Alma Mater Studiorum
Università degli Studi di Bologna

Dottorato di Ricerca in
Ingegneria Energetica, Nucleare e del Controllo Ambientale

Ciclo XXIV

Spectrum reconstruction from a scattering measurement using the adjoint Boltzmann transport equation for photons

Jonathan BARE

Coordinatore Dottorato:

Prof. Antonio Barletta

Relatore:

Prof. Jorge Fernandez

Correlatore:

Prof. François Tondeur

Settore concorsuale di afferenza: 09/C3

Settore scientifico disciplinare: ING-IND/18

Esame finale anno 2012

Acknowledgement

First, I would like to thank Prof. Jorge E. Fernandez, who introduced to me a different perspective of photon transport calculations and to an interesting and innovative approach of inverse problems. I also would like to thank him for having accepted to manage this thesis at distance.

I also would like to express my gratitude to Viviana Scot, for all the ideas and advices developed and proposed during the thesis, for the countless explanations she gave me, for the interesting discussions – about photon transport and inverse calculations, or not – and for her exceptional speed in answering all my emails!

En français maintenant, je profite de ces quelques lignes pour adresser mes plus vifs remerciements à Jean-Michel Mattens pour les très nombreuses heures passées à discuter analyse numérique, pour ses remarques et suggestions pleines de sens, pour sa patience et son intérêt dans la mise en œuvre des différentes méthodes numériques, ainsi que pour toutes les relectures du manuscrit avec son point de vue de mathématicien, parfois un peu décalé, mais toujours très pertinent!

Je souhaite ici également remercier François Tondeur, pour ses conseils avisés lors des relectures du manuscrit de thèse, mais aussi pour les nombreuses discussions scientifiques menées à l'ISIB dans le cadre de cette thèse, ainsi que dans bien d'autres circonstances.

Dans un registre quelque peu différent, j'exprime ici toute mon amitié à Damien Grobet, toujours prêt à mettre la main à la pâte (informatique, mais pâte quand même), toujours motivé à mettre ses capacités techniques à mon service, mais aussi pour la mise à disposition de son super ordinateur, ainsi que pour tous les programmes intitulés 'Matrice' qu'on a pu réaliser pendant ces trois années. J'ose d'ailleurs espérer que je n'aurai jamais à les trier, ces programmes... Enfin, un merci particulier à Nicolas Maurissen pour la mise à disposition de ses très intéressants talents de graphiste, de même que pour sa patience face à mes exigences de qualité.

Pour terminer, je tiens à remercier mon entourage proche pour ses encouragements, ainsi que pour son soutien presqu'inconditionnel!

Summary of the thesis

Today, the quality control of medical and industrial radiological systems is of fundamental importance for evident questions of safety. Therefore, efficient methods for the systematic practical and accurate evaluation of the X-ray source spectrum are required.

The straightforward measurement of an X-ray source spectrum in standard operating conditions is very complicated since the photon flux is very high. At these fluence rates, common detectors cease to work properly and a pile-up effect can be observed. This undesired effet may jeopardize the instrument ability to correctly recognize events and to assign them proper energies. This often leads to distorsions in the measured X-ray spectrum.

In order to overcome experimental difficulties, particular straightforward measurement techniques are necessary. The very high photon flux may be limited by modifying the geometry of the experimental set-up, having for example recourse to pinhole collimators and / or by increasing the distance between the X-ray source and the detector. However, such a geometrical set-up is not sufficient and the fluency has to be reduced further by keeping the current of the X-ray tube at very low values. The resulting measurement is consequently wandering away from the real operating conditions, and the design of efficient systems becomes necessary.

In the past, different techniques have been developed for this purpose. In particular, a specific indirect technique based on the measurement of the

photon beam scattered by a target inserted into the path of the X-ray source beam, was developed. Using that method, the photon fluence is reduced up to three orders of magnitude. However, when such a spectrometer is used, the pulse height distribution recorded by the detector does not reproduce the source beam spectrum because of different physical phenomena occurring during the beam scattering or the detection process. The measured spectrum then presents a lack of the original beam information for the complete characterization of the X-ray source.

When using such a spectrometer, the formal problem may be mathematically represented by a matrix equation whose solution gives the source spectrum. However, in most physical situations, the resulting algebraic system of equations is extremely ill-posed. The resolution of the matrix system can then be very complicated. Numerous attempts have already been made to solve this problem by using matrix regularization techniques. These methods are based on purely mathematical criteria to discriminate between signal and noise, not on physical criteria. In addition, the problem has always been considered as a whole, without any distinction between the different physical steps occurring during the photon transport. The physics of the problem has then not been considered in the resolution.

In this thesis, a more physical point of view has been adopted to solve the problem. The forward measurement of an X-ray spectrum by means of a spectrometer may be divided in two successive physical stages: the photon scattering on the target, and the detection process while using a classical HPGe detector. The innovative inverse method described in the following intends to solve the problem in two different steps: the unfolding from the detector response functions, and the inverse scattering on the target of the spectrometer. In the first step of the procedure, classical unfolding techniques are applied on the measured spectra to suppress the detector's influences. For the inverse scattering, our method is based on both the forward and the adjoint solutions of the BOLTZMANN transport equation for photons to generate - if necessary - a

better conditioned linear algebraic system. Both the forward and the adjoint scattering terms are computed from the analytical solution to the transport equation for finite thickness specimens

In a first part of the thesis, the theory of interactions between photons and matter is outlined for the range of energies of interest, the BOLTZMANN transport equation for photons is constructed and some emphasis is made on the signification and the formal description of the adjoint BOLTZMANN equation. The complete inverse calculation strategy is then exposed. This explanation serves as groundwork for the next chapters. Unfolding techniques and some selected numerical methods are theoretically described.

Secondly, a numerical approach of the inverse scattering problem has been considered. This numerical approach aimed at both evaluating the consistency of the numerical methods, and characterizing two possible target materials. In order to focus this study on the mathematical aspects of the problem, artificial X-ray spectra have been considered. The accuracy in the reconstruction has been evaluated on a purely mathematical basis, and a very interesting characterization of the scattering materials has been deduced. This validation on theoretical X-ray source spectra has provided excellent results.

Finally, the complete inverse strategy has been tested on two different spectra obtained by using an experimental prototype built at the OPERATIONAL UNIT OF HEALTH PHYSICS of the UNIVERSITY OF BOLOGNA (Italy). The system was based on a Tungsten X-ray tube, operated until 150 kV. A finite thickness target of graphite has been used for the scattering. The detector was a 10 mm diameter and 10 mm thickness ORTEC HPGe detector, in a POP-TOP configuration. The entering window was constituted by a thin beryllium foil. A comparison of the reconstructed source spectra with direct measurements of the X-ray tube has been performed, using a standard radiological device modified to generate a primary beam with a lower intensity, showing very good agreements.

Résumé de la thèse

Pour des raisons évidentes de protection des patients, le contrôle optimal des systèmes radiologiques médicaux et industriels est un souci permanent. Afin d'en assurer la qualité, il est nécessaire de pouvoir recourir à des méthodes efficaces permettant une évaluation précise des spectres d'émission des rayons X dans les conditions normales d'utilisation.

Les mesures directes des spectres d'émission X dans les conditions courantes d'utilisation se révèlent être particulièrement complexes, en raison du flux élevé rendant les détecteurs courants peu efficaces. A ces fluences élevées, un effet d'accumulation peut survenir dans le détecteur. Cet effet, fortement indésirable, peut compromettre plus ou moins fortement la capacité de l'instrument à discriminer les différents événements liés à la détection du rayonnement. Il en résulte généralement une distorsion du spectre mesuré.

Afin d'outrepasser cette difficulté expérimentale majeure, il est impératif de recourir à des techniques particulières de mesure. Le flux élevé de photons peut, par exemple, être limité par une modification de la géométrie du système de mesure, par l'utilisation de collimateurs étroits et / ou par une augmentation de la distance entre la source de rayons X et le détecteur. Bien que permettant une limitation significative de la fluence photonique, un tel dispositif n'est souvent pas suffisant, et la fluence doit encore être diminuée en réduisant le courant et la tension de fonctionnement du tube. Dans ces circonstances, les spectres mesurés sont néanmoins trop différents de ceux obtenus dans les conditions réelles d'utilisation. Ce type de mesure présente donc un intérêt quelque peu

limité. En outre, les effets de vieillissement du tube à rayons X occasionnent constamment des modifications du spectre d'émission tout au long du cycle de vie du tube, exigeant des contrôles fréquents et périodiques. Ces facteurs de fréquence, de périodicité et de complexité ont induit de manière naturelle la nécessité de disposer de techniques performantes pour la mesure, et ont été un appui majeur dans le développement de celles-ci.

Dans le passé, diverses techniques de mesure ont été développées. En particulier, une technique indirecte basée sur la détection des photons diffusés par une cible solide insérée dans la trajectoire du faisceau primaire a pu être mise en œuvre. L'utilisation de cette technique de diffusion du flux de photons permet une réduction de la fluence initiale pouvant aller jusqu'à trois ordres de grandeur, autorisant dès lors l'utilisation de détecteurs classiques. Cependant, lorsqu'un tel spectromètre est utilisé, le spectre différentiel ne reproduit pas le faisceau source de façon exacte en raison d'une multitude de phénomènes physiques pouvant survenir durant le transport des photons, spécifiquement lors de la diffusion sur la cible. Il en résulte que le spectre mesuré présente un manque plus ou moins conséquent d'informations pour la caractérisation complète et précise du spectre source.

Lorsqu'une mesure est réalisée au moyen d'un spectromètre, le transport des photons peut être modélisé mathématiquement par un système d'équations linéaires. Dans la plupart des situations physiques cependant, la matrice de transport des photons est extrêmement mal conditionnée, rendant la résolution du système mal aisée, voire même particulièrement complexe. De nombreuses tentatives, plus ou moins fructueuses, ont été réalisées dans le passé afin de résoudre ce problème en mettant en œuvre différentes méthodes, en particulier les techniques de régularisation matricielle. Pour la plupart, ces méthodes fournissent des solutions approchées acceptables dans la majorité des cas envisagés, sauf aux basses énergies. De plus, elles ne se basent que sur des critères purement mathématiques pour réaliser la discrimination entre le signal et le bruit, en tronquant par exemple les matrices de calcul, ou en imposant de manière

plus ou moins arbitraire des critères de stabilité. La physique du problème n'est donc pas prise en considération dans la résolution du système.

Ce travail de thèse se propose de résoudre le problème en adoptant un point de vue plus physique, innovant, afin d'améliorer la qualité de la reconstruction du spectre d'émission du tube à rayons X. Le processus de mesure au moyen d'un spectromètre peut être divisé en deux étapes physiques successives: la diffusion des photons sur la cible solide du spectromètre, et la détection au moyen d'un détecteur HPGe. La méthode inverse décrite dans la suite se base sur la différenciation de ces deux étapes afin d'apporter une solution physique au problème. Dans un premier temps, les effets du détecteur sont supprimés des spectres mesurés par les techniques classiques de déconvolution spectrale. Cette première étape de calcul permet d'obtenir le spectre incident au détecteur, avant perturbation due à l'instrument, et après diffusion sur la cible du spectromètre. Dans un second temps, la diffusion inverse du spectre incident au détecteur sur la cible solide est réalisée. La méthode développée dans la thèse pour cette diffusion inverse se base sur les solutions analytiques directes et adjointes de l'équation de Boltzmann, décrivant le transport des photons. Ces solutions permettent de générer, lorsque nécessaire, un système d'équations linéaires dont le nombre de condition spectrale est réduit par rapport à celui du système original.

Dans la première partie de la thèse, la théorie des interactions entre les photons et la matière est décrite pour les énergies présentant un intérêt pratique dans notre cadre de travail. L'équation de Boltzmann pour le transport des photons est également construite, et une attention toute particulière est donnée à la signification et à la description formelle de l'équation adjointe. La stratégie mise en place pour le calcul inverse est ensuite développée. Cette explication sert de point de référence aux chapitres suivants. Pour clôturer la partie théorique du travail, un choix de techniques de déconvolution et une sélection de méthodes numériques sont décrits.

Dans la deuxième partie du travail, une approche purement numérique

du problème a été envisagée, en ayant pour but d'évaluer la consistance mathématique des méthodes numériques choisies sur le type de spectre considéré et de caractériser deux diffuseurs soigneusement sélectionnés. Afin de focaliser l'étude sur les aspects mathématiques du problème, des spectres RX ont été numériquement construits. La précision de la reconstruction spectrale a été évaluée au moyen de critères mathématiques, et une caractérisation intéressante des matériaux cibles a pu être réalisée. Cette première validation, partielle, de la méthode a fourni d'excellents résultats.

Enfin, la stratégie de calcul inverse a été testée sur trois mesures obtenues via un spectromètre construit à l'Unité de Physique Médicale Opérationnelle de l'Université de Bologne (Italie). Le système de mesure était basé sur un tube à rayons X composé d'un filament de tungstène et d'une cible solide de graphite, d'épaisseur égale à 2 mm, assurant ainsi une faible proportion de diffusions multiples. La détection du rayonnement diffusé a été réalisée avec un détecteur HPGe de marque ORTEC (diamètre 10 mm, épaisseur 10 mm), en configuration POP-TOP, avec une fenêtre d'entrée en béryllium. Afin de vérifier la validité du modèle de calcul inverse, les spectres reconstruits ont été comparés à des mesures directes, réalisées grâce à une modification des circuits d'alimentation du tube permettant de générer un faisceau primaire avec une intensité de courant réduite. Une excellente correspondance entre les spectres mesurés et calculés a pu être observée.

Table of Contents

List of Figures

List of Tables

Chapter

1 General introduction

Since W. Röntgen accidentally discovered X-rays in the late 19^{th} century, they have been widely used in various fields, ranging material analysis and crystallography to security and border controls. Today among the very large panel of applications, X-rays are used for medical imaging, where they are of crucial importance in structural diagnosis. For evident questions of safety, the quality control of radiological systems is of fundamental importance. Therefore, efficient methods for the systematic practical and accurate evaluation of the X-ray source spectrum in normal operating conditions are necessary.

Straightforward measurements of X-ray spectra in standard operating conditions (current and voltage) require high-performance laboratory spectrometers. In this type of measurements, the number of incident photons per time unit reaching the detector must necessarily be strongly limited. Indeed, at high fluence rates typical for medical X-ray tubes, common detectors cease to work properly and a pile-up effect, i.e. the quasi simultaneous accumulation of pulses in the detection crystal, can be observed. This undesired effect may jeopardize the instrument's ability to correctly recognize events and to assign them proper energies, leading to distorsions in the measured X-ray spectrum.

In order to reach appropriate count rates, and to overcome experimental problems due to the pulse pile-up, particular straightforward measurement techniques are necessary. Some of them are based on the photon beam attenuation (e.g. [1, 2, 3]), or on the reduction of the detection efficiency by using low intrinsic efficiency detectors or very thin detectors (e.g. [4, 5]). The high photon fluence rate may also be limited by modifying the geometry of the experimental

set-up, having for example recourse to small diameter (pinhole) collimators and / or by increasing the distance between the X-ray tube and the detector. However, such a geometrical set-up does not decrease the photon flux sufficiently, and the fluence has to be reduced further by keeping the current of the tube at very low values. The resulting measurement is consequently wandering away from the real operating conditions. Since aging effects on the X-ray tube produce changes in the spectrum during the tube life, regular controls are required. Laboratory measurements are then very impractical for these controls, and the necessity of designing portable systems for in situ measurements is real.

In the past, a specific indirect technique for the X-ray source measurement, based on the measurement of the photon beam scattered by a target inserted into the path of the source beam, was developed [6]. This technique, falling into the energy-dispersive spectrometers category [7, 8], is usually known as COMPTON spectrometry [9, 10]. Using that method, the X-ray source beam fluency is reduced up to three orders of magnitude, and the device can be made portable. However, when such a spectrometer is used, the pulse height spectrum recorded by the detection system does not reproduce the source beam distribution correctly because of different physical phenomena occurring during the beam scattering or the detection process [11]. These phenomena are, for example, the photon interactions in the scatterer, the detector's physical and statistical influences, the multiple influences of the surrounding environment or the perturbations produced by the electronic devices. Consequently, the measured spectrum does not contain all the physical information available in the source spectrum [12].

When using such a spectrometer, the formal scattering problem may be mathematically represented by a matrix equation whose solution is the source spectrum. In most physical situations, the resulting algebraic system of equations is extremely ill-posed [13, 14]. Generally, the ill-posedness of the problem is a direct consequence of the ill-conditioning of the coefficient matrix (the forward transport matrix, in our case). Practically, the solution to the matrix problem

may be extremely sensitive to small variations in the data of the problem, and classical mathematical methods are often completely inefficient for solving the system of equations [15, 16]. In order to obtain a stable and physically meaningful solution, special strategies are necessary to circumvent the ill-posedness of the forward transport algebraic system. Some attempts have already been made to reconstruct the X-ray source spectrum by using matrix regularization techniques (e.g. [17, 18, 19, 20, 21]). The major drawback of these methods is that they are only based on purely mathematical criteria to discriminate between the signal and the noise, not on physical parameters. In addition, all authors have always treated this inverse problem as a whole. However, the scattering of X-ray photons on the target and their subsequent detection in the crystal are not ruled by similar principles. While the first is a physical scattering on the material, governed by the fundamental interactions between X-ray photons and matter, the second results from the convolution of the incident spectrum to the detector by the detector response function. By considering the inverse problem as a whole, the physics of the photon transport is then not respected.

In this thesis, the problem has been solved by adopting a more physical point of view. The method proposed here is based on a detailed modeling of the X-ray photon scattering in the target material of the spectrometer. This scattering problem may be equivalently represented by the direct or by the adjoint Boltzmann transport equations. In our method, both forward and adjoint scattering terms are computed from the analytical solution to the photon transport equation for finite thickness specimens, and used to generate a better conditioned linear algebraic system whose solution is the source spectrum. The method has first been validated on numerical X-ray spectra, providing excellent results. The technique has secondly been tested on two different spectra obtained with a simple experimental prototype built at the Operational Unit of Health Physics of the University of Bologna (Italy). This experimental set-up allows the measurement of X-ray photons scattered inside a narrow cone with its axis at a 90° angle with respect to the primary beam direc-

tion. The X-ray source distribution has been reconstructed starting with both measurements by using the full inverse technique, after cleaning the measured spectra from the detector response functions. A comparison of the reconstructed spectra with direct measurements of the X-ray tube has been performed, using a standard radiological device modified to generate a primary beam with a lower intensity, showing very good agreements.

1.1 Objectives of the thesis

The thesis has for main objective the complete characterization of an X-ray source spectrum from scattering measurements in normal operating conditions, by performing inverse calculation. A complete characterization of an X-ray beam should ideally include a precise evaluation of the photon fluence and provide information about the quality of the radiation, specifying its spectral energy distribution in particular. This information is of fundamental importance in medical physics for:

- assessing the dose absorbed by a patient;
- obtaining an evaluation of the kerma from the source photon distribution;
- the design and the calibration of dosimeters;
- the design of radiological equipment.

An improvement in the characterization quality of X-ray tubes spectral distribution then aims at improving both the quality of the medical imaging and the patient protection.

The characterization of the X-ray source is made through inverse calculation. An innovative inverse strategy based on the solutions of the forward and adjoint transport equations for photons has been developed. This strategy aims at taking into consideration the inherent physics of the problem, moving away from most of the actual resolution techniques. For that purpose, the following operational objectives have been defined:

- the conceptual design of a complete inverse calculation strategy, respecting the physics of the problem, deduced from the understanding of a forward measurement procedure;
- the derivation of the forward and adjoint transport operators for numerical calculations (in the general and in the monochromatic beam model);
- the computation of the forward and adjoint scattering matrices by using a deterministic code, for two different scattering materials, with a particular discretization;
- the numerical characterization of the two scattering matrices;
- the selection of a consistent numerical method for solving the inverse scattering problem;
- the validation of the complete inverse method by comparing reconstructed source spectra to direct measurements performed in non-standard operating conditions.

1.2 Organization of the thesis

After this general introduction, where the thesis has been situated in the general context of the quality controls of medical equipments, and where the objectives have been outlined, this document is organized as follows:

- in Chapter 2, the main photon-matter interactions occurring in the X-ray regime are described. Even if the interactions between matter and electromagnetic radiations form one of the most diversified classes of phenomena arising in experimental physics, it is possible to identify the three most probable characteristic interactions. First, what is considered as an interaction is explained, and the kernel of a generic interaction is given. Secondly, the three main interaction processes are described, using both physical and mathematical descriptions of the processes;
- in Chapter 3, a general form of the Boltzmann transport equation for photons is constructed. Consideration is also given to the equation that is

adjoint to the transport equation. In this context, the concept of adjoint function is formally explained. Both the forward and adjoint transport equations are then derived in the general and in the monochromatic beam model, and discretized for numerical calculations;

- in Chapter 4, the global inverse strategy for reconstructing the X-ray source vector from a scattering measurement is described. Based on the general description of the forward measurement procedure, the complete inverse technique, developed in this thesis, is outlined with some emphasis on the physics of the problem;
- in Chapter 5, the concept of inverse problems is introduced. Inverse problems are well known to be ill-posed in most physical situations, and form a very complicated class of physical problems. The main difficulties in solving ill-posed problem are due to a cluster of small singular values of the coefficient matrix (said to be ill-conditioned). The singular value decomposition (SVD) is a powerful tool to analyze the anatomy of an ill-conditioned matrix. This decomposition is linked to the condition number, a characteristic value of the scattering matrix, that gives an asymptotical evaluation of the problem stability. Both the SVD and the condition number of the matrix are studied in this chapter, because they are of major importance in practical numerical analysis;
- in Chapter 6, some selected algorithms for the cleaning of measured spectra from the detector response function are introduced theoretically. Measured spectra result from the convolution of the detection system response function with the spectrum hitting the detector. In order to suppress the effects of the detector, unfolding techniques may be applied. Three classes of unfolding methods have been selected. The algorithms are discussed, and the advantages / disadvantages of the techniques are outlined;
- in Chapter 7, numerical methods for solving the inverse scattering problem are detailed. They are classified in two major categories: direct methods and iterative methods. The choice of a particular method depends on the

characteristics of the coefficient matrix. In this chapter, some well-adapted numerical techniques are described theoretically;

– in Chapter 8, the characterization of two selected scatterers - aluminium and carbon - is made by using some artificial spectra. This characterization is a fundamental aspect of the method since the material properties are strongly related to the mathematical performances of the inverse method through the scattering matrix characteristics. The mathematical consistency of the numerical methods exposed in chapter 7 is evaluated. In order to avoid the additional difficulties coming from the unavoidable experimental noise of the measurement, numerical X-ray spectra have been constructed as a first approach to the inverse scattering problem;

– in Chapter 9, the application of the full inverse procedure is made on real experiments obtained at the Operational Unit of Health Physics of the University of Bologna. Different X-ray scattering measurements have been performed, for various current intensities. The complete inverse technique has been applied, and the reconstructed source vectors are compared with direct measurements performed in non-standard geometrical and operating conditions;

– in Chapter 10, the conclusions of the thesis are given, and some future prospects are outlined.

Chapter

2 Photon–matter interactions

Interactions between matter and electromagnetic radiation represent one of the most diversified classes of phenomena in the whole of experimental physics [22]. Even within the energy range usually associated to the X-ray regime, i.e. the energy range between 1 keV to 100 keV, many different processes can occur, all of which having their own individual characteristics. The nature of the matter with which the radiations interact offers almost as wide a range of phenomena as does the nature of radiation itself. This is also true in the relatively restricted domain of X-ray physics.

In the present work, the main interest concerns situations in which the overall behavior of an absorber or scatterer can be deduced by regarding them as a collection of individual atoms, each one absorbing or scattering independently of its surroundings. In such cases, we can assert that interactions between X-ray photons and matter are single identifiable processes, each associated with an individual atom. This is true as long as the response of an atom is not considerably distorted by chemical / molecular forces, and this condition will be implicitly accepted in the following. Such an interaction may be primarily a scattering event, or an absorption process. In the case of scattering, little or no energy is imparted to the atom in question. In the case of an absorption process, the great majority of the incident photon energy is transferred to the atom.

Most of this chapter follows the description in reference [23].

The way X-ray photons interact with matter fundamentally depends on their own energy [24, 25, 26]. In the considered range of energy, X-ray photons

interact exclusively with the electron shells surrounding the atomic nucleus. The nucleus itself does not contribute to the scattering or absorption of photons.

The interaction of a photon of energy $h\nu$ with an isolated atom A has the effect of changing the atom state. Denoting the initial state by $|i\rangle$ and the final state by $|f\rangle$, a general interaction can be expressed as:

$$A_i + h\nu_i \longrightarrow A_f + h\nu_f \tag{2.1}$$

Equation 2.1 indicates in a mathematical way the photon–atom interaction of interest in this work, having one initial photon and only one resulting photon. The term A_f in the right side of equation 2.1 denotes the atom in its final state plus all the non-photonic particles produced during the reaction.

Although a large number of possible interaction mechanisms are known for X-rays in matter, three major processes play an important role in the X-ray regime. These mechanisms of interest are:

- the photoelectric effect, during which an initial photon undergoes an interaction with an atom, causing the ejection of an electron from one of the atom's internal shells, leaving a vacancy in the electronic structure. The vacancy is quickly filled by an electron from upper energetic shells of the atom during a reorganization mechanism. This rearrangement process is accompanied by the emission of a characteristic fluorescence photon;
- the incoherent or COMPTON scattering, where the incident photon undergoes an inelastic collision with an atom external electron, causing both momentum and energy transfer to the electron in question;
- the coherent or RAYLEIGH scattering, during which the photon undergoes an elastic collision with an atomic electron, changing its momentum, but not its energy.

The term 'interaction' has not to be considered here in a restrictive way. Any sequence of physical processes occurring in rapid succession, originated by a photon and producing another (other) photon(s), can be statistically considered

as an unique interaction. This is the case with the photoelectric effect, for example. Then, what is called interaction in the following does not strictly refer to as a single process.

During photon-atom interactions, not only photons are produced: the photoelectric effect and the COMPTON scattering also produce electrons. These electrons are governed by other kinds of interaction laws, and can as well produce new photons. Since the electron contributions render the transport problem considerably more complicated because of the coupling between photons and electrons, we shall neglect in this work Bremsstrahlung of the COMPTON and photoelectric electrons [27], and also other photon sources such as anomalous scattering [28].

The photon resulting from interaction 2.1 may in turn interact with another atom of the matter, starting a multiple chain of events. However, the single-process kernel plays a very major role in the photon transport theory. They represent the probability - by wavelength, solid angle and path unit - that the process changes the phase-space variables from direction $\overrightarrow{\omega}'$ and wavelength λ' to direction $\overrightarrow{\omega}$ and wavelength λ. Therefore a kernel is directly related to the atomic double differential cross-section of the interaction. The integrated cross-section for the process T can thus be obtained from:

$$\sigma_T(\lambda', \overrightarrow{\omega}') = \int_0^\infty \int_{4\pi} k_T(\overrightarrow{\omega}, \lambda, \overrightarrow{\omega}', \lambda')\, d\overrightarrow{\omega}\, d\lambda \tag{2.2}$$

allowing the comparison with experimental or theoretical data.

Since the three processes of interest are statistically independent [22, 24] and since they constitute the main part of the total cross-section [29, 30], the total cross-section μ may be defined as the sum of the photon cross-sections of the processes of interest:

$$\mu = \sigma_C + \sigma_R + \tau \tag{2.3}$$

where:

- σ_C is the COMPTON (incoherent) integral cross-section;
- σ_R is the RAYLEIGH (coherent) integral cross-section;
- τ is the photoelectric cross-section.

The relative probabilities of the different interaction processes depend on the photon energy, as illustrated for pure carbon, pure aluminium and pure germanium in Figure 2.1, in Figure 2.2, and in Figure 2.3 respectively [31]. Graphs are given for a large range of energy (i.e. 1 keV to 1 MeV), in a logarithmic scale.

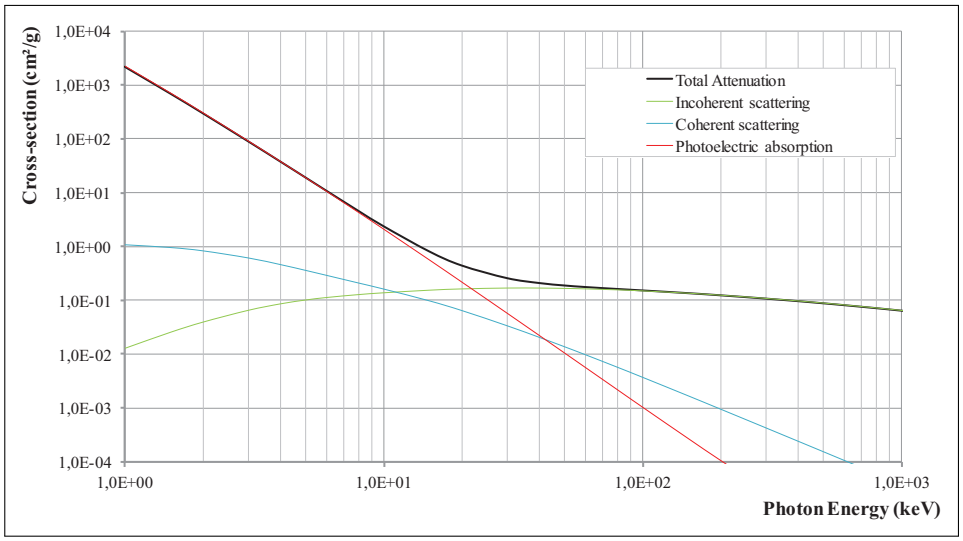

Figure 2.1: Mass attenuation coefficient for pure carbon with different contributing factors of attenuation: COMPTON scattering, RAYLEIGH scattering and photoelectric absorption. Graph data coming from NIST's XCOM database.

Considering the photon attenuation, Figures 2.1 to 2.3 indicate that the atomic photoelectric effect predominates for low energies, while the COMPTON scattering dominates for intermediate energies. Considering only scattering processes, they also show a high probability of RAYLEIGH scattering for low energies or forward angles, while the COMPTON scattering probability of interaction dominates for higher energies or larger angles.

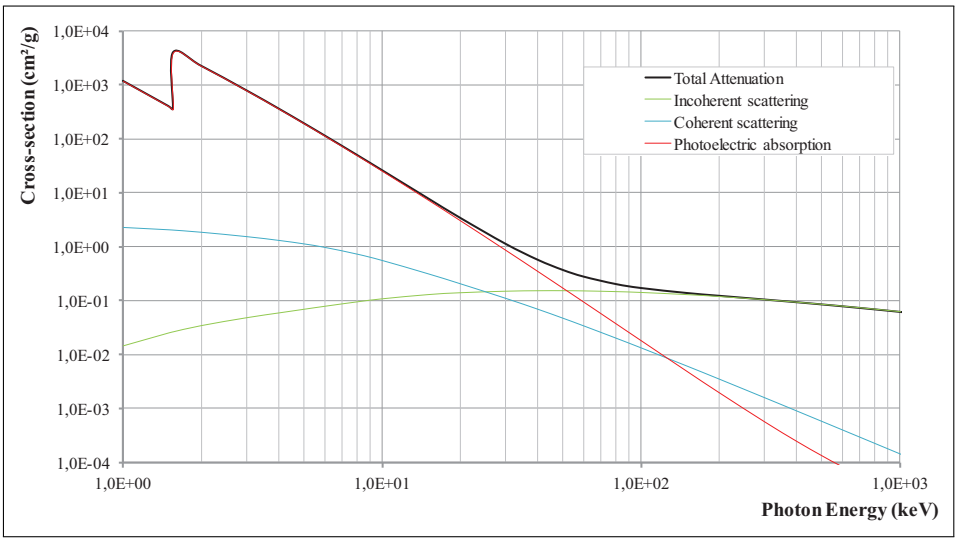

Figure 2.2: Mass attenuation coefficient for pure aluminium with different contributing factors of attenuation: COMPTON scattering, RAYLEIGH scattering and photoelectric absorption. Graph data coming from NIST's XCOM database.

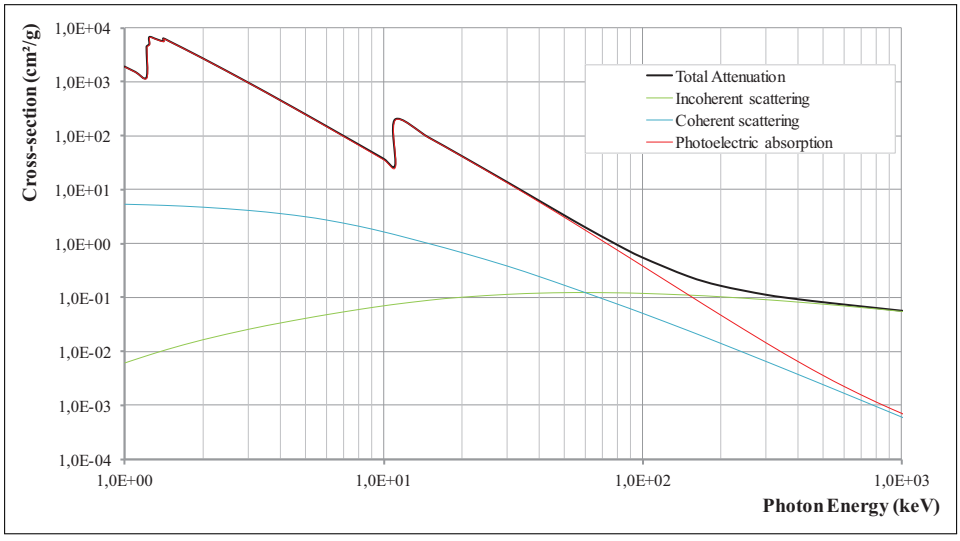

Figure 2.3: Mass attenuation coefficient for pure germanium with different contributing factors of attenuation: COMPTON scattering, RAYLEIGH scattering and photoelectric absorption. Graph data coming from NIST's XCOM database.

2.1 The photoelectric effect

During the photoelectric effect, illustrated in Figure 2.4, a photon of energy $E_\gamma = h\nu_0$ undergoes an interaction with an entire atom. This interaction process results in the emission of a photoelectron, usually from the most internal shells of the atom, leaving a vacancy in the atom electronic structure. The generic equation of this interaction is given by:

$$\gamma + (atom) \longrightarrow (atom)^* + e^- \tag{2.4}$$

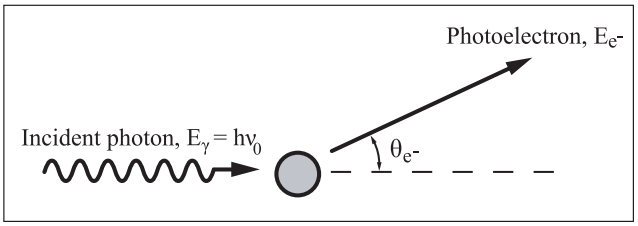

Figure 2.4: Schematic illustration of the photoelectric effect.

During the photoelectric interaction, a photon transfers almost the totality of its energy $h\nu_0$ to an atomic electron, and completely disappears. If the incident photon energy is higher than the binding energy of the electron in its original shell, B_i, the photoelectron is ejected from this particular shell, resulting in the atom ionization. The probability of photoelectric absorption is greater the more tightly bound the electron. Therefore, K-electrons are most affected, provided the X-ray energy exceeds the K-electron binding energy. As the photon energy drops below B_k, the cross-section drops discontinuously. As the energy decreases further, the cross-section increases until the first L edge is reached, at which energy the cross-section drops again, then rices once more, and so on for the remaining edges. The difference of energy between the incident photon energy $h\nu_0$ and the electron binding energy B_i is distributed between the electron (kinetic energy) and the atom (recoil energy), with respect to the energy and momentum principles. However, because of the comparatively small electron mass, it can be approximated that the entire incident photon energy is carried out as kinetic energy by the photoelectron. In a first approximation,

the recoil energy of the atom can effectively be neglected. Thanks to the energy conservation principle, the kinetic energy of the photoelectron is given by:

$$E_{e^-} = h\nu_0 - B_i \tag{2.5}$$

where E_{e^-} is the energy of the emitted photoelectron. The index i of the electron binding energy B_i stands for the different electronic shells (K, L, M,...) of the atom.

The cross-section of the photoelectric effect cannot be described by a simple general mathematical expression. However, empirical descriptions are available. The cross-section approximately varies as:

- E^{-n}, where $n \simeq 3$ for energies less than about 150 keV, and $n \simeq 1$ for energies greater than about 5 MeV;
- Z^m, where m varies from about 4 at $E = 100$ keV to 4,6 at $E = 3$ MeV.

As a crude approximation, in the energy region for which the photoelectric effect is dominant, the cross-section is proportional to:

$$\tau \propto \frac{Z^m}{E^n} \tag{2.6}$$

For light nuclei, K-shell electrons are responsible for almost all photoelectric interactions. For heavy nuclei, however, about 80 % of photoelectric interactions result in the ejection of a K-shell electron [32].

The vacancy left by the photoelectron in the electronic structure of the atom is quickly filled (time delay between 10^{-17} s to 10^{-14} s) by a reorganization mechanism in which an electron comes from an outer shell to the incomplete one. This reorganization process is accompanied by the liberation of a well defined quantity of energy, under the form of characteristic fluorescence X-ray photons. In some cases, the emission of an Auger or a Coster-Kronig electron may substitute for the fluorescence X-ray lines in carrying away the atomic excitation energy. The competition between these two processes is given by the fluorescence yield, i.e. the probability that an atom in an excited state will emit an X-ray

photon in its first transition rather than an AUGER electron. For the K-shell, fluorescence yields vary from 0.005 for $Z = 8$ to 0.965 for $Z = 90$ [33, 34], as illustrated in Figure 2.5. Although X-ray photons of various energies may be emitted, the approximation is often made that only one fluorescence photon or one AUGER electron is emitted, with an energy equal to the binding energy of the photoelectron.

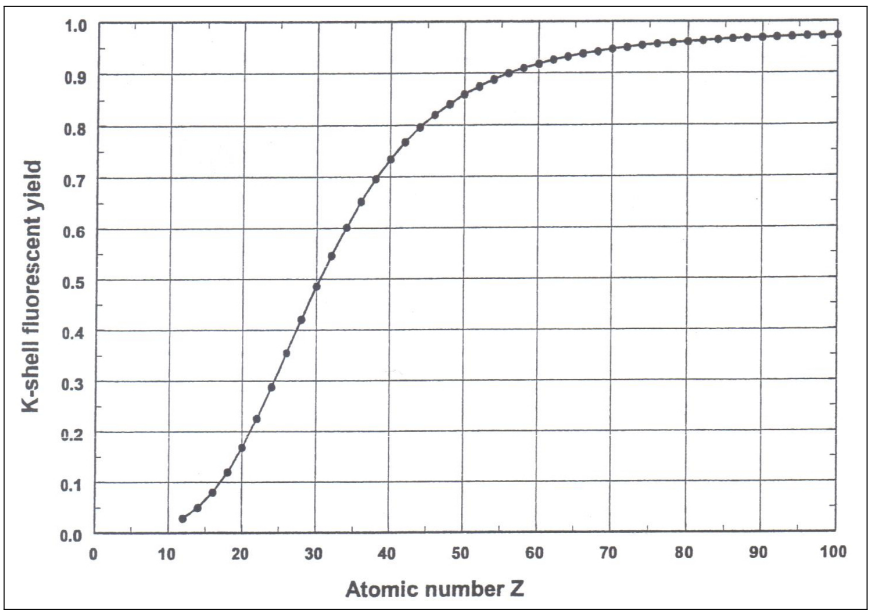

Figure 2.5: Variation of the K-shell fluorescence yield in function of the atomic number, Z.

Statistically, the two combined processes of photon absorption and subsequent X-ray or electron emission may be considered as a single interaction. It is worth noting that the interaction takes place with the atom as a whole, and cannot occur with a free electron, otherwise the energy and momentum conservation principles would not be borne out. The electron must necessarily be bound to an atom: the interaction is then global.

2.1.1 The scalar photoelectric kernel

The scalar kernel, k, for a single X-ray fluorescence characteristic line of wavelength λ emitted by a pure element target s as the consequence of the photoelectric absorption of photons with wavelength λ' may be mathematically described by [35]:

$$k_{P_{\lambda_i}}\left(\lambda' \to \lambda, \overrightarrow{\omega}' \to \overrightarrow{\omega}\right) = \frac{1}{4\pi} Q_{\lambda_i}(\lambda')\, \delta\left(\lambda - \lambda_i\right)\, \left[1 - \mathcal{H}(\lambda' - \lambda_{e_i})\right] \qquad (2.7)$$

with:

- $Q_{\lambda_i}(\lambda')$, the X-ray fluorescence emission probability density (in cm^{-1}) for the single line of wavelength λ_i;
- λ_{e_i}, the wavelength of the absorption edge;
- $\mathcal{H}$, the HEAVISIDE function.

A particular line can only be emitted when λ' is lower than the threshold wavelength of the absorption edge λ_{e_i} of the series to which the line belongs [36, 37], as expressed by the HEAVISIDE function in equation 2.7. The line is assumed to be monochromatic, and its natural width is not taken into account. The isotropy of the fluorescence in the photoelectric process is expressed by the independence of the kernel on the direction $\overrightarrow{\omega}$ and by the 4π normalization factor.

The X-ray fluorescence emission probability density of the the single wavelength λ_i, $Q_{\lambda_i}(\lambda')$, is a quantity related to the photoelectric attenuation coefficient $\tau_s(\lambda')$ of the emitter element s, to the absorption edge jump J_{e_i} [38, 36], to the fluorescence yield ω_{e_i} and to the line emission probability Γ_{λ_i} of the line at wavelegnth λ_i into its own spectral series, by the probability relationship:

$$Q_{\lambda_i}(\lambda') = \tau_s(\lambda') \left(1 - \frac{1}{J_{e_i}}\right) \omega_{e_i}\, \Gamma_{\lambda_i} \qquad (2.8)$$

where the parameters J_{e_i} and ω_{e_i} are series dependent. They will therefore be identical for all the individual lines belonging to a particular series. The radiative fraction for a given series of transitions is commonly denoted by g_{e_i}. The fraction

of vacancies produced in the K-subshell will be filled with transitions from outer shells giving:

$$g_K = \left(1 - \frac{1}{J_K}\right) \omega_K \tag{2.9}$$

The allowed transitions for the filling of the vacancy created by the photoelectron ejection can be either radiative or radiationless. Radiative transitions lead to the characteristic fluorescence photon emission. Radiationless transitions can be of two similar types: AUGER and COSTER-KRONIG. If an electron of an external shell moves to a lower energy level to fill the vacancy, an amount of energy equal to the difference in orbital energies, i.e. from the original to the final electron shells, is lost. The transition energy can be retrieved by an outer shell electron, which is subsequently ejected from the atom if the transferred energy is greater than the orbital binding energy. This is an AUGER electron, and this process produces a double ionization of the atom, without any photon emission. During the COSTER-KRONIG process, the electron makes a transition from one subshell to a vacancy in another subshell of the same electronic shell. The small difference in binding energies may be transferred to an outer-electron, in this case called a COSTER-KRONIG electron. AUGER and COSTER-KRONIG transitions are schematically illustrated and compared with radiative transitions in Figure 2.6.

