.24	−0.23	1.00	0.71	0.36	0.00	0.62	0.34	0.00	0.53	0.21	0.00	0.58	0.13	0.11

information when only eight factors are considered instead of one hundred (in Figure 4.8 you can be see that).

As a further result the same operation do until now are repeated with unique modification that a logarithmic transformation is done on body composition variable data. In this way the PC of PCA (that retain their interpretation) can be view as multiplicative model of the primary body composition variables (instead of additive representation of the "classical" PC).

Logarithmic transformation is common in treatment of problems related to biology. Indeed many variables in biology have log-normal distributions, meaning that after log-transformation, the values are normally distributed. This is because if you take a bunch of independent factors and multiply them together, the resulting product tends to a log-normal distribution.

Table 4.16: Classical (R^2) and adjusted coefficient of determination (R^2_{adj}) and the adjusted p-value q of multiple regression between factors of NMF of metabonomics data and PC of PCA of logarithmic transformed body composition variables.

	IT			UK			NE			PL			FR		
Factors	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q	R^2	R^2_{adj}	q
FAC9	0.17	0.14	< 0.01	0.15	0.11	0.01	0.18	0.15	< 0.01	0.23	0.21	< 0.01	0.22	0.19	< 0.01
FAC14	0.22	0.20	< 0.01	0.09	0.05	0.14	0.13	0.09	< 0.01	0.15	0.12	< 0.01	0.16	0.13	< 0.01
FAC16	0.10	0.07	0.01	0.13	0.09	0.03	0.10	0.07	0.02	0.18	0.16	< 0.01	0.16	0.12	< 0.01
FAC23	0.09	0.06	0.03	0.14	0.10	0.01	0.07	0.04	0.11	0.07	0.04	0.14	0.03	−0.02	0.89
FAC27	0.12	0.09	< 0.01	0.17	0.13	< 0.01	0.11	0.08	0.01	0.09	0.06	0.03	0.06	0.02	0.41
FAC31	0.10	0.07	0.01	0.14	0.10	0.01	0.09	0.06	0.03	0.10	0.07	0.02	0.08	0.04	0.20
FAC50	0.17	0.14	< 0.01	0.17	0.13	< 0.01	0.15	0.12	< 0.01	0.26	0.23	< 0.01	0.12	0.08	0.03
FAC66	0.13	0.10	< 0.01	0.10	0.06	0.08	0.17	0.14	< 0.01	0.15	0.12	< 0.01	0.17	0.13	< 0.01

As you can see by comparing the Tables 4.13 and 4.16, there are no particular improvements considering the log transformation of body composition variables in multiple regression between metabolic factors and PC, the results remain essen-

Table 4.17: Classical (R^2), adjusted coefficient of determination (R^2_{adj}) and the p of multiple regression between PC of PCA on logarithmic transformation of body composition variables and all factors of NMF.

	IT			UK			NE			PL			FR		
	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p	R^2	R^2_{adj}	p
PC0	0.50	0.18	< 0.01	0.74	0.44	< 0.01	0.57	0.24	< 0.01	0.67	0.43	< 0.01	0.70	0.37	< 0.01
PC1	0.69	0.50	< 0.01	0.70	0.34	< 0.01	0.75	0.57	< 0.01	0.71	0.50	< 0.01	0.72	0.41	< 0.01
PC2	0.49	0.18	0.01	0.55	0.02	0.43	0.46	0.06	0.22	0.42	0.02	0.40	0.58	0.13	0.11
PC3	0.47	0.13	0.03	0.53	-0.01	0.52	0.49	0.11	0.08	0.48	0.12	0.05	0.55	0.07	0.27
PC4	0.39	< 0.01	0.48	0.57	0.07	0.27	0.51	0.14	0.04	0.48	0.12	0.06	0.53	0.02	0.42
PC5	0.41	0.04	0.29	0.56	0.06	0.31	0.47	0.08	0.17	0.45	0.07	0.19	0.55	0.06	0.29
PC6	0.48	0.15	0.02	0.61	0.16	0.08	0.58	0.26	< 0.01	0.49	0.14	0.03	0.63	0.24	0.01
PC7	0.38	-0.01	0.54	0.59	0.12	0.15	0.46	0.06	0.22	0.38	-0.06	0.79	0.55	0.05	0.31

tially the same.

On the contrary there is a substantial increase of the values of the coefficients of determination R^2 (and R^2_{adj}) of multiple regression between PC of log transformation of body composition variables and all factors, respect to the same correlation without the log transformation.

We note that the first two PCs have correlations always high, reaching $R^2_{adj} = 0.44$ for the first and $R^2_{adj} = 0.57$ for the second.

Chapter 5

Conclusion

The increasing use of DXA for the study of body composition, especially in the analysis of fat distribution, opens new horizons in medical and scientific research.

In lasts decades scientific research have highlighted that proper nutrition and good physical fitness are key factors for health. Moreover the increase in the average age of the population raises a critical importance of identifying strategies able contrast the age-related diseases. These problems require high reliable measurements, both for the nutrient input, the metabolic and the physiological state of subjects. Body fat distribution is a relevant measurement in this kind of studies, especially since the fat is starting to be seen as a single functional organ instead of just a storage system.

DXA is one of the most simple and cheap techniques to study the fat distribution in a systematic and quantitative way. In this thesis work we have studied the correlation of the body composition, in term of fat and lean content, with the metabolic state, that is a direct consequence of diet, of elderly subjects living in European region.

First we have establish the proper method and preprocessing operation over a group of healthy and normal-weight subjects. Through the principal components analysis we have show emergent properties of body composition variable from DXA.

The leading PC take in account masses of lean versus fat soft tissues, the second seem to consider the upper in contrast to lower fat content and the third component is related to composition of the appendicular regions versus composition of the central body regions; together, these components contain more than 80% of the information about the body composition of the subjects.

The interpretation of these three components is really clear, indeed the first

PC can be interpreted as a version of BMI and the second as a more accurately definition of the *Waist to Hip Ratio* (WHR). Also the third PC can be view as good descriptive index of the body composition of a subject. Moreover these components are surprisingly stable across different subject status, age, gender and nationality.

In the last part of the analysis correlation between body composition variables and metabolic state of the subject are evaluated. The cohort is recruited strategically to represent different geographical areas covering as a whole a NEWS approach (Northern, Eastern, Western and Southern Europe).

A blind analysis design is chosen for the metabolites analysis. Metabolic profiles of patients, performed by Nestlé Institute of Health Sciences, obtained by *Nuclear Magnetic Resonance Spectroscopy* (MRS) on urine samples, are considered to get measurements of metabolites concentration.

Also in this database a PCA approach is used to process data from DXA, and the same interpretation of the component is found.

A factor analysis technique is used to processing metabolic data. Using the *Non-negative Matrix Factorization* (NMF) algorithm we were able express the original spectra (from MRS) as a combination of basis factors, that can be understood as single metabolite.

The factors found using this technique are "stable", meaning that although subjects from different countries are considered, their metabolic spectra are composed by the same factors (i.e. the same pattern of metabolites).

Two experimental design are considered: one consider the single factors as a function of PC, other consider the single PC as a function of an ensemble of factors.

In the first design multiple linear regression models are fitted to identify the relationship between metabonomic data and body composition variables. The predictors variables are the PC, computed by PCA on DXA data, and the response variables are the factors founded by NMF on MRS data.

Significant but low values of correlations are found when trying to describe the metabolic state trough the body composition variables.

Then we have perform new multiple linear regression models, this time the response variables are the PC from the PCA and the predictors variables are once the factors, founded previously, that are more correlated with the PC (8 factors are selected), the other time all factors (100) returned from NMF.

The result of linear regression between the PC, as response variables, and the

small sub-set of factors are very low due to loss of information when only eight factors are considered instead of one hundred.

On the other hand the high values obtained with the second design suggest us that there is the concrete possibility to derive the "body shape" from the metabolic state. In particular high value of correlation ($R^2_{adj} \simeq 0.50$) are obtained between first two PC and the metabolic state.

Appendix A

Graphs and Images

This appendix lists the graphs and figures useful to understand previous chapters but that for reasons of convenience of viewing and reading are reported out of the text.

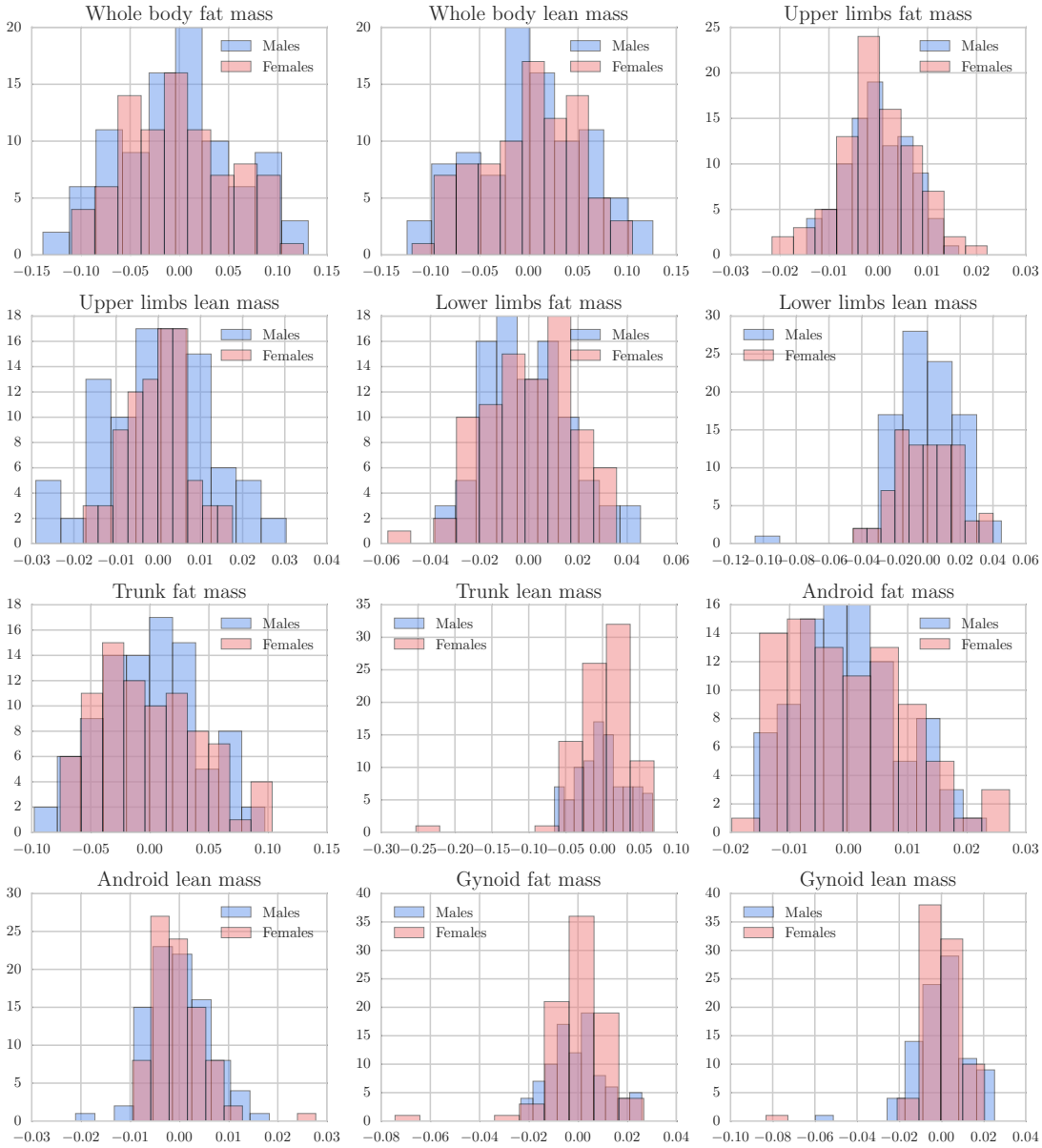

Figure A.1: Distributions of resides of variables for Healthy database after detrending operation.