The complete emission spectrum of an element s, containing many lines, is obtained by the summation of one term such as equation 2.8 for each single line belonging to the spectrum. The complete emission spectrum is then given by:

$$k_P\left(\lambda' \rightarrow \lambda, \overrightarrow{\omega}' \rightarrow \overrightarrow{\omega}\right) = \frac{1}{4\pi} \sum_i Q_{\lambda_i}(\lambda')\, \delta\left(\lambda - \lambda'\right) \left[1 - \mathcal{H}(\lambda' - \lambda_{e_i})\right] \tag{2.10}$$

In equation 2.10, each term represents the emission after absorption of an initial radiation of wavelength λ', resulting in the emission of a single characteristic XRF line.

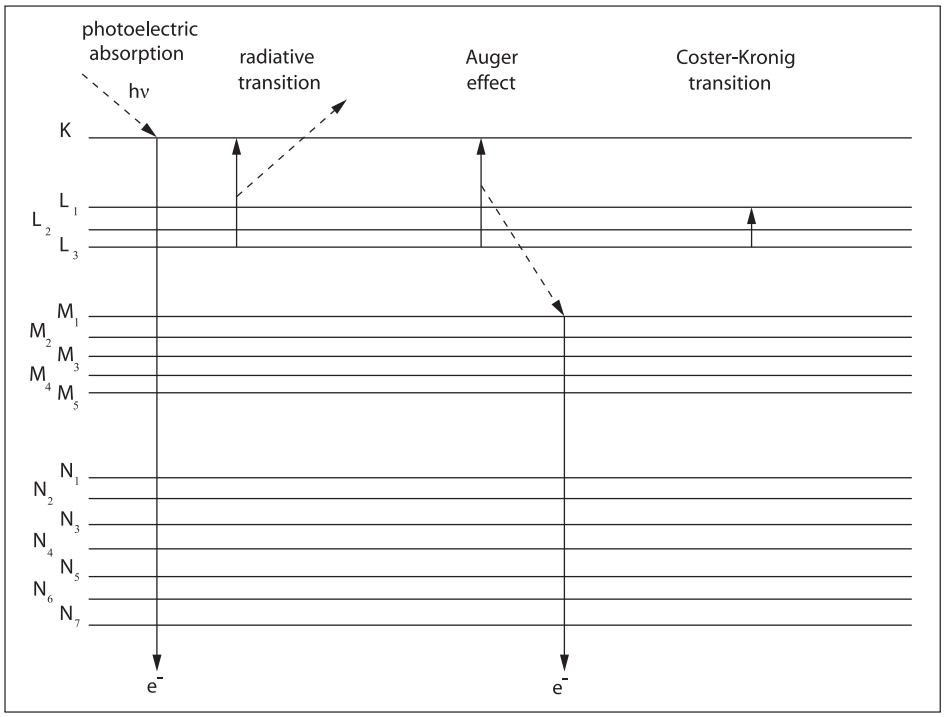

Figure 2.6: Schematic comparison between radiative, AUGER and COSTER-KRONIG transitions.

2.2 The RAYLEIGH scattering

Named for the 3^{rd} Lord RAYLEIGH, J. W. STRUTT, coherent scattering is a process where the photon changes its direction (momentum transfer) without any change of energy [39]. This scattering takes place with the more tightly bound electrons of the atom which behave rigidly during the collision, and the recoil momentum is taken up by the atom as a whole. The scattering is mainly in the forward direction, and the energy loss is then slight.

Coherent scattering has been treated in a first approximation by J. J. THOMSON, using the classical theory of radiation. The electron in the electromagnetic field of the incident radiation vibrates with the same frequency as that of the incident radiation, thereby giving rise to the emission of secondary electromagnetic radiation of the same frequency. For unpolarized radiation, the

electronic cross-section per steradian is given by [26]:

$$\left(\frac{d^2\sigma}{d\Omega\, d\lambda}\right)_{Thomson} = \frac{r_e^2}{2}\left[1 + (\overrightarrow{\omega}.\overrightarrow{\omega}')^2\right]\delta(\lambda - \lambda') \tag{2.11}$$

where r_e is the classical radius of the electron:

$$r_e = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{m_e c^2} = 2.81794\,10^{-15}\,\mathrm{m} \tag{2.12}$$

with:

- e and m_e, respectively the electric charge ($1.602\,10^{-19}$ C) and the mass of the electron ($9.109\,10^{-31}$ kg);
- c, the speed of the light;
- ε_0, the permittivity of the free space ($8.854\,10^{-14}\,\mathrm{F\,cm^{-1}}$).

The δ function in equation 2.11 expresses the monochromaticity of the scattering. In many atoms, however, a cooperative effect from all the electrons belonging to the electronic structure can be verified. Since the scattering is coherent, the amplitudes must be added before squaring to obtain the intensity. Therefore, the cross-section for electron results non-additive, and it is necessary to define an atomic differential cross-section by:

$$\begin{aligned}\left(\frac{d^2\sigma_R}{d\Omega\, d\lambda}\right)_{at.} &= F^2(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z)\left(\frac{d^2\sigma}{d\Omega\, d\lambda}\right)_{Thomson} \\ &= \frac{r_e^2}{2}\left[1 + (\overrightarrow{\omega}.\overrightarrow{\omega}')^2\right] F^2(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z)\,\delta(\lambda - \lambda')\end{aligned} \tag{2.13}$$

The square form factor $F^2(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z)$ can produce atomic contributions significantly greater than Z times the single electronic contribution. In terms of transferred momentum $\overrightarrow{q}$, the form factor for an atom containing Z electrons has been defined as the matrix element [40]:

$$F(\overrightarrow{q}, Z) = \sum_{n=1}^{Z} \langle \Psi_0 | e^{i\overrightarrow{q}.\overrightarrow{r}_n} | \Psi_0 \rangle \tag{2.14}$$

where $\overrightarrow{r}_n$ denotes the instantaneous position of the n-th electron, respectively, and Ψ_0 the ground-state wave function. By defining the momentum-transfer parameter as:

$$x = \lambda\left[\text{Å}\right]^{-1}\sin\frac{\theta}{2} \tag{2.15}$$

for given wavelength and scattering angle, form factors $F(x, Z)$ were computed for all elements of the periodic table. An exhaustive review of the form factor's computation has been done by HUBBELL *et al.* [41]. Some special limits of the form factor are $F(0, Z) = Z$ and $F(\infty, Z) = 0$. Experimental data and other tables may be found in different publications by HUBBELL and OVERBØ [42], SCHAUPP *et al.* [43], KANE *et al.* [44] or CHANTLER *et al.* [45].

2.2.1 Scalar kernel

The RAYLEIGH scalar atomic kernel for incident unpolarized photons, with space-phase coordinates $(\overrightarrow{\omega}', \lambda')$ scattered by a pure element target s of atomic number Z into the coordinates $(\overrightarrow{\omega}, \lambda)$ is described by:

$$\begin{aligned} k_R\left(\lambda' \rightarrow \lambda, \overrightarrow{\omega}' \rightarrow \overrightarrow{\omega}\right) &= \frac{\rho N Z r_e^2}{2A} \left(\frac{d^2\sigma_R}{d\Omega\, d\lambda}\right)_{at.} \\ &= \sigma\left[1 + (\overrightarrow{\omega}.\overrightarrow{\omega}')^2\right] \times \\ &\quad \frac{F^2(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z)}{Z}\, \delta\left(\lambda - \lambda'\right) \end{aligned} \tag{2.16}$$

where:

$$\sigma = \frac{\rho N Z r_e^2}{2A} \tag{2.17}$$

is the macroscopic attenuation coefficient with N, the AVOGADRO number, r_e the classical radius of the electron and A the atomic weight. The angular dependance of the scalar kernel given by equation 2.16 is due to:

- the THOMSON angular factor, representing an average polarisation state;
- the square of the atomic form factor F comprising the constructive interference from the whole charge distribution.

2.3 The COMPTON scattering

The COMPTON scattering, also known as incoherent scattering, has been discovered in 1922 by A. H. COMPTON, and described one year later in two publications [46, 47]. During the scattering, an incoming photon transfers a fraction of

its energy to an outer electron of an atom. During the interaction, the photon is deflected through an angle θ with respect to its original direction, and the electron (sometimes called recoil electron) is ejected from its orbital position with an angle θ_{e^-}, as illustrated in Figure 2.7. In normal scattering conditions, all scattering angles are possible. Therefore, the energy transferred to the electron can vary from zero to a large proportion of the initial photon energy. During this interaction mechanism, both energy and momentum of the X-ray photon are modified.

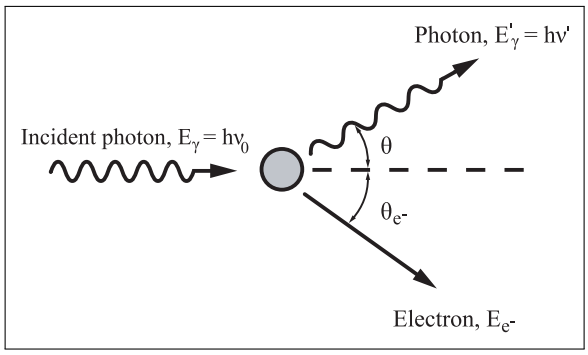

Figure 2.7: Schematic illustration of the COMPTON scattering.

In a first approximation, the atomic COMPTON cross-section for incoherent scattering may be evaluated by the product between the atomic number Z and the electronic cross-section for the scattering by a free electron. Let us consider the collision of a photon carrying an energy $h\nu'$, momentum $p' = h\nu'/c$, and with a direction $\overrightarrow{\omega}'$ against a free electron at rest (electron mass energy at rest $m_e c^2 = 511\,\text{keV}$, null momentum). Energy and momentum conservation principles during the collision establish that, if the photon scatters with a scattering angle θ, it has a wavelength:

$$\lambda = \lambda' + \lambda_C\,(1 - \cos\theta) \tag{2.18}$$

with:

- $\cos\theta = \overrightarrow{\omega}'.\overrightarrow{\omega}$;
- $\lambda_C = \dfrac{h}{m_e c} = 0.0242631\,\text{Å}$, the COMPTON wavelength.

The validity of this approximation is generally considered as a proof of the particle nature of the photon. In agreement with the approximation, scattering experiments for a narrowly defined scattering angle show a well defined peak at a higher wavelength than the incident one. The computation of the differential electronic cross-section for the described collision has been made in 1929 by O. KLEIN and Y. NISHINA, having recourse to Relativistic Quantum Mechanics theories [48]. The analytical expression for the differential cross-section for a photon in an average polarization state, based on the DIRAC's theory, is given by:

$$\frac{d^2\sigma}{d\Omega\, d\lambda} = \frac{r_e^2}{2} K_{KN}\left(\lambda, \lambda'\right) \frac{1}{\lambda_C} \delta\left(1 - \overrightarrow{\omega}.\overrightarrow{\omega}' + \frac{\lambda' - \lambda}{\lambda_C}\right) \tag{2.19}$$

where:

$$\begin{aligned} K_{KN}\left(\lambda, \lambda'\right) &= \left(\frac{\lambda'}{\lambda}\right)^2 \left[\frac{\lambda}{\lambda'} + \frac{\lambda'}{\lambda} + \frac{\lambda - \lambda'}{\lambda_C}\left(\frac{\lambda - \lambda'}{\lambda_C} - 2\right)\right] \\ &= \left(\frac{\lambda'}{\lambda}\right)^2 \left[\frac{\lambda}{\lambda'} + \frac{\lambda'}{\lambda} - \sin^2\theta\right] \end{aligned} \tag{2.20}$$

is known as the KLEIN-NISHINA function. The direction-wavelength δ in equation 2.19 fixes the integration path in the phase space along the line $(1 - \overrightarrow{\omega}.\overrightarrow{\omega}' + (\lambda' - \lambda)/\lambda_C = 0)$.

2.3.1 COMPTON kernel in the WALLER-HARTREE approximation

A departure from the KLEIN-NISHINA cross section is verified as soon as the energy of the exciting photons becomes comparable with the binding energy of the inner-shell electrons of the target. WALLER and HARTREE used non-relativistic wave mechanics to consider the effect of electron binding for the whole atom [49], laying the groundwork for extensive computations.

It is actually customary to define the WALLER-HARTREE incoherent scat-

tering function $S^{WH}(\overrightarrow{q}, Z)$ by:

$$S^{WH}(\overrightarrow{q}, Z) = \sum_{m=1}^{Z}\sum_{n=1}^{Z}\langle\Psi_0|e^{i\overrightarrow{q}.(\overrightarrow{r}_m-\overrightarrow{r}_n)}|\Psi_0\rangle - \left|\sum_{m=1}^{Z}\langle\Psi_0|e^{i\overrightarrow{q}.\overrightarrow{r}_m}|\Psi_0\rangle\right|^2 \tag{2.21}$$

where $\overrightarrow{q}$ denotes the transferred momentum during the collision, $\overrightarrow{r}_m$ and $\overrightarrow{r}_n$ the instantaneous position of the m-th and n-th electrons respectively, and Ψ_0 the ground-state wave function.

The computational techniques for obtaining the scattering function were reviewed exhaustively by Hubbell *et al.*, in 1975 [41]. By defining the transferred momentum as in equation 2.15, tables of $S^{WH}(x, Z)$ were computed for all the elements in the periodic table. Special limits of the scattering function are $S^{WH}(0, Z) = 0$ and $S^{WH}(\infty, Z) = Z$. The double differential atomic cross-section for incident photons, with phase-space coordinates $(\overrightarrow{\omega}', \lambda')$ scattered by a pure element target s of atomic number Z into the coordinates $(\overrightarrow{\omega}, \lambda)$, is expressed as:

$$\left(\frac{d^2\sigma}{d\Omega\, d\lambda}\right)_{at.}^{WH} = \frac{r_e^2}{2}K_{KN}\left(\lambda, \lambda'\right)S^{WH}(x, Z)\frac{1}{\lambda_C} \times\delta\left(1 - \overrightarrow{\omega}.\overrightarrow{\omega}' + \frac{\lambda' - \lambda}{\lambda_C}\right) \tag{2.22}$$

Therefore, the Compton kernel in the Waller-Hartree approximation is described by:

$$\begin{aligned} k_C^{WH}\left(\lambda' \to \lambda, \overrightarrow{\omega}' \to \overrightarrow{\omega}\right) &= \frac{\rho N}{A}\left(\frac{d^2\sigma}{d\Omega\, d\lambda}\right)_{at.}^{WH} \\ &= \sigma K_{KN}\left(\lambda, \lambda'\right)S^{WH}(x, Z)\frac{1}{\lambda_C} \\ &\times\delta\left(1 - \overrightarrow{\omega}.\overrightarrow{\omega}' + \frac{\lambda' - \lambda}{\lambda_C}\right) \end{aligned} \tag{2.23}$$

The statistical model of the atomic charge density developed by Thomas [50] and Fermi [51] may considerably simplify the calculation. Using this model,

known as the THOMAS-FERMI model, a simpler approximated expression for S^{WH} has been obtained by VEIGELE *et al.* [52]:

$$S^{WH}(V) = Z\left[1 - e^{-4.88V^{0.856}}\right] \tag{2.24}$$

with:

$$V = \frac{2}{3}\frac{137}{Z^{2/3}}\frac{\lambda_C}{\lambda}\sin\frac{\theta}{2} \tag{2.25}$$

More precise values of $S^{WH}(x, Z)$ can be computed using semi-empirical formulas and fitting coefficients to theoretical calculations [53].

2.3.2 COMPTON kernel in the Impulse Approximation

In the WALLER-HARTREE approximation, the pre-collision motion of the electrons has been ignored. Therefore, the kernel of equation 2.23 limits the COMPTON peak to a monochromatic line. Because of the COMPTON profile, i.e. the projection of the electron momentum distribution on the z-axis, the width of the scattered peak is larger than the instrumental width [54]. A more rigorous theoretical treatment associated with the COMPTON profile is then necessary, as given for example in [23].

Defining as $p_z = \overrightarrow{p}.\overrightarrow{q}/q$, the projection of the momentum of the interacting electron on the scattering vector $\overrightarrow{q} = \overrightarrow{k} - \overrightarrow{k}'$ where $\overrightarrow{k}$ and $\overrightarrow{k}'$ are the momenta of the scattered and incident photons, it can be demonstrated that the COMPTON shift produced by a moving electron is also a function of p_z:

$$\lambda = \lambda' + \lambda_C\left(1 - \overrightarrow{\omega}.\overrightarrow{\omega}'\right) - \frac{p_z}{mc}\sqrt{\lambda'^2 + \lambda^2 - 2\,\lambda\lambda'\,\overrightarrow{\omega}.\overrightarrow{\omega}'} \tag{2.26}$$

It is customary to use the dimensionless variable Q [55] defined as:

$$\frac{p_z}{mc} = \frac{e^2}{4\pi\varepsilon_0\hbar c}\,Q \approx \frac{Q}{137} \tag{2.27}$$

in place of p_z in equation 2.26. However, a bound electron in the atom do not hold a definite state of momentum like the one shown in 2.26, but has a momentum distribution that depends on the subshell occupied by the electron.

Denoting with an index i the subshell occupied by the electron, the COMPTON profile is related to the momentum distribution $\rho(p)$ of the scatterer before the collision through the relationship:

$$J_i(Q) = \frac{1}{2}\int_Q^\infty \rho(p) p \, dp \tag{2.28}$$

As a consequence of wave-function normalisation, the integrated profile must satisfy the normalization condition:

$$2\int_0^\infty J_i(Q)\, dQ = 1 \tag{2.29}$$

In order to deduce the COMPTON intensity in the Impulse Approximation (IA), the following relations are used:

$$\left(\frac{d\sigma}{d\Omega}\right)_{at.} = \int_0^\infty \left(\frac{d^2\sigma}{d\Omega\, d\lambda}\right)_{at.} d\lambda \tag{2.30}$$

and

$$\left(\frac{d\sigma}{d\Omega}\right)_{at.}^{WH} \simeq \left(\frac{d\sigma}{d\Omega}\right)_{at.}^{IA} \tag{2.31}$$

implying that the scattering function should be equivalent in both representations [56]:

$$S^{WH} = S^{IA} \tag{2.32}$$

From equations 2.23 and 2.29 to 2.32 we obtain:

$$\begin{aligned} \frac{\rho N}{A}\left(\frac{d\sigma}{d\Omega}\right)_{at.}^{WH} &= \sigma K_{KN}\left(\lambda_p, \lambda'\right) S^{WH}\left(\frac{1}{\lambda'}\sqrt{\frac{1-\overrightarrow{\omega}.\overrightarrow{\omega}'}{2}}, Z\right) \\ &\approx \sigma K_{KN}\left(\lambda_p, \lambda'\right) S^{IA}\left(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z\right) \end{aligned} \tag{2.33}$$

where $\lambda_p = \lambda' + \lambda_C\left(1 - \overrightarrow{\omega}.\overrightarrow{\omega}'\right)$ is the peak wavelength. Since in the Impulse Approximation the scattering function for the atom is obtained as a sum of the contributions from all the subshells, we have:

$$S^{IA}\left(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z\right) = \sum_{i=1}^{\textit{Shell Number}} n_i \int_{-\infty}^{Q_{i,\,max}} J_i(Q)\, dQ \tag{2.34}$$

where n_i is the number of electrons in the shell i, and $Q_{i,max}$ is obtained by putting $\lambda = hc/E = hc/(E' - B_i)$ (with B_i the binding energy of the subshell) in the expression:

$$Q = 137\,\frac{[\lambda' + \lambda_C\,(1 - \overrightarrow{\omega}.\overrightarrow{\omega}') - \lambda]}{\sqrt{\lambda'^2 + \lambda^2 - 2\,\lambda\lambda'\,\overrightarrow{\omega}.\overrightarrow{\omega}'}} \tag{2.35}$$

which is obtained straightforwardly from equation 2.26:

$$Q = 137\,\frac{\left[\dfrac{(E' - B_i)E'}{mc^2}\,(1 - cos\theta) - B_i\right]}{[(E' - B_i)^2 + E'^2 - 2\,(E' - B_i)\,E'\,\cos\theta]^{1/2}} \tag{2.36}$$

The integral in the right side of equation 2.34 represents the contribution of one electron in the subshell i to the scattering function. Being such a contribution upper-limited by $Q_{i,max}$, it is equivalent to integrate with a higher upper limit the subshell profile truncated at $Q_{i,max}$, i.e.:

$$\int_{-\infty}^{Q_{i,max}} J_i(Q)\,dQ = \int_{-\infty}^{\infty} J_i(Q, Q_{i,max})\,dQ \tag{2.37}$$

The sum over the occupied states in the right side of equation 2.34 can be shifted into the integral. In this way we can define the whole profile at (λ', $\overrightarrow{\omega}.\overrightarrow{\omega}'$) and Z as the overlapping of the truncated profiles of the Z electrons of the element, i.e.:

$$J_i(Q, \lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z) = \sum_{i=1}^{OCC} J_i\,[Q, Q_{i,max}] \tag{2.38}$$

Equation 2.34 can be rewritten in an other way, using a change of variable in the integral:

$$S^{IA}\left(\lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z\right) = \int_0^{\infty} J\left[Q(\lambda), \lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z\right]\,\frac{dQ}{d\lambda}\,d\lambda \tag{2.39}$$

From equations 2.26, 2.33 and 2.39 we can write the Compton kernel in the Impulse Approximation as:

$$\begin{aligned} k_C^{IA}\left(\lambda' \rightarrow \lambda, \overrightarrow{\omega}' \rightarrow \overrightarrow{\omega}\right) &= \frac{\rho N}{A}\left(\frac{d^2\sigma}{d\Omega\,d\lambda}\right)_{at.}^{IA} \\ &\simeq \sigma K_{KN}\left(\lambda_p, \lambda'\right) \times \\ &\quad J\left[Q(\lambda), \lambda', \overrightarrow{\omega}.\overrightarrow{\omega}', Z\right]\,\frac{dQ}{d\lambda} \end{aligned} \tag{2.40}$$

where:

$$\frac{dQ}{d\lambda} = -137\left(1 - \overrightarrow{\omega}.\overrightarrow{\omega}'\right)\frac{\left[\lambda'\left[\lambda' - \lambda_C(\overrightarrow{\omega}.\overrightarrow{\omega}')\right] + \lambda(\lambda' + \lambda_C)\right]}{(\lambda'^2 + \lambda' - 2\,\lambda\lambda'\,\overrightarrow{\omega}.\overrightarrow{\omega}')^{3/2}} \tag{2.41}$$

is obtained from equation 2.35. Equation 2.40 is the alternative to equation 2.23 using COMPTON profiles. Since the broadening of the COMPTON peak is considerably large, the Impulse Approximation gives a much more precise estimate of the intensity distribution of the COMPTON peak, especially in relation with spectrum build-up in the X-ray regime.

Chapter

3 Forward and adjoint Boltzmann transport equations for photons

The X-ray photon flux can be described by a scalar integro-differential transport equation, known as Boltzmann equation. In the following, the Boltzmann transport equation for photons will first be constructed in its most general formulation, and the equation related to a simplified physical beam model – an infinite thickness target irradiated with a monochromatic collimated X-ray beam – will secondly be deduced. From these mathematical expressions of the photon transport, the equation that is adjoint to the Boltzmann equation will be derived in both the general formulation and the simplified physical beam model. In the framework of photon transport theory, the adjoint transport equation has a very particular significance of importance, and will be considered as a significant help in the resolution of the inverse scattering problem previously introduced and more precisely developed in the next chapter.

3.1 The forward Boltzmann equation

An X-ray flux can be completely determined as the solution to a transport equation describing the balance between the number of photons of determined wavelength and direction entering and leaving an infinitesimal volume element. The balance may be formulated for conditions where the X-ray source is con-

stant with time and, therefore, the photon flow through the medium is also constant in time [35].

Let us denote the position of a point in a cartesian frame of reference by $\overrightarrow{r}$. Let us consider also the infinitesimal cylinder, centered in $\overrightarrow{r}$, characterized by:

- a base area dA;
- a height dl;
- a lateral surface parallel to a direction $\overrightarrow{\omega}$.

The conceptual setup of this infinitesimal cylinder is illustrated in Figure 3.1.

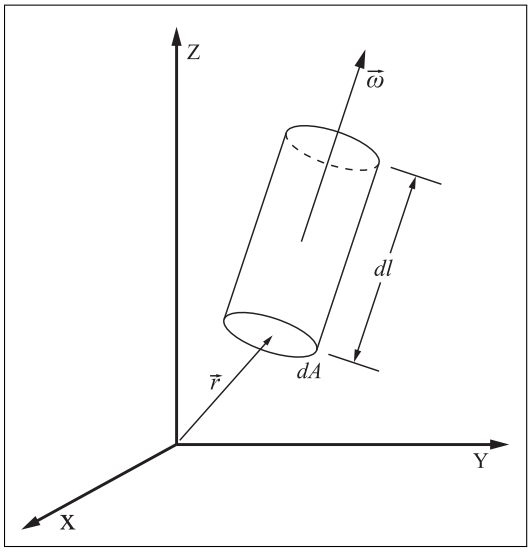

Figure 3.1: Illustration of the infinitesimal cylinder used for the construction of the BOLTZMANN transport equation.

The photon flux $f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda)\, d\lambda\, d\omega$ passing through the cylinder is defined as the number of photons with wavelength included in the interval $[\lambda, \lambda + d\lambda]$ and within a direction in the range $[\overrightarrow{\omega}, \overrightarrow{\omega} + \overrightarrow{\omega}]$, which goes through the base dA of the cylinder per unit time. The wavelength λ is used here in place of the energy E for the sake of convenience, but the use of the energy E is entirely

equivalent. The link between wavelength λ and energy E is given by:

$$\lambda = \frac{hc}{E} \tag{3.1}$$

where h is the PLANCK constant, and c the speed of light in vacuum.

The net rate of photons with specified direction and energy leaving the infinitesimal cylinder through the surface dA per unit time is given by:

$$f(\overrightarrow{r} + \overrightarrow{\omega} dl, \overrightarrow{\omega}, \lambda)\, dA - f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda) dA \tag{3.2}$$

or, under differential form:

$$\overrightarrow{\omega}.f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda)\, dA\, dl \tag{3.3}$$

Three factors contribute to this net outflow [35]:

1. the narrow-beam attenuation in the whole volume of the cylinder. This attenuation is ruled by the well-known BEER-LAMBERT law (one-dimensional exponential attenuation law), and is expressed by:

   $$-\mu(\lambda)\, f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda)\, dA\, dl \tag{3.4}$$

   where $\mu(\lambda)$ is the total attenuation coefficient of the sample. This coefficient is strongly wavelength dependent, but completely independent of the geometry. The coefficient is linear, i.e. we assume that, after absorption, the photon is simply removed from the beam.

2. the photon scattering in the cylinder, from wavelength λ' and direction ω' to wavelength λ and direction ω. If the scattering happens in the volume of the cylinder, it contributes to a positive outflow from the cylinder. The term of scattering has here to be considered in its largest sense, making reference to any atomic process sparking off a photon of a determined wavelength and direction into an other photon with different (or eventually equal) wavelength and direction. This second factor depends on the product of the flow $f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda)$ by the probability distribution function

$k\,(\lambda' \to \lambda, \overrightarrow{\omega}' \to \overrightarrow{\omega})$, i.e. the probability that a photon with initial wavelength λ' and direction $\overrightarrow{\omega}'$ has the new set λ and $\overrightarrow{\omega}$ per unit path through the medium and per unit $d\overrightarrow{\omega}$ and $d\lambda$ after the scattering. The whole scattering contribution can be obtained after integrating the product with respect to the direction $\overrightarrow{\omega}'$ and the wavelength λ'. The entire scattering contributions is given by:

$$\int_0^\infty \int_{4\,\pi} k\left(\lambda' \to \lambda, \overrightarrow{\omega}' \to \overrightarrow{\omega}\right) f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda)\, d\overrightarrow{\omega}'\, d\lambda' \tag{3.5}$$

Note that the scattering probability depends only upon the scattering angle, i.e. the scalar product $\overrightarrow{\omega}'.\overrightarrow{\omega}$, and not upon the directions before and after the event $\overrightarrow{\omega}'$ and $\overrightarrow{\omega}$.

3. the production rate, i.e. the source contribution of photons with given wavelength and direction within the infinitesimal cylinder previously defined. This contribution can be denoted by:

$$S\,(\overrightarrow{r}, \overrightarrow{\omega}, \lambda) \tag{3.6}$$

where S is a general source term given per unit volume, time, wavelength and per steradian. This factor is of course not necessarily present in all situations.

The expression of the photon transport equation is finally obtained by equating the net rate of photons passing through the surface dA per unit time (equation 3.3) with the three factors (equations 3.4, 3.5 and 3.6) contributing to this net outflow:

$$\begin{aligned} \overrightarrow{\omega}.\nabla f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda) &= -\,\mu(\lambda)\, f(\overrightarrow{r}, \overrightarrow{\omega}, \lambda) \\ &+ \int_0^\infty \int_{4\,\pi} k\left(\lambda' \to \lambda, \overrightarrow{\omega}' \to \overrightarrow{\omega}\right) f(\overrightarrow{r}, \overrightarrow{\omega}', \lambda')\, d\overrightarrow{\omega}'\, d\lambda' \\ &+ S(\overrightarrow{r}, \overrightarrow{\omega}, \lambda) \end{aligned} \tag{3.7}$$

Equation 3.7 is a very general formulation of the transport equation, containing all the possible aspects of the photon transport. The functions μ, k and S depend on the process taken into consideration, on the range of the variables

which are of interest and on the required degree of precision. The mathematical difficulty in obtaining a complete analytical solution to the integro-differential equation is in most cases extremely high. If the shapes of the functions μ, k and S do not allow significant simplifications that make the calculation of an analytical solution possible, the only way to solve it is to use approximate numerical methods. It is worth noting that, although this equation is a very general transport situation, the subtle assumption of an infinite homogeneous space is still present in it [35].

For the sake of simplicity, equation 3.7 may also be written in the following condensed way:

$$\mathbf{L}f = S \tag{3.8}$$

where f is the forward angular flux, S the forward source distribution and $\mathbf{L}$ the integro-differential operator defined by [57]:

$$\begin{aligned} \mathbf{L}f &= \overrightarrow{\omega}.\nabla f(\overrightarrow{r},\overrightarrow{\omega},\lambda) + \mu(\lambda)\, f(\overrightarrow{r},\overrightarrow{\omega},\lambda) \\ &- \int_0^\infty \int_{4\,\pi} k\left(\lambda' \rightarrow \lambda, \overrightarrow{\omega}' \rightarrow \overrightarrow{\omega}\right) f(\overrightarrow{r},\overrightarrow{\omega}',\lambda')\, d\overrightarrow{\omega}' d\lambda' \end{aligned} \tag{3.9}$$

The operator $\mathbf{L}$ is not symmetric due to both the first-order differential operator in the streaming term and the energy dependance of the integral operator kernel.

3.2 The adjoint BOLTZMANN transport equation

In this section, consideration will be given to the equation which is adjoint to the photon transport equation. The solutions to the adjoint transport equation are orthogonal to the ones of the forward transport equation. Moreover, the former has a clear significance of photon 'importance' within a particular system [58, 59], as it will be explained later.

3.2.1 The adjoint function

The first step in this development is to define certain quantities which will be used in the following. Let $\varphi(\xi)$ and $\psi(\xi)$ be two functions of the same variable,

represented by the same general symbol ξ. The inner product of these two functions is then expressed and defined by:

$$(\varphi, \psi) \equiv \int \varphi(\xi)\psi(\xi)\, d\xi \tag{3.10}$$

where the integration is carried over the whole accessible range of variables. If φ and ψ are two acceptable and well-behaved functions, in the sense that they satisfy certain boundary and smoothness conditions, then a hermitian (or self-adjoint) operator $\mathbf{M}$ is one for which the inner products $(\psi, \mathbf{M}\varphi)$ and $(\varphi, \mathbf{M}\psi)$ are equal, i.e.:

$$(\psi, \mathbf{M}\varphi) = (\varphi, \mathbf{M}\psi) \tag{3.11}$$

The eigenfunctions of hermitian operators are orthogonal, and the eigenvalues are always real.

In quantum mechanics for example, operators representing physical quantities are hermitian and they operate on the wave functions. Both the operators and the wave functions in quantum mechanics are complex, and so complex conjugates are used in defining the inner product. In the treatment of photon transport theory, the operators and the functions on which they operate are real. Complex conjugates are therefore not required. However, the operator associated with the transport equation is not self-adjoint.

If the operator $\mathbf{L}$ is no self-adjoint, it is possible to define an operator $\mathbf{L}^\dagger$ that is adjoint to $\mathbf{L}$. The operator $\mathbf{L}^\dagger$ will operate on functions $\psi^\dagger$, often called adjoint functions, which may satisfy boundary conditions different from those satisfied by the functions φ on which $\mathbf{L}$ operates. The adjoint operator, $\mathbf{L}^\dagger$, is naturally defined by the requirement that:

$$(\psi^\dagger, \mathbf{L}\varphi) = (\varphi, \mathbf{L}^\dagger\psi^\dagger) \tag{3.12}$$

for any acceptable functions φ and $\psi^\dagger$. The eigenfunctions of the adjoint operator, $\mathbf{L}^\dagger$, are then orthogonal to those of the operator $\mathbf{L}$.

3.2.2 The adjoint to the transport operator

Since in this work $\mathbf{L}$ will operate on the forward photon angular flux f, the adjoint operator $\mathbf{L}^\dagger$ will be defined by the requirement that:

$$(f^\dagger, \mathbf{L}f) = (f, \mathbf{L}^\dagger f^\dagger) \tag{3.13}$$

where $f^\dagger$ is referred to as the adjoint angular flux or as the adjoint function. The functions f and $f^\dagger$ are two functions satisfying appropriate boundary and continuity conditions for the angular flux and its adjoint, respectively.

In accordance with the definition in equation 3.13, the adjoint transport operator $\mathbf{L}^\dagger$ is given by [57]:

$$\begin{aligned} \mathbf{L}^\dagger f^\dagger &= -\,\overrightarrow{\omega}.\nabla f^\dagger(\overrightarrow{r},\overrightarrow{\omega},\lambda) + \mu(\lambda)\, f^\dagger(\overrightarrow{r},\overrightarrow{\omega},\lambda) \\ &- \int_0^\infty \int_{4\,\pi} k\left(\lambda \rightarrow \lambda', \overrightarrow{\omega} \rightarrow \overrightarrow{\omega}'\right)\, f^\dagger(\overrightarrow{r},\overrightarrow{\omega}',\lambda')\, d\overrightarrow{\omega}' d\lambda' \end{aligned} \tag{3.14}$$

The following differences should be noted between $\mathbf{L}^\dagger$ as given by equation 3.14 and $\mathbf{L}$ as defined by equation 3.9:

- the gradient terms have opposite signs;
- the initial and final states in the interaction kernel within the scattering term have been interchanged, i.e. $(\lambda', \overrightarrow{\omega}') \rightarrow (\lambda, \overrightarrow{\omega})$ in $\mathbf{L}$ is replaced by $(\lambda, \overrightarrow{\omega}) \rightarrow (\lambda', \overrightarrow{\omega}')$ in $\mathbf{L}^\dagger$.

The adjoint BOLTZMANN transport equation may then be written as:

$$\begin{aligned} -\,\overrightarrow{\omega}.\nabla f^\dagger(\overrightarrow{r},\overrightarrow{\omega},\lambda) &= -\,\mu(\lambda)\, f^\dagger(\overrightarrow{r},\overrightarrow{\omega},\lambda) \\ &+ \int_0^\infty \int_{4\,\pi} k\left(\lambda \rightarrow \lambda', \overrightarrow{\omega} \rightarrow \overrightarrow{\omega}'\right) f^\dagger(\overrightarrow{r},\overrightarrow{\omega}',\lambda')\, d\overrightarrow{\omega}'\, d\lambda' \\ &+ S^\dagger(\overrightarrow{r},\overrightarrow{\omega},\lambda) \end{aligned} \tag{3.15}$$

or in its condensed form:

$$\mathbf{L}^\dagger f^\dagger = S^\dagger \tag{3.16}$$

where $f^\dagger$ is the adjoint angular flux, $S^\dagger$ the adjoint source distribution and $\mathbf{L}^\dagger$ the adjoint operator defined by equation 3.14.

Since their integral responses are equal, i.e.:

$$\left(f^\dagger, S\right) = \left(f, S^\dagger\right) \tag{3.17}$$

linear neutral particle transport can then be described either by the adjoint transport equation 3.16 or by the forward transport equation 3.8. Both description are the identical, even if a different point of view is adopted for the description.

3.3 Forward and adjoint transport equations in the monochromatic beam model

Let us consider now the simple backscattering model for a plane monochromatic X-ray source on an infinitely thick sample, as shown in Figure 3.2.

This model assumes that photons only interact in the target, i.e. the photons escaping towards the empty half-space may only be subjected to a total absorption, and cannot be sent back to the target. It represents fairly well the radiation behavior in two media of different density (the density of the sample being much greater than the density of the surrounding environment). The scalar BOLTZMANN equation for this model is [35]:

$$\begin{aligned} \eta \frac{\partial f(z, \overrightarrow{\omega}, \lambda)}{\partial z} &= -\mu(\lambda)\, f(z, \overrightarrow{\omega}, \lambda) \\ &+ \int_0^\infty \int_{4\,\pi} k\left(\lambda' \rightarrow \lambda, \overrightarrow{\omega}' \rightarrow \overrightarrow{\omega}\right) \mathcal{H}(z)\, f(z, \overrightarrow{\omega}', \lambda')\, d\overrightarrow{\omega}'\, d\lambda' \\ &+ I_0\, \delta(z)\, \delta\left(\overrightarrow{\omega} - \overrightarrow{\omega}'\right)\, \delta\left(\lambda - \lambda_0\right) \end{aligned} \tag{3.18}$$

where η is used for the directional cosine ω_z, $d\omega' = d\eta' d\phi'$ is the differential of solid angle in the direction of the unitary vector $\overrightarrow{\omega}'$ and $\mathcal{H}(z)$ is the unitary step HEAVISIDE function. The term $I_0\, \delta(z)\, \delta(\overrightarrow{\omega} - \overrightarrow{\omega}_0)\, \delta(\lambda - \lambda_0)$ represents a plane slant source of intensity I_0 (photons/cm^2.s) uniformly distributed and producing

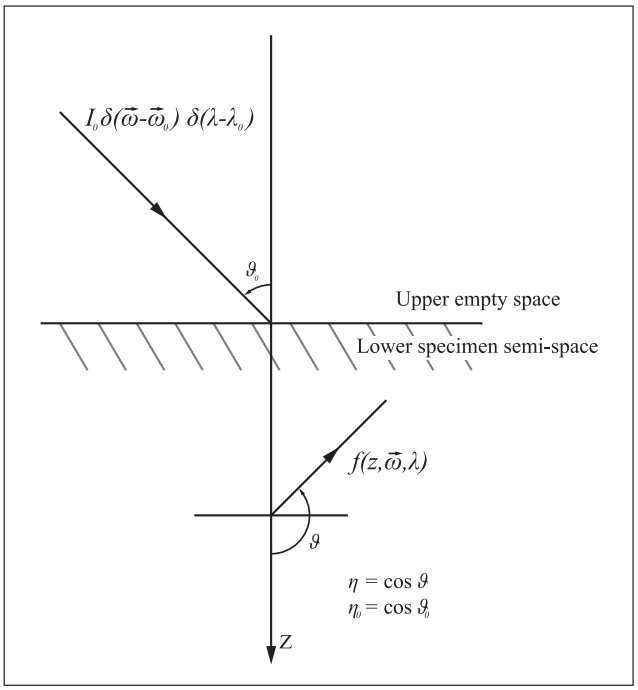

Figure 3.2: Photon backscattering model for a plane monochromatic X-ray beam in a homogeneous infinitely thick specimen.

an incident beam of parallel rays with flight direction $\overrightarrow{\omega}'$ and wavelength λ', hitting the infinite sample surface at $z = 0$.

Although the scalar transport equation 3.18 is one-dimensional in the space coordinates, the flux maintains all the angular information through its dependence on $\overrightarrow{\omega}$. According to the model, the empty semi-space in equation 3.18 is figured through its non-restitution property, rather than by a change in the density or in the absorption coefficient. This choice allows us to consider the attenuation coefficient $\mu(\lambda)$ independent from $\overrightarrow{r}$ in the transport equation.

Thanks to the linearity of equation 3.18, analytical solutions may be obtained by computing separately the contributions $f^k(\overrightarrow{r}, \overrightarrow{\omega}, \lambda)$ for different orders of collisions ($k = 1, 2, \ldots, n$). The complete solution is obtained by adding all the individual scattering contributions corresponding to each order of collision.

The forward transport operator $\mathbf{L}$ in this model of an infinitely thick target excited with a monochromatic and collimated X-ray beam becomes:

$$\begin{aligned}\mathbf{L}f(z,\overrightarrow{\omega},\lambda) &= \eta\,\frac{\partial f(z,\overrightarrow{\omega},\lambda)}{\partial z} + \mu(\lambda)\,f(z,\overrightarrow{\omega},\lambda)\\ &- \int_0^{\infty}\int_{4\,\pi} k\left(\lambda'\rightarrow\lambda,\overrightarrow{\omega}'\rightarrow\overrightarrow{\omega}\right)\mathcal{H}(z)\\ &\times\quad f(z,\overrightarrow{\omega}',\lambda')\,d\overrightarrow{\omega}'\,d\lambda'\end{aligned} \tag{3.19}$$

with a source density S (photons/cm^2.s) given by:

$$S\left(z,\overrightarrow{\omega},\lambda\right) = \delta(z)\,\delta\left(\overrightarrow{\omega}-\overrightarrow{\omega}_0\right)\,\delta\left(\lambda-\lambda_0\right) \tag{3.20}$$

The adjoint operator $\mathbf{L}^{\dagger}$ is therefore given by:

$$\begin{aligned}\mathbf{L}^{\dagger}f^{\dagger}(z,\overrightarrow{\omega},\lambda) &= -\eta\,\frac{\partial f^{\dagger}(z,\overrightarrow{\omega},\lambda)}{\partial z} + \mu(\lambda)\,f^{\dagger}(z,\overrightarrow{\omega},\lambda)\\ &- \int_0^{\infty}\int_{4\,\pi} k\left(\lambda\rightarrow\lambda',\overrightarrow{\omega}\rightarrow\overrightarrow{\omega}'\right)\mathcal{H}(z)\\ &\times\quad f^{\dagger}(z,\overrightarrow{\omega}',\lambda')\,d\overrightarrow{\omega}'\,d\lambda'\end{aligned} \tag{3.21}$$

As already mentioned in the general model, the gradient terms have opposite signs, and the initial / final states in the interaction kernel within the scattering terms are interchanged.

Let us assume now a point detector looking at the photons with direction $\overrightarrow{\omega}_1$ and wavelength λ_1. Let us suppose also a unitary adjoint source expressed by:

$$S^{\dagger}\left(z,\overrightarrow{\omega},\lambda\right) = \delta(z)\,\delta(\overrightarrow{\omega}-\overrightarrow{\omega}_1)\,\delta(\lambda-\lambda_1) \tag{3.22}$$

The sources defined by equation 3.20 and 3.22 may be replaced in equation 3.17. From this operation results the following equality:

$$f^{\dagger}\left(0,\overrightarrow{\omega}_0,\lambda_0;\overrightarrow{\omega}_1,\lambda_1\right) = f\left(0,\overrightarrow{\omega}_1,\lambda_1;\overrightarrow{\omega}_0,\lambda_0\right) \tag{3.23}$$

Equation 3.23 stresses the equivalence between the forward albedo angular flux at $(\overrightarrow{\omega}_1,\lambda_1)$ produced by a unitary source at $(\overrightarrow{\omega}_0,\lambda_0)$, and the adjoint albedo angular flux at $(\overrightarrow{\omega}_0,\lambda_0)$ produced by an analogous source at $(\overrightarrow{\omega}_1,\lambda_1)$.

The adjoint flux $f^{\dagger}(0, \overrightarrow{\omega}_0, \lambda_0)$ then has a meaning of importance with which a beam of λ_0-wavelength monochromatic photons, oriented along $\overrightarrow{\omega}_0$, contributes to a reading of one photon/cm.s in a counter placed at $z = 0$ and sensitive to photons with phase-space variables $\overrightarrow{\omega}_1$ and λ_1. Equation 3.23 allows immediate knowledge of the importance function to be obtained once the albedo angular flux is known.

3.4 Discretization of the forward and adjoint transport equations for numerical calculations

For a unitary monochromatic excitation of wavelength λ_k, the forward transport equation 3.8 becomes:

$$\mathbf{L} f_{(k)} = \delta(\lambda - \lambda_k) \tag{3.24}$$

For all practical purposes, the continuous transport equation should be discretized in order to allow its computation. In the discrete monochromatic problem, the function $f_{(k)}$ becomes:

$$f_{(k)} = (f_1, \ldots, f_n)^T \tag{3.25}$$

and the monochromatic source may be defined by:

$$\overrightarrow{s}_{(k)} = (0, \ldots, \underbrace{1}_{k}, \ldots, 0)^T \tag{3.26}$$

Since the BOLTZMANN transport equation is linear, a polychromatic source can be expressed as the linear overlapping of monochromatic unitary excitations $\overrightarrow{s}_{(k)}$, weighted by coefficients $\alpha_{(k)}$:

$$\overrightarrow{s} = \sum_k \alpha_k \overrightarrow{s}_{(k)} \tag{3.27}$$

The corresponding scattering vector is:

$$\overrightarrow{f} = \sum_k \alpha_k \overrightarrow{f}_{(k)} \tag{3.28}$$

Finally, for polychromatic excitation, the forward discretized system is given by:

$$F\overrightarrow{s} = \overrightarrow{f} \tag{3.29}$$

where F is the discrete forward matrix. It is important to note that the k-th column of the forward matrix F represents the k-th discrete solution to the forward problem corresponding to a unitary monochromatic excitation $\overrightarrow{s}_{(k)}$. The size of the matrix F is strongly linked to the discretization of the scattered and source vectors. For the sake of simplicity, the same even discretization for both vectors will be considered in the following.

For a unitary monochromatic reading at the wavelength λ_k, the adjoint transport equation 3.16 becomes:

$$\mathbf{L}^{\dagger} f^{\dagger}_{(k)} = \delta\left(\lambda - \lambda_k\right) \tag{3.30}$$

Using the notation introduced above, the adjoint discretized system is:

$$F^{\dagger}\overrightarrow{s}^{\dagger} = \overrightarrow{f}^{\dagger} \tag{3.31}$$

where $\overrightarrow{f}^{\dagger}$ represents the discretized importance, $\overrightarrow{s}^{\dagger}$ is the discretized reading and $F^{\dagger}$ is the discrete adjoint matrix. Again, it is worthwhile noting that the k-th column of the adjoint matrix $F^{\dagger}$ represents the k-th discrete solution to the adjoint problem, which corresponds to a unitary monochromatic reading $\overrightarrow{s}^{\dagger}$.

Broadly speaking, the adjoint matrix of an $n \times m$ matrix F is the $m \times n$ matrix $F^{\dagger}$ obtained by taking the transpose and then the complexe conjugate of each entries of the matrix. In our situation, the discrete adjoint matrix simply corresponds to the transpose of the discrete forward matrix:

$$F^{\dagger} = F^{T} \tag{3.32}$$

Chapter 4

The complete inverse calculation strategy

In this chapter, the full inverse strategy for reconstructing the unknown X-ray source spectrum is designed. A general description of a forward measurement carried out when using a spectrometer is first given. In this description, the photon course is voluntarily divided into two well-marked parts. This division is related to the physical phenomena occurring during the photon transport: the scattering on the solid target inside the spectrometer, and the detection of the scattered photon beam. From these considerations, an inverse calculation strategy respecting the physics of the problem is proposed. The inverse problem is then also divided into two successive physical steps: the cleaning of the measurement from the detector influence, and the inverse scattering on the target.

4.1 Description of the forward measurement procedure

The complete forward procedure for the measurement of an X-ray source spectrum by using a Compton spectrometer may be broken down in two major distinguishable physical steps.

Photons are initially emitted from the X-ray tube, and their starting energy distribution forms the *source vector* $\overrightarrow{s}$. They secondly hit a solid target of light material interposed at $45\,^{\circ}$ in the path of the primary photon beam,

undergoing scattering on it, through RAYLEIGH (elastic scattering) and COMPTON (inelastic scattering) interactions with the outer electrons of the atoms. The photons scattered at a $(90 \pm 1.5)^\circ$ angle with respect to the initial photon beam axis are selected by passing through the spectrometer. Photons that are not scattered within this particular direction are assumed to be completely absorbed in the spectrometer. They consequently completely disappear, having no influence on the detection process. The resulting secondary photon beam description, just after the scattering on the target, is called in the following *scattered vector* $\overrightarrow{b}$.

The scattered photon beam passes through the spectrometer and ends its flight by hitting a detector placed inside the spectrometer. Inside the detector, the scattered photons undergo different kinds of fundamental interactions leading to their detection. During the detection process, the energy distribution of the scattered beam suffer modifications in two different ways. The first one is due to the detector's inherent physical response to an external excitation. The second modification arises from the statistical uncertainty associated to the detection process, leading to a global broadening of the detector response. Both these effects are included in the so-called detector response function. The final pulse height distribution results from the convolution of the incident energy distribution hitting the detector, i.e. the scattered vector, with the characteristic response function of the detector. This pulse height distribution is the only measurable information of the whole process, and is referred to as *measured vector* $\overrightarrow{m}$ in the following.

This description of the complete forward measurement procedure is schematically illustrated in Figure 4.1. A complete cross-section view of a COMPTON spectrometer is given in Figure 4.2. The different vectors of interest are situated on the schema, and the different materials are symbolized by a color code (see caption for the color code description).

Using a COMPTON spectrometer for the measurement of an X-ray tube, the measured vector $\overrightarrow{m}$ does not correctly reproduce the X-ray source vector

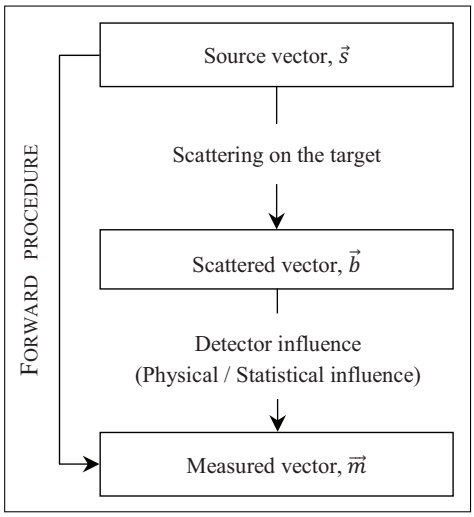

Figure 4.1: Schematic view of the forward measurement procedure, including the specific notations of the source, scattered and measured vectors.

$\overrightarrow{s}$ because of a multitude of physical phenomena occurring during the photon transport: scattering on the target, detector influence and the influence of the surrounding environment.The measured spectrum then presents a lack of information, and is not convenient to fully characterize the source vector.

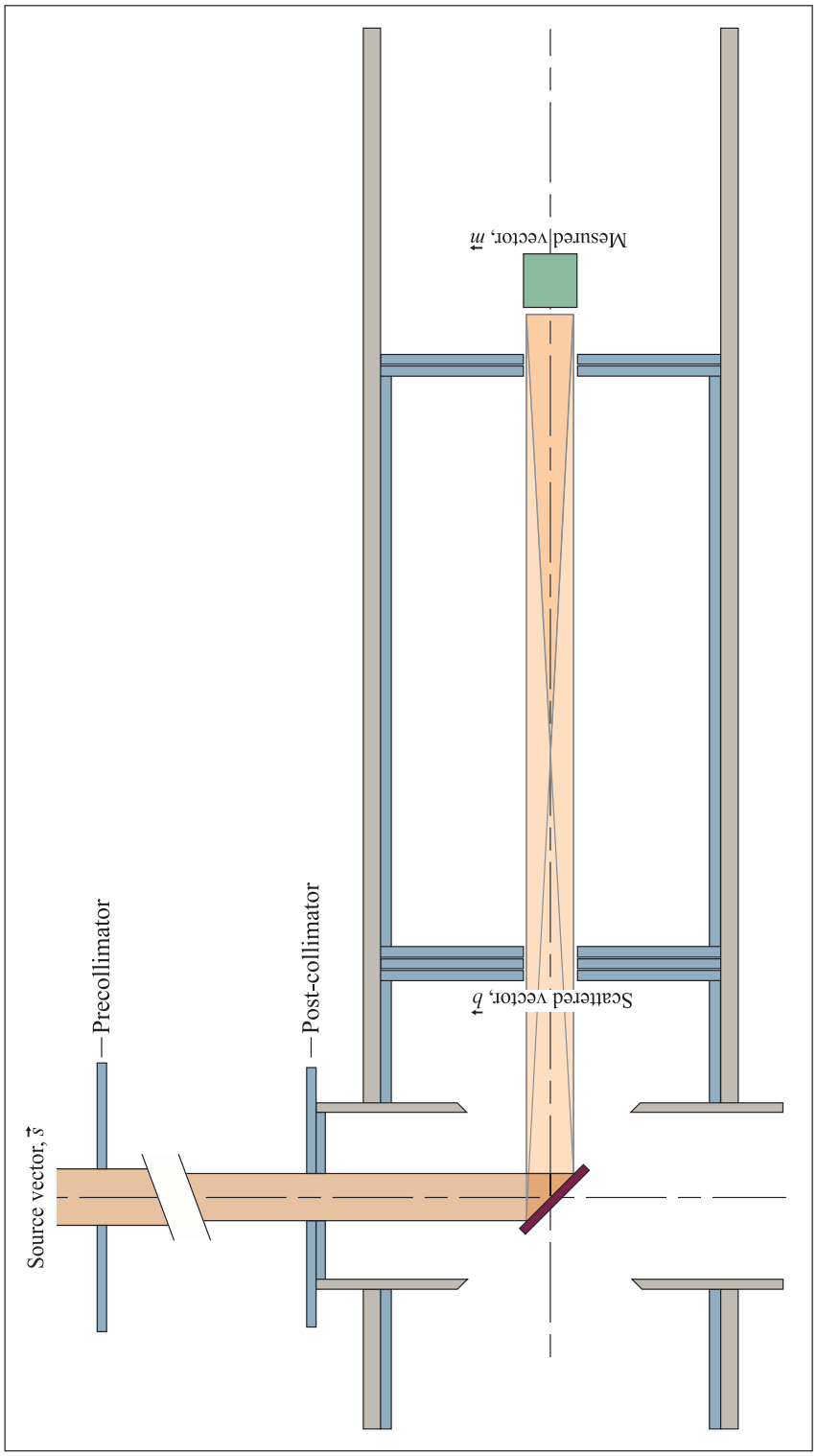

Figure 4.2: Cross section view of a COMPTON spectrometer, with indication of the source vector $\vec{s}$, the scattered vector $\vec{b}$ and the measured vector $\vec{m}$. The materials are represented by a color code: blue for the absorbing layer of lead, grey for the external polymer envelop, dark red for the solid scattering target and green for the HPGe detector.

4.2 Description of the complete inverse procedure

In order to fully and accurately characterize the X-ray source spectrum, it is fundamental to find a way to reconstruct the source spectrum starting from the measurement. The inverse procedure described in the following aims to recover the source vector $\overrightarrow{s}$ starting from the measured vector $\overrightarrow{m}$, by reversing the photon course.

Similarly to the division in physical steps of the forward measurement procedure, the inverse procedure may also be divided in two different successive stages:

- the cleaning of the measured spectrum from both physical and statistical detector influences, i.e. from the detector response functions. The aim of this step is to obtain the scattered vector $\overrightarrow{b}$ starting from the measured vector $\overrightarrow{m}$;
- the reverse scattering of the photons on the target, through the resolution of the linear system of equations that models the scattering process. Starting from the scattered vector $\overrightarrow{b}$, the result of this second step is the reconstructed source vector $\overrightarrow{s}$.

First, the cleaning of the measured spectrum from the detector influences is typically an ill-posed problem. The concept of ill-posed problem will be rigorously explained in section 5. It is however interesting to already introduce them as a class of mathematical problems whose solutions are potentially extremely sensitive to small variations in the data of the problem. Successful solutions of inverse ill-posed problems then require particular methods and specially designed algorithms that can support errors in the measured data, in order to circumvent the ill-posedness of the problem. Very often, these methods involve including additional assumptions to the initial problem, such as the smoothness, the positivity or the minimal entropy of the expected solution. The unfolding problem (and its resolution mechanisms) will be discussed in Chapter 6.

Secondly the inverse scattering on the target may also lead to a highly ill-posed system of equations, due to the ill-conditioning of the coefficient matrix. For numerical treatment, this part of the problem has often to be reformulated in a way allowing a reduction of the ill-conditioning (cf. section 5.2), by preconditioning the matrix system of equations. Many preconditioners may be chosen in order to reduce ill-conditioned character of the coefficient matrix, and to dampen the ill-posed character of the system, making it more suitable to support numerical operations. The most classical preconditioners are probably the diagonal or the incomplete lower-upper (ILU) factorization of the coefficient matrix. However, among the very large amount of possibilities of system preconditioning, our choice is to use the adjoint scattering matrix presented in Chapter 3, because of its very strong conceptual sense related to the physics of the problem.

Chapter

5 The concept of ill-posed problem

Numerous inverse problems arising in many physical fields are ill-posed. In most cases, the ill-posedness of a problem is a consequence of the ill-conditioning of the coefficient matrix (matrix of the coefficients of the variables in a set of linear equations). Both these concepts of ill-posedness and ill-conditioning will be carefully discussed in the next sections, since they occupy a central place in the resolution of inverse problems. A major and powerful numerical tool - namely, the singular value decomposition - for the analysis of ill-conditioned systems of equations is also explained in details. Using this numerical tool, the anatomy of an ill-posed problem may be investigated, revealing all the difficulties associated with the resolution of the matrix system of equations, and giving useful information about its stability with respect to successive numerical operations. For completing the chapter, this fundamental concept of stability will be defined by using the vector and matrix norms definitions.

In the beginning of 20^{th} century, the concept of ill-posed problem has been defined by Hadamard as a problem whose solution is not unique, or if it is not a continuous function of the data. For Hadamard, as for many mathematicians, ill-posed problems were mainly artificial, in the sense that they do not correctly describe a real physical system. He was wrong, however, and today the multiplicity of situations arising in many fields of science and engineering that can be described by ill-posed problems has generated a vast amount of literature (see, for instance, [60] for ill-posed problems in astronomy,

[61] in medical physics, [62] in wave study or [63] in meteorology).

A typical example of ill-posed problem is the FREDHOLM equation of the first kind, with a square integrable kernel [64]. This equation is a convolution equation, as the one expressing the detection processes by a detector. This equations may be written as:

$$g(s) = \int_a^b K(s,t) f(t)\, dt, \quad c \leq s \leq d \tag{5.1}$$

where the left-hand side $g(s)$ and the kernel $K(s,t)$ are given, and where $f(t)$ is the unknown solution to the equation. If the solution is disturbed by a small variation as, for example:

$$\Delta f(t) = \varepsilon \, \sin(2\pi p t), \quad p = 1, 2, \ldots \tag{5.2}$$

the corresponding perturbation of the right-hand side $g(s)$ is given by:

$$\Delta g(s) = \varepsilon \int_a^b K(s,t) \, \sin(2\pi p t)\, dt, \quad p = 1, 2, \ldots \tag{5.3}$$

and, due to the RIEMANN-LEBESGUE theorem, it follows that $\Delta g(s) \to 0$ as $p \to \infty$ [64]. Hence, the ratio $\|\Delta f\|/\|\Delta g\|$ becomes arbitrarily large by choosing p large enough. This shows that 5.1 is an ill-posed problem.

Strictly speaking, ill-posed linear problems must be infinite-dimensional, otherwise the ratio $\|\Delta f\|/\|\Delta g\|$ of course stays bounded, although it may become very large. However, certain finite-dimensional discrete problems have properties very similar to those of ill-posed problems, such as being highly sensitive to high-frequency perturbations. It is consequently natural to associate the term discrete ill-posed problems also with this kind of situations [65].

It is worthwhile noting that the ill-posedness of the problem does not signify the non-existence of a meaningful approximate solution. Rather, it implies that the standard methods frequently used in numerical linear algebra cannot be applied straightforwardly to compute such a solution [16]. Instead, more sophisticated and powerful computation methods must be applied in order to ensure physically meaningful solutions.

Before going further into details about some particular methods for solving ill-posed problems, it is first appropriate to introduce the most important and convenient numerical tool for the analysis of discrete ill-posed problems: the singular value decomposition.

5.1 Ill-posed problem analysis tool: the singular value decomposition

The singular value decomposition (SVD) [16, 66, 67, 68] of the coefficient matrix is one of the major numerical tool for discrete ill-posed problem analysis. This structural analysis of the discrete coefficient matrix may help in revealing all the difficulties associated with the ill-conditioning of the coefficient matrix (and then with the probable ill-posed character of the problem), giving information about the stability of the reconstructed solution.

Let us define a rectangular matrix $F \in \Re^{m \times n}$, with $m \geq n$, which maps vectors in $\Re^n$ to vectors in $\Re^m$. The singular value decomposition of F is a decomposition of the form:

$$F = U \Sigma V^T = \sum \overrightarrow{u}_i \sigma_i \overrightarrow{v}_i^T \tag{5.4}$$

where $U = (\overrightarrow{u}_1, \overrightarrow{u}_2, \ldots, \overrightarrow{u}_m) \in \Re^{m \times m}$ and $V = (\overrightarrow{v}_1, \overrightarrow{v}_2, \ldots, \overrightarrow{v}_n) \in \Re^{n \times n}$ are matrices with orthonormal columns, i.e. $U^T U = U U^T = I_m$ and $V^T V = V V^T = I_n$ (with I, the identity matrix), and where $\Sigma = diag\,(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \Re^{m \times n}$ with $p = min(m, n)$, has non-negative diagonal elements appearing in non-increasing order such that:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0 \tag{5.5}$$

The scalar quantities σ_i are called the singular values of the matrix F, while the vectors $\overrightarrow{u}_i$ and $\overrightarrow{v}_i$ are known as the i-th left and i-th right singular vectors of F, respectively. Under the conditions expressed by equations 5.4 and 5.5, the singular values matrix Σ is uniquely determined for a given matrix F, except

for singular vectors associated with multiple singular values. This uniqueness is not observable for the U and V matrix.

In connection with discrete ill-posed problems, two main characteristic features of the singular value decomposition are very often found:

- the singular values σ_i decay gradually to zero with no particular gap in the spectrum. An increase of the dimensions of the matrix F will only increase the number of small singular values;
- the left and right singular vectors $\overrightarrow{u}_i$ and $\overrightarrow{v}_i$ tend to have more sign changes in their elements as the index i increases, i.e. as the singular values σ_i decrease.

Although these characteristics are found in many discrete ill-posed problems arising in practical applications, they are unfortunately very difficult – or perhaps impossible – to prove in general [65].

The singular value decomposition also gives important insight into the smoothing effect typically associated with a square integrable kernel. As σ_i decreases, the singular vectors $\overrightarrow{u}_i$ and $\overrightarrow{v}_i$ become more and more oscillatory. Consider now the mapping $F\overrightarrow{x}$ of an arbitrary vector $\overrightarrow{x}$. Using the singular value decomposition, we get:

$$\overrightarrow{x} = \sum_{i=1}^{n} (\overrightarrow{v}_i^T \overrightarrow{x})\, \overrightarrow{v}_i \tag{5.6}$$

and

$$F\overrightarrow{x} = \sum_{i=1}^{n} \sigma_i (\overrightarrow{v}_i^T \overrightarrow{x})\, \overrightarrow{u}_i \tag{5.7}$$

This shows that, due to the multiplication with the σ_i, the high-frequency components of $\overrightarrow{x}$ are more damped in $F\overrightarrow{x}$ than the low-frequency components, creating smoothing of the signal. Of course, the inverse problem of computing $\overrightarrow{x}$ from $F\overrightarrow{x} = \overrightarrow{m}$ must have the opposite effect: it amplifies the high-frequency oscillations of the right-hand side $\overrightarrow{m}$. The resulting reconstruction of

$\overrightarrow{x}$ is consequently very poor, essentially composed of important oscillations. In our measurements, high frequency components may be mostly associated to the experimental noise, i.e. the undesired fluctuations that appear superimposed on the signal source.

It is currently generally accepted that a linear system of equations is ill-posed if both following conditions are satisfied:

- the singular values of the coefficient matrix decay continuously to zero, without any particular gap in their spectrum;
- the ratio between the largest and the smallest non-zero singular values, called the (spectral) condition number of the matrix, is (very) large.

The gradual decrease of the singular values implies that there is no nearby problem with a well-conditioned coefficient matrix. The singular value decomposition of a matrix will be discussed in details in section 5.1. The second condition, i.e. the high value of the condition number associated to the coefficient matrix, reflects the potentially very high sensitivity of the solution to variations in the data of the problem. The condition number will be discussed in section 5.2.

5.2 Stability of a linear system of equations: the condition number

The analysis of the potential effects of round-off errors on solutions of linear systems of equations requires an appropriate way of quantification. This measure is fundamentally provided by the concept of norm, and quantified by the condition number associated to a coefficient matrix. Broadly speaking, a norm is a function that assigns a strictly positive size to all vectors in a vector space, other than the zero vector. They are many ways for defining vector or matrix norms, as they are several different norms. The most famous is the euclidian norm, associated to the geometrical length of the vector (also called magnitude). However, the definition is more general. In the following, the concepts of

vector and matrix norms are defined and discussed, and linked to the condition number in a subordinate matrix norm.

5.2.1 Vector norm: definitions

Suppose that V is a linear space over the field K, equipped with an absolute value. The nonnegative real-valued application $\|\cdot\|$ is said to be a vector norm provided that it satisfies the following properties:

- $\|\overrightarrow{v}\| = 0$ if and only if $\overrightarrow{v} = \overrightarrow{0}$ in V;
- $\|\alpha\overrightarrow{v}\| = |\alpha|\|\overrightarrow{v}\|$, $\forall\ \alpha \in K$ and $\forall\ \overrightarrow{v} \in V$;
- $\|\overrightarrow{u} + \overrightarrow{v}\| \leq \|\overrightarrow{u}\| + \|\overrightarrow{v}\|$, $\forall\ \overrightarrow{u}, \overrightarrow{v} \in V$;

where $|\alpha|$ is the absolute value of α if $K = \Re$. A linear space V equipped with a norm, i.e. $(V, \|\cdot\|)$, is called a normed linear space.

Any norm defined on the linear space $V \in \Re^n$ will be called a vector norm. A useful class of vector norms is the so-called p-norm. This norm class may be defined by:

$$\|\overrightarrow{v}\|_p = \left\{ \sum_{i=1}^{n} |v_i|^p \right\}^{1/p},\ p \geq 1 \tag{5.8}$$

Among the possible set of norms, three particular vector norms are in common use in numerical analysis: the 1-norm, the 2-norm and the ∞-norm. These are special cases of the p-norms, for $p = 1$, $p = 2$ and $p \to \infty$.

- the 1-norm, usually noted as $\|\cdot\|_1$, of the vector $\overrightarrow{v} = (v_1, v_2, \ldots, v_n)^T \in \Re^n$ is defined by:

$$\|\overrightarrow{v}\|_1 = \sum_{i=1}^{n} |v_i| \tag{5.9}$$

- the 2-norm, usually noted as $\|\cdot\|_2$, of the vector $\overrightarrow{v} = (v_1, v_2, \ldots, v_n)^T \in \Re^n$ is defined by:

$$\|\overrightarrow{v}\|_2 = \left(\overrightarrow{v}^T \overrightarrow{v}\right)^{1/2} \tag{5.10}$$

or, in other words:

$$\|\overrightarrow{v}\|_2 = \left\{ \sum_{i=1}^{n} |v_i|^2 \right\}^{1/2} \tag{5.11}$$

– the ∞-norm, usually noted as $\|\cdot\|_\infty$, of the vector $\overrightarrow{v} = (v_1, v_2, \ldots, v_n)^T \in \Re^n$ is defined by:

$$\|\overrightarrow{v}\|_\infty = \max_{1 \le i \le n} |v_i| \tag{5.12}$$

It is easy to show that each of these norms verify the properties of the norm definition. The demonstrations may for example be found in [69].

5.2.2 Vector norm: some fundamental properties

All vector norms on finite dimensional vector spaces on $\Re^n$ are topologically equivalent, i.e. if $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two norms of $\Re^n$, then there exists real positive constants, c_1 and c_2, such that:

$$c_1\|\overrightarrow{v}\|_\alpha \le \|\overrightarrow{v}\|_\beta \le c_2\|\overrightarrow{v}\|_\alpha \tag{5.13}$$

for all $\overrightarrow{v}$ on $\Re^n$. In particular, if $\overrightarrow{v} \in \Re^n$, then:

$$\|\overrightarrow{v}\|_2 \le \|\overrightarrow{v}\|_1 \le \sqrt{n}\,\|\overrightarrow{v}\|_2 \tag{5.14}$$

$$\|\overrightarrow{v}\|_\infty \le \|\overrightarrow{v}\|_2 \le \sqrt{n}\,\|\overrightarrow{v}\|_\infty \tag{5.15}$$

$$\|\overrightarrow{v}\|_\infty \le \|\overrightarrow{v}\|_1 \le n\,\|\overrightarrow{v}\|_\infty \tag{5.16}$$

5.2.3 Matrix norm: definitions

Matrix norms are natural extensions of vector norms to matrices. In general, any norm on the linear space $\Re^{n\times n}$ of $n \times n$ matrices with real entries will be referred to as a matrix norm. We shall now consider matrix norms which are induced by vector norms in a sense that will be made precise in the next definition.

Given any vector norm $\|\cdot\|$ on the space $\Re^n$ of n-dimensional vectors with real entries, the subordinate (or associated) matrix norm on the space $R^{n\times n}$ of $n \times n$ matrices with real entries is defined by:

$$\|F\| = \max_{\overrightarrow{v}\neq 0} \frac{\|F\overrightarrow{v}\|}{\|\overrightarrow{v}\|} = \max_{\|\overrightarrow{v}\|=1} \|F\overrightarrow{v}\| \tag{5.17}$$

This norm is sometimes referred to as natural matrix norm, or matrix norm induced by the vector norm $\|\cdot\|$. It can be shown that a subordinate matrix norm satisfies the properties listed in the vector norm definition.

The most frequently used matrix norms in numerical linear algebra are the p-norms, that may be defined in their most general sense by:

$$\|F\|_p = \max_{\overrightarrow{v}\neq 0} \frac{\|F\overrightarrow{v}\|_p}{\|\overrightarrow{v}\|_p} \tag{5.18}$$

The matrix p-norms are defined in terms of the vector p-norms already discussed in the previous section. In particular:

– the matrix norm subordinate to the vector norm $\|\cdot\|_1$ can be expressed, for an $n \times n$ matrix $F = (f_{ij}) \in \Re^{n\times n}$, as:

$$\|F\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |f_{ij}| \tag{5.19}$$

The 1-norm of a matrix is its largest absolute column-sum.

– the matrix norm subordinate to the vector norm $\|\cdot\|_2$ can be expressed, for an $n \times n$ matrix $F = (f_{ij}) \in \Re^{n\times n}$, as:

$$\|F\|_2 = \sqrt{\rho(F^*F)} = \sqrt{\rho(FF^*)} = \sigma_{\max}(F) \tag{5.20}$$

where F^* is the adjoint (conjugate transpose) to the matrix F, ρ is the spectral radius[1] of the matrix (FF^*) or (F^*F), and $\sigma_{\max}(F)$ is the highest singular value of F. This matrix norm is often called the spectral norm, because of its relation with the spectral radius of the matrix.

[1]Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of a matrix $F \in \Re^{n\times n}$, then its spectral radius $\rho(F)$ is defined by: $\rho(F) = \max_i |\lambda_i|$.

– the matrix norm subordinate to the vector norm $\|\cdot\|_\infty$ can be expressed, for an $n \times n$ matrix $F = (f_{ij}) \in \Re^{n\times n}$, as:

$$\|F\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |f_{ij}| \tag{5.21}$$

The ∞-norm of a matrix is its largest absolute row-sum.

The calculation of the 2-norm requires a precise evaluation of the highest singular value of the coefficient matrix (special algorithms have been developed for that purpose, the complete singular value decomposition of the coefficient matrix is then not necessary), and is much costlier in terms of computer time than the 1-norm or the ∞-norm. It is important to note that all these matrix norms are topologically equivalent. Consequently, if an estimation of the 2-norm is sufficient, the properties expressed in section 5.2.4 may be used.

5.2.4 Matrix norms: some properties

The p-norms applied on matrices satisfy certain inequalities that are frequently used in the analysis of matrix computations, especially for $p = 1$, $p = 2$ and $p \to \infty$. For a matrix $F \in \Re^{n\times n}$, the following properties hold:

$$\max_{i,j} |f_{ij}| \le \|F\|_2 \le n \max_{i,j} |f_{ij}|, \quad 1 \le i, j \le n \tag{5.22}$$

$$\frac{1}{\sqrt{n}} \|F\|_1 \le \|F\|_2 \le \sqrt{n}\, \|F\|_1 \tag{5.23}$$

$$\frac{1}{\sqrt{n}} \|F\|_\infty \le \|F\|_2 \le \sqrt{n}\, \|F\|_\infty \tag{5.24}$$

$$\|F\|_2 \le \sqrt{\|F\|_1 \|F\|_\infty} \tag{5.25}$$

5.2.5 The condition number in a subordinate matrix norms

The condition number of a coefficient matrix $F \in \Re^{n\times n}$, $\kappa(F)$, is a characteristic property that quantifies the asymptotically worst case of how much the solution $\overrightarrow{s}$ of a linear system of equations $F\overrightarrow{s} = \overrightarrow{b}$ can vary with respect to small variations in the data vector $\overrightarrow{b}$ of the problem. This number reflects the sensitivity of a particular solution, before the rounding effects are taken into

account. As a general rule, if the condition number $\kappa(F) = 10^k$, then one can expect to lose up to k digits of accuracy in solving the system of equations [70], in addition of what would be lost due to loss of precision arising from numerical operations (or approximative floating point numbers).

The condition number of a nonsingular coefficient matrix $F \in \Re^{n \times n}$ is defined by:

$$\kappa(F) = \|F\|.\|F^{-1}\| \tag{5.26}$$

where $\|\cdot\|$ is a subordinate matrix norm. By convention, $\kappa(F) = \infty$ if F is singular. In general, the value of $\kappa(F)$ is strongly dependent on the choice of the norm taken into consideration. This choice is usually denoted by introducing an index in the notation. In the p-norm, for example, the condition number of the matrix F is denoted by $\kappa_p(F)$. The condition number calculated in the 1-norm, $\kappa_1(F)$, and in the ∞-norm, $\kappa_\infty(F)$, are blatantly evaluated by using straightforwardly the definition 5.26. However, the condition number calculated in the 2-norm requires more attention. It can be demonstrated [69] that, in the case $p = 2$ and if F is a nonsingular coefficient matrix, $\kappa_2(F)$ may be characterized by:

$$\kappa(F) = \frac{\sigma_1(F)}{\sigma_n(F)}, \tag{5.27}$$

where $\sigma_1(F)$ is the highest singular value of the matrix F and $\sigma_n(F)$ is the smallest singular value of F. This number arises very often in numerical analysis, and is referred to as the spectral condition number of the matrix F.

Whatever the norm, the condition number $\kappa_p(F)$ is higher than 1, for every matrix because:

$$1 = \|FF^{-1}\| \leq \|F\|\|F^{-1}\| = \kappa_p(F) \tag{5.28}$$

If the condition number is exactly one, the relative accuracy of the solution is expected to be similar to the accuracy of the source data. However, it does not signify that the algorithm will converge rapidly to this solution, but only that it

won't diverge arbitrarily because of inaccuracy in the source data (provided that the forward error introduced by the algorithm does not diverge as well because of accumulating intermediate rounding errors). In general, matrices with a low condition number, i.e. $\kappa_p(F) \approx 1$, are said to be well-conditioned, while matrices with a high condition number, i.e. $\kappa_p(F) \gg 1$, are said to be ill-conditioned. It is important to note that a low condition number does not necessarily indicate that a solution will accurately be computed. The choice of stable resolution algorithms is crucial at this point. Inversely, the fact of having a matrix with a very high condition number does not automatically prevent of having excellent and accurate solutions for particular independent term. Evidently the condition number of a matrix is unaffected by scaling all its elements by multiplying by a nonzero constant.

Finally, it is possible to assess the sensitivity of the solution to the linear system of equations to changes in the independent vector. Let us start from the following linear system of equations:

$$F\overrightarrow{s} = \overrightarrow{b} \tag{5.29}$$

where $F \in \Re^{n \times n}$ and $\overrightarrow{b} \in \Re_0^n$. The matrix F is assumed to be nonsingular. Suppose now that the system 5.29 is subject to the perturbations $\delta\overrightarrow{s}$ and $\delta\overrightarrow{b} \in \Re^n$. Both these δ-vectors contain very small elements. The perturbated system is then defined by:

$$F\,(\overrightarrow{s} + \delta\overrightarrow{s}) = \overrightarrow{b} + \delta\overrightarrow{b} \tag{5.30}$$

In this situation, $\overrightarrow{s} \in \Re_0^n$ and the following relation is respected:

$$\frac{\|\delta\overrightarrow{s}\|}{\|\overrightarrow{s}\|} \leq \kappa(F)\,\frac{\|\delta\overrightarrow{b}\|}{\|\overrightarrow{b}\|} \tag{5.31}$$

The proof of this relation may be found, for example, in [69]. The conclusion of relation 5.31 is that, owing to the effect of rounding errors during the calculation, the numerical solution to the matrix system $F\overrightarrow{s} = \overrightarrow{b}$ will never be exact. The numerical solution may be written as $(\overrightarrow{s} + \delta\overrightarrow{s})$, and the relative error of the

solution vector $\|\delta\overrightarrow{s}\|/\|\overrightarrow{s}\|$ is bounded by the product between the relative error in the data vector $\|\delta\overrightarrow{b}\|/\|\overrightarrow{b}\|$ and the condition number $\kappa(F)$. Consequently, if the coefficient matrix F has a large condition number, the elements of $\delta\overrightarrow{s}$ may not be small.

In order to insure as maximum the stability of the solution, the condition number of the coefficient matrix should be decreased as much as possible. This can be done by using preconditioning techniques.

Chapter

6 Unfolding from the detector response

The objective of any spectrometric measurement is to access the complete and detailed information carried out by the radiation beam. However, since the detection is performed by a radiation detector, unavoidable modifications may considerably alter this information. These modifications are mainly due to the finite size of the detector, and to the statistics associated to the detection processes. Both these physical and statistical effects are included into the detector response functions, as explained in Chapter 4. The measured spectrum is then a modified version of the incident spectrum, and it has to be cleaned from the detector influence to collect precise and correct information. In our global inverse strategy for the X-ray source vector reconstruction, the first step of the procedure aims to rebuilt the energy distribution of the incident photon beam, i.e. the scattered vector $\overrightarrow{b}$, starting from the measured vector $\overrightarrow{m}$. This process, where an experimental measurement is cleaned from the detector response functions, is usually known as deconvolution - or unfolding - of the measurement.

In their most general sense, unfolding methods are a set of powerful mathematical algorithm-based techniques used to reverse the effects of convolution on recorded data, and to reconstruct the detector incident spectrum starting from a measurement. The methods are based on a detailed knowledge of the system response functions in well known geometrical conditions. Historically, response functions were determined experimentally by successive measurements of monoenergetic sources [71, 72]. In the last couple of decades, with the de-

velopment of computers (specifically the exponential increase of the processing power), simulation codes have been enjoying a growing importance in this field. Monte Carlo codes are today the main tool for calculating the response functions of a detection system.

The aim of this chapter is to theoretically introduce the concept of unfolding and to give an overview of some selected deconvolution techniques. The convolution equation that models the detection process is first outlined, and the discrete model used for numerical calculations is then deduced. Some common unfolding methods are then described. First, regularization techniques are discussed in a very general way, and the TIKHONOV and Truncated Singular Value Decomposition methods are explained in details. Generally speaking, regularization techniques require additional information about the expected solution, such as restrictions for smoothness or on the vector space norm, in order to regularize the solution vector and to prevent overfitting. The other two methods are based on a physical consideration about the physics of the problem: the non-negativity of the expected solution spectrum. This additional criterion is included into the system of equations to solve under the form of constraints.

6.1 The convolution equation and its discretization

A measured spectrum, recorded from any radiation detector, results from the convolution of the incident radiation distribution with the detector response functions. The measured differential pulse height spectrum $m(H)$, with H the pulse height, may be expressed as the convolution equation [73]:

$$\int R(H,E)\, b(E)\, dE = m(H) \tag{6.1}$$

with:

- $R(H,E)$, the detector response function;
- $b(E)$, the unknown incident energy distribution hitting the radiation detector.

In equation 6.1, the quantity $R(H, E)\, dH\, dE$ represents the differential probability that a quantum of energy within dE about E leads to a pulse with amplitude dH about H, and $b(E)\, dE$ is the differential number of incident photons with energy within dE about E.

Since the spectrum is recorded by means of a multi-channel analyzer (MCA), equation 6.1 must be discretized for numerical applications. The measured differential pulse height distribution may be discretized in m pulse height intervals:

$$m_i = \int_{H_{i-1}}^{H_i} m(H)\, dH, \quad i = 1, 2, \ldots, m \tag{6.2}$$

where m_i is the number of pulses recorded in the i-th interval. Similarly, the incident energy distribution may be divided into n intervals of energy:

$$b_j = \int_{E_{j-1}}^{E_j} b(E)\, dE, \quad j = 1, 2, \ldots, n \tag{6.3}$$

with b_j, the number of photons hitting the detector with energy included in the interval $[E_{j-1}, E_j]$. The detector response function $R(H, E)$ can be approximated, assuming low variations of the function in the energy interval $[E_{j-1}, E_j]$, by a $m \times n$ matrix, denoted by R_{ij}, whose elements are defined by:

$$R_{ij} = \frac{1}{E_j - E_{j-1}} \int_{H_{i-1}}^{H_i} \int_{E_{j-1}}^{E_j} R(H, E)\, dE\, dH, \tag{6.4}$$

with $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. The quantity R_{ij}, defined by equation 6.4, represents the probability that a pulse in the i-th pulse height interval will be recorded when a photon in the j-th energy interval penetrates into the detector.

Considering the discretization of the different quantities expressed by equations 6.2, 6.3 and 6.4, the integral equation 6.1 may equivalently be expressed by the following expression:

$$\sum_{j=1}^{n} R_{ij} b_j = m_i, \quad i = 1, 2, \ldots, m \tag{6.5}$$

The discrete equation 6.5 may also be rewritten in terms of a matrix equation:

$$R\overrightarrow{b} = \overrightarrow{m}, \quad R \in \Re^{m \times n} \tag{6.6}$$

with:

- R, the discretized response matrix of the detector;
- $\overrightarrow{b} = (b_1, b_2, \ldots, b_n)^T$, the incident photon energy distribution to the detector (the incident vector);
- $\overrightarrow{m} = (m_1, m_2, \ldots, m_m)^T$, the measured pulse height distribution (the measured vector).

Equation 6.6 models the convolution process arising from the detection of an incident energy distribution $\overrightarrow{b}$ with any radiation detector, using a MCA. Starting with a real measurement $\overrightarrow{m}$, it forms the system of equations to solve in the first step of the inverse procedure.

Equation 6.6 however only represents an ideal model for the evaluation of the incident energy distribution, $\overrightarrow{b}$. Indeed, it does not consider the experimental uncertainties associated to each energy bin of the measurement. From a practical point of view, it is nearly impossible to have an access to exact data: all measured quantities are a corrupted version of the quantities to be measured. The difference between the true and the measured quantities constitutes the noise, ε. Equation 6.6 must then be completed, for a real measurement, by adding this quantity:

$$\overrightarrow{m} + \overrightarrow{\varepsilon} = \overrightarrow{\tilde{m}} = R\overrightarrow{b}, \quad R \in \Re^{m \times n} \tag{6.7}$$

where $\overrightarrow{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)^T$ is an unknown fluctuating vector. The most attracting and intuitive way for solving equation 6.7 consists in the direct inversion of the discretized response matrix, R. Under certain mathematical conditions of nonsingularity, the solution to equation 6.7 may be given by:

$$\overrightarrow{b} = R^{-1}(\overrightarrow{m} + \overrightarrow{\varepsilon}) \tag{6.8}$$

Due to the high-frequency components of the detector response functions, the reconstruction of the scattered vector by direct inversion of the discretized response matrix R is very often excessively poor. In most cases, the reconstructed energy distribution does not fit the data in a reasonable way and the physical information of the reconstructed spectrum is totally drown into large oscillations. In addition, this inversion method is out of sense in the framework of numerical analysis, because it implies very heavy numerical calculations. In order to single out a significant physical solution, it is then required to resort to more robust methodologies.