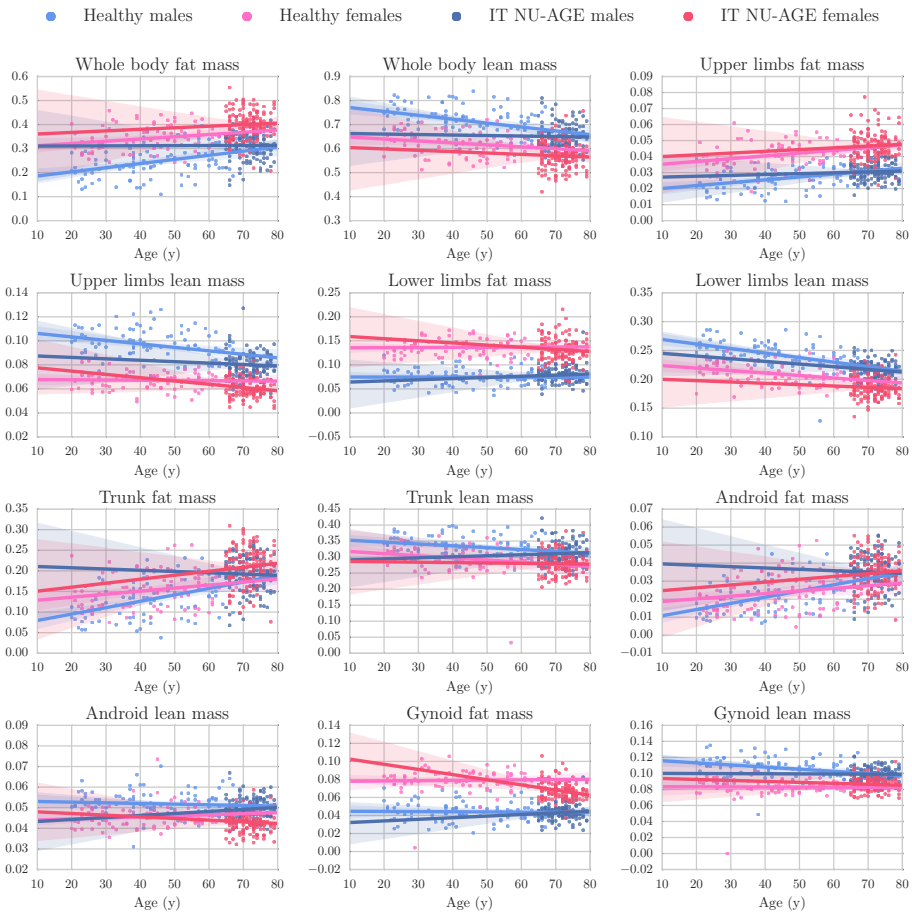

Figure A.2: Regression plot of DXA variables for Italian subjects of NU-AGE database and Healthy database. The shaded indicates the 95% of confidence interval for the regression line.

Table A.1: Coefficients of linear regression between DXA variables and age of subjects separately for males and females. In the NU-AGE database only the Italian subjects are reported. Standard deviation is reported as error.

		Male		Female	
		Healthy	IT NU-AGE	Healthy	IT NU-AGE
Whole body fat	Inter	0.17 ± 0.02	0.31 ± 0.09	0.30 ± 0.02	0.35 ± 0.09
	Coeff	0.0017 ± 0.0004	0.0001 ± 0.0010	0.001 ± 0.001	0.001 ± 0.001
Whole body lean	Inter	0.79 ± 0.02	0.66 ± 0.09	0.66 ± 0.02	0.61 ± 0.09
	Coeff	-0.0016 ± 0.0004	-0.0002 ± 0.0012	-0.0010 ± 0.0005	-0.0006 ± 0.0012
Upper limbs fat	Inter	0.018 ± 0.002	0.027 ± 0.012	0.034 ± 0.003	0.040 ± 0.012
	Coeff	$0.000\,18 \pm 0.000\,05$	$0.000\,05 \pm 0.000\,16$	$0.000\,17 \pm 0.000\,06$	$0.000\,11 \pm 0.000\,16$
Upper limbs lean	Inter	0.110 ± 0.004	0.089 ± 0.014	0.068 ± 0.004	0.080 ± 0.014
	Coeff	$-0.000\,29 \pm 0.000\,08$	$-0.000\,12 \pm 0.000\,19$	$-0.000\,02 \pm 0.000\,09$	$-0.000\,27 \pm 0.000\,20$
Lower limbs fat	Inter	0.075 ± 0.006	0.062 ± 0.033	0.135 ± 0.007	0.16 ± 0.03
	Coeff	$-0.000\,01 \pm 0.000\,13$	$0.000\,24 \pm 0.000\,45$	$0.000\,03 \pm 0.000\,15$	$-0.000\,44 \pm 0.000\,47$
Lower limbs lean	Inter	0.277 ± 0.007	0.25 ± 0.03	0.228 ± 0.007	0.20 ± 0.03
	Coeff	$-0.000\,78 \pm 0.000\,14$	$-0.000\,46 \pm 0.000\,37$	$-0.000\,43 \pm 0.000\,17$	$-0.000\,23 \pm 0.000\,38$
Trunk fat	Inter	0.064 ± 0.015	0.21 ± 0.07	0.12 ± 0.02	0.14 ± 0.07
	Coeff	0.0016 ± 0.0003	$-0.000\,30 \pm 0.000\,96$	$0.000\,74 \pm 0.000\,37$	$0.000\,97 \pm 0.000\,99$
Trunk lean	Inter	0.360 ± 0.013	0.29 ± 0.05	0.324 ± 0.013	0.28 ± 0.05
	Coeff	$-0.000\,58 \pm 0.000\,27$	$0.000\,32 \pm 0.000\,71$	$-0.000\,67 \pm 0.000\,31$	$-0.000\,12 \pm 0.000\,73$
Android fat	Inter	0.007 ± 0.003	0.040 ± 0.015	0.017 ± 0.004	0.023 ± 0.015
	Coeff	$0.000\,34 \pm 0.000\,07$	$-0.000\,08 \pm 0.000\,21$	$0.000\,15 \pm 0.000\,08$	$0.000\,16 \pm 0.000\,22$
Android lean	Inter	0.053 ± 0.002	0.042 ± 0.008	0.044 ± 0.002	0.049 ± 0.008
	Coeff	$-0.000\,03 \pm 0.000\,04$	$0.000\,10 \pm 0.000\,11$	$0.000\,05 \pm 0.000\,05$	$-0.000\,08 \pm 0.000\,12$
Gynoid fat	Inter	0.045 ± 0.004	0.03 ± 0.01	0.078 ± 0.004	0.108 ± 0.015
	Coeff	$-0.000\,02 \pm 0.000\,09$	$0.000\,18 \pm 0.000\,20$	$0.000\,03 \pm 0.000\,10$	$-0.000\,57 \pm 0.000\,21$
Gynoid lean	Inter	0.118 ± 0.004	0.100 ± 0.013	0.084 ± 0.005	0.095 ± 0.014
	Coeff	$-0.000\,26 \pm 0.000\,09$	$-0.000\,02 \pm 0.000\,19$	$-0.000\,01 \pm 0.000\,10$	$-0.000\,12 \pm 0.000\,19$

Table A.2: Table shows the most similar factors across the centres, in parentheses are the values of the correlation. As you can see factors are substantially the same, less than changes in position.