6.2 Regularization techniques: general introduction

Discrete ill-posed problems form a complex class of problems owing to the multiplicity of small singular values of the coefficient matrix. As already discussed in section 5.1, this situation leads to extremely oscillating (and improbable) solutions. In order to stabilize the problem and to obtain a useful and significant physical solution, the point of view adopted in regularization techniques is to incorporate additional information to the underlying least square problem under the form of a penalty for complexity, such as restriction for smoothness on the vector space norm for example. A theoretical justification for regularization is that it attempts to impose OCKHAM's razor principle[1] on the solution.

Let us consider the convolution problem expressed by equation 6.7, and let us assume that the response function matrix R and the measured spectrum $\overrightarrow{m} + \overrightarrow{\varepsilon}$ are known. In the least-squares sense, the best approximate solution to this problem is given by the minimization of the following quantity:

$$\overrightarrow{\Lambda} = \left\| R\,\overrightarrow{\widehat{b}} - (\overrightarrow{m} + \overrightarrow{\varepsilon}) \right\|_2, \quad R \in \Re^{m \times n} \tag{6.9}$$

where $\overrightarrow{\Lambda}$ is the residual vector norm, i.e. the vector containing the differences

[1]The OCKHAM's razor principle is a general rule recommending that, from among a large set of competing hypothesis, selecting the one that makes the fewest new assumptions usually provides the correct one.

between the measured data ($\overrightarrow{m} + \overrightarrow{\varepsilon}$) and our best estimate model of the reconstructed spectrum $\overrightarrow{\widehat{b}}$, and where $\|\cdot\|_2$ is the matrix 2-norm (the so-called EUCLIDIAN norm).

Many types of additional information about the expected solution are possible. However, the dominating and most efficient approach in discrete ill-posed problems regularization is to require that the 2-norm (or an appropriate norm) of the solution be small. The subsequent constraint, often called the *side constraint* and denoted by Ω, may also include an initial estimate $\overrightarrow{b}^*$ of the solution, for example. In this situation, the side constraint involves the minimization of the quantity:

$$\Omega(\overrightarrow{\widehat{b}}) = \left\| L(\overrightarrow{\widehat{b}} - \overrightarrow{b}^*) \right\|_2 \tag{6.10}$$

The matrix L used in equation 6.10 is typically either the identity matrix I_n or a $p \times n$ discrete approximation of the $(n-p)$-th derivative operator, in which case L is a banded matrix with full row rank. By means of this side constraint, the smoothness of the regularized solution may be controlled.

As soon as the side constraint $\Omega(\overrightarrow{\widehat{b}})$ is introduced in the problem, the requirement that $R\overrightarrow{b}$ is equal to $\overrightarrow{m} + \overrightarrow{\varepsilon}$ (cf. equation 6.7) must obviously be given up, and replaced by an appropriate balance between minimizing the constraint $\Omega(\overrightarrow{\widehat{b}})$ and minimizing the residual norm $\overrightarrow{\Lambda}$. The underlying idea is that a regularized solution with a small least-square norm and a suitably small residual norm is not too far from the exact, unknown and unaccessible solution to the unperturbed problem that models the physical situation. The initial problem is then replaced by a second one, containing sufficient additional information to obtain an accurate approximate solution.

Many regularization methods have been developed in the past. Each of them has properties that makes it better suited to certain problems or certain computers. In the following subsections, two common and successful regularization techniques are outlined.

6.2.1 The TIKHONOV regularization method

The TIKHONOV method [74, 75] is perhaps the most commonly used regularization technique for solving ill-posed problems, and is a typical application of the matrix regularization principles. In this method, the best approximate solution $\overrightarrow{\widehat{b}}_{\Gamma}$ results from of a weighted combination between the side constraint $\Omega(\overrightarrow{\widehat{b}})$ and the residual norm $\overrightarrow{\Lambda}$ using a factor Γ^2, i.e.:

$$\overrightarrow{\widehat{b}}_{\Gamma} = argmin\left\{\left\|R\overrightarrow{\widehat{b}} - (\overrightarrow{m} + \overrightarrow{\varepsilon})\right\|_2 + \Gamma^2\left\|L(\overrightarrow{\widehat{b}} - \overrightarrow{b}^*)\right\|_2\right\} \tag{6.11}$$

where Γ is the so-called regularization parameter. This parameter controls the weight given to the minimization of the side constraint $\Omega(\overrightarrow{\widehat{b}})$ relative to the minimization of the residual norm $\overrightarrow{\Lambda}$. If the regularization parameter is large, a small solution norm is favored at the cost of a large residual norm: the level of oscillations in the solution is low, but the data are not very much taken into consideration. On the other hand, a small Γ, i.e. a small amount of regularization, has the opposite effect: the solution is less regularized and more sensitive to oscillations, but the importance given to the initial data is high. Obviously, if the regularization parameter is set to zero, the problem is reduced to the simple least-square case considered earlier, keeping its extreme sensitivity to random fluctuations affecting the measurement. The parameter Γ also controls the sensitivity of the regularized solution $\overrightarrow{\widehat{b}}_{\Gamma}$ to eventual perturbations in the discretized matrix R. It can be demonstrated that the perturbation bound is proportional to Γ^{-1} [65, 76].

6.2.2 Truncated singular value decomposition, TSVD

The truncated singular value decomposition (TSVD) [77, 78, 79] is based on a simple observation: for the larger singular values of the coefficient matrix R, the components of the reconstruction along the corresponding singular vector are well-determined by the data. This is not the case for the smaller singular values, where the reconstruction components are very oscillatory, causing large perturbations in the solution vector. In the TSVD regularization method, the part of the Σ matrix containing the smaller singular values is then discarded.

Since the singular values are ordered decreasingly in the Σ matrix, the process of removing the smallest ones is straightforward: it can be done by simple truncation.

The underlying idea in the treatment of the ill-conditioning by the TSVD method is to circumvent the problem by deriving a new problem with a well-conditioned rank deficient coefficient matrix. A fundamental result about deficient matrices, which can be obtained from the singular value decomposition of R, is that the closest rank-k approximation R_k to R measured in the 2-norm, is obtained by truncating the SVD expansion at k. Therefore, an integer $k \leq n$ is chosen for which the singular values are deemed to be significant. The regularized matrix R_k is then given by:

$$R_k = \sum_{i=1}^{k} \overrightarrow{u}_i \sigma_i \overrightarrow{v}_i^T, \quad k \leq n \tag{6.12}$$

The truncated SVD regularization method is based on this observation, in that ones solves the problem of minimizing the 2-norm of the best estimate model of the reconstructed spectrum $\overrightarrow{\widehat{b}}$ subject to minimizing the residual vector of the underlying least-square problem given by:

$$\overrightarrow{\Lambda}_k = \left\| R_k \overrightarrow{\widehat{b}} - (\overrightarrow{m} + \overrightarrow{\varepsilon}) \right\|_2 \tag{6.13}$$

From expression 6.12, the MOORE-PENROSE pseudo-inverse matrix may be derived:

$$R_k^+ = \sum_{i=1}^{k} \frac{\overrightarrow{v}_i \, \overrightarrow{u}_i^T}{\sigma_i}, \quad k \leq n \tag{6.14}$$

Using the MOORE-PENROSE inverse matrix, the solution to the regularized problem $\overrightarrow{\widehat{b}}$ may be found by:

$$\overrightarrow{\widehat{b}}_k = R_k^+ \overrightarrow{m} \tag{6.15}$$

or, in an explicit formulation:

$$\overrightarrow{\widehat{b}}_k = \sum_{i=1}^{k} \frac{\overrightarrow{u}_i^T \overrightarrow{m}}{\sigma_i} \overrightarrow{v}_i \tag{6.16}$$

The vector $\overrightarrow{\widehat{b}}$ is the best approximate solution to equation 6.7.

6.2.3 Selection of the truncation order

A very convenient graphical tool for analysis of discrete ill-posed problems is the so-called L-curve. For any valid regularization parameter, this curve is a plot of the regularized solution norm versus the corresponding residual norm. In this way, the L-curve displays the balance between the minimization of these quantities, which is the crucial point of the regularization approach. The use of such plots in connection with ill-posed problems goes back to Miller [80] and Lawson & Hanson [81].

For discrete ill-posed problems, it turns out that the L-curve, when plotted in a logarithmic scale, almost always has a characteristic L-shaped appearance, with a distinct corner separating the vertical and the horizontal parts of the curve. The vertical part of the L-curve corresponds to solutions where the norm of the regularized solution is very sensitive to changes in the regularization parameter, and the horizontal one to solutions where it is the residual norm that is most sensitive to the regularization parameter.

The L-curve is a continuous curve when the regularization parameter is continuous, as it is the case in Tikhonov regularization. For regularization methods with a discrete regularization parameter, such as the TSVD method, the L-curve is plotted as a finite set of points. The method to transform such a discrete L-curve in a continuous form is discussed in [82].

There is always an optimal regularization parameter that trades off the perturbation error and the regularization error in the regularized solution. An essential feature of the L-curve is that this optimal regularization parameter - defined in the above sense - is not far from the regularization parameter that corresponds to the L-curve's corner [83]. In other words, by locating the corner of the L-curve it is possible to obtain a good approximation of the optimal regularization parameter and thus, in turn, to compute a regularized solution presenting a good compromise between the two types of error.

6.3 Non-linear least square method

Linear least-square methods (and associated techniques) are recommended for use when good prior information and consistent measurements are available [84]. However, the main disadvantage of these methods is that the solution spectra may be physically inconsistent since negative fluence values may not be excluded. In order to introduce the condition of non-negative fluence, an algorithm first developed in the SandII code [85] and re-used later in the Gravel code [86] can be applied. In this case, instead of determining the particle fluence, its natural logarithm is calculated by a special iteration procedure minimizing a chi-square function.

In the Gravel code, the set of admissible spectra is defined using two restrictions. The first one expresses the link between the measured and the unknown quantities, i.e. between the measured vector $(\overrightarrow{m} + \overrightarrow{\varepsilon})$ and the incident energy distribution to the detector $\overrightarrow{b}$. This relation is given by the convolution equation 6.7, that may equivalently be written under the following form, using an element-based notation:

$$m_i + \varepsilon_i = \sum_{j=1}^{n} R_{ij} \, \exp\left[\ln b_j\right], \quad R \in \Re^{m \times n} \tag{6.17}$$

The mathematical trick in the right-hand side of equation 6.17 ensures the non-negativity of the incident vector values, b_j. The second restriction is defined by a chi-square expression, in which the logarithms of the pulse height spectrum values m_i are used with a diagonal relative covariance matrix. Using equation 6.17 for expressing the unknown errors ε_i, the chi-square criterion may be defined by:

$$\chi^2 = \sum_{i=1}^{m} \frac{\varepsilon_i^2}{\rho_i^2} = \sum_{i=1}^{m} \frac{\left(m_i - \sum_{j=1}^{n} R_{ij} \, \exp\left[\ln b_j\right] \right)^2}{\rho_i^2} \tag{6.18}$$

where χ^2 is familiar chi-square statistic criterion, and the ρ_i are the relative standard deviations [86].

The set of equations to minimize has then form:

$$\begin{cases} \widetilde{m}_i = \sum_{j=1}^{n} R_{ij} \, \exp\left[\ln b_j\right], \quad R \in \Re^{m \times n} \\ \\ \chi^2 = \sum_{i=1}^{m} \dfrac{\left(m_i - \sum_{j=1}^{n} R_{ij} \, \exp\left[\ln b_j\right] \right)^2}{\rho_i^2} \end{cases} \tag{6.19}$$

Assuming that there exists a solution $\ln b_j^{(1)}$ already known in the vicinity of the exact solution, $\ln \widetilde{m}_i$ can be expanded into a TAYLOR series truncated after the second term of the development:

$$\ln \widetilde{m}_i = \ln \widetilde{m}_i^{(1)} + \sum_{j=1}^{n} w_{ij}^{(1)} \left(\ln b_j - \ln b_j^{(1)} \right) \tag{6.20}$$

where:

$$\widetilde{m}_i^{(1)} = \sum_{j=1}^{n} R_{ij} \, \exp\left[\ln b_j^{(1)} \right] \tag{6.21}$$

and:

$$w_{ij}^{(1)} = \frac{R_{ij} \exp\left[\ln b_j^{(1)} \right]}{\widetilde{m}_i^{(1)}} \tag{6.22}$$

An optimal solution to the system of equations can be obtained by minimizing the chi-square value thanks to a special iteration procedure [87]. In each iteration step, the current solution $(k+1)$ is obtained from the previous solution (k) via an iteration algorithm given in first order by [88]:

$$b_j^{(k+1)} = b_j^{(k)} \exp \left(\frac{\sum_{j=1}^{n} w_{ij}^{(1)} \ln \left(m_i / \sum_{j'=1}^{n} R_{ij'} b_{j'}^{(k)} \right)}{\sum_{j=1}^{n} w_{ij}^{(1)}} \right) \tag{6.23}$$

with j', a summation index. For the iteration procedure, a first input spectrum (default spectrum) is required when the iterations are started. This default spectrum is modified during the iterations according to the different constraints

imposed to the system of equations. Given a non-negative default spectrum, this iterative procedure always leads to non-negative solution spectrum which tends to have a lower χ^2.

The GRAVEL code makes part of the HEPRO [89] / HEPROW [90] program systems and is currently distributed by the PHYSIKALISCH-TECHNISCHE BUNDESANSTALT (PTB), Braunschweig, Germany.

6.4 The maximum entropy method

In Bayesian probability theory, the principle of maximum entropy is a general-purpose axiom for determining positive and additive distributions starting from defined but incomplete constraints or information. It states that, subject to precisely stated *a priori* data expressing testable information, the probability distribution which best represents the current state of knowledge is the one with largest information theoretical entropy S. In this context, *a priori* information are not only limited to measured data and to their experimental uncertainties, but may also for example include correlations between variables, physical and mathematical constraints or physically meaningful features of the distribution [91]. In our case, the positive and additive distribution to be determined is the differential incident photon fluence, and the constraints are the measurements and their experimental uncertainties. The only available *a priori* information is the nonnegativity of the *a posteriori* distribution.

In this framework, the term entropy usually refers to the SHANNON entropy, a mathematical function that intuitively quantify the amount of information contained in a data source.

6.4.1 The SHANNON and the cross entropy

Let us consider the discrete random variable X with n possible outcomes $x = \{x_1, x_2, \ldots, x_n\}$ with probabilities $p = \{p_1, p_2, \ldots, p_n\}$ such that $p_i \geq 0$ for

$i = 1, 2, \ldots, n$ giving partial information on the variable X. The quantity:

$$I_i = \ln\left(\frac{1}{p_i}\right) \tag{6.24}$$

may be defined as the quantity of information I_i gained by each event x_i of the variable X. The logarithm is used to provide the additivity feature for independent uncertainties.

The entropy in the sense of SHANNON of a discrete random process is defined as [92]:

$$S(p) = \sum_{i=1}^{n} p_i \ln\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{n} p_i \ln p_i \tag{6.25}$$

This entropy may be seen as the measure of the uncertainty expressed in the distribution $p = \{p_1, p_2, \ldots, p_n\}$. The SHANNON entropy concept may be generalized by defining the logarithmic term as the information gain on an *a priori* probability q_i offered by the knowledge of the probability p_i given by the realization of an event x_i. It is then possible to define the quantity:

$$S_{CE} = -\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right) \tag{6.26}$$

as the cross-entropy of the probability distribution $p = \{p_1, p_2, \ldots, p_n\}$ with respect to the *a priori* distribution $q = \{q_1, q_2, \ldots, q_n\}$. The cross entropy satisfies the condition $S_{CE} \leq 0$, and it equals to zero only if $q(x) = p(x)$.

In the original derivation by E. JAYNES [93, 94], the use of the maximum entropy principle is justified on the basis of the cross-entropy's unique properties as uncertainty measure. Arguments originating in information theory show that the magnitude $|S_{CE}|$ of the cross entropy, as a generalization of the SHANNON entropy, is the appropriate measure of the amount of information necessary to change $p(x)$, which by assumption contains all the *a priori* information, into $q(x)$ [95]. It seems therefore reasonable to use $|S_{CE}|$ as a measure of how much $q(x)$ differs from $p(x)$. From among all the spectra that fit the data, the maximum entropy method chooses the one for which $|S_{CE}|$ is a minimum, and,

therefore, the one that is closest to the default spectrum. Under this form, the maximum entropy principle gives a selection rule for a particular solution. In our case:

- the solution $p(x)$ to be determined is the photon fluence;
- the constraints are the measures and the associated experimental uncertainties.

The maximum entropy principle argue that, among the possible set of solutions, the best solution is the one having both the maximum entropy, and being as few compromising as possible compared to all the other unknown information.

6.4.2 The MAXED algorithm

The set of admissible spectra is defined using two restrictions. The first one expresses the link between the measured and the unknown quantities, i.e. between the measured vector $\overrightarrow{m}$ and the incident vector $\overrightarrow{b}$. This relation is defined by the following element-based equation:

$$m_i + \varepsilon_i = \sum_{j=1}^{n} R_{ij}\, b_j, \quad i = 1, 2, \ldots, m \tag{6.27}$$

with $R_{ij} \in \Re^{m \times n}$, the discretized response matrix. The second restriction takes into account the (unknown) experimental errors ε_i in each bin i of the measured vector, and assumes a normal distribution of probability with zero mean and variance σ_i^2:

$$\chi^2 = \sum_{i=1}^{m} \frac{\varepsilon_i^2}{\sigma_i^2}, \quad i = 1, 2, \ldots, m \tag{6.28}$$

where χ^2 stand for the familiar chi-square statistic. From this set of admissible spectra, it is supposed that the best estimate of the incident vector $\overrightarrow{b}$ is the one maximizing the entropy S of the distribution. The entropy of the distribution is considered in a general form of the cross-entropy, given by Skilling [96]:

$$S_{CE} = -\sum_{j=1}^{n} \left[b_j \ln \left(\frac{b_j}{b_j^{(0)}} \right) + b_j^{(0)} - b_j \right] \tag{6.29}$$

where $b_j^{(0)}$ is the discretized default spectrum. It contains all the *a priori* information available. Therefore, any deviations from the default spectrum that results from the deconvolution should be driven by the new information provided by the measurements, otherwise we would be introducing structure that neither agrees with our *a priori* information nor is justified by the measurements. In other words, the aim is to take the default spectrum, and to change it into a spectrum that fits the data but remains 'as close as possible' to the default spectrum [97, 98]. The cross-entropy function 6.29 satisfies $S_{CE} \leq 0$, and the particular case $S_{CE} = 0$ arises only if $b_j = b_j^{(0)}$. The unfolding problem can then be reduced in searching the solution to the system composed by equation 6.27 and 6.28, under constraint of maximizing the cross-entropy function defined by equation 6.29. The solution to this optimization problem is found by using the LAGRANGE multipliers method.

The MAXED algorithm explained in the following is a modified form of the one presented by WILCZEK and DRAPATZ [99]. The LAGRANGE function associated with the optimization problem is of the form:

$$\begin{aligned} L\left(b_j, \varepsilon_i, \lambda_i, \mu\right) = &- \sum_{j=1}^{n} \left[b_j \ln\left(\frac{b_j}{b_j^{(0)}}\right) + b_j^{(0)} - b_j \right] \\ &- \sum_{i=1}^{m} \lambda_i \left[\sum_{j=1}^{n} R_{ij} b_j - m_i - \varepsilon_i \right] \\ &- \mu \left[\sum_{i=1}^{m} \left(\frac{\varepsilon_i^2}{\sigma_i^2}\right) - \chi^2 \right] \end{aligned} \tag{6.30}$$

where λ_i and μ are the LAGRANGE multipliers. Variation with respect to b_j, ε_i lead to the following set of $(n+m)$ equations:

$$\frac{\partial L}{\partial b_j} = 0 \quad \Longrightarrow \quad -\ln\left(\frac{b_j}{b_j^{(0)}}\right) - \sum_{i=1}^{m} \lambda_i R_{ij} = 0 \quad j = 1, 2, \ldots, n \tag{6.31}$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \quad \Longrightarrow \quad \lambda_i - 2\,\mu\,\frac{\varepsilon_i}{\sigma_i^2} = 0, \quad i = 1, 2, \ldots, m \tag{6.32}$$

Equations 6.28, 6.31 and 6.32 are solved for the variables b_j, the ε_i and μ in terms of the λ_i:

$$b_j = b_j^{(0)} \exp\left(-\sum_{i=1}^{m} \lambda_i R_{ij}\right), \quad j = 1, 2, \ldots, n \tag{6.33}$$

$$\varepsilon_i = \frac{\lambda_i \, \sigma_i^2}{2\mu}, \quad i = 1, 2, \ldots, m \tag{6.34}$$

$$\mu = \sqrt{\frac{\sum_{i=1}^{m} \lambda_i^2 \, \sigma_i^2}{4\chi^2}} \tag{6.35}$$

Using equations 6.33 to 6.35, the m equations 6.27 may be rewritten as:

$$m_i + \lambda_i \, \sigma_i^2 \sqrt{\frac{\chi^2}{\sum_{j=1}^{n} \lambda_j^2 \, \sigma_j^2}} - \sum_{j=1}^{n} R_{ij} b_j^{(0)} \exp\left(-\sum_{l=1}^{m} \lambda_l R_{lj}\right) = 0 \tag{6.36}$$

The initial optimization problem has then been reduced to a system of m equations with m unknowns $\lambda_1, \lambda_2, \ldots, \lambda_m$. The resolution of the system 6.36 is equivalent to the maximization of the following potential function Z:

$$Z = -\sum_{j=1}^{n} b_j^{(0)} \, exp\left[-\sum_{i=1}^{m} \lambda_i R_{ij}\right] - \sqrt{\chi^2 \sum_{i=1}^{m} \lambda_i^2 \, \sigma_i^2} - \sum_{i=1}^{m} m_i \, \lambda_i \tag{6.37}$$

with respect to the λ_i. Therefore, we can reformulate the problem in terms of the maximization of Z. In MAXED, the potential function 6.37 is maximized using an optimization subroutine based on the L-BFGS-B algorithm (version 2.3) [100, 101], a quasi-NEWTON code for solving large nonlinear optimization problems that uses a limited memory variation of the BROYDEN-FLETCHER-GOLDFARB-SHANNO method. The L-BFGS-B algorithm is particularly suited to problems with very large numbers of variables (often greater than 1000).

The MAXED code makes part of the UMG 3.3 package [88, 102] program systems and is currently distributed by the PHYSIKALISCH-TECHNISCHE BUNDESANSTALT (PTB), Braunschweig, Germany.

Chapter

7 Inverse scattering in the spectrometer

During the first step of the inverse procedure, the measured spectrum $\overrightarrow{m}$ is cleaned from the detector influence using the different methods discussed in the previous chapter. The result of the unfolding procedure is the energy distribution hitting the detector.

The second stage of the inverse strategy is to compute the inverse of the photon scattering on the target. Starting from the scattered vector representing the photon flux just after the scattering of the primary beam, the aim is to recover the initial X-ray source spectrum.

The forward discretized photon scattering on the spectrometer target can be expressed as a general matrix equation:

$$F\overrightarrow{s} = \overrightarrow{b}, \tag{7.1}$$

where:

- $F \in \Re^{n \times n}$ is the discretized forward scattering matrix;
- $\overrightarrow{s} = (s_1, s_2, \ldots, s_n)^T$ is the photon pulse height distribution emitted from the source (the unknown source vector);
- $\overrightarrow{b} = (b_1, b_2, \ldots, b_n)^T$ is the photon distribution after scattering on the target (the scattered vector).

It is assumed that all the scattered photons reach the detector without loss during their transport in the spectrometer. Then, the vector $\overrightarrow{b} = (b_1, b_2, \ldots, b_n)^T$ also represents the photon flux arriving at the detector.

In some physical situations, the direct scattering system may be extremely ill-conditioned, and the algebraic system can not necessarily be solved directly. The stability problems arising with such types of algebraic systems may again be studied by having recourse to the singular value decomposition of the coefficient matrix and by its condition number. This number quantifies the asymptotically worst case of how much the solution to the linear system of equations can vary with respect to small variations in the source data of the problem. Since the stability of a particular solution to successive numerical operations is not obvious, more robust methodologies may be required to obtain a significant and accurate physical solution to the scattering problem.

In the case expressed by equation 7.1, unfolding methods like regularization techniques may be used. However, it can be shown that, in our case, no acceptable solutions are obtained using the unfolding methods explained in the previous chapter.

In the next sections of this chapter, some methods adapted to the particular structure of the scattering matrix are theoretically described. All these methods will be applied in the next chapter, where their performances for the spectrum reconstruction will be evaluated. Since the ill-conditioning of the matrix may be a strong limitation into the achievement of a stable and accurate solution to the scattering problem, it could be mathematically interesting to transform the system in a form that is more adapted to numerical calculations. This operation is referred to as preconditioning of the matrix system. In the following, the two main system preconditioning ways will also be explained.

7.1 Computation of the forward matrix

The deterministic code FPCSHAPE has been used to compute the discrete forward scattering matrix F of equation 7.1. This code is a modern powered version of the SHAPE code, developed at the LABORATORY OF MONTECUCCOLINO (UNIVERSITY OF BOLOGNA), for computing the deterministic solution of the BOLTZMANN transport equation. More information about the code are supplied in [103]. Two specific scatterers (carbon and aluminium) of finite thickness have been considered, with the following approximations:

- both RAYLEIGH and COMPTON scattering in the solid target have been considered up to the first order of collision. This assumption can be justified by considering the very thin target used in the spectrometer;
- the model includes the form factors for RAYLEIGH scattering and the scattering function for the COMPTON scattering. For COMPTON scattering, the DOPPLER broadening of the COMPTON profile has not been considered;
- A DIRAC-δ interaction model has been considered for both RAYLEIGH and COMPTON scatterings, allowing their corresponding scattering intensities to be stored in only one bin of the matrix.

The problem has been solved in the wavelength regime, using a constant bin width. The choice of this bin width is fundamental in order to completely isolate the RAYLEIGH contribution from the COMPTON one, in the main and in a secondary diagonal of the forward matrix respectively. In these conditions the problem becomes more easily solvable, thanks to the particular structure of the scattering matrix (the matrix is said to be well-structured). If required, the final result of the calculation can be converted again to the energy regime at the end of the whole inverse procedure.

Denoting by i the index of the row and by j the index of the column, a diagonal matrix R containing the terms of RAYLEIGH scattering has been

defined as:

$$R_{(i,j)} = \begin{cases} r_{i,j} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{7.2}$$

and a single minor diagonal matrix C, containing the COMPTON scattering contributions:

$$C_{(i,j)} = \begin{cases} c_{i,j} & \text{if } i = j + \delta \\ 0 & \text{if } i \neq j + \delta \end{cases} \tag{7.3}$$

where δ is a strictly positive integer defining the COMPTON shift in units of the wavelength discretization interval. The forward scattering matrix, F, is defined by the sum of the R and C matrices:

$$F = R + C \tag{7.4}$$

Under these computation conditions, the forward scattering matrix F has the following properties:

- it is a square sparse matrix;
- the main diagonal only contains the RAYLEIGH contributions;
- the terms corresponding to the COMPTON scattering are placed in an lower diagonal, that is not directly adjacent to the main diagonal (the position of the minor diagonal depends on the wavelength discretization and, specifically, on the δ value);
- except for these 2 diagonals, all the other matrix elements are null.

The forward matrix obtained by the specific wavelength discretization has then a very simple structure: it is a positive semi-definite square matrix, whose elements are concentrated on the main diagonal and on a secondary lower diagonal.

7.2 Direct numerical methods for the resolution of linear systems

Direct numerical methods form a first class of efficient methods for solving linear systems of equations. A particularity of direct methods is that they compute

the solution to a linear system of equations in a finite number of steps, entirely determined by the size of the coefficient matrix. These methods would give the precise answer if they were performed in infinite precision arithmetic. In practice, however, this situation is very theoretical, because finite precision is used for the computation. Consequently, assuming numerical stability of the algorithm, the resulting vector is an approximation of the true solution.

Different direct numerical methods will be used in the next chapters to solve the inverse scattering problem, in particular:

- the substitution and the bidiagonal elimination techniques;
- the GAUSS elimination technique, with / without partial pivoting;
- the *LU* factorization;
- the CHOLESKY decomposition.

Since these methods are well-known, they are not described in this part of the document. The reader interested in some details about these methods is invited to read the Appendix A.

7.3 Iterative numerical methods for the resolution of linear systems

In contrast to direct methods, iterative techniques are not expected to terminate within a defined number of steps. Theoretically, iterative methods may require an infinite number of iterations to converge to the solution to a linear system of equations. The basic underlying ideas developed in iterative methods is to form a convergent sequence of vectors $\overrightarrow{s}^{(k)}$ such that:

$$\overrightarrow{s} = \lim_{k \to \infty} \overrightarrow{s}^{(k)} \tag{7.5}$$

where $\overrightarrow{s}$ is the solution to the matrix system of equations. Practically, starting from an initial estimation $\overrightarrow{s}^{(0)}$ of the solution, successive approximations

converging to the exact solution are formed. A convergence test is often specified in order to decide when a sufficiently accurate estimate of the solution has been found. This convergence test may take different forms. The most common consists in the comparison between the norm of the difference between two successive estimations of the reconstructed vector $\overrightarrow{s}$ (at iterations k and $(k+1)$) and a user-defined fixed stability tolerance ε. The stopping condition may then take a form such as:

$$\|\overrightarrow{s}^{(k+1)} - \overrightarrow{s}^{(k)}\| < \varepsilon \tag{7.6}$$

with $\|\cdot\|$, a given norm. The stopping condition is evaluated after each step of the iteration procedure. Denoting the error at the k-th iteration as:

$$\overrightarrow{e}^{(k)} = \overrightarrow{s}^{(k)} - \overrightarrow{s} \tag{7.7}$$

the condition given in 7.5 amounts to $\lim_{k\to\infty} \overrightarrow{e}^{(k)} = 0$ for each starting vector $\overrightarrow{s}^{(0)}$.

The convergence of iterative algorithms is a crucial aspect of these methods. Let us consider the matrix system $F\overrightarrow{s} = \overrightarrow{b}$ and its associated iterative method of the general following form:

$$\overrightarrow{s}^{(k+1)} = B\overrightarrow{s}^{(k)} + \overrightarrow{f}, \quad k \geq 0 \tag{7.8}$$

with a given initial estimate $\overrightarrow{s}^{(0)}$ of the solution. The matrix B is a square $n \times n$ matrix usually referred to as iteration matrix, and $\overrightarrow{f}$ a vector that depends on the independent term $\overrightarrow{b}$. An iterative method of the form 7.8 is said to be consistent with 7.1 if B and $\overrightarrow{f}$ are such that $\overrightarrow{s} = B\overrightarrow{s} + \overrightarrow{f}$, $\overrightarrow{s}$ being the solution to the matrix system 7.1, or equivalently if B and $\overrightarrow{f}$ satisfy:

$$\overrightarrow{f} = (I - B)\, F^{-1}\overrightarrow{b} \tag{7.9}$$

This single property of consistency does not suffice to ensure the convergence of an iterative method. However, if the method 7.8 is consistent, it can be demonstrated (see, for example, [104] for a formal demonstration) that the

vector sequence $\{\overrightarrow{s}^{(k)}\}$ obtained from 7.8 converges on the solution to the matrix system 7.1 for each starting vector $\overrightarrow{s}^{(0)}$ if and only if $\rho(B) < 1$.

A general technique for defining a consistent linear iterative methods is based on the decomposition (or splitting) of the coefficient matrix F under the form $F = P - N$, where P is a nonsingular matrix. Given a starting solution $\overrightarrow{s}^{(0)}$, the k-th solution $\overrightarrow{s}^{(k)}$ for $k \geq 1$ is obtained by solving:

$$P\overrightarrow{s}^{(k+1)} = N\overrightarrow{s}^{(k)} + \overrightarrow{b}, \quad k \geq 0 \tag{7.10}$$

The iteration matrix of the method 7.10 is $B = P^{-1}N$, et $\overrightarrow{f} = P^{-1}\overrightarrow{s}$. Equation 7.10 may also be written under the form:

$$\overrightarrow{s}^{(k+1)} = \overrightarrow{s}^{(k)} + P^{-1}\overrightarrow{r}^{(k)} \tag{7.11}$$

where:

$$\overrightarrow{r}^{(k)} = \overrightarrow{b} - F\overrightarrow{s}^{(k)} \tag{7.12}$$

is the residual at iteration k. Equation 7.11 shows the structure of the system to be solved for each iteration. In addition to be nonsingular, P should be easy to inverse in order minimize the calculation costs. Note that if P is equal to F and $N = 0$, the method defined by 7.11 converges in one iteration with the same computer cost than any direct method.

The method defined by equation 7.11 is guaranteed to converge if the spectral radius, i.e. the supremum among the absolute values of the elements in its spectrum, of the iteration matrix B is less than 1, i.e. if:

$$\rho(B) < 1 \tag{7.13}$$

In addition, the convergence is monotone. The demonstrations of these properties may be found in [104]. This property of convergence will be detailed for the different methods considered in the following.

The arithmetic cost of iterative methods on full matrices is of the order of n^2 floating point operations at each step. Most direct methods require a total of

$2n^3/3$ floating point operations. Iterative methods may then be time efficient for solving large systems of equations only if they converge in a number of iterations that is independent of n, or increasing under-linearly with n. For large sparse matrices, general direct methods usually turn out to have a high computational cost because of the fill-in, and iterative methods offer an interesting alternative.

Finally, the main advantage of iterative methods is that, due to their intrinsic nature, they are very well adapted to support round-off errors.

7.3.1 The JACOBI method

The JACOBI iterative algorithm is a method for determining the solutions of a linear system of equations, and is well adapted for matrices having absolute values in each row and column dominated by the diagonal elements.

Let us define a square system of linear equations with unknown $\overrightarrow{s}$:

$$F\overrightarrow{s} = \overrightarrow{b}, \quad F \in \Re^{n\times n} \tag{7.14}$$

where:

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}, \overrightarrow{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix}, \overrightarrow{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{7.15}$$

The coefficient matrix F may be decomposed into a diagonal coefficient matrix D, and a remainder R. The decomposition of F may be defined by:

$$F = D + R \tag{7.16}$$

with:

$$D = \begin{bmatrix} f_{11} & 0 & \cdots & 0 \\ 0 & f_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_{nn} \end{bmatrix}, \quad \text{and R} = \begin{bmatrix} 0 & f_{12} & \cdots & f_{1n} \\ f_{21} & 0 & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & 0 \end{bmatrix}$$

Using the decomposition expressed by equation 7.16, the system of equations 7.14 may be rewritten in the following form:

$$(D+R)\overrightarrow{s} = \overrightarrow{b} \tag{7.17}$$

and consequently:

$$\overrightarrow{s} = D^{-1}\left(\overrightarrow{b} - R\overrightarrow{s}\right) \tag{7.18}$$

The matrix $B_J = D^{-1}R$ is the iteration matrix of the JACOBI method. The iterative formulation of equation 7.18 may then be written in the following form:

$$\overrightarrow{s}^{(k+1)} = D^{-1}\left(\overrightarrow{b} - R\overrightarrow{s}^{(k)}\right) \tag{7.19}$$

and the corresponding element-based formula is given by:

$$s_i^{(k+1)} = \frac{1}{f_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} f_{ij}\, s_j^{(k)}\right), \quad i = 1, 2, \ldots, n \tag{7.20}$$

For the JACOBI method, the standard convergence condition may be expressed by:

$$\rho(D^{-1}R) < 1 \tag{7.21}$$

Consequently, the JACOBI method is guaranteed to converge if the matrix F is strictly diagonally dominant. Strict row diagonal dominance means that, for each row, the absolute value of the diagonal term is greater than the sum of absolute values of other terms, i.e.:

$$\|f_{ii}\| > \sum_{\substack{j=1 \\ j\neq i}}^{n} \|f_{ij}\|, \quad i = 1, 2, \ldots, n \tag{7.22}$$

Considering this relation, we may write:

$$\|B_J\|_\infty = \max_{i=1,2,\ldots,n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \|\frac{f_{ij}}{f_{ii}}\| < 1 \tag{7.23}$$

proving the convergence condition. However, the JACOBI method may converge even if these conditions are not fully satisfied.

7.3.2 The Gauss-Seidel method

The Gauss-Seidel method, also known as Liebmann method, is similar to the Jacobi method but takes advantage of the matrix decomposition to continuously give an update of the current solution. The coefficient matrix F of the linear system is again subdivided into two components. The first is a lower triangular component L, and the second is a strictly upper triangular component U. The matrix F may then be written as:

$$F = L + U \tag{7.24}$$

where:

$$L = \begin{bmatrix} f_{11} & 0 & \cdots & 0 \\ f_{21} & f_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}, \quad \text{and U} = \begin{bmatrix} 0 & f_{12} & \cdots & f_{1n} \\ 0 & 0 & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

The system of equations may be rewritten as:

$$\overrightarrow{s} = L^{-1}\left(\overrightarrow{b} - U\overrightarrow{s}\right) \tag{7.25}$$

or, in iterative formulation:

$$\overrightarrow{s}^{(k+1)} = L^{-1}\left(\overrightarrow{b} - U\overrightarrow{s}^{(k)}\right) \tag{7.26}$$

The matrix $B_{GS} = L^{-1}U$ is the iteration matrix. By taking advantage of the triangular form of L, the elements of $\overrightarrow{s}^{(k+1)}$ can be computed sequentially using forward substitution. This operation gives an update to the current solution, speeding up the convergence compared with the Jacobi method. Practically, when computing the i-th element $\overrightarrow{s}_i^{(k+1)}$, the first k-th elements $\overrightarrow{s}_1^{(k+1)}, \overrightarrow{s}_2^{(k+1)}, \ldots, \overrightarrow{s}_{i-1}^{(k+1)}$ are already known, and presumably more precise than $\overrightarrow{s}_1^{(k)}, \overrightarrow{s}_2^{(k)}, \ldots, \overrightarrow{s}_{i-1}^{(k)}$. The corresponding element-based formula is given by:

$$s_i^{(k+1)} = \frac{1}{f_{ii}}\left(b_i - \sum_{j=i+1}^{n} f_{ij}s_j^{(k)} - \sum_{j=1}^{i-1} f_{ij}s_j^{(k+1)}\right), \quad i = 1, 2, \ldots, n \tag{7.27}$$

The GAUSS-SEIDEL method converges for any starting vector if:

$$\rho\left(L^{-1}U\right) < 1 \tag{7.28}$$

The procedure is guaranteed to converge for strictly diagonally dominant matrices as well as for symmetric positive definite coefficient matrices. If the coefficient matrix F of the system to solve is positive definite, the method is in addition known to converge monotonically. Of course, the GAUSS-SEIDEL method sometimes converges even if these conditions are not completely satisfied.

7.3.3 The method of successive over-relaxation

The successive overrelaxation (SOR) iterative algorithm is a variant of the GAUSS-SEIDEL method for solving linear system of equations, resulting in possible faster convergence. The method is basically devised by applying extrapolation to the GAUSS-SEIDEL method. This extrapolation takes the form of a weighting factor. Starting with a square system of linear equations $F\overrightarrow{s} = \overrightarrow{b}$ with $F \in \Re^{n\times n}$, the coefficient matrix F may be decomposed into a diagonal matrix D, a strictly upper triangular matrices U and a strictly lower triangular matrix L. The matrix F may then be defined by:

$$F = D + L + U \tag{7.29}$$

where:

$$D = \begin{bmatrix} f_{11} & 0 & \cdots & 0 \\ 0 & f_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_{nn} \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ f_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & 0 \end{bmatrix}$$

and

$$U = \begin{bmatrix} 0 & f_{12} & \cdots & f_{1n} \\ 0 & 0 & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

The system of linear equations may be rewritten as:

$$\omega\,(D+L+U)\overrightarrow{s} = \omega\,\overrightarrow{b} \tag{7.30}$$

where ω is a constant called the relaxation parameter. By developing equation 7.30, the system becomes:

$$(D+\omega\,L)\overrightarrow{s} = \omega\,\overrightarrow{b} - [\omega\,U + (\omega-1)D]\overrightarrow{s} \tag{7.31}$$

The method of successive over-relaxation solves the left-hand side of expression 7.31 for $\overrightarrow{s}$, using previous value for $\overrightarrow{s}$ on the right-hand side. Analytically, the iterative equation is defined by:

$$\overrightarrow{s}^{(k+1)} = (D+\omega\,L)^{-1}(\omega\,\overrightarrow{b} - [\omega\,U + (\omega-1)D]\overrightarrow{s}^{(k)}) \tag{7.32}$$

By taking advantage of the triangular form of $(D+\omega L)$, the elements of $\overrightarrow{s}^{(k+1)}$ can be computed sequentially using forward substitution with the following element-based expression:

$$s_i^{(k+1)} = (1-\omega)\,s_i^{(k)} + \frac{\omega}{f_{ii}}\left(b_i - \sum_{j=1}^{n} f_{ij}s_j^{(k+1)} - \sum_{j=i+1}^{n} f_{ij}\,s_j^{(k)}\right) \tag{7.33}$$

for $i = 1, 2, \ldots, n$. The method is perfectly consistent for $\omega \neq 0$, and corresponds to the Gauss-Seidel method if $\omega = 1$. Though technically the term underrelaxation should be used when $\omega \in]0,1[$, for convenience the term overrelaxation is commonly used for every value of the relaxation parameter.

The choice of the relaxation parameter ω significantly affect the rate at which the method converges. However, the question of the choice of an optimal parameter ω_{opt} (i.e. for which the convergence rate is the highest) is still open: except for some very specific cases (see, for example, [105] or [106]), it is not possible to compute it in advance. In addition, even in the case when such an estimation is possible, the expense of such computation is usually prohibitive. Some general rules may however be mentioned. First, the choice of the relaxation parameter is strongly dependent on the properties of the coefficient matrix. It can be demonstrated that, without any particular hypothesis on the coefficient

matrix F, the successive-overrelaxation method always fails to converge if ω is outside the interval $[0, 2]$ [104, 107]. Secondly, if the coefficient matrix is symmetric and positive-definite, the SOR iterative algorithm will necessarily lead to convergence for any $\omega \in]0, 2[$. In addition, the convergence will be monotone [108]. Finally, is the coefficient matrix F is strictly diagonally dominant, the SOR method always converges for $0 < \omega \leq 1$.