	IT									
	0	1	2	3	4	5	6	7	8	9
FR	0 (0.97)	1 (1.00)	2 (1.00)	19 (0.74)	5 (0.99)	-	14 (0.90)	7 (0.98)	8 (0.97)	-
PL	0 (0.99)	2 (1.00)	15 (0.61)	16 (0.94)	4 (0.99)	5 (0.63)	-	8 (0.97)	10 (0.99)	13 (0.68)
NE	0 (0.94)	1 (1.00)	2 (1.00)	9 (0.98)	12 (0.68)	6 (0.66)	11 (0.96)	10 (0.95)	8 (0.97)	-
UK	0 (0.98)	1 (0.99)	2 (0.75)	4 (0.98)	19 (0.58)	5 (0.62)	17 (0.93)	9 (0.90)	15 (0.95)	-

	UK									
	0	1	2	3	4	5	6	7	8	9
FR	0 (0.99)	1 (1.00)	2 (0.77)	5 (0.48)	19 (0.76)	-	-	9 (0.37)	8 (0.55)	10 (0.85)
PL	0 (0.99)	2 (1.00)	4 (0.66)	16 (0.84)	16 (0.94)	5 (0.99)	13 (0.98)	7 (0.99)	10 (0.61)	12 (0.86)
NE	0 (0.97)	1 (1.00)	2 (0.74)	3 (0.83)	9 (0.97)	6 (0.99)	-	-	8 (0.62)	10 (0.96)

	NE									
	0	1	2	3	4	5	6	7	8	9
FR	0 (0.99)	1 (1.00)	2 (1.00)	5 (0.99)	-	16 (0.87)	-	15 (0.93)	8 (0.96)	19 (0.77)
PL	0 (0.96)	2 (1.00)	15 (0.61)	4 (0.99)	13 (0.97)	-	17 (0.34)	19 (0.94)	10 (0.97)	16 (0.96)

	PL									
	0	1	2	3	4	5	6	7	8	9
FR	0 (0.99)	2 (1.00)	18 (0.52)	16 (0.87)	5 (0.99)	-	-	9 (0.33)	10 (0.44)	12 (0.93)

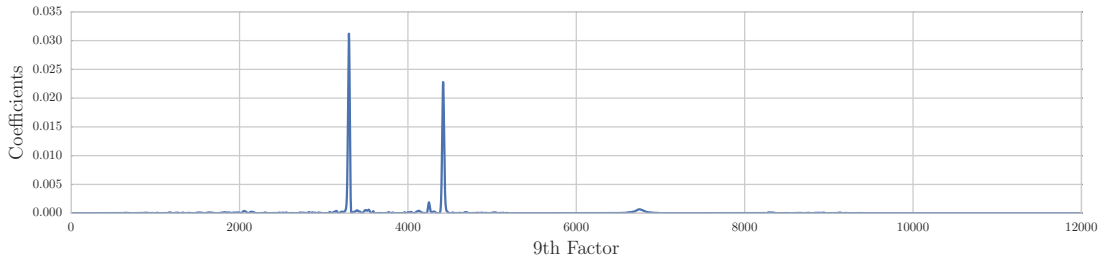

Figure A.3: Plot of the ninth factor. Each factor correspond to a linear combination of the primary pecks, on the vertical axis the weight are indicates, on horizontal axis the corresponding channel of the original spectrum are indicates.

Table A.3: Detail of linear model between the PC from PCA of DXA variables and the ninth factor. σ is the standard deviation, $p - value$ is the probability of obtaining a result equal to or more extreme than that observed under the supposed true null hypothesis is $H_0 : \beta = 0$, where β is the coefficient of the predictor. Useless predictors are sequentially removed.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC0	−0.201	0.067	0.003
	PC2	−0.114	0.059	0.055
	PC3	−0.172	0.060	0.005
	PC4	−0.192	0.064	0.003
	PC5	−0.189	0.063	0.003
United Kingdom	PC0	−0.491	0.116	0.000
	PC2	−0.130	0.080	0.106
	PC3	−0.117	0.075	0.119
	PC4	−0.257	0.082	0.002
	PC5	−0.201	0.082	0.015
	PC7	−0.487	0.125	0.000
Netherlands	PC0	−0.143	0.071	0.044
	PC1	−0.202	0.066	0.003
	PC2	−0.093	0.061	0.128
	PC6	−0.144	0.064	0.025
	PC7	−0.215	0.075	0.005
Poland	PC0	−0.248	0.065	0.000
	PC3	−0.116	0.058	0.047
	PC4	−0.285	0.058	0.000
	PC5	−0.140	0.069	0.042
	PC6	−0.154	0.068	0.024
France	PC0	−0.442	0.107	0.000
	PC1	−1.099	0.174	0.000
	PC2	0.365	0.099	0.000
	PC3	0.505	0.104	0.000
	PC5	1.134	0.224	0.000
	PC6	−0.761	0.144	0.000
	PC7	0.245	0.111	0.028

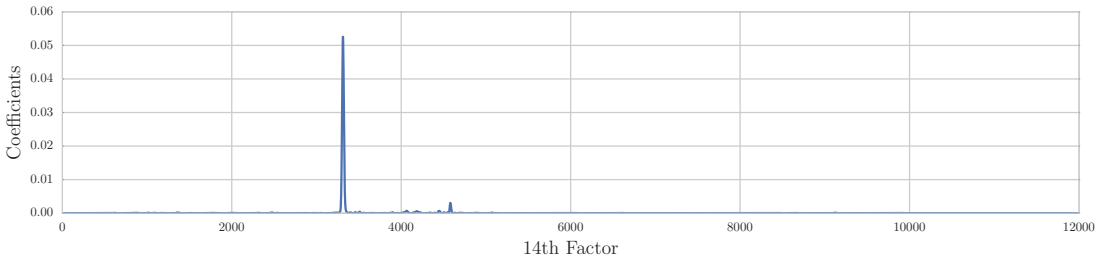

Figure A.4: Plot of the 14th factor..

Table A.4: Detail of linear model between the PC from PCA of DXA variables and the 14th factor.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC1	−0.160	0.067	0.018
	PC3	0.189	0.062	0.003
	PC4	0.129	0.062	0.038
	PC5	0.252	0.070	0.000
	PC6	0.171	0.062	0.007
United Kingdom	PC0	0.253	0.124	0.043
	PC4	0.146	0.078	0.064
	PC6	0.174	0.082	0.035
	PC7	0.183	0.120	0.130
Netherlands	PC0	−0.403	0.080	0.000
	PC1	−0.176	0.075	0.020
	PC3	0.132	0.068	0.054
	PC4	0.122	0.065	0.061
	PC5	0.126	0.072	0.081
	PC7	0.285	0.083	0.001
Poland	PC2	−0.099	0.061	0.107
	PC6	0.277	0.061	0.000
	PC7	0.218	0.061	0.000
France	PC0	0.416	0.105	0.000
	PC1	0.749	0.155	0.000
	PC2	−0.368	0.083	0.000
	PC3	−0.419	0.104	0.000
	PC5	−0.998	0.203	0.000
	PC6	0.655	0.139	0.000

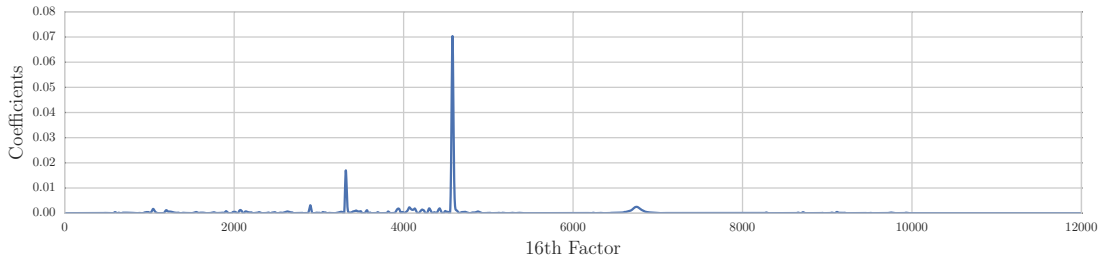

Figure A.5: Plot of the 16th factor.

Table A.5: Detail of linear model between the PC from PCA of DXA variables and the 16th factor.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC1	−0.117	0.068	0.088
	PC3	0.228	0.062	0.000
	PC4	0.117	0.063	0.064
	PC5	0.185	0.071	0.009
United Kingdom	PC0	0.291	0.117	0.014
	PC1	0.201	0.072	0.006
	PC6	0.123	0.080	0.125
	PC7	0.277	0.109	0.012
Netherlands	PC0	−0.310	0.082	0.000
	PC1	−0.209	0.077	0.007
	PC3	0.156	0.069	0.024
	PC4	0.107	0.066	0.105
	PC5	0.170	0.073	0.021
	PC7	0.242	0.084	0.005
Poland	PC0	0.095	0.071	0.181
	PC2	−0.100	0.062	0.107
	PC3	0.255	0.061	0.000
	PC6	0.270	0.063	0.000
	PC7	0.197	0.068	0.004
France	PC0	0.398	0.107	0.000
	PC1	0.661	0.157	0.000
	PC2	−0.145	0.085	0.088
	PC3	−0.195	0.106	0.068
	PC5	−0.776	0.207	0.000
	PC6	0.564	0.141	0.000

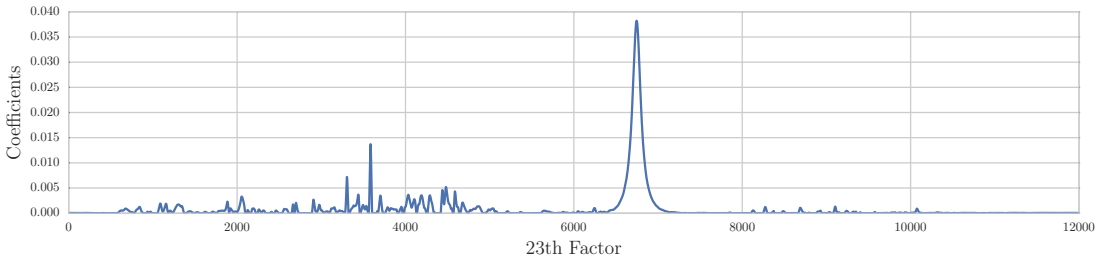

Figure A.6: Plot of the 23th factor.

Table A.6: Detail of linear model between the PC from PCA of DXA variables and the 23th factor.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC0	0.156	0.065	0.018
	PC2	−0.110	0.061	0.071
	PC3	−0.150	0.062	0.017
	PC5	−0.154	0.065	0.018
United Kingdom	PC0	−0.240	0.113	0.035
	PC2	−0.269	0.076	0.001
	PC4	−0.314	0.081	0.000
	PC7	−0.433	0.119	0.000
Netherlands	PC0	0.127	0.076	0.096
	PC1	0.143	0.072	0.048
	PC3	0.089	0.068	0.193
	PC4	−0.138	0.068	0.043
	PC6	0.099	0.071	0.162
	PC7	−0.152	0.084	0.070
Poland	PC1	0.109	0.063	0.086
	PC6	−0.178	0.063	0.005
	PC7	−0.126	0.062	0.045
France	PC0	−0.188	0.094	0.047
	PC1	0.196	0.100	0.051
	PC5	−0.109	0.099	0.273
	PC7	−0.150	0.096	0.121

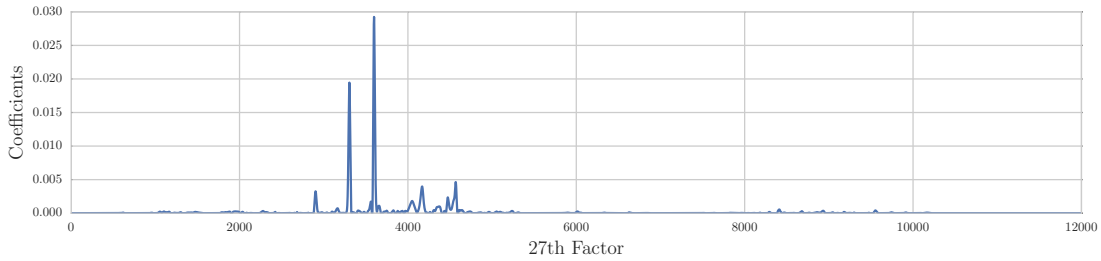

Figure A.7: Plot of the 23th factor.