7.4 Preconditioning of the system of equations

In the previous sections, different numerical methods for solving linear systems such as the forward scattering problem 7.1 have been discussed. Each of these methods is characterized by its stability with respect to round-off errors propagation and by particular convergence conditions. However, even with stable algorithms, the solution may be very sensitive to small variations in the data of the problem. As already seen, this instability is related to the conditioning of the coefficient matrix.

7.4.1 Obtention of a better conditioned system of equations

In order to condition the problem into a form that is more suitable for numerical calculations, having more favorable properties for direct / iterative resolution procedure, a preconditioning of the matrix system of equations may be applied. Generally speaking, preconditioning attempts to improve the spectral properties of the coefficient matrix. In other words, the preconditioning procedure acts in reducing the condition number of the problem, whatever the subordinate norm in which the condition number is evaluated.

The preconditioner P^{-1} of the coefficient matrix F is a matrix such that the matrix products $\left(FP^{-1}\right)$ or $\left(P^{-1}F\right)$ has a smaller condition number than F. The system preconditioning may be performed into two different principal ways: the right and the left preconditioning, according to the position of the preconditioner towards the coefficient matrix. Instead of solving the original

system, one may solve the following right preconditioned system:

$$FP^{-1}\overrightarrow{y} = \overrightarrow{b}, \quad \overrightarrow{s} = P^{-1}\overrightarrow{y} \tag{7.34}$$

The resolution of this system is done in two successive steps: the preconditioned system is first solved for any $\overrightarrow{y}$, and the solution vector is secondly found by solving the residual equation for $\overrightarrow{s}$. One may also solve the left preconditioned system:

$$P^{-1}F\overrightarrow{s} = P^{-1}\overrightarrow{b} \tag{7.35}$$

Both the left and right preconditioned systems give the solution to the original problem, as long as the preconditioner is non-singular. However, the preconditioned systems are easier to solve.

Since the preconditioner acts on the spectral radius of the coefficient matrix, it should be useful to determine an optimal preconditioner, making the number of iterations necessary to the convergence independent of the system's size. Unfortunately, the construction of optimal preconditioner is not possible: there is no general purpose preconditioner and there are no selecting rules for choosing the best appropriate preconditioner. The choice of a particular preconditioner turns out to be an extremely delicate task, since an unadapted preconditioning may considerably deteriorate the system. In general, a good preconditioner should meet the following obvious requirements:

- the preconditioner should be cheap to construct and to apply;
- the preconditioned system should be easier to solve than the unpreconditioned one.

In other words, the first property means that the preconditioning procedure should not be too expensive in terms of computer time, while the second implies that the convergence of any resolution technique should be rapid, i.e. that the preconditioned matrix is quasi normal and its singular values are contained in a reduced region of the space. Both these requirements are in competition

with each other. The cheapest preconditioner is, of course, the identity matrix I because $P = P^{-1} = I$. Obviously, the preconditioning of the system with the identity matrix results in the original linear system itself and the preconditioner does nothing. At the other extreme, the choice $P = F$ gives $P^{-1}F = FP^{-1} = I$, which has an optimal condition number, requiring a single iteration for convergence. However, this preconditioning is trivially inefficient. It therefore is necessary to strike a fair balance between these two extreme requirements, and to choose the matrix P as a compromise, in an attempt to achieve a minimal number of iterations while keeping the operator P^{-1} as simple as possible. With a good preconditioner, the condition number associated to the matrix system of equations should be considerably reduced.

In Chapter 3, the adjoint scattering matrix has been derived from the BOLTZMANN transport equations for photons. Because of its particular meaning in the framework of transport theory, the adjoint matrix is a particular preconditioner for the scattering problem 7.1. Therefore, the following study does not concern the selection of the best possible preconditioner among a large variety of mathematically acceptable preconditioners: the one which has a strong physical link with the problem to solve has been selected.

7.4.2 Physical signification of the adjoint transport matrix as a left preconditioner

Using the adjoint matrix F^T as a left preconditioner, the system defined by equation 7.1 becomes:

$$F^T F \overrightarrow{s} = F^T \overrightarrow{b} \tag{7.36}$$

Denoting $F^T F = A$, and $F^T \overrightarrow{b} = \overrightarrow{i}$, the system 7.36 can be rewritten as:

$$A \overrightarrow{s} = \overrightarrow{i} \tag{7.37}$$

The vector $\overrightarrow{i}$ is a measure of a photon's importance in the measured spectrum in contributing to a reading equivalent to the scattered vector.

Chapter

8 Simulated analysis of the scattering problem using two target materials

In practice, the scattering matrix system of equations may be very difficult to solve in most physical situations, because of its intrinsic instability. Basically, the difficulties in performing the inverse scattering calculation are linked to the mathematical structure of the scattering matrix, which also affects the convergence properties of the numerical method. The matrix structure is itself determined by the way X-ray photons interact with the matter, and then by the physical properties of the target material. Among these properties, the atomic number Z of the atoms making up the target plays a key role. The characterization of the target material from a physical and a mathematical point of view is then a fundamental aspect of the proposed inverse technique. In the following, two different materials – pure carbon (Z = 6) and pure aluminium (Z = 13) – are considered for their different physical behavior with respect to X-ray scattering. The scattering matrices associated to these scatterers are in consequence also different from a mathematical point of view. The first objective of this chapter is to analyse the mathematical convergence conditions of the numerical methods theoretically described in Chapter 7, and to link these conditions with the fundamental physics of the scattering problem (and then to the mathematical structure of the scattering matrix). Starting from this analysis, general rules regarding the use of particular target materials for the optimization of the inverse

scattering method will be deduced. An other key point to expect an accurate estimation of the source vector, and to correctly characterize the target material, are the mathematical performances of the different numerical methods used for solving the inverse scattering problem. In particular, the mathematical stability of the numerical algorithms and their ability in giving accurate reconstructions are two fundamental aspects of the inverse method developed in the following. A second objective of the chapter is then to appraise the stability and the accuracy of some carefully selected numerical methods in solving the inverse scattering problem. These two aspects have been studied starting from a sharp-peaked source spectrum. They are quantified by characteristic p-norms taking into account the differences between the theoretical source spectrum and the reconstructed solution.

In order to focus this study only on the consistency of the numerical methods and to avoid the additional difficulties arising from the unavoidable experimental noise of the measurement, numerical X-ray spectra have been considered in a first approach to the inverse scattering problem. Artificial scattered vectors for carbon, $\overrightarrow{b}_{(carbon)}$, and aluminium, $\overrightarrow{b}_{(aluminium)}$, have been computed by multiplying a numerical source spectrum $\overrightarrow{s}$ representing a Tungsten X-ray tube operating at 50 kV by the forward scattering matrix $F_{(carbon)}$ / $F_{(aluminium)}$, respectively. The scattering matrices and the numerical X-ray source spectrum have been constructed according to the following conditions:

– the forward scattering matrices have been computed for the two scatterers from analytical transport calculation, as detailed in section 7.1. Let us recall from this section that only RAYLEIGH and COMPTON interactions are taken into account for the computation. The photoelectric absorption and the subsequent fluorescence emission are then neglected. In the energy interval considered for the numerical experiment, approximately ranging from 2 keV to 50 keV, this approximation is straightforwardly acceptable for both scatterers because of their very low fluorescence energies, as indicated in table 8.1.

Material	$K\alpha_1$ (keV)	$K\alpha_2$ (keV)	$K\beta_1$ (keV)
Carbon	0.2770	/	/
Aluminium	1.4867	1.4827	1.5575

Table 8.1: Fluorescence X-ray lines for carbon and aluminium.

These energies are out of the energy interval considered, and the fluorescence approximation makes sense. Finally, it is worthwhile noting that the carbon scattering case has been chosen for its practical interest in the following step of the development. Indeed, a graphite target has been used as scatterer for the measurements performed in the real application (cf. Chapter 9);

– the theoretical X-ray source spectrum $\overrightarrow{s}$ has been built using *X-ray_ tube*, a computer code developed at the Laboratory of Montecuccolino (University of Bologna) for computing general X-ray tube spectral distributions. This code has been developed on the basis of two theoretical algorithms (Pella's algorithm [109, 110], and Schossmann's algorithm [111]). Both these algorithms appear to be accurate in the description of the continuum, but not for the characteristic X-ray lines emission. The program *X-ray_ tube* has been developed to compute both the continuous and the discrete part of the spectrum with a high level of accuracy. It includes the calculation of the continuum and the ratio(s) of the characteristic line(s) to the underlying continuum intensity at the wavelength(s) of the characteristic line(s). Practically, twelve X-ray lines, visible at the operating high-voltage of the tube, have been considered. X-ray transition wavelengths and energies have been taken from [112] and [113]. They are listed in Table 8.2, and the theoretical X-ray source spectrum is shown in Figure 8.1.

Using the theoretical source spectrum and the scattering matrices just described, the resulting scattered vectors $\overrightarrow{b}_{(carbon)}$ and $\overrightarrow{b}_{(aluminium)}$ present multiple strong discontinuities (sharp peaks), all located between 1.0689 Å and

1.7170 Å. The scattered vectors are shown in the wavelength regime in Figure 8.2 for carbon, and in Figure 8.3 for aluminium. In both scattered vectors, a multitude of spectral structures are observed.

Starting from the numerical scattered vectors $\overrightarrow{b}_{(aluminium)}$ and $\overrightarrow{b}_{(carbon)}$, all direct and iterative numerical methods introduced in Chapter 7 have been implemented to solve the system of equations describing the two scattering problems, eventually using the adjoint matrix as left / right preconditioner because of its conceptual importance in the physics of the problem.

Denomination	Energy (keV)	Wavelength (Å)
$L_1\beta_4$	9.5252	1.3016
$L_1\beta_3$	9.8189	1.2627
$L_1\gamma_2$	11.6105	1.0681
$L_1\gamma_3$	11.6805	1.0620
$L_2\eta$	8.7244	1.4211
$L_2\gamma_5$	10.9489	1.1323
$L_2\gamma_1$	11.2860	1.0985
L_3l	7.3878	1.6782
$L_3\alpha_2$	8.3353	1.4874
$L_3\alpha_1$	8.3982	1.4764
$L_3\beta_6$	9.6082	1.2899
$L_3\beta_2$	9.9641	1.2446

Table 8.2: X-ray lines taken into consideration for the theoretical Tungsten source spectrum computation.

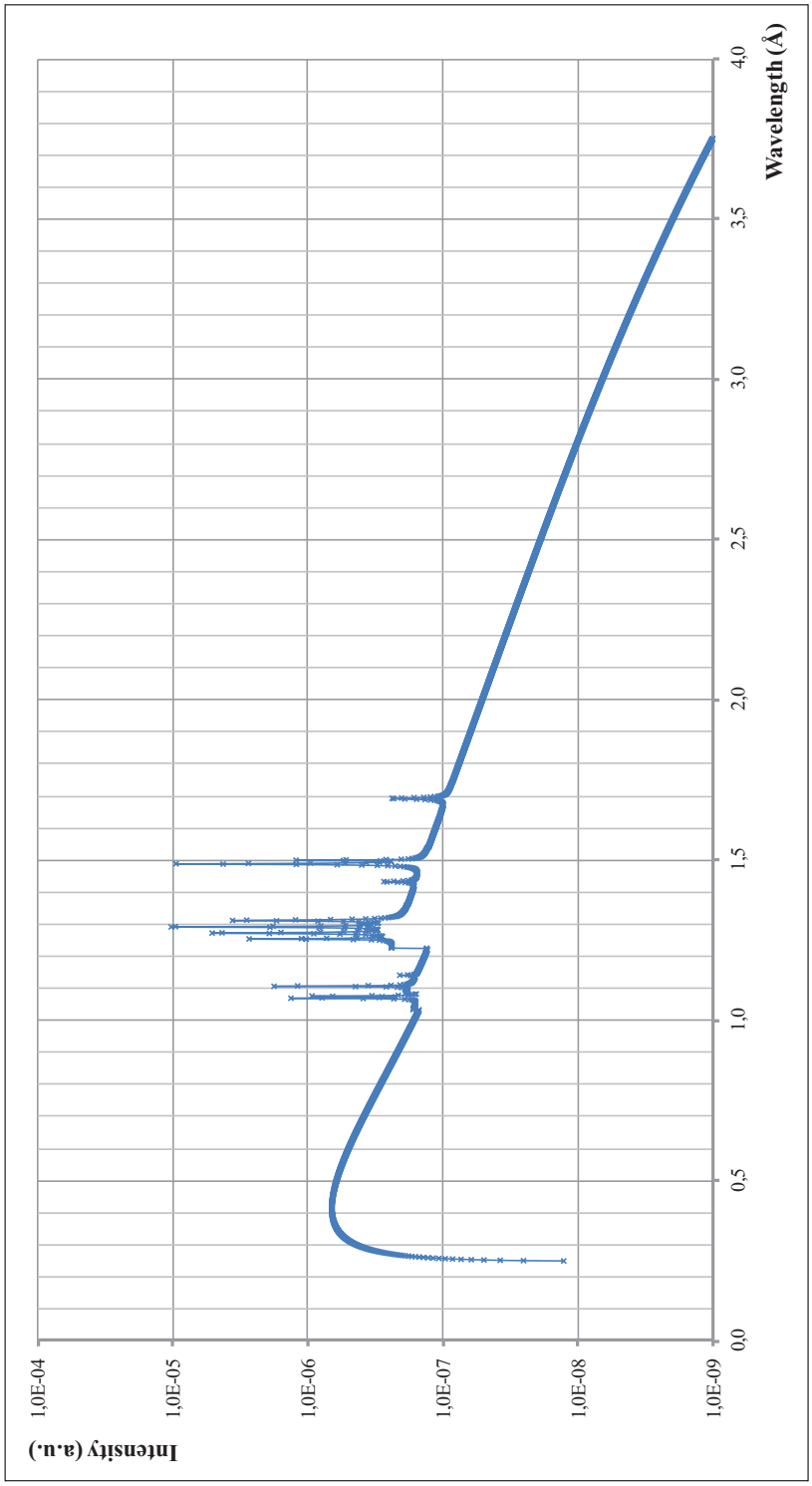

Figure 8.1: Numerical source spectrum used for the target material characterization, in arbitrary units. The source spectrum represents a numerical Tungsten X-ray tube operating at 50 kV, computed using the PELLA's algorithm.

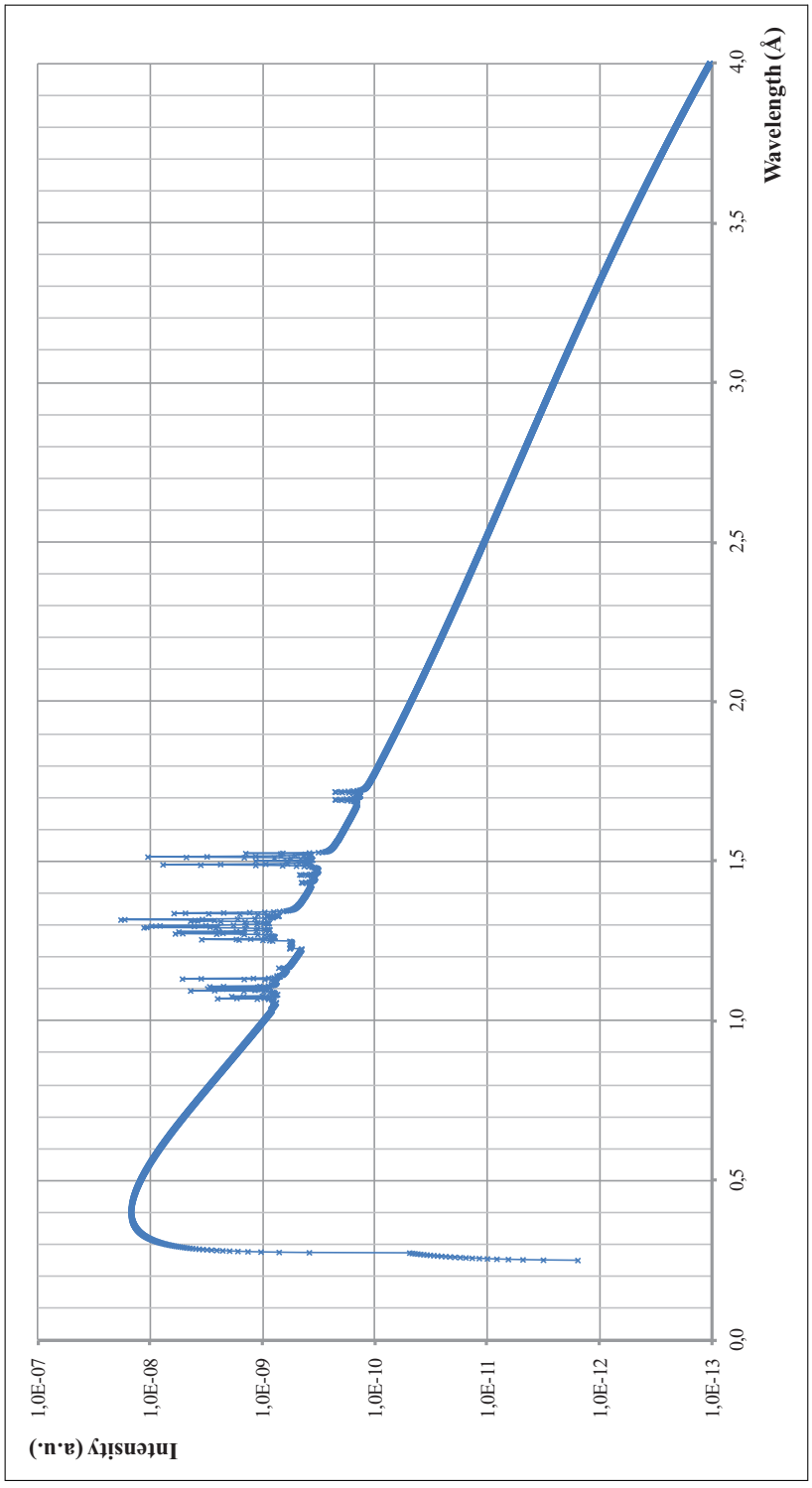

Figure 8.2: Scattered vector $\overrightarrow{b}_{(carbon)}$, obtained by multiplying the numerical source vector $\overrightarrow{s}$ by the carbon scattering matrix, in arbitrary units.

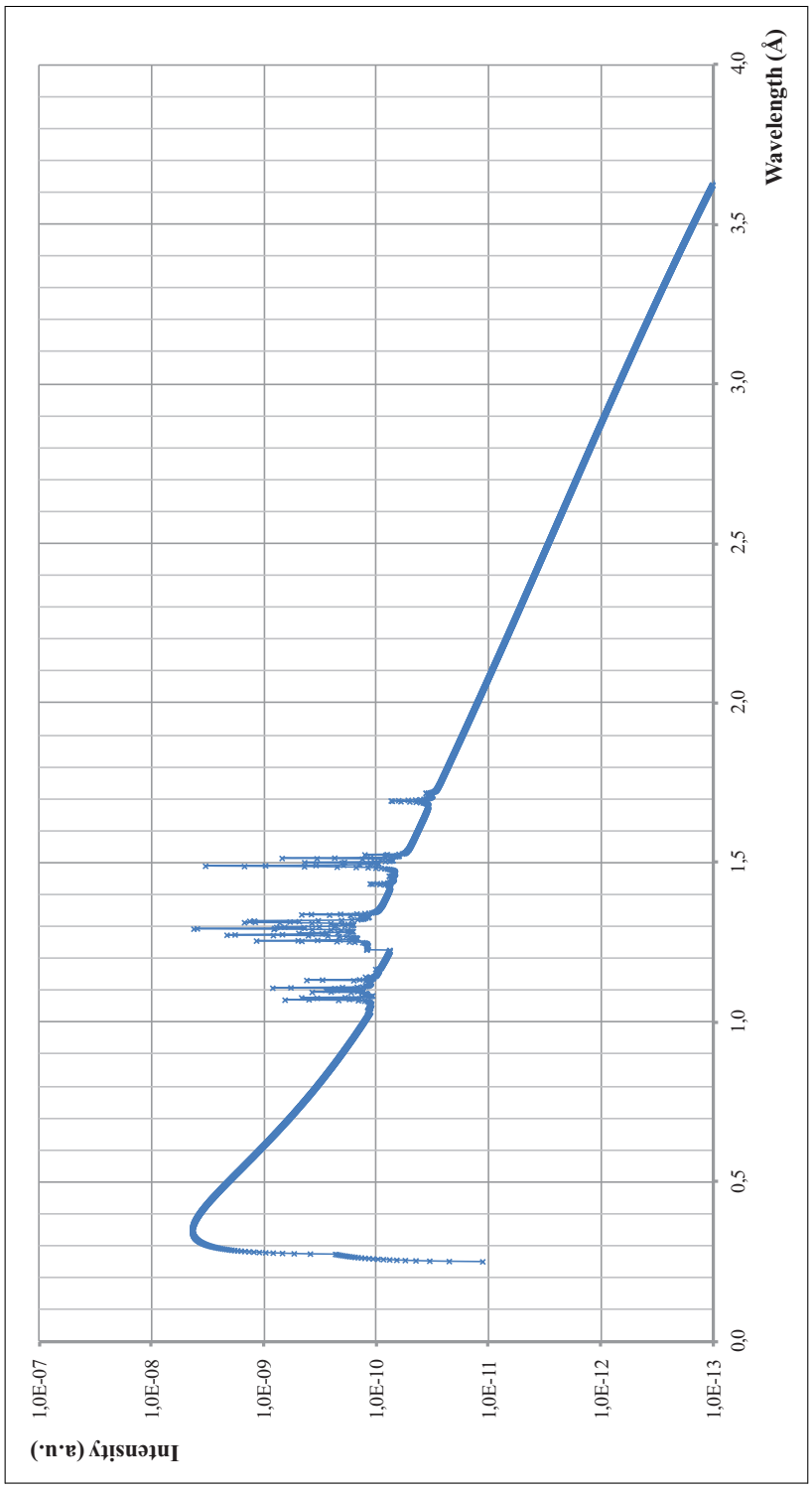

Figure 8.3: Scattered vector $\overrightarrow{b}_{(aluminium)}$, obtained by multiplying the numerical source vector $\overrightarrow{s}$ by the aluminium scattering matrix, in arbitrary units.

Practically, the photon scattering on the target material is represented by a 6000×6000 matrix, for wavelengths between 0.248000 Å (49.99 keV) and 6.312989 Å (1.9639 keV), with a constant wavelength step equal to 0.001011 Å. The choice of this discretization is fundamental to isolate the COMPTON contribution in only one diagonal of the forward scattering matrix. Using this particular wavelength discretization, the COMPTON contribution is separated from the RAYLEIGH contribution by 24 wavelength bins in each single response function of both scattering matrices. The scattered spectra $\overrightarrow{b}_{(carbon)}$ and $\overrightarrow{b}_{(aluminium)}$ are also represented by 6000 bins vectors, within the same wavelength range and with the same wavelength discretization. In these conditions, the system of equations is mathematically compatible, and may be expressed by:

$$F\overrightarrow{s} = \overrightarrow{b} \tag{8.1}$$

Equation 8.1 forms the matrix system to solve in the numerical experiment.

In the next sections of this chapter, the following notations are used for designating the reconstructed source vector recovered with a specific numerical method:

- solution using the singular value decomposition: $\overrightarrow{s}_{\mathrm{rec,\,SVD}}$;
- solution using the CHOLESKY factorization: $\overrightarrow{s}_{\mathrm{rec,\,Chol}}$;
- solution using the *LU* factorization: $\overrightarrow{s}_{\mathrm{rec,\,LU}}$;
- solution using the substitution technique: $\overrightarrow{s}_{\mathrm{rec,\,Sub}}$;
- solution using the bidiagonal elimination: $\overrightarrow{s}_{\mathrm{rec,\,BE}}$;
- solution using the GAUSS method: $\overrightarrow{s}_{\mathrm{rec,\,G}}$;
- solution using the GAUSS method with partial pivoting: $\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$;
- solution using the JACOBI iterative method: $\overrightarrow{s}_{\mathrm{rec,\,J}}$;
- solution using the SOR method: $\overrightarrow{s}_{\mathrm{rec,\,SOR}}$.

The special notation $\overrightarrow{s}_{\mathrm{rec},\star}$ will sometimes be used in the following to designate the complete set of reconstructed source vectors obtained using the different numerical methods for a particular scattering matrix system.

It has been seen in Chapter 7 that the relaxation parameter ω of the SOR method plays an important role in the convergence of the algorithm. There are no applicable rules for calculating a priori the optimal value of this parameter. However, it can be demonstrated that the convergence of the method is ensured for values of ω included in the interval $[0, 2]$ in a general case (e.g. the forward scattering system), and in the interval $]0, 2[$ if the matrix is symmetric and positive-definite (e.g. the right / left preconditioned scattering systems). In order to find the best SOR reconstruction, solution vectors have been calculated for different values of the relaxation parameter, i.e.:

- for $\omega \in [0.00, 2.00]$ in the forward scattering system case;
- for $\omega \in [0.05, 1.95]$ in the preconditioned scattering system case;

with a constant step of $\omega = 0.05$. For the sake of conciseness, only the best reconstructed vector in the sense of the residual 2-norm will be presented in this part of the document, and the correspondant value of the relaxation parameter ω will be given.

Both the JACOBI and the SOR methods have been implemented with a fixed stability tolerance ε equal to 10^{-20}, and a maximum number of iterations of 10.000.

The characterization of the carbon and aluminium scattering matrices is organized as follows. For each target material, the corresponding unpreconditioned, right preconditioned and left preconditioned matrix systems of equations have successively been solved. For each system considered, the condition number of the coefficient matrix has been evaluated in the 1-, 2- and ∞-norms, giving an a priori degree of complexity of the system resolution. The evaluation of the source reconstruction quality is made through a residual vector $\Delta\overrightarrow{s} = \overrightarrow{s} - \overrightarrow{s}_{\text{rec},\star}$, i.e. the difference between the theoretical source vector $\overrightarrow{s}$ and the reconstructed source vectors $\overrightarrow{s}_{\text{rec},\star}$. This characteristic vector is considered here as a standard evaluation of the spectrum reconstruction accuracy. The residual vector has been calculated in the 2-norm and in the ∞-norm, $\|\Delta\overrightarrow{s}\|_2$

and $\|\Delta\vec{s}\|_\infty$, respectively. These norms have been chosen for their particular physical signification: the 2-norm represents the length of the residual vector, while the ∞-norm is the maximum distance (or the maximum discrepancy) between the theoretical and reconstructed source vectors. The computer time for the reconstruction (CT) is also given for information. Numerical calculations have been performed on an Intel Core(TM)i7 CPU 950 at 3.07 GHz, with 16.00 Go of RAM memory, under a 64 bits Windows 7 Entreprise Operating System.

In this part of the document, it has been chosen to show exclusively the best reconstructed source vector in the sense of the 2-norm for each matrix system considered. For the sake of completeness, all the spectra are shown in Appendix B for the carbon scattering system and in Appendix C for the aluminium scattering system.

8.1 Pure carbon scattering matrix case

8.1.1 Evaluation of the systems ill-conditioning

The condition numbers of the unpreconditioned, right preconditioned and left preconditioned coefficient matrices of the corresponding carbon scattering systems have been evaluated in different subordinate p-norms, for $p = 1$, $p = 2$ and $p \to \infty$. The following notations have been used for denoting the different coefficient matrices:

- $F_{(carbon)}$ is the coefficient matrix of the unpreconditioned system, i.e. the forward scattering matrix;
- $P_{(right,\,carbon)} = F_{(carbon)}\,F^T_{(carbon)}$ is the coefficient matrix of the right preconditioned system;
- $P_{(left,\,carbon)} = F^T_{(carbon)}\,F_{(carbon)}$ is the coefficient matrix of the left preconditioned system.

The estimated condition numbers and their computer calculation times are reported in Table 8.3.

	$F_{(carbon)}$ (CT)	$P_{(right,\,carbon)}$ (CT)	$P_{(left,\,carbon)}$ (CT)
κ_1	$4.985\,10^{37}$ (14.21 s)	$4.537\,10^{21}$ (39.43 s)	$2.556\,10^{21}$ (39.84 s)
κ_2	$1.058\,10^{37}$ (409.31 s)	$7.399\,10^{18}$ (526.80 s)	$4.968\,10^{18}$ (528.04 s)
κ_∞	$3.225\,10^{37}$ (15.56 s)	$4.537\,10^{21}$ (40.30 s)	$2.556\,10^{21}$ (40.49 s)

Table 8.3: Condition number estimates (and computer time) for the unpreconditioned, right preconditioned and left preconditioned carbon matrix systems of equations.

All these values are consistent with the matrix p-norm properties given in section 5.2.4. Some conclusions may be given about the potential instability of the expected reconstructed source vectors by considering these condition numbers. Firstly, the condition numbers of the forward scattering matrix $F_{(carbon)}$ (e.g. $\kappa_2 = 1.058\,10^{37}$) are very high taking into account the dimensions of the matrix. The matrix $F_{(carbon)}$ is then extremely ill-conditioned, and the system may potentially be very ill-posed. It is then expected that the solution vectors reconstructed by solving the unpreconditioned system be very sensitive to small variations in the data. Secondly, the condition numbers of the right preconditioned coefficient matrix $P_{(right,\,carbon)}$ (e.g. $\kappa_2 = 7.399\,10^{18}$) are significantly reduced compared to those of the unpreconditioned matrix system (19 orders of magnitude). The right preconditioning of the system has then a positive effect on the ill-conditioning of the coefficient matrix. Finally, in the left preconditioning case, the condition numbers are similar to those of the right preconditioned matrix system, although slightly reduced (e.g. $\kappa_2 = 4.968\,10^{18}$). The right and left preconditioned systems are consequently significantly less sensible to small variations in the data of the problem than the unpreconditioned system. The solution vectors are consequently expected to be considerably more stable when solving the system preconditioned by the adjoint matrix (right / left) than the one without preconditioning.

8.1.2 Resolution of the three matrix systems

In this section, the unpreconditioned / preconditioned carbon matrix systems of equations are solved, using the numerical methods explained in Chapter 7. For each solution vector considered, an evaluation of the reconstruction quality is given through the residual vector norms. For the sake of readability and continuity of the paragraphs, the comparisons between the numerical source vector and the best estimates of the reconstructed vectors are shown, for each matrix system, at the end of the section.

Unpreconditioned system

The residual norms of the unpreconditioned system solution vectors are given in Table 8.4. In this table, it may be seen that all the reconstructed vectors $\overrightarrow{s}_{\text{rec},\star}$ very inaccurately reproduce the original source vector: the 2-norms of their respective residual vectors are ranging between $\|\Delta\overrightarrow{s}\|_2 = 1.4133\,10^{-3}$ for the SVD method and $\|\Delta\overrightarrow{s}\|_2 = 3.418\,10^{11}$ for the GAUSS technique with partial pivoting, i.e. between $5.770\,10^{3}\,\%$ and $1.395\,10^{18}\,\%$ of the theoretical source vector 2-norm, respectively.

The best reconstruction in the sense of the residual 2-norm has been obtained by using the SVD method. The reconstructed vector $\overrightarrow{s}_{\text{rec,SVD}}$ is given in Figure 8.4 p. 105 for information, and compared to the theoretical source vector $\overrightarrow{s}$. As expected from the extremely high ill-conditioning of the coefficient matrix, the forward carbon scattering system is very complicated to solve since it is numerically unstable. This instability only leads to an extremely high level of oscillations in the reconstructed vector. A detailed analysis of the oscillations position shows that they are situated in the wavelength range between 1.301 Å – 2.382 Å, while the rest of the spectrum reconstruction is good (even in the region of strong discontinuities between 1.062 Å and 1.301 Å). The region of high level oscillations may be associated to a wavelength interval where the scattering matrix is not diagonally dominant, creating instabilities. In this situation, most numerical algorithms usually fail to converge to a stable solution.

Method	$\|\Delta\overrightarrow{s}\|_2$	$\|\Delta\overrightarrow{s}\|_\infty$	CT (s)
$\overrightarrow{s}_{\text{rec, SVD}}$	$1.41329954\,10^{-3}$	$1.84306597\,10^{-4}$	$3.5220\,10^{3}$
$\overrightarrow{s}_{\text{rec, LU}}$	$2.59861590\,10^{11}$	$4.81432902\,10^{10}$	3.9041
$\overrightarrow{s}_{\text{rec, Sub}}$	$1.90134257\,10^{11}$	$4.56887893\,10^{10}$	$1.4136\,10^{-2}$
$\overrightarrow{s}_{\text{rec, BE}}$	$1.90134257\,10^{11}$	$4.56887893\,10^{10}$	$1.2550\,10^{-2}$
$\overrightarrow{s}_{\text{rec, G}}$	$2.94728911\,10^{11}$	$4.57058683\,10^{10}$	$1.4937\,10^{1}$
$\overrightarrow{s}_{\text{rec, Gpp}}$	$3.41752996\,10^{11}$	$4.87810469\,10^{10}$	$1.9779\,10^{1}$
$\overrightarrow{s}_{\text{rec, J}}$	$2.84529308\,10^{11}$	$4.62978796\,10^{10}$	4.4339
$\overrightarrow{s}_{\text{rec, SOR}}(\omega = 1.00)$	$1.90134257\,10^{11}$	$4.56887893\,10^{10}$	1.1732

Table 8.4: Differences between the source vector $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec},\star}$, calculated in the 2-norm and in the ∞-norm. Unpreconditioned system. Carbon case.

Right preconditioned system

The residual norms of the right preconditioned carbon matrix system solution vectors are presented in Table 8.5. Results are considerably less oscillatory than those obtained by solving the unpreconditioned matrix system. However, the reconstructed vectors obtained by solving the right preconditioning system are very particular: the 2-norms of the residual vectors $\Delta\overrightarrow{s}$ are equal until the 15^{th} decimal ($\|\Delta\overrightarrow{s}\|_2 = 2.299\,10^{-7}$, i.e. 0.939 % of the theoretical source vector 2-norm), except for the JACOBI method ($\|\Delta\overrightarrow{s}\|_2 = 7.807\,10^{-7}$, i.e. 3.187 % of the theoretical source vector 2-norm). For all the different methods, the reconstructed source vectors are very similar and present oscillations in the same region of the spectrum, included in the wavelength interval between 1.534 Å and 2.274 Å. This part of the spectrum corresponds to the non-diagonally dominant section of the right preconditioned carbon coefficient matrix $P_{(right,\,carbon)}$. It is finally very interesting to underline that the oscillations are periodic, with a 24 bins periodicity. This periodicity is directly linked to the discretization wavelength step of the coefficient scattering matrix $P_{(right,\,carbon)}$.

Method	$\|\Delta\overrightarrow{s}\|_2$	$\|\Delta\overrightarrow{s}\|_\infty$	CT (s)
$\overrightarrow{s}_{\text{rec, SVD}}$	$2.29917209\,10^{-7}$	$5.73526965\,10^{-8}$	$3.0954\,10^{2}$
$\overrightarrow{s}_{\text{rec, Chol}}$	$2.29917209\,10^{-7}$	$5.73526965\,10^{-8}$	$5.0968\,10^{1}$
$\overrightarrow{s}_{\text{rec, LU}}$	$2.29917203\,10^{-7}$	$5.73526963\,10^{-8}$	3.6807
$\overrightarrow{s}_{\text{rec, Sub}}$	$2.29917208\,10^{-7}$	$5.73526965\,10^{-8}$	3.5901
$\overrightarrow{s}_{\text{rec, G}}$	$2.29917204\,10^{-7}$	$5.73526965\,10^{-8}$	$1.4930\,10^{1}$
$\overrightarrow{s}_{\text{rec, Gpp}}$	$2.29917206\,10^{-7}$	$5.73526964\,10^{-8}$	$1.9637\,10^{1}$
$\overrightarrow{s}_{\text{rec, J}}$	$7.80762072\,10^{-7}$	$2.48978232\,10^{-8}$	$2.1433\,10^{2}$
$\overrightarrow{s}_{\text{rec, SOR}}$ ($\omega = 1.00$)	$2.29917208\,10^{-7}$	$5.73526965\,10^{-8}$	$3.4610\,10^{2}$

Table 8.5: Differences between the source vector $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec},\star}$, calculated in the 2-norm and in the ∞-norm. Right preconditioned system. Carbon case.

The best reconstruction in the sense of the smallest 2-norm residual vector has been obtained by the substitution method, with a residual vector $\|\Delta\overrightarrow{s}\|_2$ equal to $2.299\,10^{-7}$, i.e. 0.939 % of the theoretical source vector 2-norm. The reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$ is compared to the theoretical source vector $\overrightarrow{s}$ in Figure 8.5, p. 106. Except in the wavelength interval between 1.534 Å and 2.274 Å, the reconstruction appears very good in terms of both the intensity and the wavelength position of the peaks, even in the region of the spectrum where strong discontinuities are observed. In the wavelength interval between 1.534 Å and 2.274 Å, the reconstruction presents an important level of periodic oscillations, as mentioned in the previous paragraph.

Left preconditioned system

The residual 2-norms of the left preconditioned system solution vectors are reported in Table 8.6. As in the right preconditioned case, the vectors reconstructed by solving the left preconditioned system are considerably less oscillatory than those obtained by solving the unpreconditioned matrix system. The 2-norms of the residual vectors $\Delta\overrightarrow{s}$ are included between $\|\Delta\overrightarrow{s}\|_2 = 2.159\,10^{-4}$

for the SVD method and $\|\Delta\overrightarrow{s}\|_2 = 6.725\,10^{-8}$ for the SOR method, i.e. between 881.4 % and 0.275 % of the numerical source vector 2-norm, respectively.

Method	$\|\Delta\overrightarrow{s}\|_2$	$\|\Delta\overrightarrow{s}\|_\infty$	CT (s)
$\overrightarrow{s}_{\text{rec, SVD}}$	$2.15901037\,10^{-4}$	$3.14859176\,10^{-5}$	$3.0788\,10^{2}$
$\overrightarrow{s}_{\text{rec, Chol}}$	$5.55854700\,10^{-5}$	$7.35416856\,10^{-6}$	$4.7846\,10^{1}$
$\overrightarrow{s}_{\text{rec, LU}}$	$1.03987502\,10^{-5}$	$2.21498744\,10^{-6}$	4.3229
$\overrightarrow{s}_{\text{rec, Sub}}$	$9.90998568\,10^{-6}$	$4.3071334\,10^{-6}$	$1.5225\,10^{-2}$
$\overrightarrow{s}_{\text{rec, G}}$	$1.51535580\,10^{-5}$	$3.05894736\,10^{-6}$	$1.4937\,10^{1}$
$\overrightarrow{s}_{\text{rec, Gpp}}$	$1.16907463\,10^{-5}$	$2.71342237\,10^{-6}$	$1.9854\,10^{1}$
$\overrightarrow{s}_{\text{rec, J}}$	$2.58493606\,10^{-6}$	$4.60459488\,10^{-6}$	$2.1195\,10^{2}$
$\overrightarrow{s}_{\text{rec, SOR}}$ ($\omega = 1.50$)	$6.72473120\,10^{-8}$	$1.07805313\,10^{-8}$	$3.4610\,10^{2}$

Table 8.6: Differences between the source vector $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec},\star}$, calculated in the 2-norm and ∞-norm. Left preconditioned system. Carbon case.

The best reconstruction has been obtained by the SOR method with a relaxation parameter ω equal to 1.50. The reconstructed source vector $\overrightarrow{s}_{\text{rec, SOR}}$ is compared to the numerical source vector $\overrightarrow{s}$ in Figure 8.6, p. 107. The reconstruction appears excellent on the entire spectrum in both terms of continuum agreement between the reconstructed and theoretical source vectors, and in peak position / intensity correspondance. However, the reconstructed vector $\overrightarrow{s}_{\text{rec, SOR}}$ appears slightly oscillating, again with the periodicity of 24 wavelength bins linked to the discretization of the forward scattering matrix, in the region of the spectrum between 1.534 Å and 2.274 Å. The disrupted wavelength interval is then identical to the disturbed one identified in the right preconditioned case, in a lesser extent. As already explained in the right preconditioned case, this part of the spectrum corresponds to a non-diagonally dominant interval of the coefficient matrix $P_{(left,\,carbon)}$. These oscillations give the essential contribution to the residual vector 2-norm.

8.1.3 Conclusions for the carbon scattering system

The condition numbers estimated in section 8.1.1 indicate that the unpreconditioned system is numerically extremely unstable. Obviously, this form of the scattering system is not adapted to support multiple numerical operations. After preconditioning the matrix system by the adjoint matrix, the condition numbers of the coefficient matrices are significantly reduced. The scattering system is then transformed in a more suitable form. Right and left preconditioned systems have a very similar sensitivity to variations in the data of the problem.

The best reconstruction is obtained by applying the SOR method on the left preconditioned scattering system. In this case, the residual vector 2-norm represents 0.275 % of the theoretical vector 2-norm (see section 8.1.2, part 3). The reconstruction appears to be of high quality over the whole spectrum: the agreement between the continuum of the reconstructed and the numerical source vectors is excellent, all the peak positions and intensities are perfectly recovered after performing the inverse scattering calculation. In the reconstruction, some oscillations ranging from 1.534 Å to 2.274 Å can however be observed. Their wavelength position corresponds to a non-diagonally part of the coefficient matrix. These oscillations give the main contribution to the norm of the residual vector, and they are at the origin of a non perfect correspondance between the continuum of the source vectors.

In order to improve the quality of the reconstruction, an other scattering material can be used. For the example, the improvement of the reconstruction will be performed in the following with aluminium. Because of the higher atomic number of aluminium, the potential non-diagonally dominant part of the aluminium scattering matrix shall be shifted to the lower energies.