Table A.7: Detail of linear model between the PC from PCA of DXA variables and the 27th factor.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC1	−0.231	0.066	0.001
	PC4	0.183	0.063	0.004
	PC5	0.289	0.068	0.000
	PC6	0.126	0.061	0.039
United Kingdom	PC0	0.552	0.116	0.000
	PC1	0.217	0.069	0.002
	PC4	0.146	0.073	0.047
	PC6	0.334	0.076	0.000
	PC7	0.376	0.112	0.001
Netherlands	PC0	−0.200	0.078	0.011
	PC5	−0.167	0.106	0.116
	PC6	0.284	0.105	0.008
	PC7	0.361	0.073	0.000
Poland	PC4	0.079	0.065	0.225
	PC6	0.180	0.063	0.004
	PC7	0.101	0.065	0.121
France	PC0	0.254	0.119	0.034
	PC1	0.285	0.192	0.140
	PC2	−0.188	0.110	0.089
	PC3	−0.219	0.115	0.060
	PC5	−0.498	0.248	0.046
	PC6	0.402	0.160	0.013
	PC7	−0.140	0.122	0.253

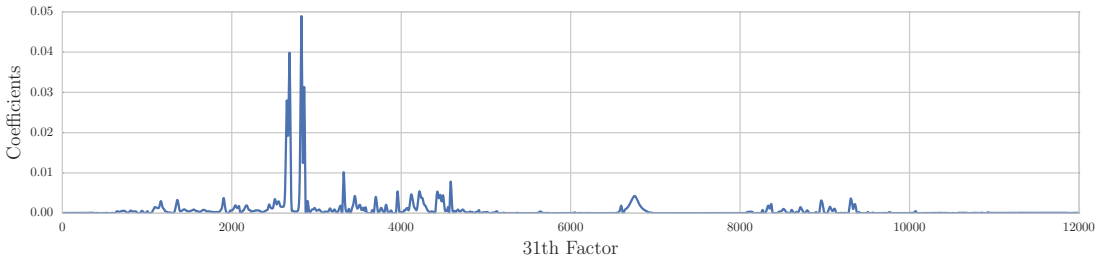

Figure A.8: Plot of the 31th factor.

Table A.8: Detail of linear model between the PC from PCA of DXA variables and the 31th factor.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC0	0.155	0.066	0.019
	PC1	0.218	0.071	0.002
	PC2	−0.194	0.062	0.002
	PC5	−0.244	0.075	0.001
	PC7	−0.108	0.063	0.088
United Kingdom	PC0	−0.410	0.122	0.001
	PC1	−0.158	0.072	0.029
	PC2	−0.218	0.077	0.005
	PC4	−0.216	0.082	0.010
	PC6	−0.254	0.079	0.002
	PC7	−0.315	0.120	0.009
Netherlands	PC0	0.309	0.103	0.003
	PC1	0.173	0.099	0.083
	PC4	−0.109	0.069	0.114
	PC5	−0.280	0.157	0.075
	PC6	0.210	0.153	0.170
	PC7	−0.297	0.081	0.000
Poland	PC3	−0.083	0.061	0.179
	PC6	−0.295	0.061	0.000
	PC7	−0.160	0.061	0.009
France	PC0	−0.358	0.104	0.001
	PC1	−0.482	0.141	0.001
	PC3	0.154	0.093	0.101
	PC5	0.691	0.184	0.000

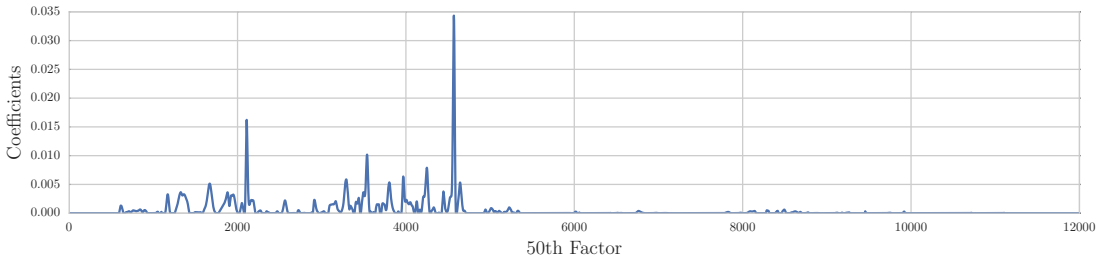

Figure A.9: Plot of the 50th factor.

Table A.9: Detail of linear model between the PC from PCA of DXA variables and the 50th factor.

Country	Regressor	Coefficent	σ	$p - value$
Italy	PC3	0.184	0.062	0.003
	PC4	0.158	0.062	0.011
	PC5	0.083	0.062	0.180
	PC6	0.148	0.062	0.019
United Kingdom	PC0	0.253	0.126	0.046
	PC3	0.200	0.076	0.009
	PC4	0.239	0.075	0.002
	PC5	0.146	0.085	0.087
	PC6	0.191	0.084	0.024
	PC7	0.338	0.118	0.005
Netherlands	PC0	−0.321	0.069	0.000
	PC2	−0.110	0.064	0.089
	PC4	0.118	0.066	0.077
	PC7	0.347	0.070	0.000
Poland	PC6	0.283	0.057	0.000
	PC7	0.370	0.057	0.000
France	PC0	0.305	0.107	0.005
	PC1	0.545	0.159	0.001
	PC2	−0.251	0.085	0.004
	PC3	−0.350	0.107	0.001
	PC5	−0.683	0.208	0.001
	PC6	0.575	0.142	0.000

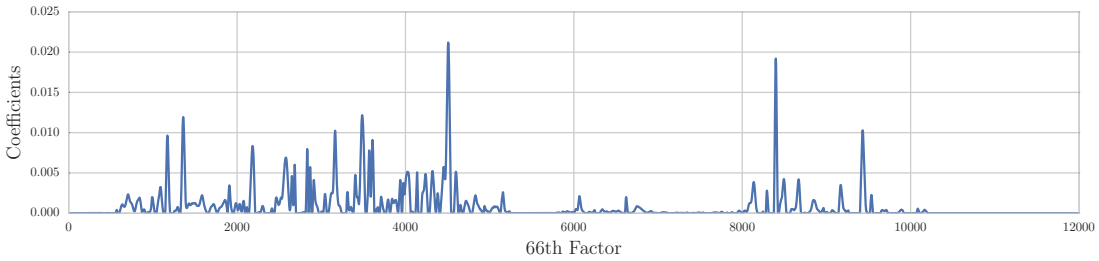

Figure A.10: Plot of the 66th factor.

Table A.10: Detail of linear model between the PC from PCA of DXA variables and the 66th factor.

Country	Regressor	Coefficent	σ	$p-value$
Italy	PC1	−0.138	0.070	0.051
	PC2	0.150	0.061	0.015
	PC3	0.189	0.062	0.002
	PC4	0.109	0.062	0.081
	PC5	0.314	0.073	0.000
	PC7	0.156	0.063	0.014
United Kingdom	PC0	0.417	0.121	0.001
	PC1	0.089	0.072	0.216
	PC4	0.096	0.077	0.210
	PC6	0.284	0.080	0.000
	PC7	0.266	0.117	0.024
Netherlands	PC0	−0.247	0.073	0.001
	PC3	0.133	0.063	0.036
	PC4	0.199	0.063	0.002
	PC5	0.156	0.065	0.017
	PC7	0.298	0.072	0.000
Poland	PC0	0.162	0.070	0.021
	PC4	0.144	0.061	0.020
	PC6	0.257	0.062	0.000
	PC7	0.156	0.069	0.026
France	PC0	0.356	0.106	0.001
	PC1	0.648	0.157	0.000
	PC2	−0.219	0.084	0.010
	PC3	−0.175	0.106	0.100
	PC5	−0.820	0.206	0.000
	PC6	0.610	0.140	0.000

Appendix B

Magnetic Resonance Spectroscopy

Magnetic Resonance Spectroscopy (MRS) is an advanced clinical and research application, based on Nuclear Magnetic Resonance (NMR), which guarantees detection and quantification of metabolites for diagnosis and disease staging.

Basically, this technique is based on the chemical shift phenomenon: nuclei in different chemical environments experience shielding of the static magnetic field by the electron clouds of the neighbouring atoms. Consequently, these nuclei will exhibit different resonance frequencies, which can be identified by the peaks in the spectrum after the Fourier transform of the time-domain signal.

Nuclear Magnetic Resonance (NMR) is a physical phenomenon in which a nucleus with a non-zero spin placed in a magnetic field B_0, submitted to a radio frequency (RF) field B_1 at the Larmor resonance frequency ν_0:

$$\nu_0 = \frac{\gamma}{2\pi} B_0 \quad \text{where } \gamma \text{ is the magnetogyric ratio of the nucleus} \tag{B.1}$$

absorbs and emits electromagnetic radiation.

The resonance frequency is dependent on the chemical environment that surrounding nucleus. The electronic cloud has a shielding effect on the nucleus, since electrons generate a secondary induced magnetic field which opposes to B_0. This is because electrons rotate about B_0 in the opposite sense to nucleus spin precession and consequently their magnetic moment μ_e is aligned against B_0.

This behaviour, called *chemical shift*, is expressed as $B_{eff} = B_0(1 - \sigma)$ where B_{eff} is the effective field and σ is the *shielding constant* which is dependent on the chemical environment of the nucleus and its relative position within the molecule (typical values for σ are 10^{-5} for protons and 10^{-2} for heavier nuclei).

Due to this shielding effect the energy difference ΔE of two adjacent Zeeman levels is lower than in absence of chemical shift:

$$\Delta E = -\hbar\gamma B_{eff}. \tag{B.2}$$

Different effective magnetic field will lead to different energies required for the nucleus to flip between Zeeman levels.

A common phenomenon which is observed in spectroscopic analysis is the splitting of resonance peaks into several smaller peaks, called *J coupling*.

The interactions of magnetic moments of the nuclei can occur through space (dipolar coupling) or indirectly through electron shared in chemical bonds (scalar coupling).

Dipolar interactions between nuclei in liquids are average to zero due to the rapid molecular tumbling. Meanwhile, scalar coupling interactions depend on the intramolecular distance. Therefore, the nearest chemical groups strongly determine the peak displacement observed in the spectrum.

This characteristic can be exploited to identify the molecular species and their covalent structure.

Therefore in NMR spectroscopy the nuclei do not resonate at the same frequency. The RF pulse frequency ω_{RF} slightly differs from the Larmor frequency of the nucleus ω_0, implies that the perturbed magnetization in the rotating frame preceeds both around B_1 and an apparent magnetic field B_{app}:

$$B_{app} = B_0 - \frac{\omega_{RF}}{\gamma} \tag{B.3}$$

resulting from the imperfect resonance. The combination of these two precessions implies that the nuclear spin rotates around an effective field B_{eff} (see Figure B.1).

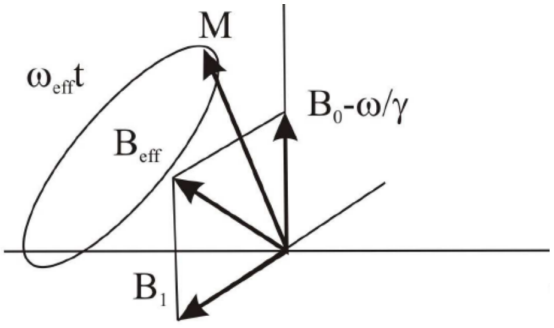

Figure B.1: Precession of magnetization around B_{eff} when $\omega_{RF} \neq \omega_0$

If B_1 is such that the magnetization vector flips onto the transverse (xy) plane

(called 90^o RF pulse), the spin precession on the transverse plane induces an oscillatory electromotive force in the receiving coil by electromagnetic induction, originating thus an induced current called *Free Induction Decay* (FID), which has an oscillating and exponentially decaying trend, and it is originated by the photons in the radio-wave range emitted by the nuclei returning to the equilibrium.

The signal coming from the returning to the equilibrium of nuclear magnetization M:

$$M_{xy}(t) = M_{xy}(0)e^{-\frac{t}{T_2^*}} \tag{B.4}$$

where the decay is enhanced by T_2^* which leads to a faster transverse relaxation because of the inhomogeneities of the magnetic field and to a multi-exponential decay.

Precisely the detected signal $s(t)$ is proportional to M:

$$s(t) = s(0)e^{i\omega t}e^{i\phi}e^{-\frac{t}{T_2^*}}. \tag{B.5}$$

The returning FID is a composite signal of many different contributions from metabolites in the volume of interest, which is resolved into individual resonance frequencies and their relative amplitudes by the Fourier transform of the signal $s(t)$, Figure B.2.