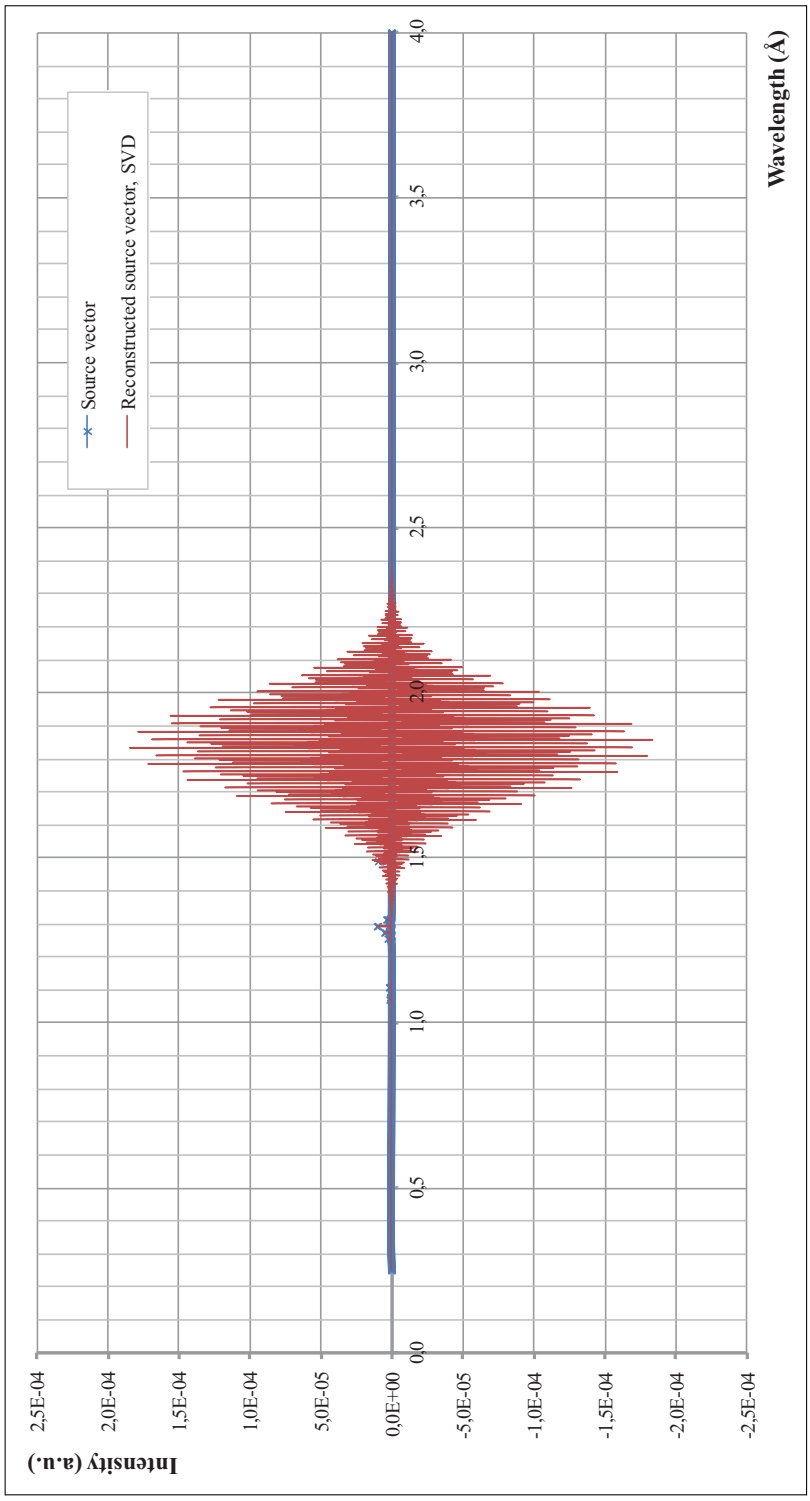

Figure 8.4: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vectors $\overrightarrow{s}_{\text{rec, SVD}}$. Unpreconditioned system. Carbon case.

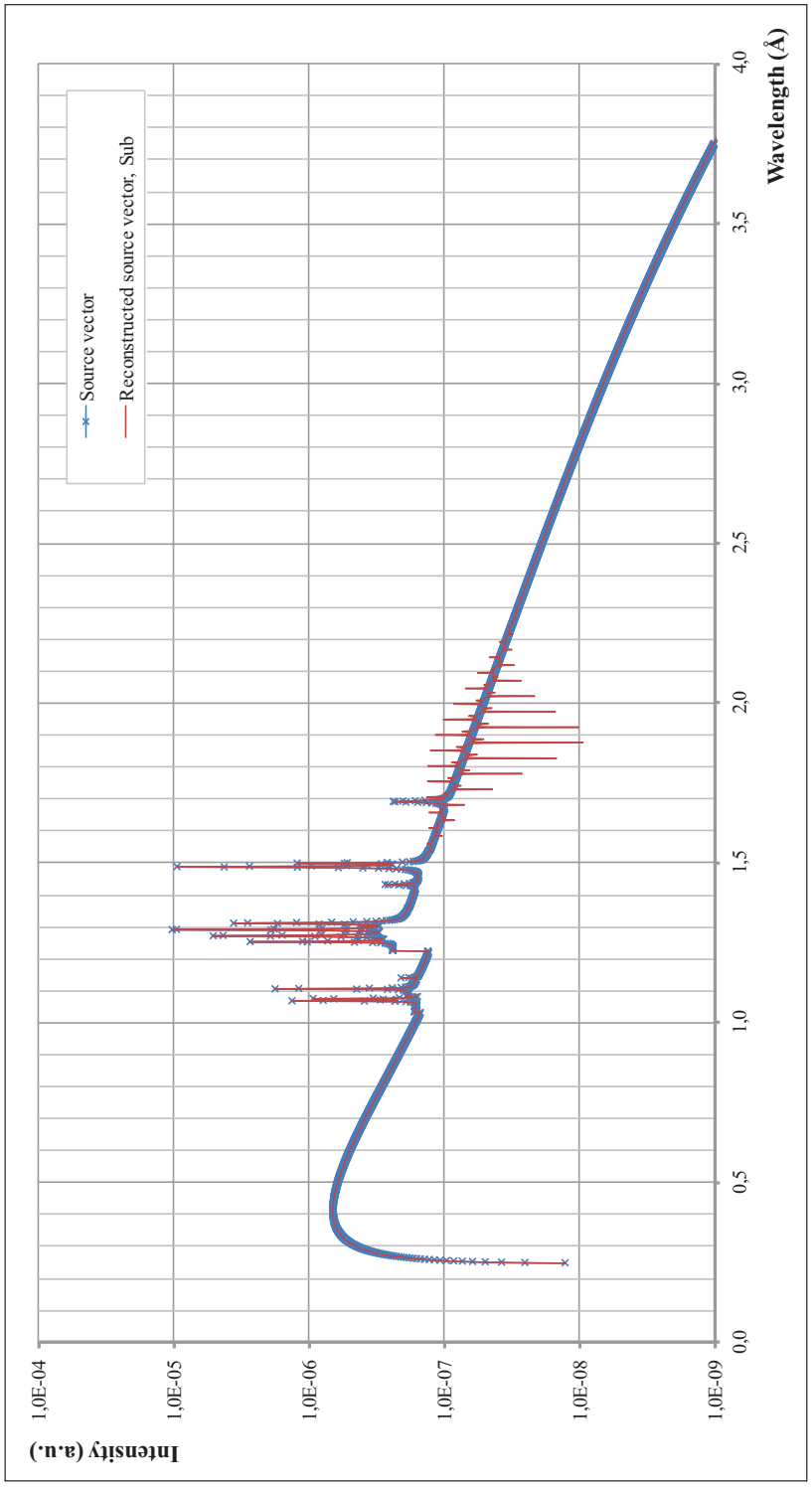

Figure 8.5: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. Right preconditioned system. Carbon case.

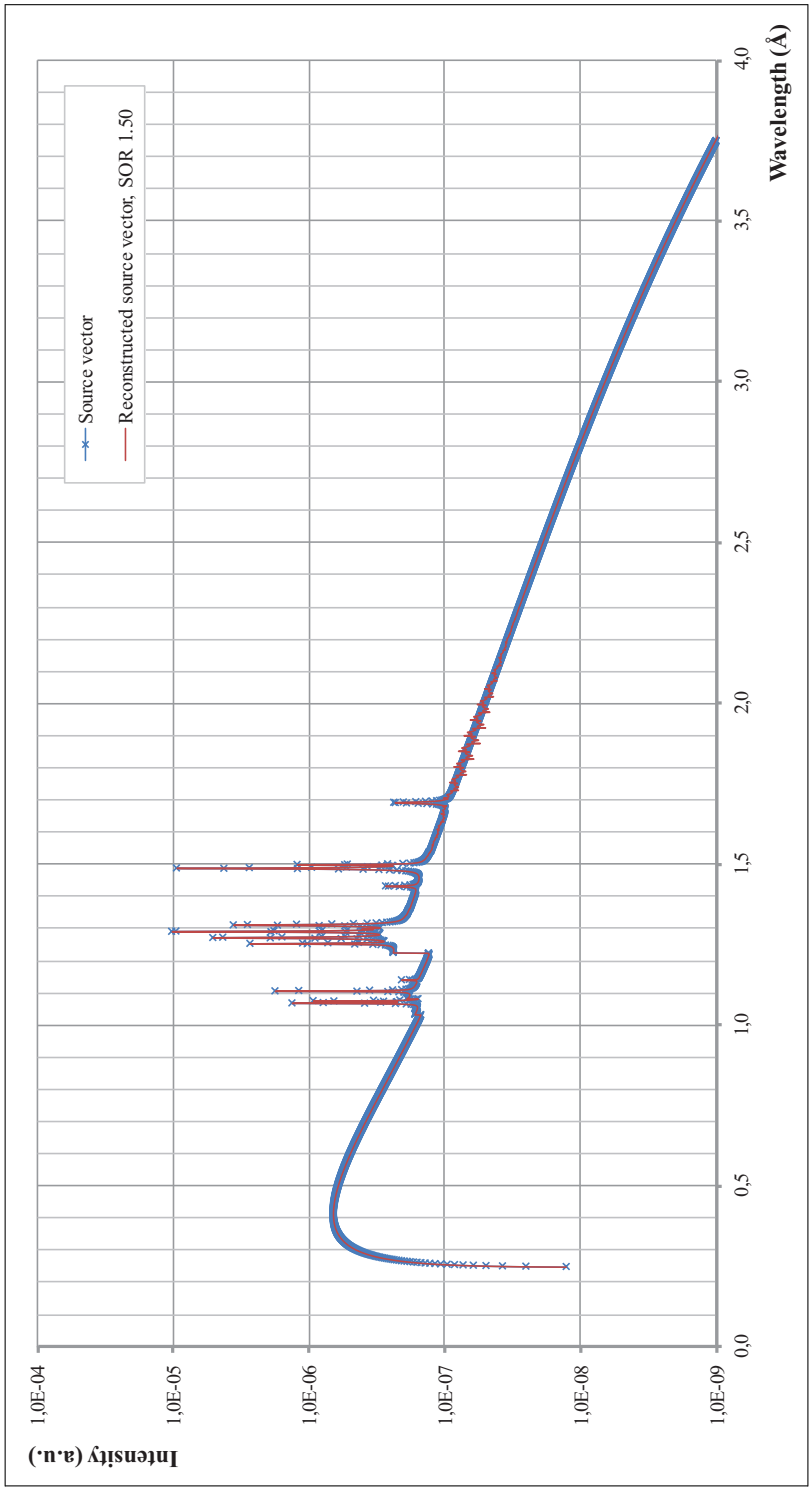

Figure 8.6: Comparison between the source spectrum $\vec{s}$ and the reconstructed source vectors $\vec{s}_{\mathrm{rec,SOR}}$. Left preconditioned system. Carbon case.

8.2 Pure aluminium scattering matrix case

8.2.1 Evaluation of the systems ill-conditioning

The condition numbers of the unpreconditioned, and right / left preconditioned coefficient matrices of the corresponding aluminium scattering systems have been evaluated in different p-norms, for $p = 1$, $p = 2$ and $p \to \infty$. The following notations have been used for denoting the different coefficient matrices:

- $F_{(aluminium)}$ is the coefficient matrix of the unpreconditioned system, i.e. the forward scattering matrix;
- $P_{(right,\, aluminium)} = F_{(aluminium)}\, F^T_{(aluminium)}$ is the coefficient matrix of the right preconditioned system;
- $P_{(left,\, aluminium)} = F^T_{(aluminium)}\, F_{(aluminium)}$ is the coefficient matrix of the left preconditioned system.

The estimated condition numbers and their computer calculation times are given in Table 8.7.

	$F_{(aluminium)}$ (CT)	$P_{(right,\, aluminium)}$ (CT)	$P_{(left,\, aluminium)}$ (CT)
κ_1	$5.998\,10^{15}$ (14.86 s)	$2.258\,10^{19}$ (39.81 s)	$2.119\,10^{22}$ (39.27 s)
κ_2	$1.541\,10^{15}$ (413.27 s)	$1.704\,10^{17}$ (662.06 s)	$1.467\,10^{19}$ (657.92 s)
κ_∞	$7.848\,10^{14}$ (15.88 s)	$2.258\,10^{19}$ (40.47 s)	$2.119\,10^{22}$ (40.22 s)

Table 8.7: Condition number estimates (and computer time) for the unpreconditioned, right preconditioned and left preconditioned aluminium matrix systems of equations.

All the condition number estimates are consistent with the matrix p-norm properties, detailed in section 5.2.4. Taking into account the very large dimensions of the coefficient matrices, the conditioning of the forward scattering matrix $F_{(aluminium)}$ (e.g. $\kappa_2 = 1.541\,10^{15}$) is not so bad (and significantly reduced than the one of the unpreconditioned carbon scattering system). The condition numbers of the right and left preconditioned coefficient matrices,

$P_{(right,\,aluminium)}$ and $P_{(left,\,aluminium)}$, are some orders of magnitude higher than those of the unpreconditioned system (2 orders of magnitude for the right preconditioning; 4 orders of magnitude for the left preconditioning). The preconditioning (right / left) by the adjoint matrix has then no positive effects on the scattering matrix system, since it increases the numerical sensitivity of the solution vectors to small variations in the data of the problem.

From the point of view of the numerical stability of the reconstructed solution, the unpreconditioned matrix system presents the most appropriate characteristics since it has the lowest condition numbers.

8.2.2 Resolution of the three matrix systems

In this section, the unpreconditioned and preconditioned aluminium matrix systems of equations are solved using the numerical methods detailed in Chapter 7, and the quality evaluation of the reconstruction is given through the residual vector norms. As in the carbon case, the comparisons between the numerical source vector and the best estimates of the reconstructions are shown at the end of the section for each matrix system, for a question of readability of the section.

Unpreconditioned matrix system

The residual norms of the unpreconditioned system solution vectors are given in Table 8.8. Except for the SVD method, very accurate results are obtained. The 2-norms of the residual vectors are included between $\|\Delta\overrightarrow{s}\|_2 = 2.459\,10^{-6}$ for the SVD method and $\|\Delta\overrightarrow{s}\|_2 = 1.434\,10^{-9}$ with a direct substitution technique, i.e. between $10.043\,\%$ and $5.854\,10^{-3}\,\%$ of the theoretical source vector 2-norm, respectively.

It is interesting to note that the SOR method is reduced to a substitution method in this particular case since the best relaxation parameter ω is equal to 1.00, and since the matrix has a particular bidiagonal structure. Similarly, due

to the bidiagonal structure of the matrix, the elimination and the substitution methods are similar techniques. The equality of the residual vector norms is then obvious.

The best solution vector has been obtained using a simple substitution technique. The reconstructed source vector $\overrightarrow{s}_{\mathrm{rec,\,Sub}}$ is compared to the theoretical source vector $\overrightarrow{s}$ in Figure 8.7, p. 114. Two main observations may be pointed out:

- the continuum parts of the reconstructed and theoretical source spectra are in perfect agreement, on the whole wavelength range;
- the reconstruction is excellent also in the regions of the spectrum where strong discontinuities (sharp peaks) are observed (between 1.037 Å and 1.705 Å), both in terms of wavelength position and peak intensity.

In this source spectrum reconstruction, no residual oscillations – even extremely small – can be observed, and the superimposition of the spectra is of the highest quality.

Method	$\|\Delta\overrightarrow{s}\|_2$	$\|\Delta\overrightarrow{s}\|_\infty$	CT (s)
$\overrightarrow{s}_{\mathrm{rec,\,SVD}}$	$2.45861138\,10^{-6}$	$5.07150691\,10^{-7}$	$3.2506\,10^{3}$
$\overrightarrow{s}_{\mathrm{rec,\,LU}}$	$2.59574409\,10^{-9}$	$5.40724035\,10^{-10}$	4.2438
$\overrightarrow{s}_{\mathrm{rec,\,Sub}}$	$1.43383388\,10^{-9}$	$2.66284460\,10^{-10}$	$1.1288\,10^{-2}$
$\overrightarrow{s}_{\mathrm{rec,\,BE}}$	$1.43383388\,10^{-9}$	$2.66284460\,10^{-10}$	$1.2186\,10^{-2}$
$\overrightarrow{s}_{\mathrm{rec,\,G}}$	$2.37056838\,10^{-9}$	$5.08043010\,10^{-10}$	$1.4883\,10^{1}$
$\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$	$2.44480955\,10^{-9}$	$5.40724035\,10^{-10}$	$1.9687\,10^{2}$
$\overrightarrow{s}_{\mathrm{rec,\,J}}$	$2.56246549\,10^{-9}$	$4.56080805\,10^{-10}$	2.5330
$\overrightarrow{s}_{\mathrm{rec,\,SOR}}(\omega = 1.00)$	$1.43383388\,10^{-9}$	$2.66284460\,10^{-10}$	1.2332

Table 8.8: Differences between the source vector $\overrightarrow{s}$ and the reconstructed source vectors $\overrightarrow{s}_{\mathrm{rec},\star}$, calculated in the 2-norm and in the ∞-norm. Unpreconditioned system. Aluminium case.

Right preconditioned system

The residual norms of the right preconditioned system solution vectors are reported in Table 8.9. The 2-norms of the residual vectors are equal for all the techniques ($\|\Delta\overrightarrow{s}\|_2 = 2.709\,10^{-8}$, i.e. 0.111 % of the numerical vector 2-norm), except for the JACOBI method ($\|\Delta\overrightarrow{s}\|_2 = 8.849\,10^{-8}$, i.e. 0.361 % of the numerical vector 2-norm). Except for this last method, all the vector norms are one order of magnitude higher than those of the unpreconditioned matrix system.

The best reconstruction in the sense of the smallest 2-norm residual vector has been obtained using the substitution technique (taking into account the calculation time). The reconstructed source vector $\overrightarrow{s}_{\mathrm{rec,\,Sub}}$ is compared to the theoretical source vector $\overrightarrow{s}$ in Figure 8.8, p. 115. The continuum parts of the reconstructed and theoretical source vectors are in very good agreement on the entire range of wavelengths considered. Slight oscillations, common to all reconstructions, can however be observed in the wavelength interval between 0.7039 Å and 1.0295 Å. This interval corresponds to a non diagonally dominant part of the coefficient matrix.

Method	$\|\Delta\overrightarrow{s}\|_2$	$\|\Delta\overrightarrow{s}\|_\infty$	CT (s)
$\overrightarrow{s}_{\mathrm{rec,\,SVD}}$	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	$3.5246\,10^{3}$
$\overrightarrow{s}_{\mathrm{rec,\,Chol}}$	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	$5.27638\,10^{1}$
$\overrightarrow{s}_{\mathrm{rec,\,LU}}$	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	3.6412
$\overrightarrow{s}_{\mathrm{rec,\,Sub}}$	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	2.6145
$\overrightarrow{s}_{\mathrm{rec,\,G}}$	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	$1.4940\,10^{1}$
$\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	$1.9687\,10^{1}$
$\overrightarrow{s}_{\mathrm{rec,\,J}}$	$8.84920264\,10^{-8}$	$7.99520481\,10^{-9}$	$2.0977\,10^{2}$
$\overrightarrow{s}_{\mathrm{rec,\,SOR}}$ ($\omega = 0.10$)	$2.70898913\,10^{-8}$	$7.99520481\,10^{-9}$	$3.7136\,10^{2}$

Table 8.9: Differences between the source vector $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\mathrm{rec},\star}$, calculated in the 2-norm and in the ∞-norm. Right preconditioned system. Aluminium case.

Left preconditioned system

The residual norms of the left preconditioned system solution vectors are reported in Table 8.10 for the different methods. As expected from the condition numbers estimates in section 8.2.1, the vectors reconstructed by solving the left preconditioned system are globally more inaccurate than those calculated with the unpreconditioned scattering system. It is worthwhile noting that the solution vectors present a high discrepancy in the reconstruction (4 orders of magnitude between the 2-norms of the different residual vectors), illustrating the instability level of the matrix system. The 2-norms of the residual vectors $\Delta\overrightarrow{s}$ are ranging between $\|\Delta\overrightarrow{s}\|_2 = 6.933\,10^{-4}$ for the singular value decomposition method and $\|\Delta\overrightarrow{s}\|_2 = 1.552\,10^{-8}$ for the SOR method, i.e. between $2.830\,10^{3}\,\%$ and $6.336\,10^{-2}\,\%$ of the numerical source vector 2-norm, respectively.

The best source vector reconstruction has been obtained with the left preconditioned system using the SOR method, with a relaxation parameter ω equal to 0.80. The reconstructed source vector $\overrightarrow{s}_{\mathrm{rec,\,SOR}}$ is compared to the numerical source vector $\overrightarrow{s}$ in Figure 8.9, p. 116. This reconstructed vector corresponds very well to the numerical source vector in both the continuum part and the discontinuity regions of the spectrum (as well in terms of intensity than in wavelength position of the different peaks). As in the right preconditioning case, slight oscillations can be observed in the region of the spectrum corresponding to the non diagonally dominant part of the matrix, located between 0.7039 Å and 1.0295 Å. These oscillations have a similar aspect than those observed in the right preconditioned case, even if they are slightly accentuated. A periodicity of 24 wavelength bins, associated to the discretization choice of the scattering matrix, is observed (as in the carbon case).

Method	$\|\Delta\overrightarrow{s}\|_2$	$\|\Delta\overrightarrow{s}\|_\infty$	CT (s)
$\overrightarrow{s}_{\text{rec, SVD}}$	$6.93263315.10^{-4}$	$1.74883449\,10^{-4}$	$3.4858\,10^{3}$
$\overrightarrow{s}_{\text{rec, Chol}}$	$1.68329476\,10^{-4}$	$4.01276405\,10^{-5}$	$5.0433\,10^{1}$
$\overrightarrow{s}_{\text{rec, LU}}$	$2.52564067\,10^{-5}$	$6.64001353\,10^{-6}$	4.1878
$\overrightarrow{s}_{\text{rec, Sub}}$	$5.75934406\,10^{-6}$	$2.68287066\,10^{-6}$	$1.5635\,10^{-2}$
$\overrightarrow{s}_{\text{rec, G}}$	$1.90162926\,10^{-4}$	$7.72220372\,10^{-5}$	$1.4903\,10^{1}$
$\overrightarrow{s}_{\text{rec, Gpp}}$	$9.34008103\,10^{-5}$	$3.33457860\,10^{-5}$	$1.9692\,10^{1}$
$\overrightarrow{s}_{\text{rec, J}}$	$4.77387912\,10^{-6}$	$4.04589258\,10^{-7}$	$2.1226\,10^{2}$
$\overrightarrow{s}_{\text{rec, SOR}}$ $(\omega = 0.80)$	$1.55228435\,10^{-8}$	$4.63669618\,10^{-9}$	$1.1484\,10^{2}$

Table 8.10: Differences between the source vector $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec},\star}$, calculated in the 2-norm and in the ∞-norm. Left preconditioned system. Aluminium case.

8.2.3 Conclusions for the aluminium scattering system

In the aluminium case, the condition numbers evaluated in section 8.2.1 indicate that the unpreconditioned system has the lowest sensitivity to small fluctuations in the data of the problem. The right or left system preconditioning by the adjoint matrix has then an undesirable harmful effect on the numerical stability of the solution vectors. The unpreconditioned system is then the most suitable form for numerical operations.

For solving the unpreconditioned scattering aluminium system, the substitution technique is the most appropriate. In this case, the residual vector 2-norm is equal to 5.854 10^3 % of the numerical source vector 2-norm (see section 8.2.2, part 1). The reconstruction appears to be of the highest quality on the entire spectrum. First, the agreement is perfect between the continuum of the reconstructed spectrum and the numerical source vector. Secondly, all the peaks positions and intensities are perfectly recovered after the inverse scattering. The reconstruction is free from any oscillations.

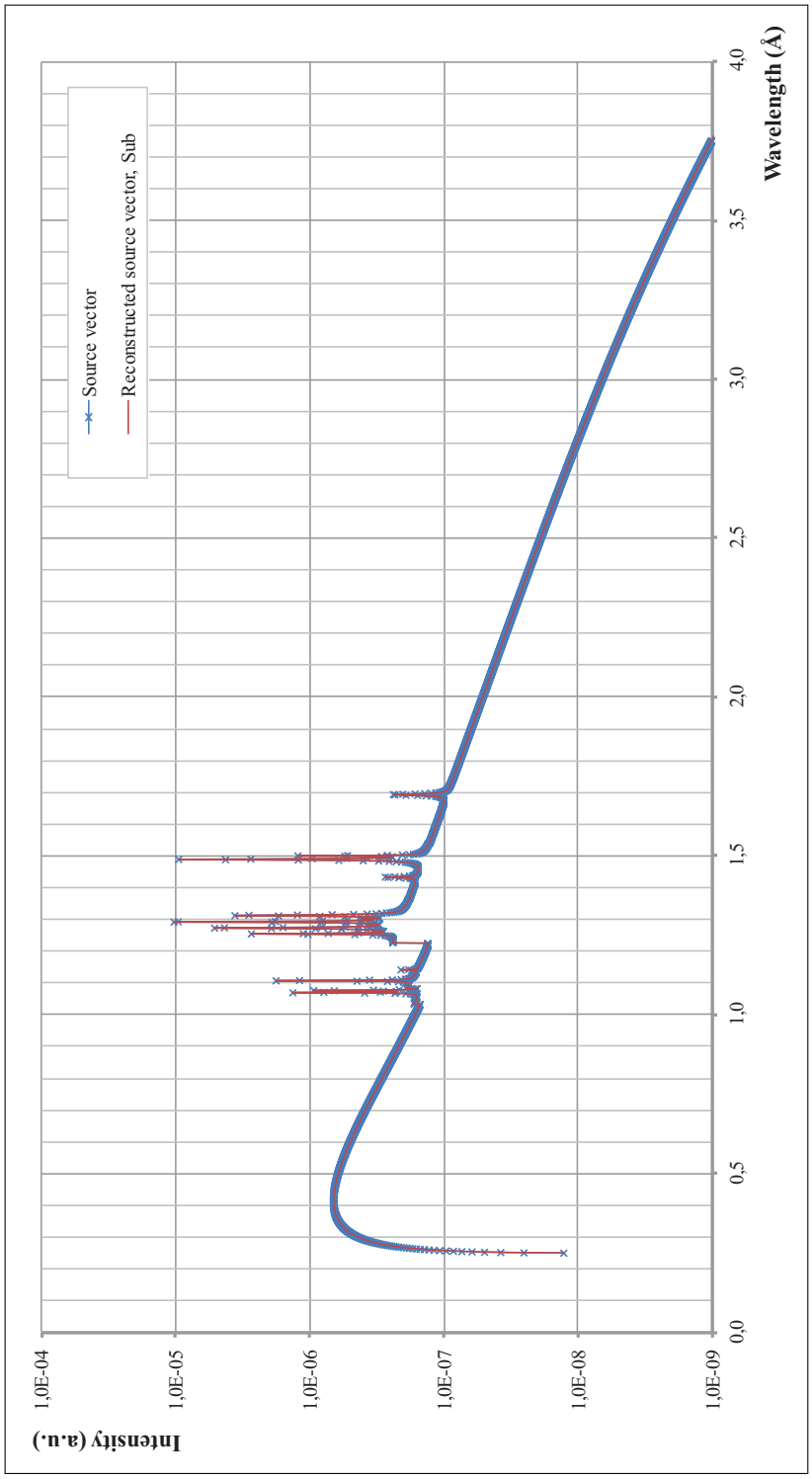

Figure 8.7: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vectors $\overrightarrow{s}_{\text{rec, Sub}}$. Unpreconditioned system. Aluminium case.

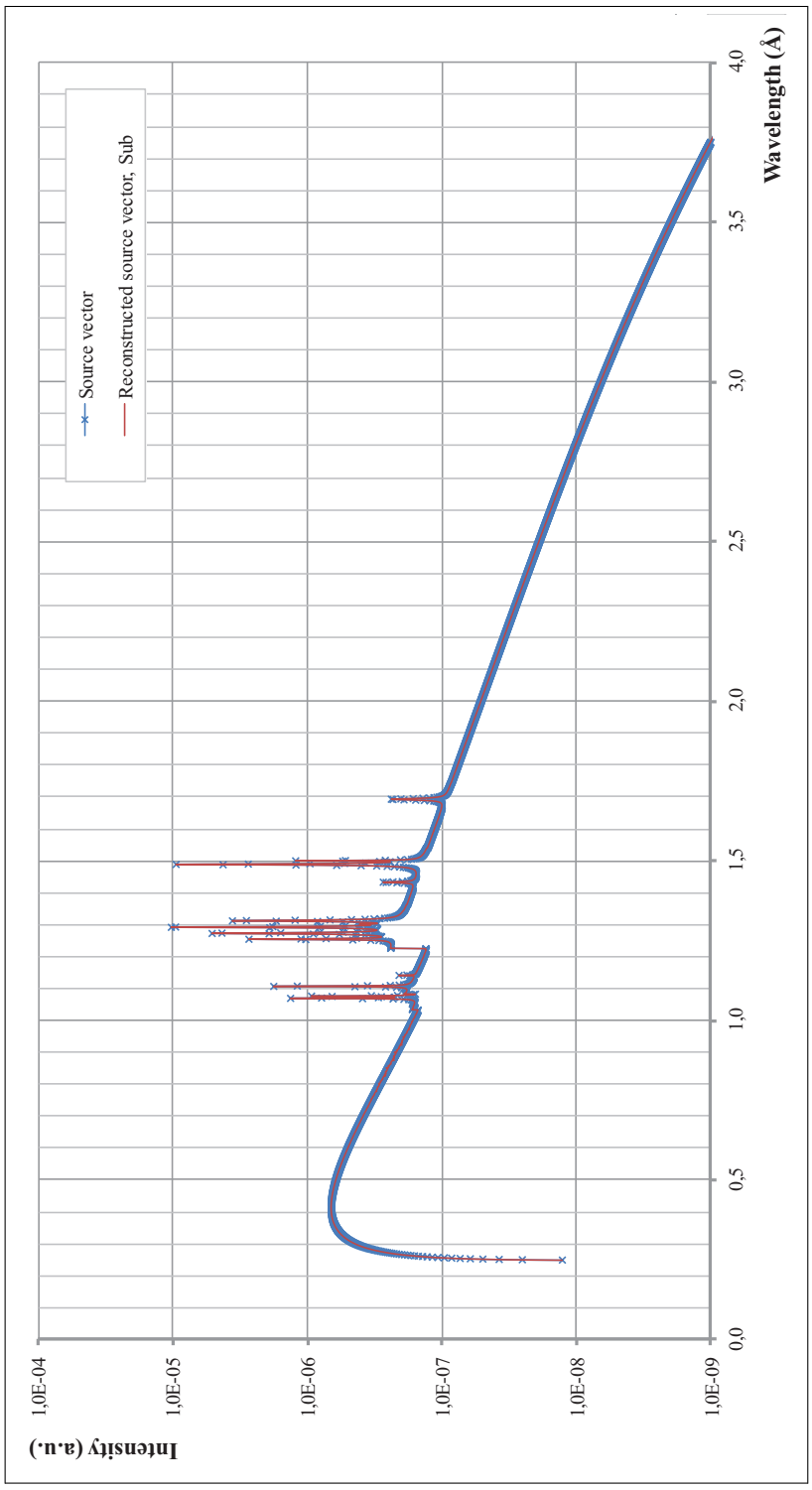

Figure 8.8: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. Right preconditioned system. Aluminium case.

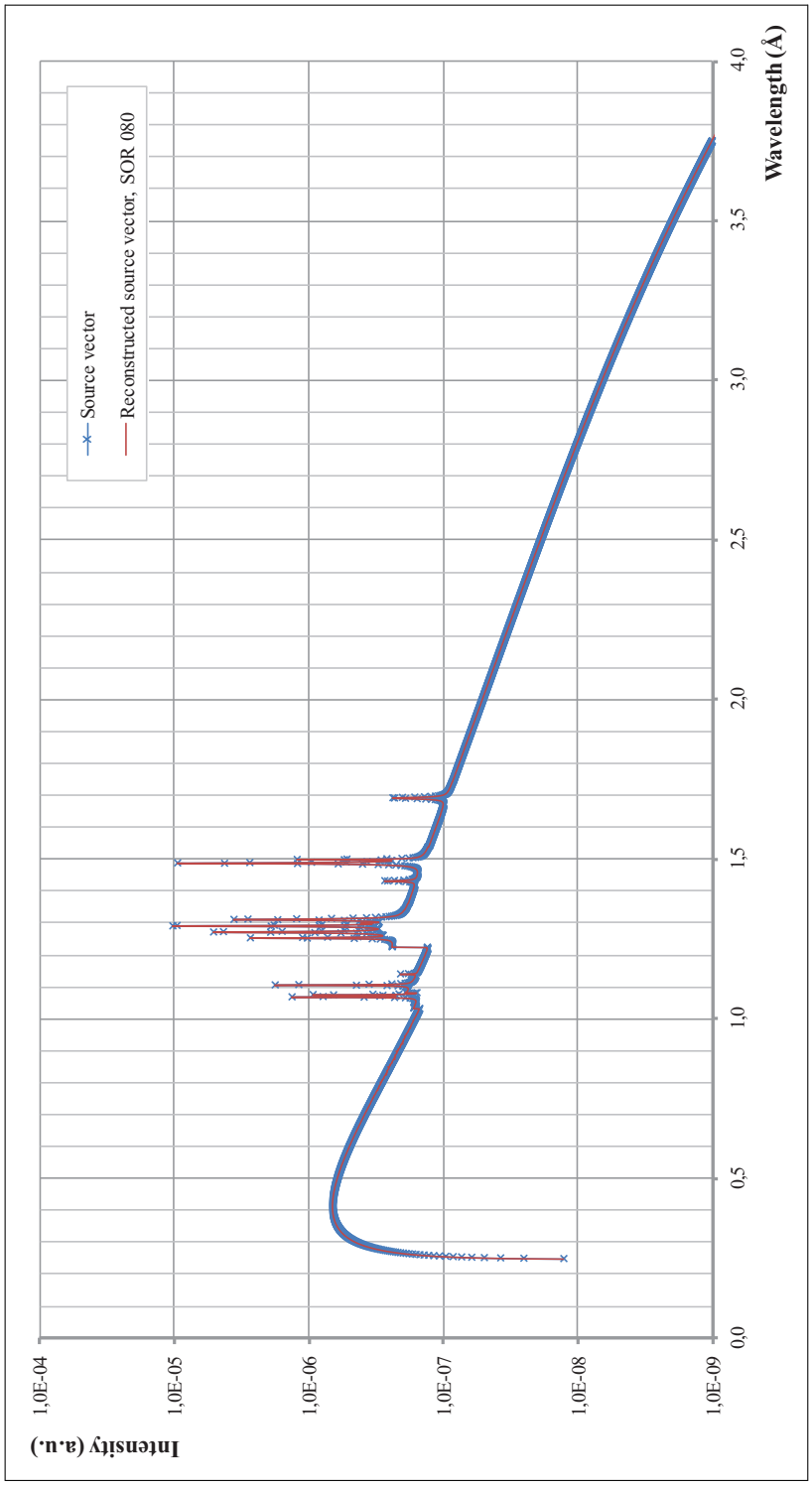

Figure 8.9: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vectors $\overrightarrow{s}_{\text{rec, SOR}}$. Left preconditioned system. Aluminium case.

8.3 Conclusions about the numerical experiments

The developments carried out in this chapter were aiming to solve the inverse scattering problem with a first numerical approach. Two aspects were investigated in particular:

- the numerical suitability of the scattering system, with and without preconditioning by the adjoint matrix;
- the assessment of the quality of reconstructed source vectors obtained by different numerical methods.

In the previous pages, different mathematical behaviors were observed when considering the carbon and the aluminium matrix systems. Before entering further into the conclusions, however, it is worthwhile to underline that the aluminium case was very theoretical. Regarding questions of target efficiency in producing photon scattering, aluminium targets are never used in practice, and this material has only been selected here to demonstrate the validity of the proposed procedure. We will then exclusively focus the conclusions on the carbon scattering system.

The unpreconditioned carbon matrix system was very intricate to solve, since the matrix was extremely ill-conditioned (whatever the subordinate matrix norm), indicating a very high sensitivity of the solution vector to small variations in the data of the problem. The preconditioning was then essential for transforming the system in a more suitable form, well-adapted to numerical operations. The condition number estimates were similar for both right and left preconditioning. Among the different techniques used for solving the inverse scattering problem, the SOR method applied on the left preconditioned system gave the best source vector reconstruction. The reconstructed vector was in a very good agreement with the numerical source vector in terms of continuum correspondance, peak intensity and peak position. Slight oscillations were however observed in a limited part of the spectrum. This disturbed interval was associated to a non diagonally dominant section of the coefficient matrix.

From these considerations, it becomes obvious that two different levels of analysis should be considered in order to get a physically meaningful solution to the inverse scattering problem:

1. the well-/ill-posed character of the matrix system should be estimated by means of the matrix condition numbers. This estimate gives a survey of the system sensitivity to small fluctuations in the data of the problem, then also on the stability of the solution vector. If necessary, the condition numbers may be reduced by the preconditioning technique, improving the spectral properties of the system. With the carbon target material, the left preconditioning with the adjoint matrix, i.e. the computation of the importance vector, was shown to be an efficient technique;
2. the diagonal dominance of the matrix is a crucial criterium for the convergence of the methods. A detailed analysis of the matrix structure then appears to be a key point for the deduction of relevant information concerning this convergence. Physically, the diagonal dominance of the forward scattering matrix is closely linked to the ratio between the RAYLEIGH interactions (main diagonal) and the COMPTON interactions (minor diagonal, whose position in the matrix depends on the wavelength discretization and on the δ parameter, as explained in section 7.1) with the target material. For solving the carbon matrix system, the SOR method has been identified as the most stable method, well adapted to the source vector reconstruction.

By denoting the RAYLEIGH and the COMPTON contributions in a row of the forward scattering matrix by $r_{i,i}$ and $c_{i,i-\delta}$ respectively, one may conclude that:

– the carbon scattering matrix is not diagonally dominant, i.e. $r_{(i,i)} \leq c_{(i,i-\delta)}$ in a great part of the matrix, and the forward system is highly ill-posed;

– the left preconditioning of the forward scattering matrix by the adjoint scattering matrix and the computation of the importance vector significantly improve the spectral properties of the system;

– a new system to solve, more suitable to numerical operations, is generated after preconditioning. Considering this preconditioned matrix system:

  - the convergence of the method to a physically meaningful solution is ensured for $r^2_{(i,i)} + c^2_{(i+\delta,i)} > r_{(i,i)}\ c_{(i,i-\delta)} + r_{(i+\delta,i+\delta)}\ c_{(i+\delta,i)}$, with $i = 1, 2, ..., n$;
  - the SOR method appears to be the most appropriate technique for solving the preconditioned system.

In the next chapter, the inverse procedure introduced in this chapter is applied for recovering the X-ray source spectrum from a set of experimental measurements, performed with a thin scattering graphite target.

Chapter

9 Application of the inverse procedure on real measurements

For the application of the inverse procedure on real measurements, a Compton spectrometer built at the Operational Unit of Health Physics of the University of Bologna was used. A cross section view of this spectrometer has already been shown, with the location of the different vectors used in the development and with the materials, in Figure 4.2, page 43. With this spectrometer, three measurements were performed at different high-voltages. This set of measurements forms the different starting points for the inverse procedure, i.e. the $\overrightarrow{m}$ vectors.

Practically, the system was based on a tungsten rotating anode tube, where electrons were produced by thermionic emission from a tungsten filament heated by an electric current, with a $0.8 \times 0.8\,\mathrm{mm}^2$ focus. The system was designed to operate until 150 kV. In the spectrometer, a 2 mm thick graphite target placed at a 45 ° angle in the primary photon beam has been used for the scattering. Graphite is an efficient scatterer that can easily be obtained in a pure state. For an energy of 10 keV, the photon mean-free-path in graphite is approximately 2 mm, while it is of course larger for higher energies. Consequently, the choice of this particular material and its thickness ensure a low content of multiple photon scattering in the target. It is worthwhile noting that the scattering target has been fixed on an extra thin mylar foil, to avoid extra collisions our of

the target. In order to isolate the photons scattered at a 90 ° angle with respect to the initial photon beam axis and within a narrow cone of 0.75 ° half-opening angle, two lead diaphragms have been inserted into the spectrometer. These internal structures were acting as perfect absorbers for all photons scattered outside the specified cone. Finally, the whole spectrometer was surrounded by a plexiglas tube, acting essentially as a structural guide for the detector alignment with the scattered beam.

The detector was a 10 mm diameter and 10 mm thick ORTEC HPGe detector, in a POP-TOP configuration. It was connected to a small DEWAR tank containing the liquid nitrogen necessary to cool the detector and the front-end electronics of the preamplifier to an optimal operating temperature. The entering window was made up by a thin beryllium foil, making possible to operate also at low photon energies with a high efficiency.

For additional information about the full experimental set-up, an extensive and detailed explanation can be found in [114].

A first set of three scattering measurements, illustrated in Figure 9.1, has been performed with the spectrometer at different high-voltages: 80 kV, 100 kV and 110 kV. The first step of the method, devoted to remove the detector effects from the spectrum, has been performed with the deconvolution codes described in Chapter 6. The different scattered vectors have been compared, and our best estimate of the scattered vector $\overrightarrow{b}$ selected for each measurement. The inverse scattering method has secondly been applied on the best estimates of the three scattered vectors $\overrightarrow{b}$, in order to reconstruct the X-ray source vectors $\overrightarrow{s}$. The quality of the source vector reconstruction has been estimated by comparison with a set of three direct measurements, corrected by the detector influences. These measurements have been performed with the same radiological device, at identical high-voltages than those of the measurement, but with a lower current intensity, decreasing then significantly the photon flux of the primary beam.

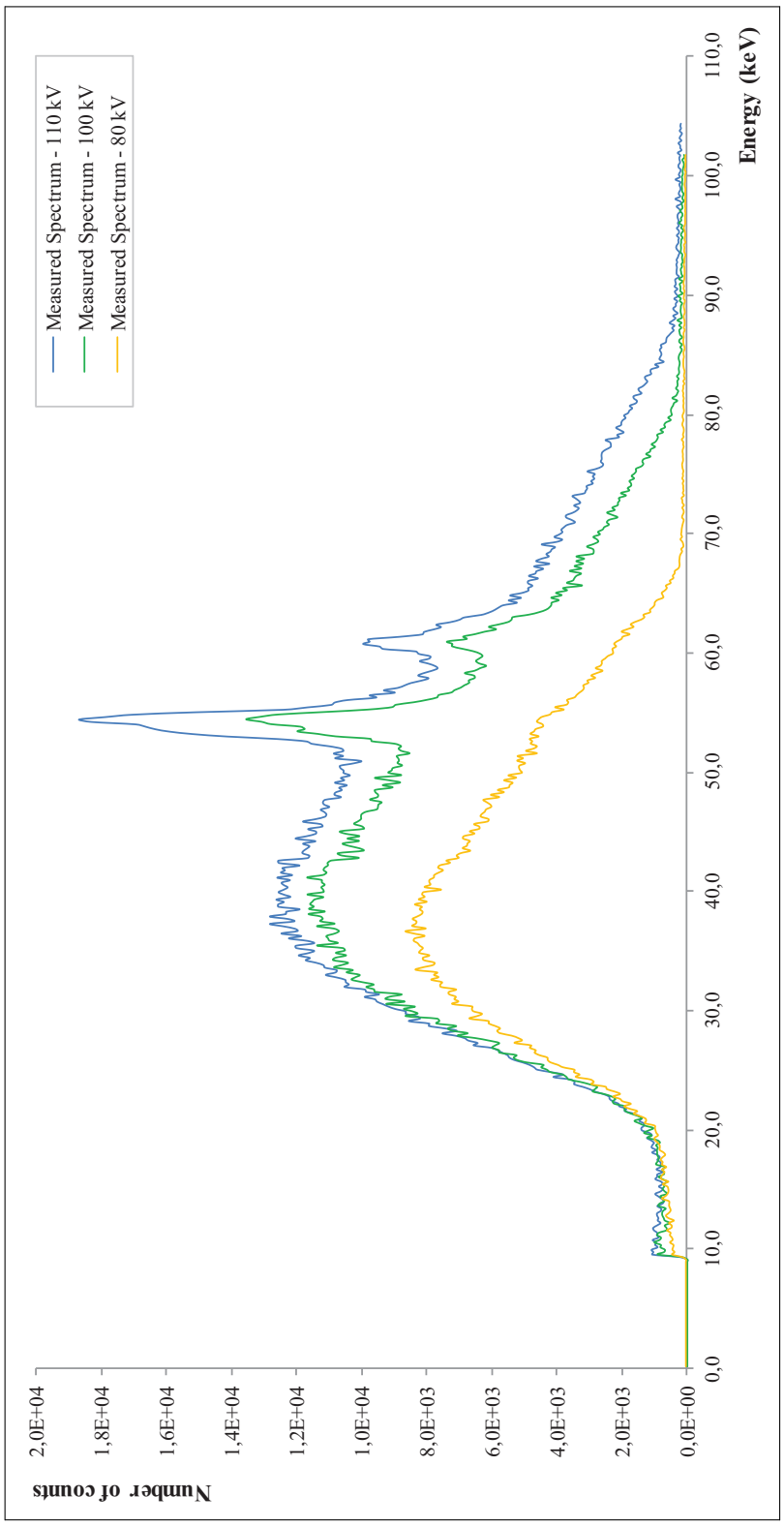

Figure 9.1: Measured spectra after scattering on the graphite target, and detected by the HPGe detector. The orange-colored spectrum has been measured at 80 kV, the green-colored spectrum at 100 kV and the blue-colored spectrum at 110 kV.

9.1 Unfolding of the measured vectors

In this section, the four deconvolution techniques described in Chapter 6 – namely, the TIKHONOV method, the truncated singular value decomposition (TSVD), an algorithm based on the maximum entropy principle (MAXED) and an iterative algorithm (GRAVEL) – are used to unfold the three measured spectra. For the application of the TIKHONOV and the TSVD methods, the '*Regularization Tools*' package [65] developed at the TECHNICAL UNIVERSITY OF DENMARK (Lyngby, Denmark) by P.-C. HANSEN has been used. This package consists in different routines to analyze and to compute stabilized solutions to discrete ill-posed problems. Both the GRAVEL (*grv_mc33*) and the MAXED (*mxd_mc33*) algorithms make part of the UMG package version 3.3 [88], and were developed at the PHYSIKALISH-TECHNISCHE BUNDESANSTALT (Braunschweig, Germany) by M. MATZKE and M. REGINATTO, respectively. The quality of these four methods has already been demonstrated in many papers, mostly focussed on neutron spectrometry. The application of unfolding techniques on X-ray spectra is here innovative, in particular for the GRAVEL and MAXED codes.

This section aims at comparing the unfolded spectra computed with the different codes, and at selecting our best estimate of the unknown scattered vectors $\overrightarrow{b}$ for the inverse scattering calculation.

9.1.1 Computation of the discretized response function

An important part of the spectrum calculation lies in the computation of the detector response matrix. As already mentioned, the response matrix contains all the information concerning the physics of the photon detection. In the discretized form used for numerical applications, each column of the matrix represents the response function of the detector to a monoenergetic excitation. The detector response matrix has been calculated using two different codes: PENELOPE and RESOLUTION [115].

Very broadly, the computer code system PENELOPE [116] (an acronym for PENetration and Energy LOss of Positrons and Electrons) performs Monte Carlo simulation of coupled electron–photon transport in arbitrary materials for a wide energy range, from a few hundreds of eV to about 1 GeV. Photon transport is simulated by means of the standard conventional simulation scheme, and is simulated through analytical differential cross-sections, derived from simple physical models and renormalized to reproduce accurate attenuation coefficients available from the literature [117]. In particular, the code takes into consideration the most important interactions in the X-ray regime: the photoelectric effect, the RAYLEIGH scattering and the COMPTON scattering. These interactions, explained in details in Chapter 2, are modelized with a sufficient level of details for most practical purposes. The effect of the velocity distribution of the electrons, i.e. the DOPPLER broadening, is also considered through the one-electron COMPTON profile. Electron and positron histories are generated on the basis of a mixed procedure (class II), which combines detailed simulation of hard events with condensed simulation of soft interactions. An additional geometry package permits the generation of random-photon showers in material systems consisting of different assemblies of arbitrary density homogeneous bodies limited by quadratic surfaces, i.e. planes, spheres, cylinders,... The code runs under supervision of a certain number of additional descriptive user-defined information for controlling the evolution of the tracks simulated by the code and keeping score of relevant quantities.

Practically, the measurements have been performed in an energy interval ranging from 0.13878 keV to 104.42449 keV, with an energy discretization step of 0.20408 keV. The detector response matrix has been obtained by simulating 512 monoenergetic response functions, within an energy interval similar than the one of the measurement, and with a constant energy increase of 0.20408 keV between each response function. The energy bin discretization in a single response function corresponds to the energy interval separating the different response functions. The resulting matrix system of equations is consequently mathemat-

ically compatible. Ideal response functions of the detector are computed in these conditions, without considering yet the energy resolution of the detector.

In order to reproduce the real response function of the HPGe detector, the spectral broadening has been added to the monoenergetic response functions. This step has been made by using the RESOLUTION code that performs the convolution between each theoretical response function obtained via PENELOPE and a GAUSSIAN function modeling the energy resolution of the detector. The energy resolution has been calculated on the basis of the full-width at half maximum (FWHM) variation in the measured spectra, in function of the energy, using the following calibration equation [73]:

$$E_{FWHM} = \sqrt{8\,(WFE + aE^b)\ln 2 + \Delta E_{elec}^2} \tag{9.1}$$

with:

- W, the average energy to produce an ion pair (keV);
- F, the FANO factor;
- a, a semi-empirical constant (keV);
- b, a semi-empirical constant;
- ΔE_{elec}, the electronic noise contribution (keV).

The analysis of some measured spectra leads to the following parameters:

- $W = 2.96\,10^{-3}\,\text{keV}$;
- $F = 6\,10^{-2}$;
- $a = 2\,10^{-6}\,\text{keV}$;
- $b = 2.0$;
- $\Delta E_{elec} = 0.17\,\text{keV}$.

This set of parameters has been implemented in equation 9.1 in order to apply the energy broadening. With these parameters, the FWHM follows the trend illustrated in Figure 9.2, in the energy range of the measurement.

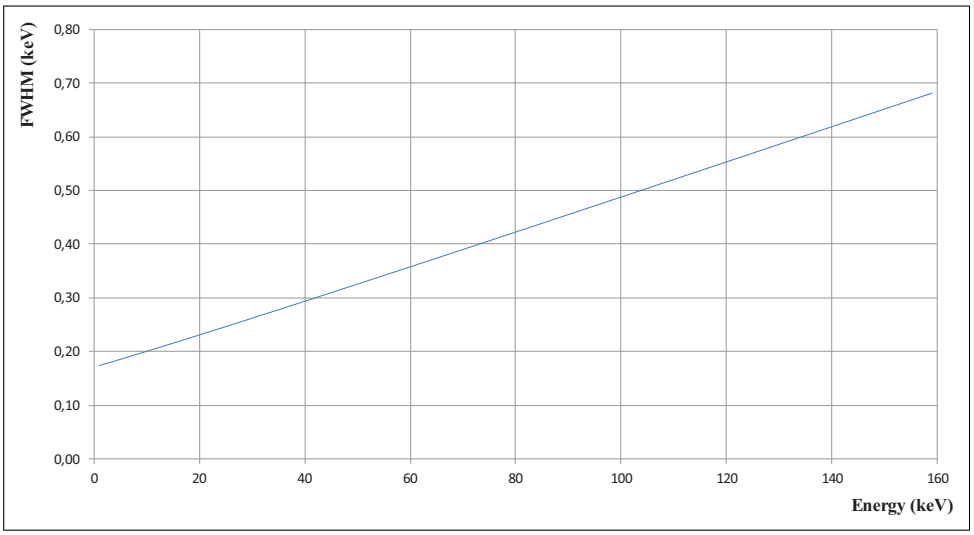

Figure 9.2: Calibration curve of the HPGe detector used for the scattering measurements.

Thanks to both the PENELOPE and the RESOLUTION codes, accurate response functions have been calculated. As an illustration, a response function obtained for an incident photon energy of 100 keV is shown in Figure 9.3 (semi-logarithmic graph). From the higher to the lower energies:

- the photopeak can first be identified at 100 keV, i.e. at the energy of the incident photon. This peak corresponds to the total absorption of the photon energy in the detector;
- the two peaks at lower energies are usually termed as escape peaks. These peaks arise whenever a fixed amount of energy is lost from the detector with a significant probability. If the energy of the incoming X-ray photon is greater than the absorption edges of the detector material, it can produce characteristic X-rays, called fluorescence photons. If these photons escape from the active part of the detector without undergoing any interaction, their energy is lost during the detection process. The escape of the characteristic X-rays from germanium following photoelectric absorption can be significant, especially for small detectors with a large surface-to-volume ratio like our HPGe. Two peaks are consequently often found in

the spectrum with an energy lower than the photopeak energy, and with an energy difference corresponding to the characteristic K X-ray energies for germanium, i.e. 9.87 keV (K_α) and 10.98 keV (K_β). In the case of a 100 keV photon, the corresponding X-ray escape peaks then have energies of 91.13 keV and 89.02 keV, as observed in Figure 9.3. It is worthwhile noting that escape peaks will be more prominent for incident low-energy photons, because photoelectric absorption is then most probable, and all interactions will tend to occur near to the detector surface.

– at low energies (between 0 and approximately 38.12 keV in the case described here), the COMPTON continuum is observed.

Some selected response functions are shown in Figure 9.4 (semi-logarithmic graph), for incident photon energies of approximately 20, 40, 60, 80 and 100 keV, i.e. columns 100, 200, 300, 400 and 500 of the detector response matrix, respectively. The description made for a single response function is of course transposable on each of these distributions. The response matrix is partly represented in 3-D in Figure 9.5, for incident photon energies between 11.1592 keV and 21.1592 keV (50 response functions have been included in the graph). This figure gives a very interesting glimpse to the response matrix.

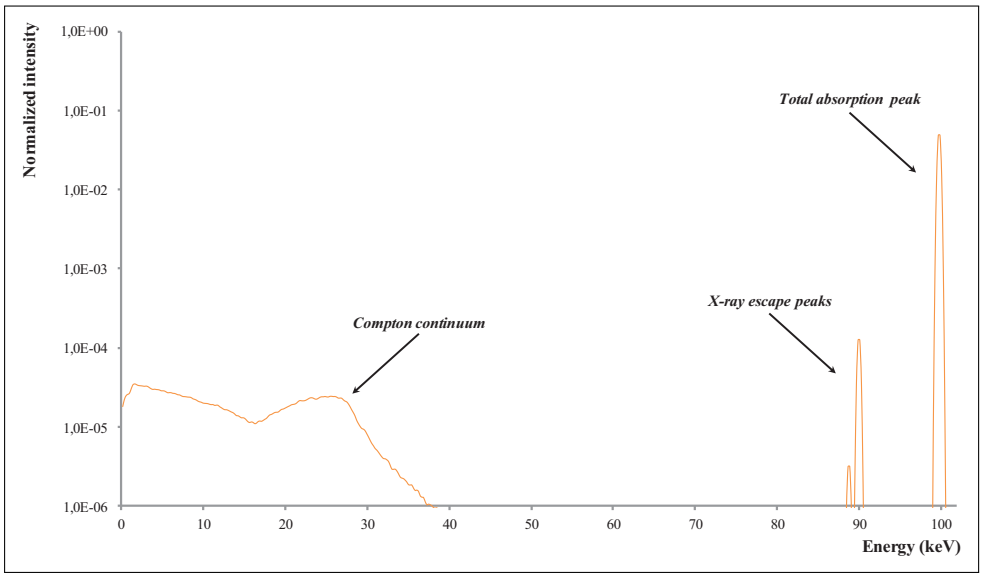

Figure 9.3: Description of the detector response function computed with the PENELOPE and RESOLUTION codes, for a 100 keV incident photon.

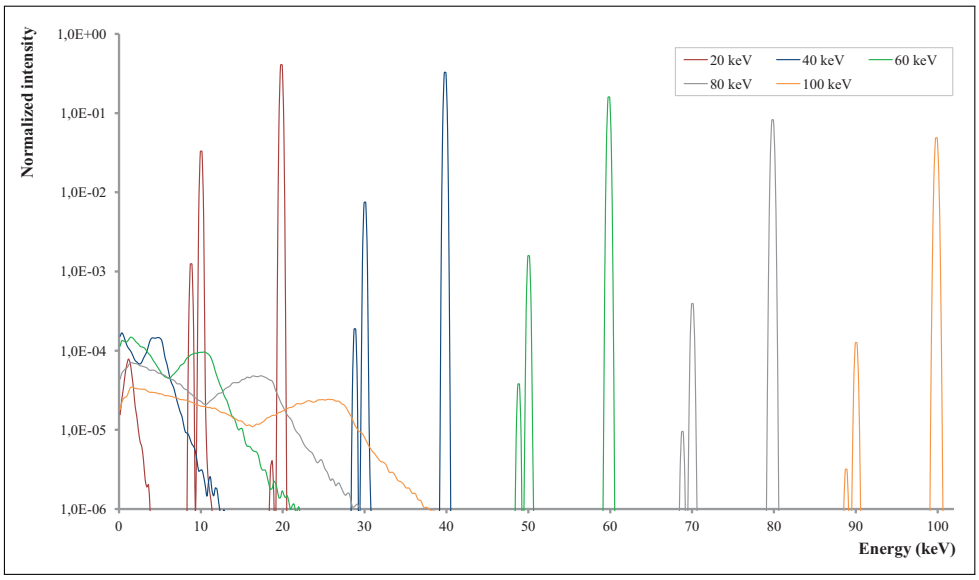

Figure 9.4: Response functions computed with the PENELOPE and RESOLUTION codes, for monoenergetic excitations of approximately 20, 40, 60, 80 and 100 keV (columns 100, 200, 300, 400 and 500 of the response matrix, respectively).

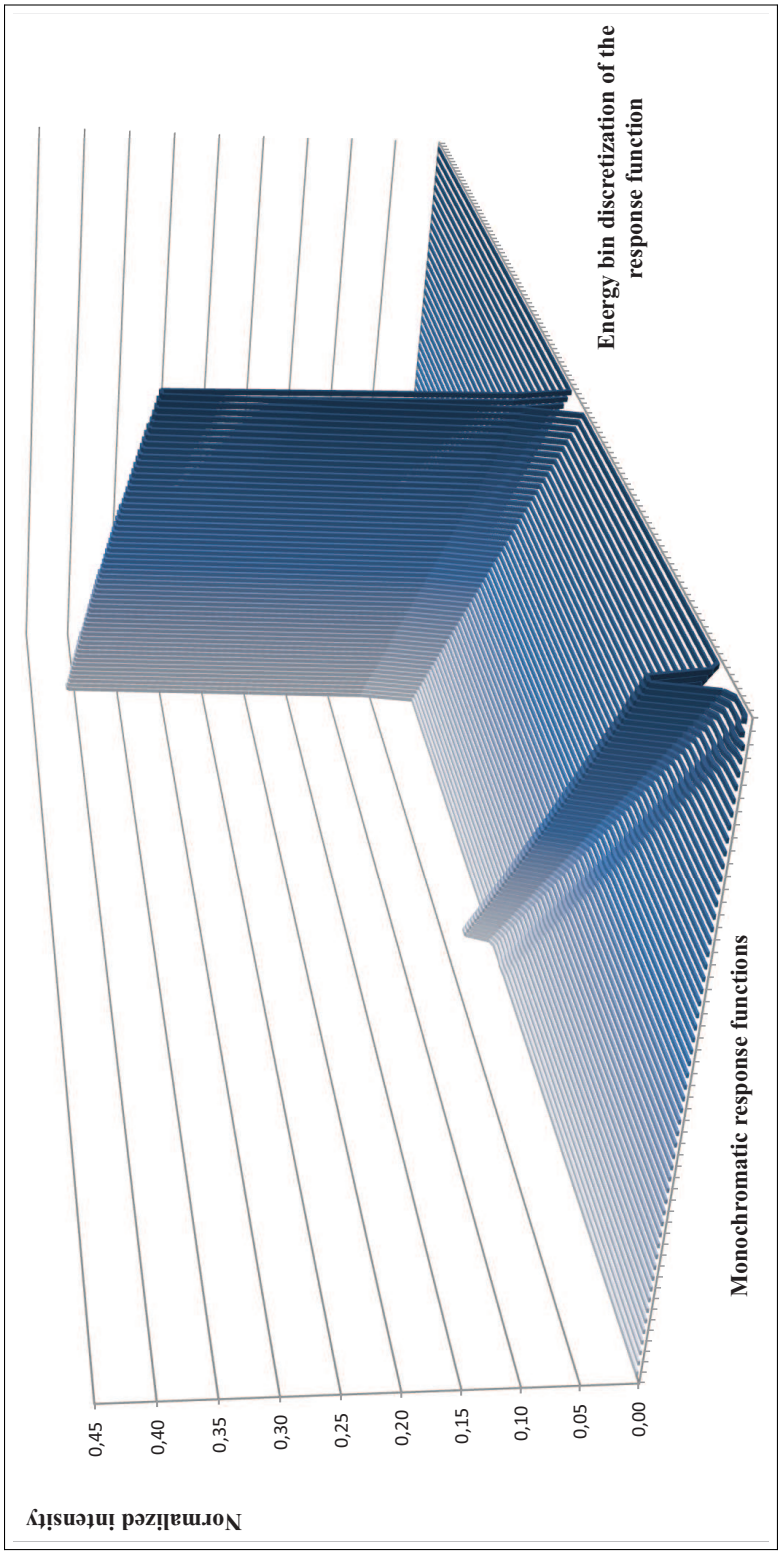

Figure 9.5: 3-D illustration of the discrete unfolding matrix. Each response function corresponds to a monoenergetic excitation. The illustration is made for energies between 11.1592 keV and 21.1592 keV (50 response functions are included).

9.1.2 Smoothing of the measured spectrum

Among other control parameters, the success of any unfolding procedure relies on the statistical quality of the detector response functions. For that purpose, simulations have been run with a high number of particles. The response matrix then has an excellent statistical behavior, since all the response functions contain very few statistical fluctuations. In order to optimize the quality of the unfolded spectra, a good statistical behavior of the input spectra is also expected. Should the opposite occur, the unfolding usually leads to perturbations in the calculated spectra in the form of artificial oscillations, even with additional criteria like positivity or minimal residual norm, and even with a well-behaved response matrix. Two different approaches are possible to ensure the statistical quality of a measured spectrum: longtime measurements, or filtering procedures. Since longtime measurements are often not possible, a filtering procedure has been favored in the next development.

In order to set the problem in sufficiently good conditions for the unfolding, oscillations in the measured spectra $\overrightarrow{m}$ (cf. Figure 9.1) have been reduced by using a Savitzky-Golay smoothing procedure [118, 119] (sometimes also referred to as DISPO - digital smoothing polynomial - filter). This filter is based on a local polynomial regression of degree k on a series of values (at least $k+1$ points which are considered to be equally spaced in the series) to determine the smoothed value of each point. The advantage of this smoothing procedure is that the distribution features such as relative maxima, minima and width, which are usually flattened by other adjacent averaging techniques, are preserved.

In our case, the best smoothing of the measured spectra has been obtained with a third degree polynomial, applied on a 7 points window. With these particular parameters, the original shape of the measured spectrum $\overrightarrow{m}$ is very well preserved, both in the continuum and in the peaks regions. The smoothed measured spectra $\overrightarrow{\widetilde{m}}$ are compared to their corresponding measured spectra $\overrightarrow{m}$ in Figure 9.6 for the 80 kV measurement, in Figure 9.7 for the 100 kV measurement and in Figure 9.8 for the 110 kV measurement.

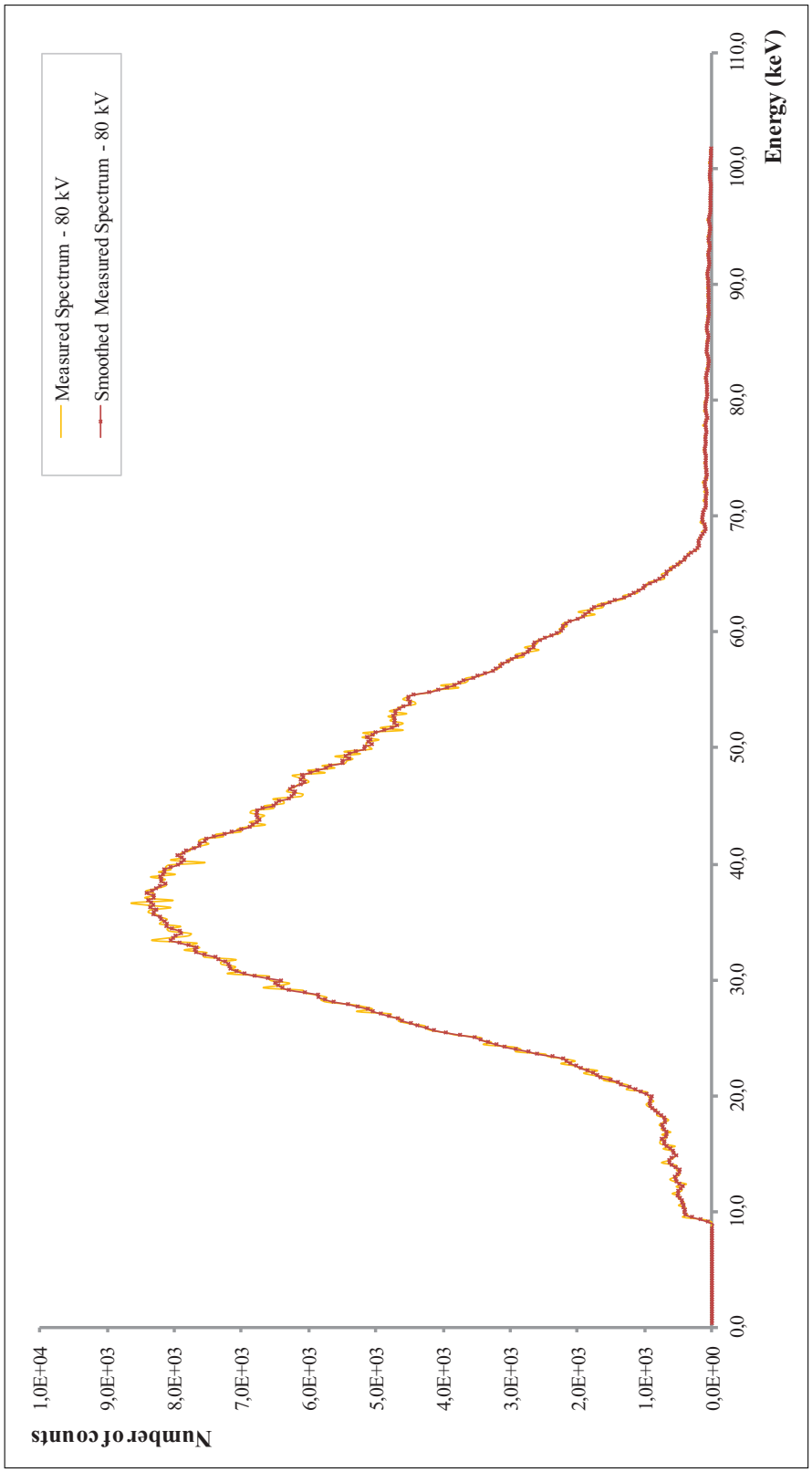

Figure 9.6: Smoothed measured spectrum $\widetilde{\overrightarrow{m}}$ by the Savitzky-Golay filter and measured spectrum $\overrightarrow{m}$ at 80 kV.

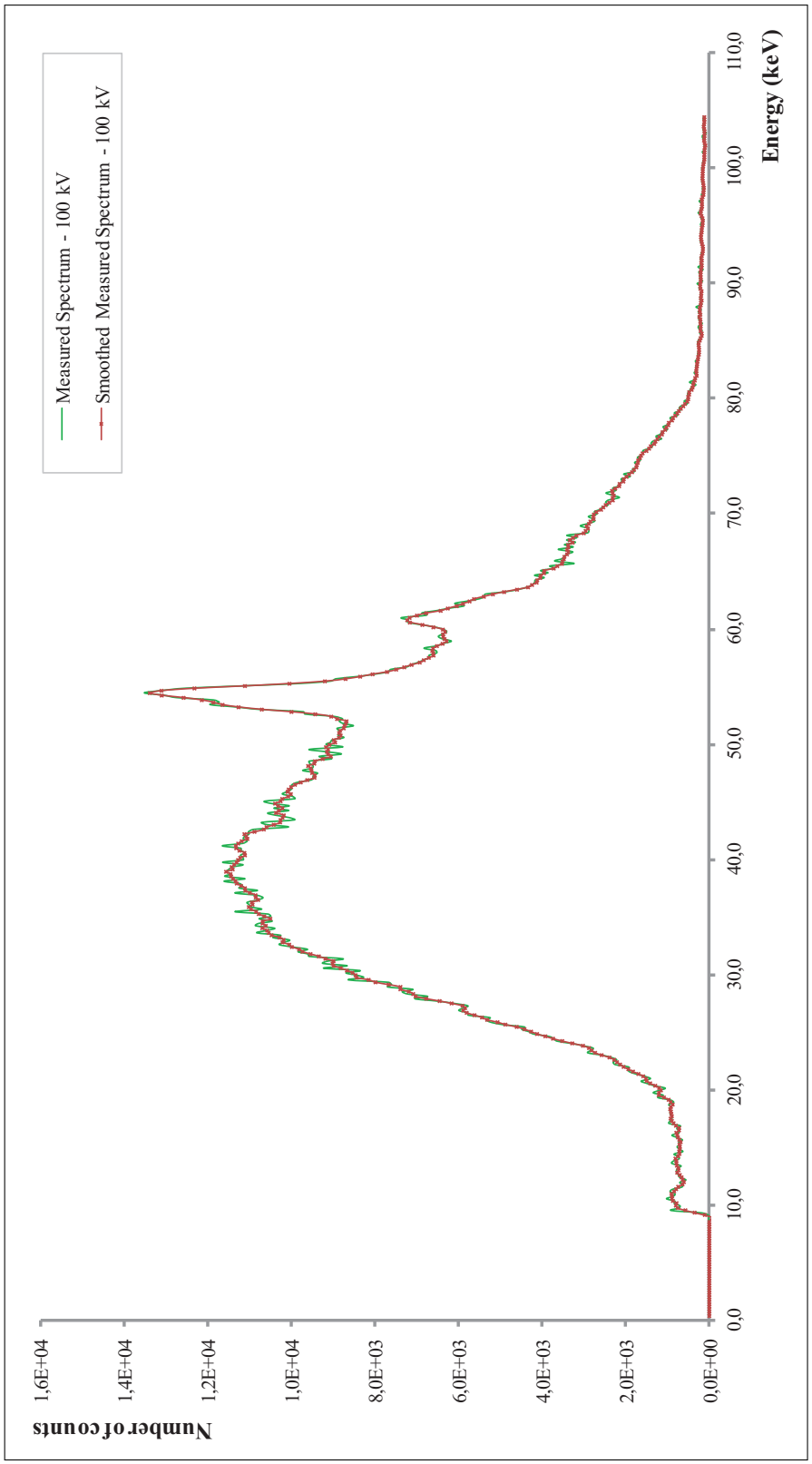

Figure 9.7: Smoothed measured spectrum $\overrightarrow{\tilde{m}}$ by the Savitzky-Golay filter and measured spectrum $\overrightarrow{m}$ at 100 kV.

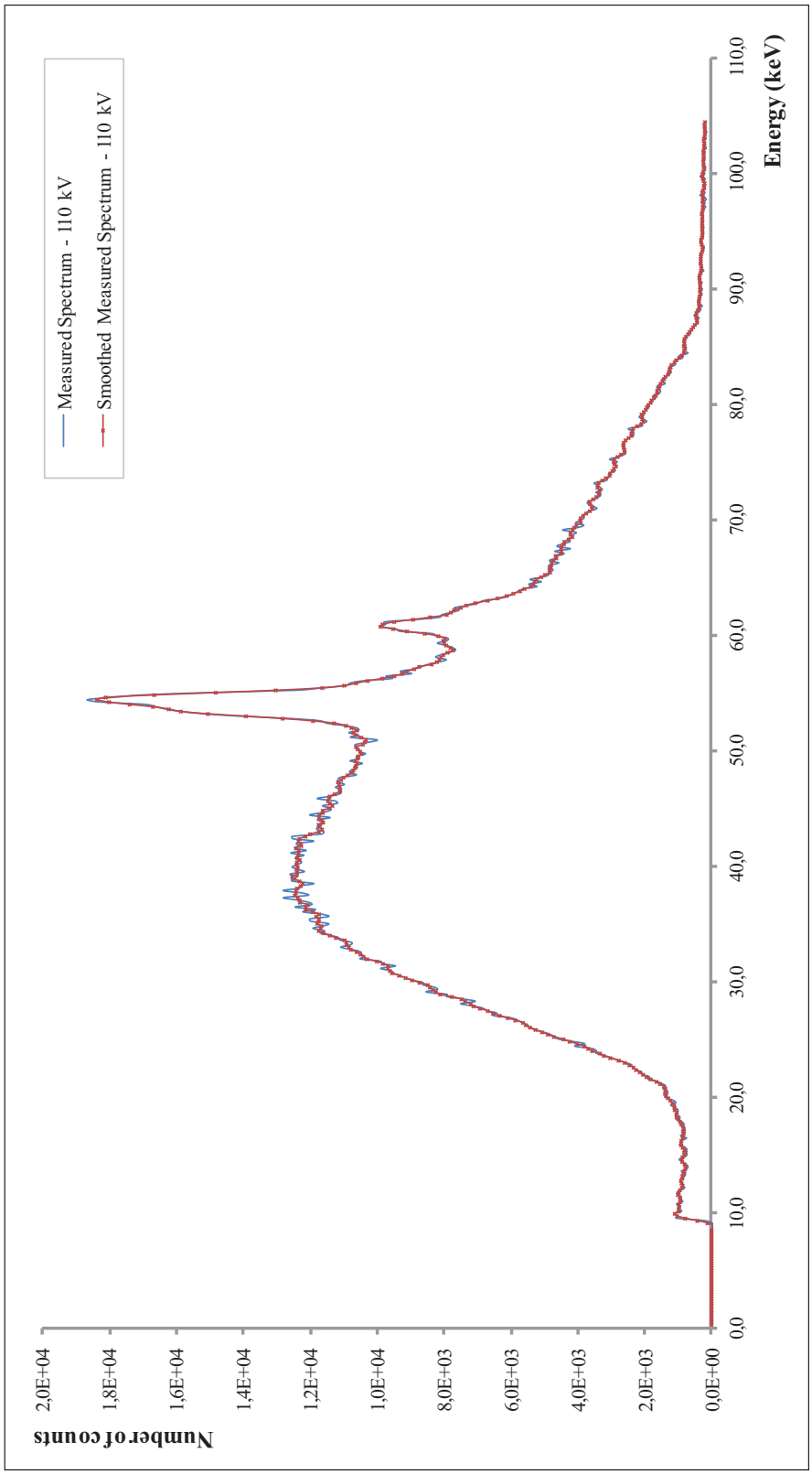

Figure 9.8: Smoothed measured spectrum $\widetilde{\overrightarrow{m}}$ by the Savitzky-Golay filter and measured spectrum $\overrightarrow{m}$ at 110 kV.

9.1.3 Comparison of the unfolding methods: selection of the scattered vector

The three smoothed measured spectra $\overrightarrow{\widetilde{m}}$ have been unfolded with the four methods previously mentioned, aiming at evaluating the best estimate of the unknown scattered vectors $\overrightarrow{b}$. Let us denote the scattered vector obtained from the measurement at 80 kV by $\overrightarrow{b}_{(80)}$, at 100 kV by $\overrightarrow{b}_{(100)}$ and the one obtained from the measurement at 110 kV by $\overrightarrow{b}_{(110)}$.

For the TIKHONOV and the TSVD regularization techniques, the regularization and truncation parameters (Γ and k, respectively) have been computed by using the L-curve algorithm proposed in the '*Regularization Tools*' package. For the TIKHONOV method, a default spectrum set to 0 has been used, as suggested in the package manual. With the GRAVEL code, the scattered vectors have been computed starting from a default spectrum equal to the measurement, a χ^2 per degree of freedom equal to 1, and a maximum number of iterations set to 10. It has been observed that, beyond this number of iterations, the spectrum is not anymore modified in its general shape, but high frequency oscillations appear in the unfolded spectra. This has been noticed visually, and mathematically by periodic variations in the calculated χ^2. The MAXED algorithm has been run starting from the same default spectrum, with a maximum number of iterations equal to 1000, and with a χ^2 factor per degree of freedom equal to the minimum GRAVEL χ^2 for each measurement. In these conditions, the best estimates of the scattered vectors $\overrightarrow{b}$ were obtained in less than 50 iterations.

The best estimates of the scattered vectors $\overrightarrow{b}_{(80)}$, $\overrightarrow{b}_{(100)}$ and $\overrightarrow{b}_{(110)}$ are compared in Figure 9.9, in Figure 9.10 and in Figure 9.11 respectively. As a first general observation, the four methods give very similar results. However, some differences in the reconstructions may be underlined:

– for the scattered vector $\overrightarrow{b}_{(80)}$:

  - in the continuum part of the different unfolded spectra, the MAXED estimate presents a high level of oscillations, mainly situated between

35.5 keV and 64.8 keV. The two regularization methods give rise to very similar vectors, with oscillations included in the energy range between 35.2 keV and 54.8 keV. The GRAVEL estimate is the most satisfactory, since it has the lowest level of oscillations.

– for the scattered vector $\overrightarrow{b}_{(100)}$:

- in the continuum part of the spectra, the TIKHONOV and the MAXED algorithms give rise to oscillatory unfolded spectra, mainly for energies between 28.32 keV and 74.35 keV. The behaviors of the TSVD and the GRAVEL estimates are more satisfactory from this statistical point of view;
- the positions of the peaks are in perfect agreement for all the methods (first peak at 54.45 keV, second peak at 60.84 keV);
- except for the MAXED code where oscillations on the right part of the second peak disturb the calculation, the FWHM of the peaks are identical for all the methods (first peak: 2.04 keV, second peak: 1.84 keV);
- the intensity of the peak at 53.46 keV is very well estimated by all the methods, although small variations (less than 3 % between the lower and the higher intensities) may be observed;
- the intensity of the peak at 60.75 keV is also very well estimated, and the level of fluctuations is smaller than that of the first peak.

– for the scattered vector $\overrightarrow{b}_{(110)}$:

- in the continuum part of the spectra, the TIKHONOV and the MAXED estimates of the scattered vectors show a considerable level of oscillations in the energy intervals ranging from 56.70 keV to 59.99 keV and from 65.64 keV to 70.09 keV. The continuum of the TSVD and the GRAVEL algorithms are satisfactory on the whole energy range, and in very good agreement between each other;

- in all the estimates of the scattered vectors, the two peaks are exactly at the same energy after unfolding (first peak at 53.46 keV, second peak at 60.95 keV);
- the FWHM of the peaks are identical for the TIKHONOV, TSVD and GRAVEL algorithms (first peak: 2.04 keV, second peak: 1.63 keV). The poor result of the MAXED code is again a consequence of the oscillations in the spectrum;
- the intensities of both peaks are very similar for all the methods, even if small variations in their intensities may still be observed. In particular, the TSVD seems to overestimate both peaks, while the GRAVEL code seems to underestimate them slightly.

From these considerations, it becomes apparent that all the methods are in very good agreement between each other concerning the shape of the continuum and the peaks positions. However, an important level of oscillations has been observed when unfolding the smoothed measured spectra with the TIKHONOV method and the MAXED algorithm. In most cases, the TSVD technique and the GRAVEL code give rise to extremely similar results. Because of its interesting underlying physical meaning, it has been decided to continue the inverse procedure with the GRAVEL estimates of the scattered vectors $\overrightarrow{b}$.

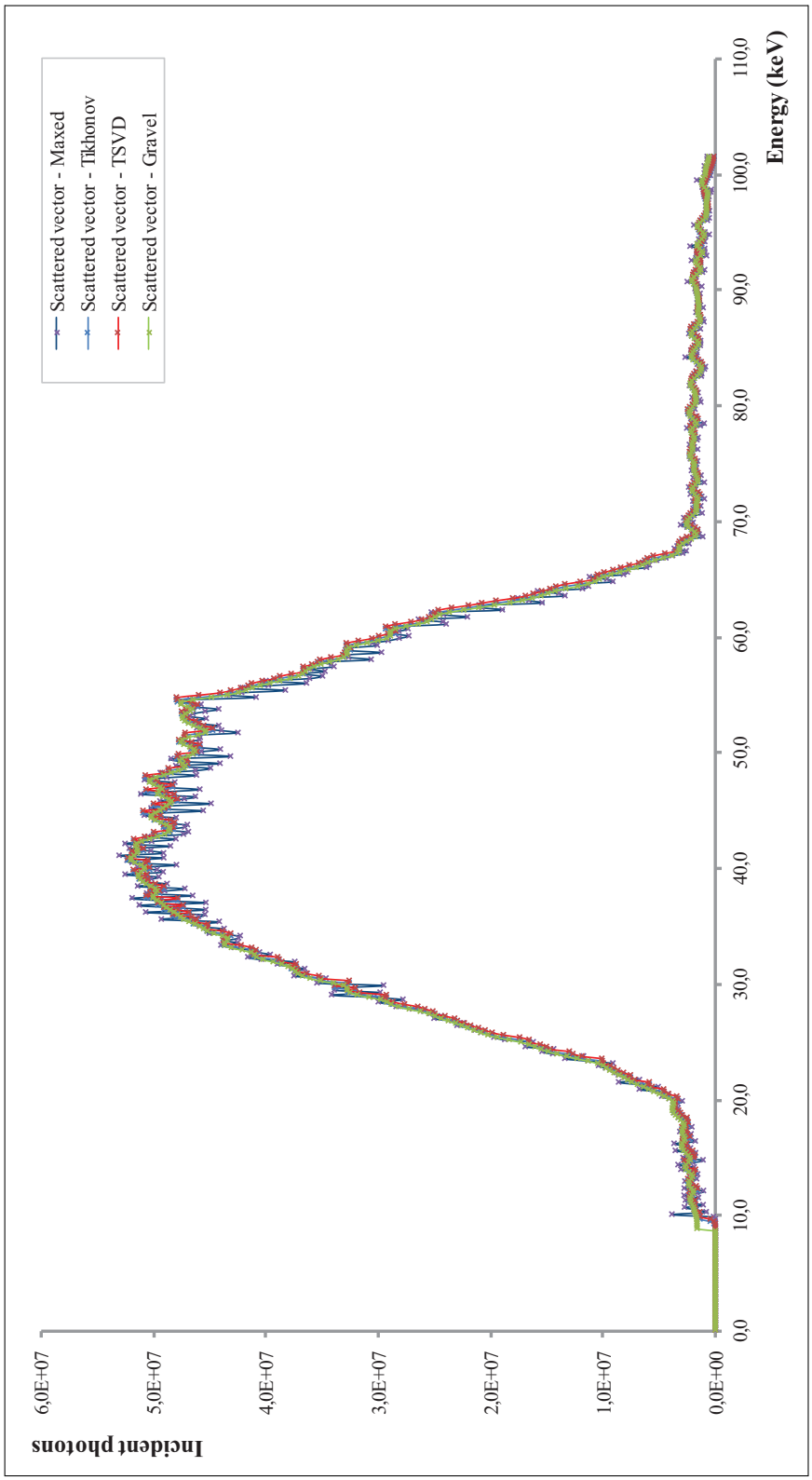

Figure 9.9: Comparison of the estimated scattered vectors $\overrightarrow{b}_{(80)}$ after unfolding of the smoothed measured spectrum $\overrightarrow{\widetilde{m}}$ at 80 kV by the four methods.