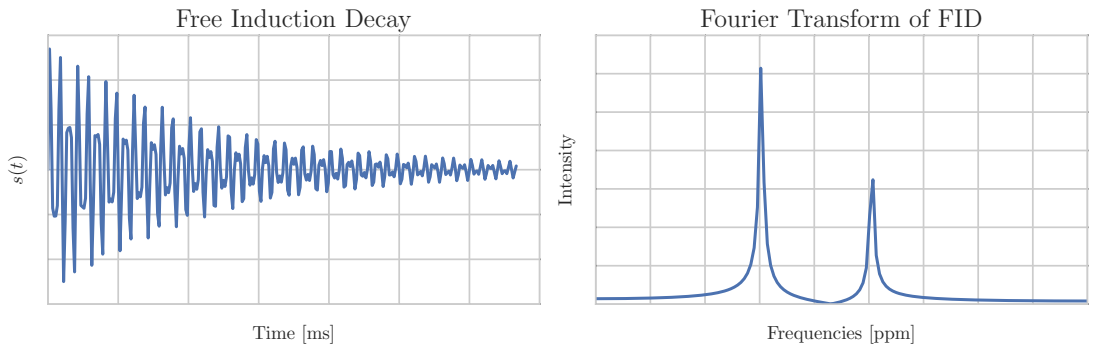

Figure B.2: FID signal and resolved resonance frequencies by the Fourier transform.

It is common practice not to express chemical shifts in Hertz units since this choice would make chemical shifts dependent on the applied magnetic field strength. Therefore, chemical shift δ are conventionally expressed in terms of ppm in function of the displacement from the frequency of a reference compound ν_{ref} measured by the spectrometer:

$$\delta = \frac{\nu - \nu_{ref}}{\nu_{ref}} 10^6. \tag{B.6}$$

Bibliography

[1] K. Giannakouris. "Ageing characterises the demographic perspectives of the European societies". *Statistics in focus* 72 (2008).

[2] International Atomic Energy Association. *Dual energy X ray absorptiometry for bone mineral density and body composition assessment*. IAEA human health series Number 15. Vienna, 2010.

[3] M.A. Laskey. "Dual-energy X-ray absorptiometry and body composition". *Nutrition* 12.1 (1996), pp. 45–51.

[4] E.B. Podgorsak. *Radiation Physics for Medical Physicists*. Springer, 2010.

[5] A.P. Dhawan. *Medical Image Analysis*. Wiley, 2011.

[6] J.T. Bushberg et al. *The Essential Physics of Medical Imaging*. Wolters Kluwer Health, 2011.

[7] A. Pietrobelli et al. "Dual-energy X-ray absorptiometry body composition model: review of physical concepts". *American Journal of Physiology-Endocrinology And Metabolism* 271.6 (1996), E941–E951.

[8] L.A. Lehmann et al. "Generalized image combinations in dual KVP digital radiography". *Medical physics* 8.5 (1981), pp. 659–667.

[9] G.M. Blake and I. Fogelman. "Technical principles of dual energy x-ray absorptiometry". 27.3 (1997), pp. 210–228.

[10] R.H. Nord and R.K. Payne. "Body composition by dual-energy X-ray absorptiometry: a review of the technology". *Asia Pac J Clin Nutr* 4 (1995), pp. 167–171.

[11] Lunar GE Healthcare. *enCORE-based X-ray Bone Densitometer User Manual*. English. May 15, 2014.

[12] T.L. Kelly, N. Berger, and T.L. Richardson. "DXA body composition: theory and practice". *Applied Radiation and Isotopes* 49.5 (1998), pp. 511–513.

[13] N.J. Crabtree, M.B. Leonard, and B.S. Zemel. "Dual-energy X-ray absorptiometry". *Bone densitometry in growing patients*. Springer, 2007, pp. 41–57.

[14] J.P. Soriano et al. "Pencil-beam vs fan-beam dual-energy X-ray absorptiometry comparisons across four systems: body composition and bone mineral". *Journal of Clinical Densitometry* 7.3 (2004), pp. 281–289.

[15] J.M. Boone and J.A. Seibert. "An accurate method for computer-generating tungsten anode x-ray spectra from 30 to 140 kV". *Medical physics* 24.11 (1997), pp. 1661–1670.

[16] International Commission on Radiological Protection. *ICRP Publication 60: 1990 Recommendations of the International Commission on Radiological Protection*. 1990 recommendations of the International Commission Radiological Protection: adopted by the Commission in November 1990. SAGE Publications, 1991.

[17] G.M. Blake, M. Naeem, and M. Boutros. "Comparison of effective dose to children and adults from dual X-ray absorptiometry examinations". *Bone* 38.6 (2006), pp. 935–942.

[18] C.F. Njeh et al. "Radiation exposure in bone mineral density assessment". *Applied radiation and isotopes* 50.1 (1999), pp. 215–236.

[19] H.K. Genant et al. "Universal standardization for dual X-ray absorptiometry: patient and phantom cross-calibration results". *Journal of bone and mineral research* 9.10 (1994), pp. 1503–1514.

[20] Ian Jolliffe. *Principal component analysis*. Springer, 2002.

[21] P.O. Hoyer. "Non-negative matrix factorization with sparseness constraints". *The Journal of Machine Learning Research* 5 (2004), pp. 1457–1469.

[22] D.D. Lee and H.S. Seung. "Algorithms for non-negative matrix factorization". *Advances in neural information processing systems*. 2001, pp. 556–562.

[23] D.D. Lee and H.S. Seung. "Learning the parts of objects by non-negative matrix factorization". *Nature* 401.6755 (1999), pp. 788–791.

[24] B. Schölkopf et al. "New support vector algorithms". *Neural computation* 12.5 (2000), pp. 1207–1245.

[25] N.S. Altman. "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician* 46.3 (1992), pp. 175–185.

[26] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[27] K. Karastergiou et al. "Sex differences in human adipose tissues-the biology of pear shape". *Biol Sex Differ* 3.1 (2012), p. 13.

[28] S.R. Daniels et al. "Association of body fat distribution and cardiovascular risk factors in children and adolescents". *Circulation* 99.4 (1999), pp. 541–545.

[29] O. Beckonert et al. "Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts". *Nature protocols* 2.11 (2007), pp. 2692–2703.

[30] P.F. Harrison. "Blind analysis". *Journal of Physics G: Nuclear and Particle Physics* 28.10 (2002), pp. 2679–2692.

[31] W.B. Dunn et al. "Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics". *Metabolomics* 9.1 (2013), pp. 44–66.

[32] R.T. Mckay. "How the 1D-NOESY suppresses solvent signal in metabonomics NMR spectroscopy: An examination of the pulse sequence components and evolution". *Concepts in Magnetic Resonance Part A* 38.5 (2011), pp. 197–220.

[33] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B (Methodological)* (1995), pp. 289–300.

[34] D. Piccolo. *Statistica*. Strumenti. Economia. Il Mulino, 2000. ISBN: 9788815075963.