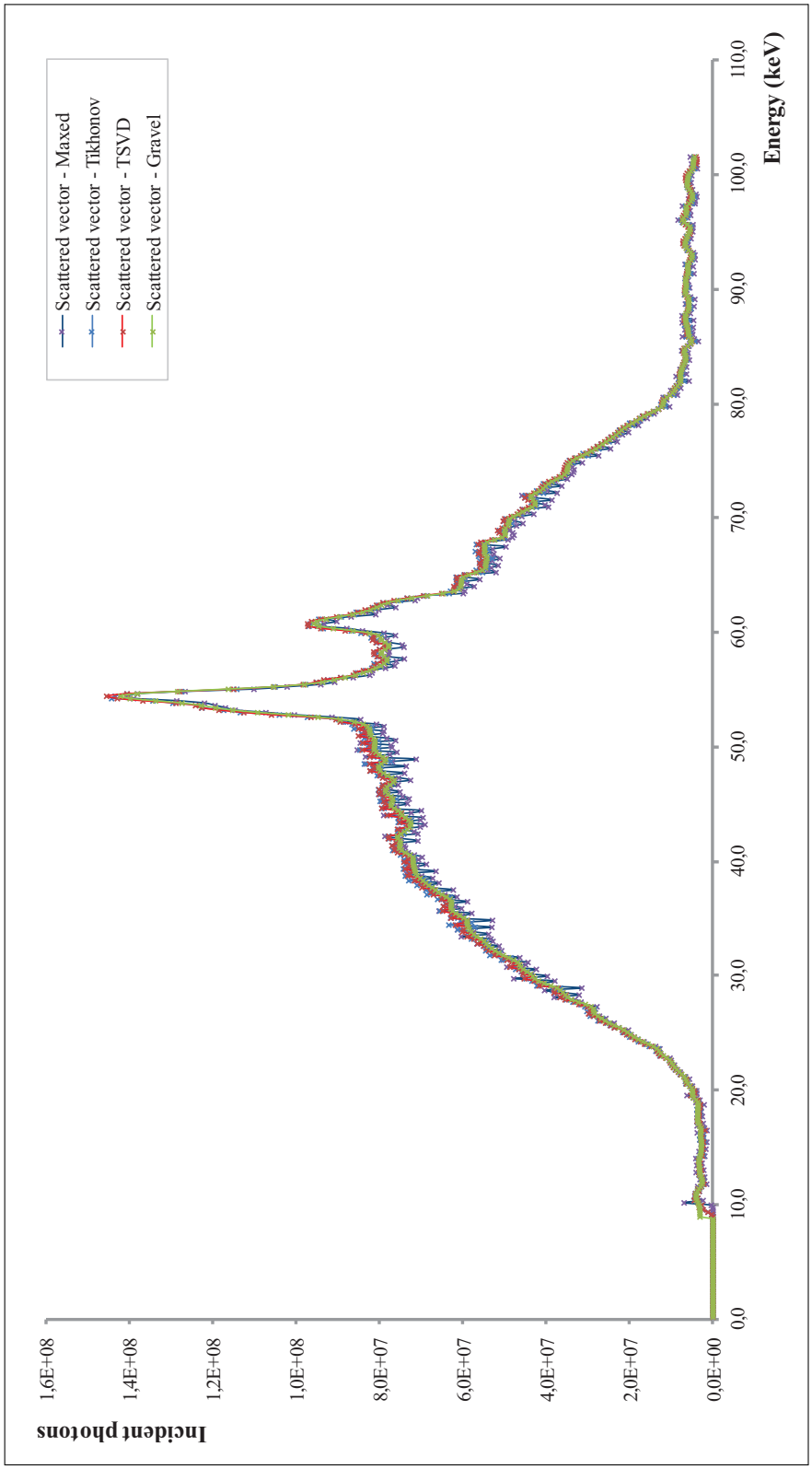

Figure 9.10: Comparison of the estimated scattered vectors $\overrightarrow{b}_{(100)}$ after unfolding of the smoothed measured spectrum $\overrightarrow{m}$ at 100 kV by the four methods.

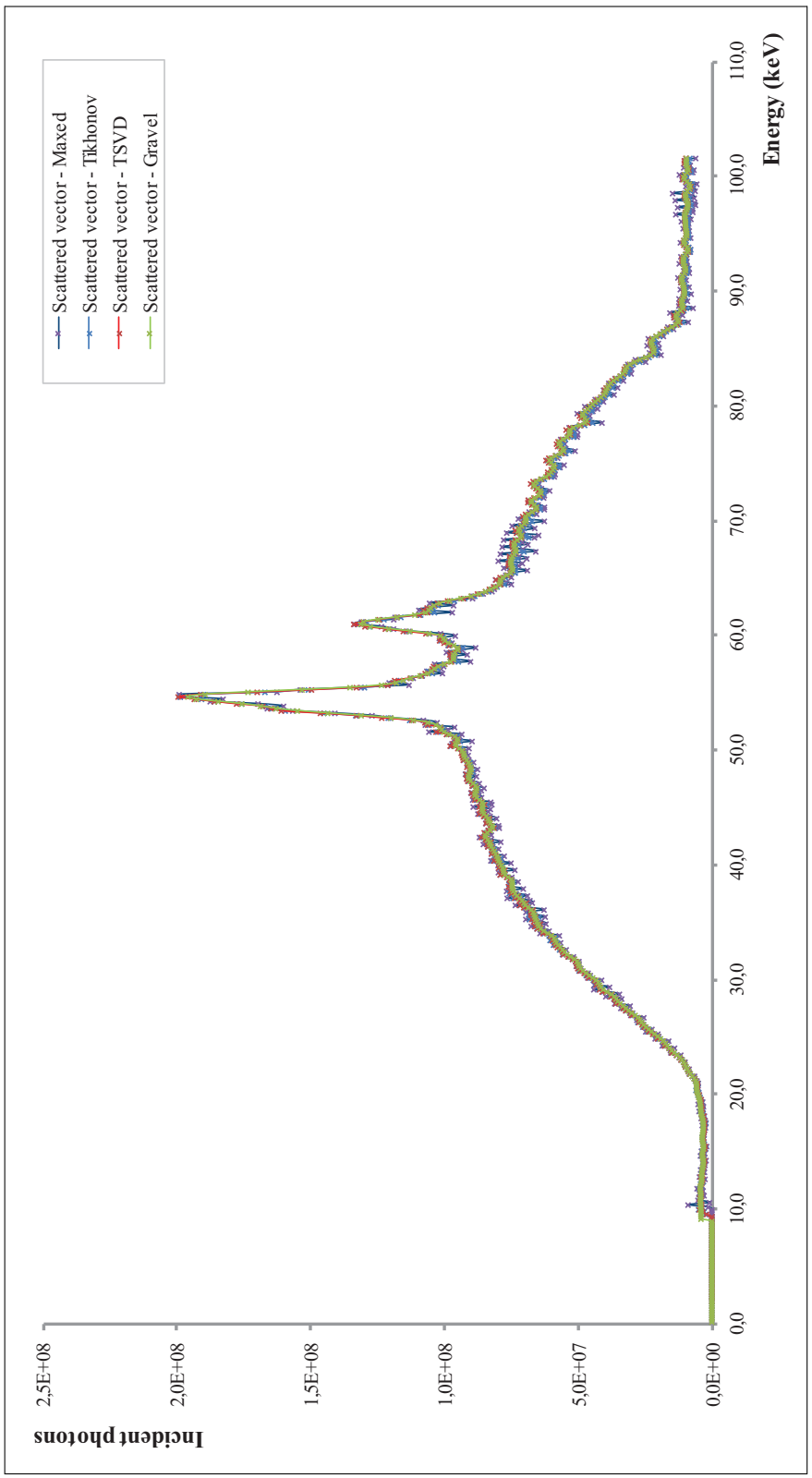

Figure 9.11: Comparison of the estimated scattered vectors $\overrightarrow{b}_{(110)}$ after unfolding of the smoothed measured spectrum $\overrightarrow{\widetilde{m}}$ at 110 kV by the four methods.

9.2 Inverse scattering in the spectrometer: calculation of the source vector

In order to apply the inverse technique developed in Chapter 8, the scattered vectors $\overrightarrow{b}_{(80)}$, $\overrightarrow{b}_{(100)}$ and $\overrightarrow{b}_{(110)}$ have to be expressed in terms of wavelength, not in energy. The infinitesimal variation of energy dE, is linked to the infinitesimal variation of wavelength $d\lambda$, by:

$$dE = -\frac{hc}{\lambda^2}\, d\lambda \tag{9.2}$$

where h is the PLANCK constant, and c the speed of the light. Using this conversion, the inverse systems for 80 kV, 100 kV and 110 kV have been solved in the wavelength regime, in which the forward scattering matrix has the very particular and attractive bi-diagonal structure, under computation approximations similar to those in section 7.1. For the sake of convenience, the source vectors have finally been converted back into the energy regime.

In the following sections, the stability of the forward transport matrix is first estimated with the condition numbers. Depending on the matrix ill-conditioning, and then taking into account the potential ill-posedness of the system of equations, a numerical method is selected following the rules deduced from the numerical experiments in Chapter 8. The three systems of equations (for 80 kV, 100 kV and 110 kV) are then solved, aiming at reconstructing the X-ray source vector. In order to estimate the quality of the reconstruction, a comparison of the reconstructed source vectors with direct measurements is made.

9.2.1 Spectral conditioning of the coefficient matrix

The condition numbers of the unpreconditioned forward scattering matrix F_{scatt} have been calculated in different p-norms, for $p = 1$, $p = 2$ and $p \to \infty$:

- $\kappa_1(F_{scatt}) = 6.89077\,10^{51}$;
- $\kappa_2(F_{scatt}) = 3.60386\,10^{51}$;

– $\kappa_\infty(F_{scatt}) = 2.95327\,10^{51}$.

These condition numbers are in perfect agreement with the matrix norms properties. They are extremely high, indicating that the forward scattering matrix is extremely ill-conditioned. In these conditions, the unpreconditioned system is highly ill-posed, i.e. very sensitive to small variations in the data of the problem. The solution vectors are extremely oscillating, and the oscillation level is significantly amplified from the low to the high energies.

As concluded in Chapter 8, the left preconditioning by the adjoint matrix and the importance vector computation should significantly improve the numerical stability of the system, and permit to generate a more adapted form of the system to solve with respect to numerical operations. The condition numbers of the left preconditioned coefficient matrix by the adjoint matrix $P_{left} = F_{scatt}^T F_{scatt}$ are:

– $\kappa_1(P_{left}) = 5.82399\,10^{17}$;

– $\kappa_2(P_{left}) = 2.44328\,10^{17}$;

– $\kappa_\infty(P_{left}) = 5.82386\,10^{17}$.

Again, all these estimations are consistent with the matrix p-norm properties. As expected, the condition numbers are significantly reduced (factor of approximately 10^{34} for all the p-norms considered), making the matrix system of equations considerably more suitable to numerical operations. The left preconditioned system has better numerical properties, and is less sensitive to fluctuations in the data of the problem. This system is solved in the following.

9.2.2 Inverse scattering on the graphite target

Using graphite as target material, the coefficient matrix of the left preconditioned scattering system is not diagonally dominant. As suggested in Chapter 8, the SOR method has to be used for solving the matrix system of equations. The different reconstructed source spectra at 80 kV, 100 kV and 110 kV are shown in Figure 9.12.

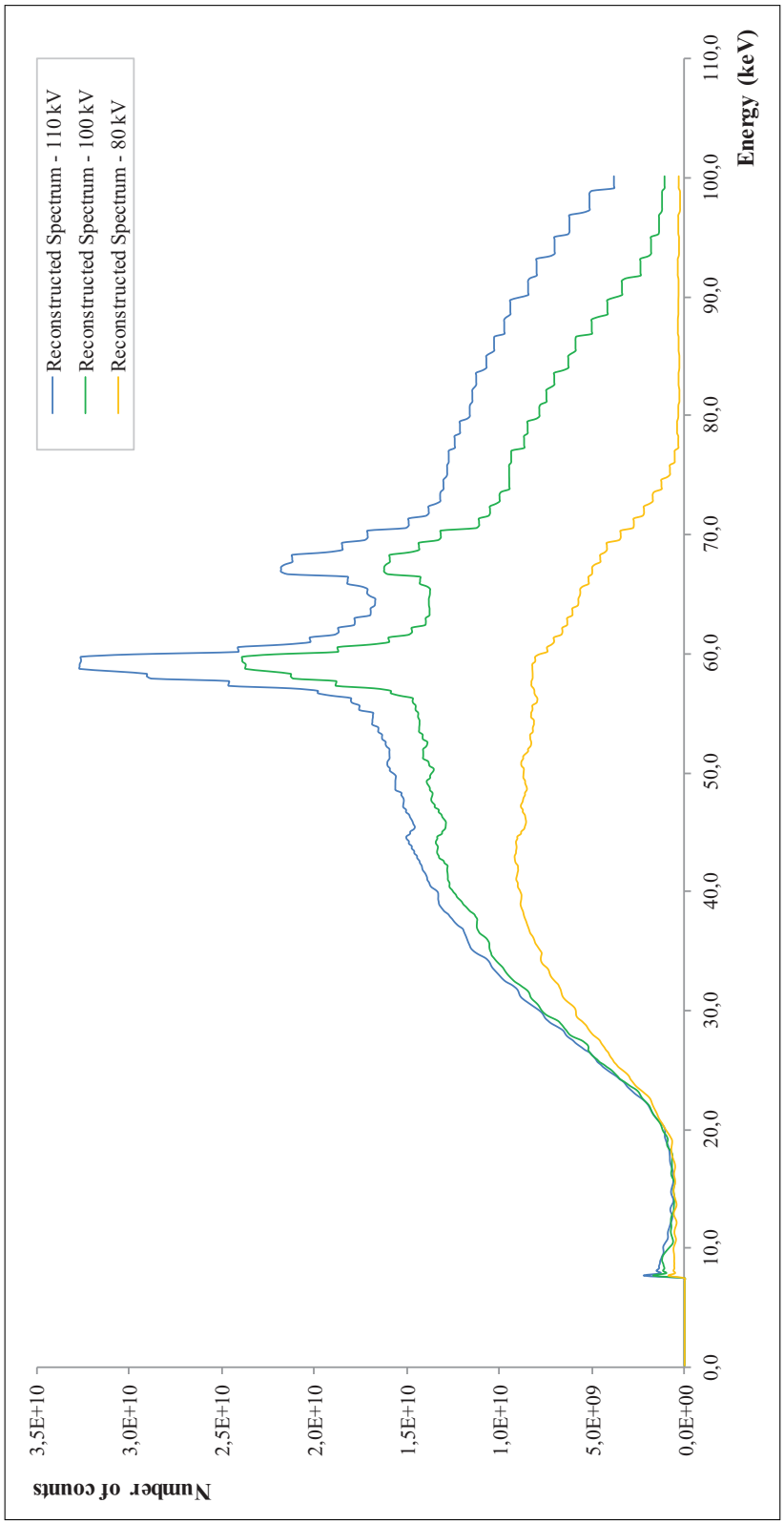

Figure 9.12: Reconstructed source vectors $\overrightarrow{s}_{(80)}$ (orange-colored), $\overrightarrow{s}_{(100)}$ (green-colored) and $\overrightarrow{s}_{(110)}$ (blue-colored) after the full inverse procedure.

As a first observation, the three reconstructed spectra have the general shape expected for a Tungsten X-ray tube, considering their own operating voltages. Let us denote the reconstructed vectors at 80 kV, 100 kV and 110 kV by $\overrightarrow{s}_{(80)}$, $\overrightarrow{s}_{(100)}$ and $\overrightarrow{s}_{(110)}$, respectively. Analyzing the reconstructed vectors, the following observations may be outlined:

- for the reconstructed source vector $\overrightarrow{s}_{(80)}$:
  - no particular structure is observed (except an emerging fluorescence peak at 59.45 keV), as expected considering the low operating voltage used for the measurement.
- for the reconstructed source vector $\overrightarrow{s}_{(100)}$:
  - the characteristic fluorescence lines of the Tungsten may be observed at 59.22 keV for the K_α line, and at 67.49 keV the K_β line;
  - the FWHM of the peaks are equal to 2.65 keV for the K_α line, and to 2.04 keV for the K_β line;
  - the ratios between the peaks and the continuum part of the spectrum are equal to $4.72\,10^{-2}$ for the K_α line, and to $1.67\,10^{-2}$ for the K_β line.
- for the reconstructed source vector $\overrightarrow{s}_{(110)}$:
  - the characteristic fluorescence lines of the Tungsten may be observed at 59.45 keV for the K_α line, and at 67.49 keV the K_β line;
  - the FWHM of the peaks are equal to 3.26 keV for the K_α line, and to 2.04 keV for the K_β line;
  - the ratios between the peaks and the continuum part of the spectrum are equal to $2.84\,10^{-2}$ for the K_α line, and to $2.31\,10^{-2}$ for the K_β line.

From these observations, it appears that the full inverse method gives very good results also with real measurements. In particular, the energy of the peaks on

the $\overrightarrow{s}_{(100)}$ and $\overrightarrow{s}_{(110)}$ source vectors are very similar, indicating a coherent source vector reconstruction.

In order to illustrate the accuracy of the different reconstructions, the source vectors $\overrightarrow{s}_{(80)}$, $\overrightarrow{s}_{(100)}$ and $\overrightarrow{s}_{(110)}$ are shown in Figure 9.13, in Figure 9.14 and in Figure 9.15, together with directly measured spectra (corrected by the detector influences with the GRAVEL code) used as references. The direct measurements were performed with the same germanium detector than that used for the scattering measurements, with the same operating high-voltages, and modified for working with a special low current configuration (of the order of 1 μA, while the operating amperage of the unmodified device is equal to 3 mA). The decrease of the current passing through the tube generates a significant decrease of the emitted photon flux, allowing then the HPGe detector to be placed in front of the X-ray tube and avoiding a pile-up effect.

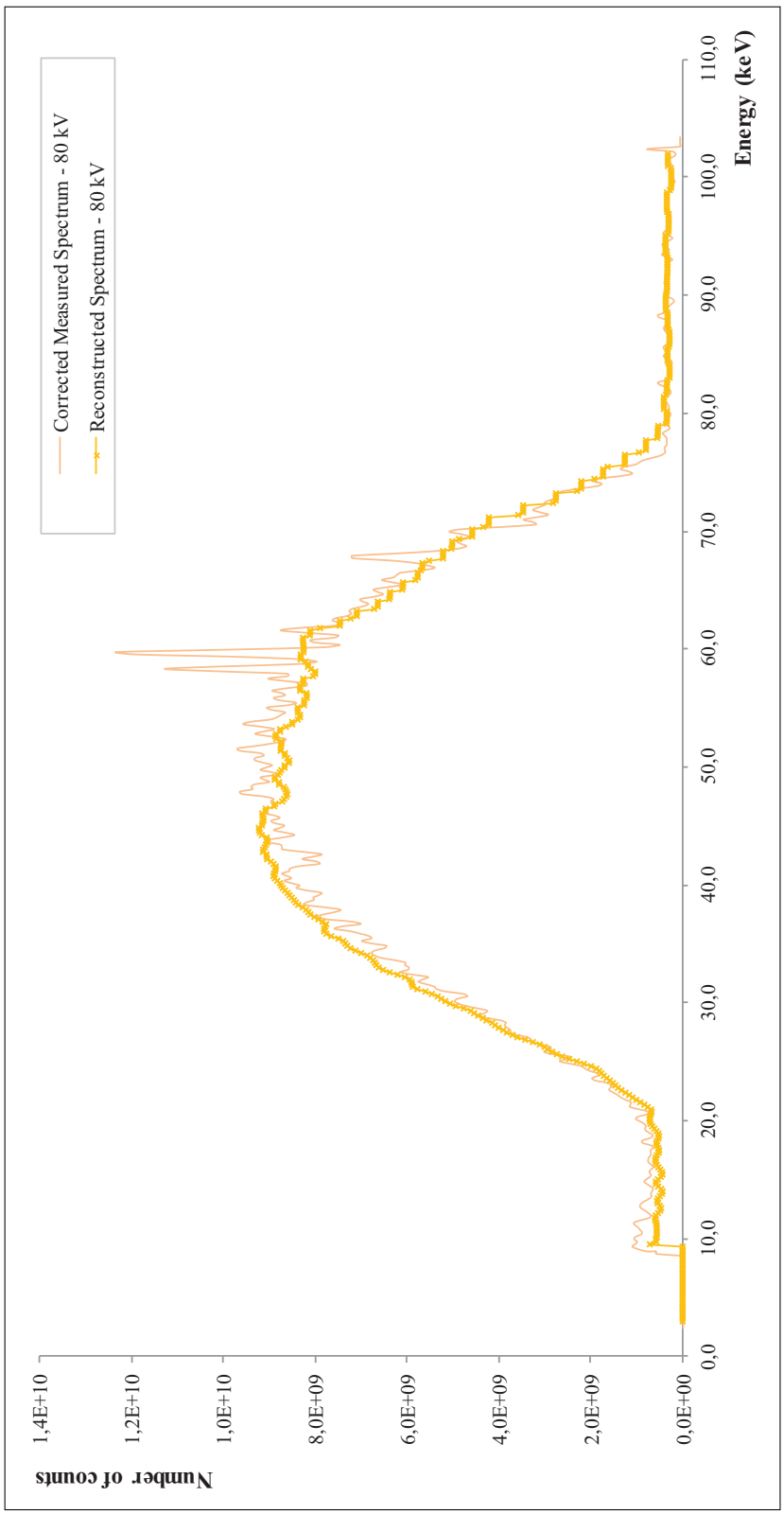

Figure 9.13: Comparison between the source $\overrightarrow{s}_{(80)}$ spectrum obtained from the full inversion procedure (crossed-line) and the direct measurement corrected by the detector influence (continuous line).

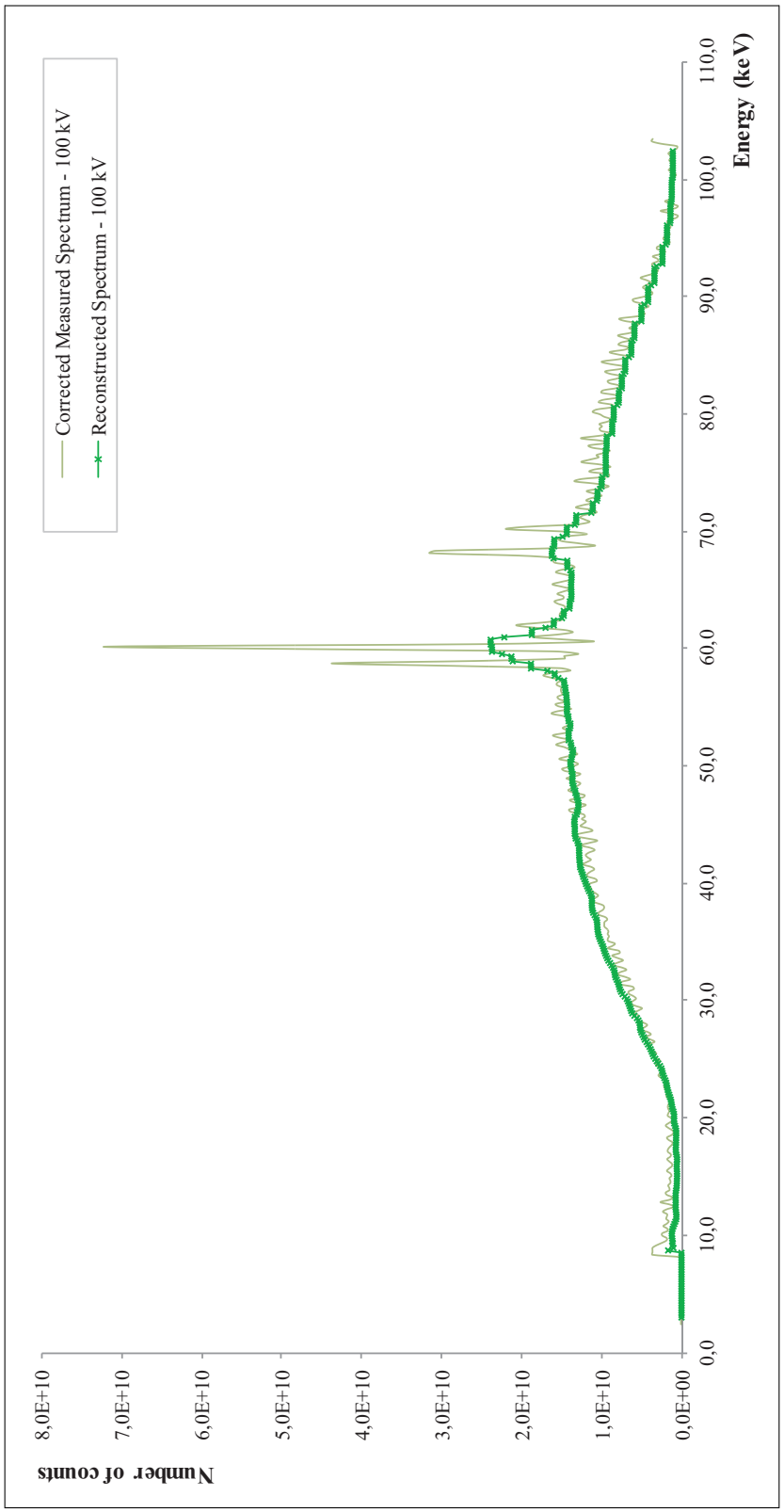

Figure 9.14: Comparison between the source $\overrightarrow{s}_{(100)}$ spectrum obtained from the full inversion procedure (crossed-line) and the direct measurement corrected by the detector influence (continuous line).

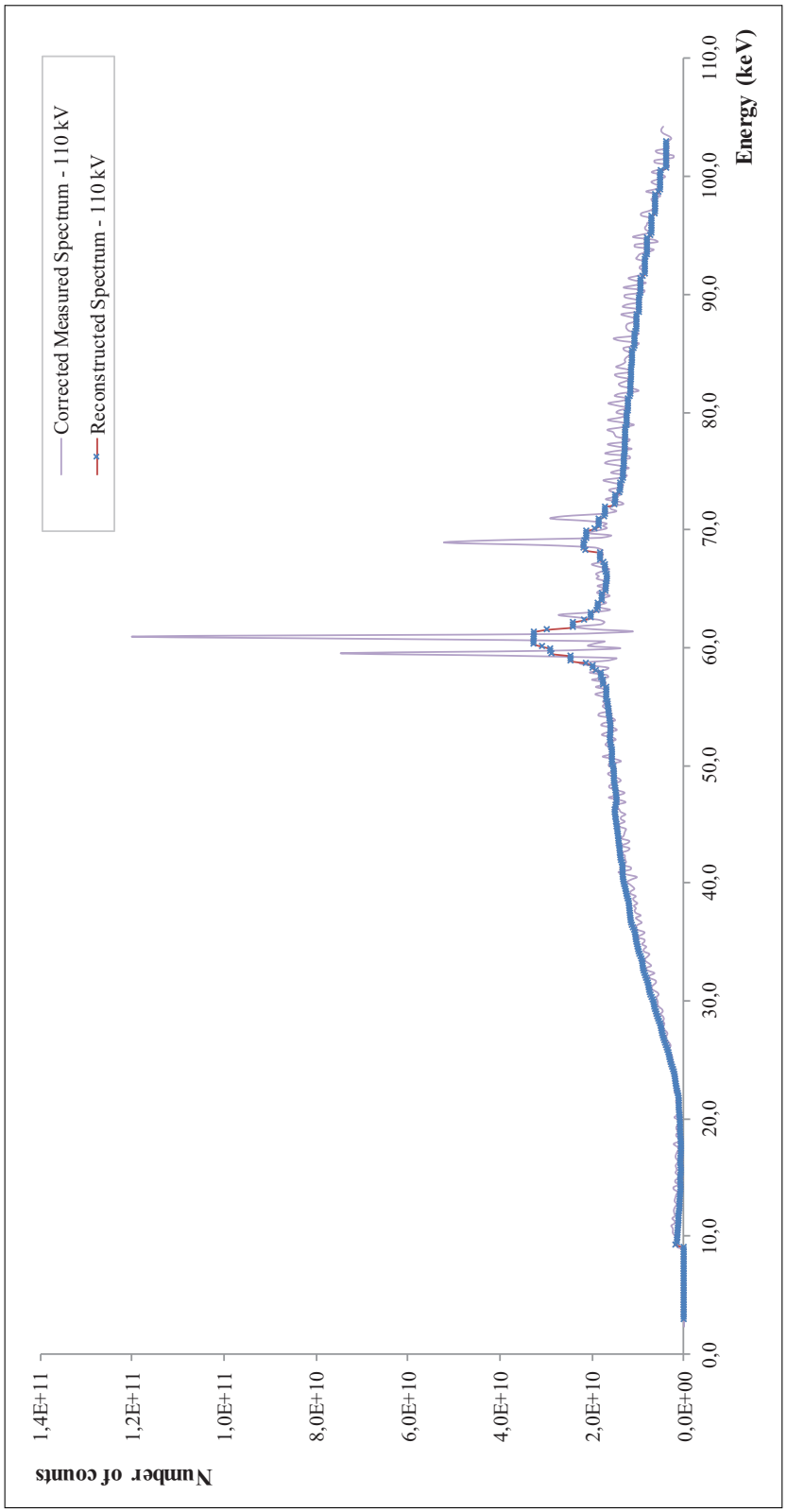

Figure 9.15: Comparison between the source $\overrightarrow{s}_{(110)}$ spectrum obtained from the full inversion procedure (crossed-line) and the direct measurement corrected by the detector influence (continuous line).

9.3 Comments on the reconstructions, and comparison with the direct measurements

For each measurement, the continuum parts of the reconstructed spectra and the direct measurements are in very good agreement with each other. In the 100 kV and the 110 kV measurements, the ratios between the peaks of the K_α / K_β lines and the continuum parts of the different spectra are compared in Table 9.1, showing good correspondance. In the 100 kV and the 110 kV reconstructed spectra, the energies of the two peaks perfectly correspond with each other. This observation tends to indicate an excellent coherence of the reconstruction technique. The energies of the K_α and K_β lines are compared with the theoretical positions of these lines in Table 9.2. An excellent correspondance is observed.

Line, vector	Reconstruction	Direct measurement	Difference (%)
K_α, $\overrightarrow{s}_{(100)}$	$4.72\,10^{-2}$	$4.46\,10^{-2}$	5.83
K_β, $\overrightarrow{s}_{(100)}$	$1.67\,10^{-2}$	$1.82\,10^{-2}$	8.37
K_α, $\overrightarrow{s}_{(110)}$	$3.84\,10^{-2}$	$3.56\,10^{-2}$	7.86
K_β, $\overrightarrow{s}_{(110)}$	$2.31\,10^{-2}$	$2.17\,10^{-2}$	6.45

Table 9.1: Comparison of the ratios between the characteristic lines and the continuum of the spectra, for the reconstructions and the direct measurements.

Line, vector	Theory (keV)	Reconstruction (keV)	Difference (%)
K_α, $\overrightarrow{s}_{(100)}$	59.31	59.22	0.15
K_β, $\overrightarrow{s}_{(100)}$	67.23	67.49	0.39
K_α, $\overrightarrow{s}_{(110)}$	59.31	59.45	0.24
K_β, $\overrightarrow{s}_{(110)}$	67.23	67.49	0.39

Table 9.2: Comparison between the theoretical energies ('Theory') of the K_α and K_β lines of Tungsten and their energies after reconstruction ('Reconstruction').

It may be observed that the peaks found with the full inverse technique are larger than those of the directly measured spectra. This effect may be assigned, at least partially, to the COMPTON profile in the scatterer, which has not been included into the actual calculation of the scattering matrix F_{scatt}. In addition, it should be mentioned that the resolution may also be deteriorated by different geometrical factors like the opening of the collimators or the thickness of the scattering target, for example. These effects have not been investigated with a great level of details in the previous sections, since the aim of this chapter was to develop the method and to test its applicability on real measurements.

In conclusion, both the continuum and the characteristic parts of X-ray spectra can be recovered from scattering measurements with a very good level of details.

Chapter

10 Conclusions and future prospects

The main objective of this thesis was to completely and accurately characterize an X-ray source spectrum from experimental scattering measurements in normal operating conditions, by performing inverse calculations. Ideally, the complete characterization of an X-ray beam should include a precise evaluation of the photon fluence and provide information about the quality of the radiation, specifying its spectral distribution in particular. This information is of crucial importance in medical physics, since it aims at improving both the quality of the medical imaging and the patient protection.

To meet this requirement of source spectrum characterization with the highest level of quality, an innovative inverse procedure based on a detailed modeling of the photon transport in the scattering target has been developed. The technique proposed in this thesis relies on the consideration of the physics inherent to the scattering problem. In that particular sense, the inverse procedure proposed here moves away from most actual resolution techniques, only based on purely mathematical criteria, and constitutes a new and innovative approach of the scattering problem.

Formally, the scattering problem can be mathematically represented by a matrix system of equations whose solution is the source spectrum. In most physical situations, however, the resulting algebraic system is extremely ill-posed. This is the case in particular when light elements, like graphite, are used as target materials. Due to the ill-posed character of the problem, its solutions

are extremely sensitive to small variations in the data, and may be unstable and / or meaningless. In order to circumvent the ill-posed character typically associated with inverse problems, specific strategies are required. In this thesis, the spectral properties of the scattering problem have been significantly improved by using the adjoint scattering term, acting as a preconditioner. This preconditioner, selected for its particular physical sense in the framework of photon transport, has been used to transform the scattering matrix system in a more suitable form to support numerical operations. Starting from the preconditioned matrix system, different numerical methods have been applied to get a meaningful reconstructed X-ray source vector. Among these methods, the successive over-relaxation technique has been identified as the most stable, well adapted to the source vector reconstruction. The effectiveness and the quality of this procedure have been demonstrated in two different stages.

In a first approach to the inverse scattering problem, numerical experiments have been considered to investigate two major aspects of the problem: the numerical suitability of the scattering systems, and the quality assessment of the reconstructed source vectors. For the tests, an X-ray source spectrum representing a Tungsten X-ray tube operating at 50 kV and presenting 12 X-ray lines, has been computed numerically. The photon scattering has been simulated with carbon and aluminium targets. The associated scattering matrices have been computed from analytical transport calculation, taking into account only RAYLEIGH and COMPTON interactions. When solving the two inverse problems, starting from the scattered vector, very different behaviors were observed. Regarding evident questions of target efficiency in producing photon scattering, our attention has been focused mainly on the carbon case. The unpreconditioned matrix system was extremely ill-conditioned, and the solution vectors recovered with different numerical methods from this system were meaningless (the residual vectors 2-norms were included between $5.770\,10^{3}$ % and $1.398\,10^{18}$ % of the numerical source vector 2-norm). The preconditioning was necessary to improve the spectral characteristics of the problem, and to stabilize the solution. The

best reconstruction was obtained with the SOR method ($\omega = 1.50$) on the left preconditioned matrix system (residual vector 2-norm equal to 0.275 % of the numerical source vector 2-norm). The reconstructed vector was in excellent agreement with the numerical source, even in the region of strong discontinuities, but slight oscillations were observed in a small part of the spectrum continuum (between 1.534 Å and 2.274 Å). The disturbed interval has been associated to a non diagonally dominant section of the coefficient matrix. From these observations, it became obvious that two levels of analysis should be considered in order to bring a physically meaningful solution to the inverse problem:

1. the ill-posed character of the scattering matrix system has to be estimated, and reduced if necessary by the preconditioning technique. The adjoint scattering matrix has been shown to be an efficient preconditioner;
2. the matrix structure appears to be a crucial point in the application of the method, and stable methods (such as the successive over-relaxation) are required to get meaningful solutions.

This approach of the problem was a first test to demonstrate the validity and the consistency of the procedure.

In a second stage, the technique has been applied on a set of experimental measurements performed with a spectrometer built at the Operational Unit of Health Physics of the University of Bologna. A 2 mm graphite target, fixed on an extra thin mylar foil, has been placed at a 45 ° angle in the primary beam for the photon scattering. The target thickness has been chosen to reduce as much as maximum the multiple photon scattering contributions into the target. The detector used for the measurements was a 10 mm diameter and 10 mm thick HPGe detector. Three scattering spectra have been measured at different high voltages. Starting from these raw measurements, the following procedure has been applied:

1. the smoothing of the experimental measurements with a Savitzky-Golay filter;

2. the cleaning of the detector effects (physical and statistical influences) from the measured spectra with the GRAVEL unfolding code;
3. the conversion of the spectrum into the wavelength regime, in order to take advantage from the bidiagonal structure of the scattering matrix;
4. the generation of the forward scattering matrix from detailed transport calculation (and the estimate of the matrix ill-conditioning);
5. the computation of the left preconditioned coefficient matrix and the importance vector (and the estimate of the matrix ill-conditioning);
6. the iterative numerical solution of the most adapted algebraic linear system of equations by means of a successive overrelaxation method (SOR), an accelerated convergence technique.

The reconstructed spectra have been successfully compared to straightforward measurements performed at the same high-voltages in particular conditions. The continuum parts of the reconstructed spectra and the direct measurements were in very good agreements with each other. The position of the peaks were in perfect correspondance. These observations indicated an excellent coherence of the reconstruction technique, and permitted the validation of the procedure.

The inverse scattering procedure presented in this thesis then appears very promising as a tool for characterizing the intensity distribution of an X-ray tube spectrum with a high level of accuracy, even at low energies where other methods usually fail. In particular, this procedure should allow to estimate the parameters essential to the quality control of medical radiological systems. Further investigation is still necessary in order to take advantage, for example, of the characteristics of different targets.

In the future, the technique could be improved by introducing further refinements in the computation of the scattering matrix, implementing for example the COMPTON profile in the scatterer, partly responsible of the peaks width, or multiple scattering effects. For second and higher orders of scattering, it will be necessary to also include the effects of the photon polarization.

Bibliography

[1] Archer B. and Wagner L. A laplace transform pair model for spectral reconstruction. *Medical Physics*, 9:844 – 847, 1982.

[2] Tominaga S. The estimation of the X-ray spectral distributions from attenuation data by means of iterative computation. *Nuclear Instruments and Methods*, 192:415 – 421, 1982.

[3] Rubio M. and Mainardi R. Determination of X-ray spectra including characteristic line intensities from attenuation data. *Physics in Medicine and Biology*, 11:1372 – 1376, 1984.

[4] Pani R., Laitano R., and Pellegrini R. Diagnostic of X-ray spectra measurements using a silicon surface barrier detector. *Physics in Medicine and Biology*, 32:1135 – 1149, 1987.

[5] Di Castro E., Pani R., Pellegrini R., and Bacci C. The use of cadmium telluride detectors for the qualitative analysis of diagnostic X-ray spectra. *Physics in Medicine and Biology*, 29:1117 – 1131, 1984.

[6] Yaffe M., Taylor K., and Johns H. Spectroscopy of diagnostic x-rays by a compton scatter method. *Medical Physics*, 3:328 – 334, 1976.

[7] Millan M. and Llopis J. *Técnicas instrumentales de rayos X. Fluorescencia, diffraccion y microanalisis.* Universidad Politécnica de Valencia, 2000.

[8] Jenkins R. *Enyclopedia of Analytical Chemistry.* John Wiley and Sons Ltd, Chichester, 2000.

[9] Matscheko G. *A Compton spectrometer for measurements of primary photon energy spectra from clinical X-ray units under working conditions*. PhD thesis, Department of Radiation Physics, Linköping University, Linköping, Sweden, 1988.

[10] Matscheko G. and Ribberfors R. A compton scattering spectrometer for determining x-ray photon energy spectra. *Physics in Medicine and Biology*, 32:577 – 594, 1987.

[11] Debertin K. and Helmer R. *Gamma- and x-ray spectrometry with semiconductor detectors*. North-Holland (Amsterdam and New York and New York, NY, USA), 1988.

[12] Carlsson A., Carlsson C., Berggen K.-F., and Ribberfors R. Calculation of scattering cross-sections for increased accuracy in diagnostic radiology. *Medical Physics*, 9:868 – 879, 1982.

[13] Gallardo S., Ginestar D., Verdu G., Rodenas J., Puchades V., and Villaescusa J. X-ray spectrum unfolding using a regularized truncated SVD method. *X-ray Spectrometry*, 35:63 – 70, 2006.

[14] Tikhonov A., Goncharsky A., Stepanov V., and Yagola A. *Numerical methods for the solution of ill-posed problems*. Kluwer Academic Publishers, Dodrecht / Boston / London, 1995.

[15] Tikhonov A. and Arsenin V. *Solution of ill-posed problems*. Vh. Winston and Sons, 1977.

[16] Golub G. and Van Loan C. *Matrix computations*. The John Hopkins University Press, London - Baltimore, 1996.

[17] Zhengming L. A numerical method for solving the Fredholm integral equation of the first kind and its application to restore the folded radiation spectrum. *Nuclear Instruments and Methods*, 255:152 – 155, 1987.

[18] Gallardo S. *Determinacion del espectro primario de rayos X para radiodiagnostico mediante espectrometria Compton, aplicando tecnicas de deconvolucion y simulacion por Monte Carlo.* PhD thesis, Departamento de Ingenieria Quimica y Nuclear, Universidad Politecnica de Valencia, 2004.

[19] Querol A., Gallardo S., Rodenas J., and Verdu G. Application of Tikhonov and MTSVD methods to unfold experimental X-ray spectra in the radiodiagnostic energy range. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 536 – 539, 2010.

[20] Gallardo S., Verdu G., Rodenas J., and Villaescusa J. Application of unfolding techniques to obtain an x-ray primary spectrum. *X-ray Spectrometry*, 35:63 – 70, 2006.

[21] Gallardo S., Rodenas J., Querol A., and Verdu G. Application of the MTSVD unfolding method for reconstruction of primary X-ray spectra using semiconductor detectors. *Nuclear Instruments and Methods*, 53:1136 – 1139, 2011.

[22] Dyson N. *X-rays in atomic and nuclear physics.* Cambridge University Press, second edition, 1990.

[23] Fernandez J. *Interaction of X-rays with matter, in Microscopic X-ray Fluorescence Analysis.* Janssens K., Adams F., and Rindby A. (Eds), John Wiley and Sons, Ltd, 2000.

[24] Evans R. *The atomic nucleus.* McGraw-Hill Book Company, New York, 1955.

[25] Jauch J. and Rohrlich F. *The theory of photons and electrons.* Springer-Verlag, Berlin, 1976.

[26] Agarwal B. *X-ray spectrometry.* Springer-Verlag, Berlin, second edition edition, 1991.

[27] Alexandropoulos N., Chatzigeogiou T., Evangelakis G., Cooper M., and Manninen S. Bremsstrahlung and its contribution to the gamma-ray spectra of solids. *Nuclear Instruments and Methods in Physics Research A*, 271:543 – 545, 1988.

[28] Bui C. and Milazzo M. Measurements of anomalous dispersion in rayleigh scattering of characteristic X-ray fluorescence. *Nuovo Cimento*, D-11:655 – 686, 1989.

[29] Saloman E., Hubbell J., and Scofield J. X-ray attenuation cross-sections for energies 100 eV to 1 GeVfor elements z=1 to z=92. *Atomic Data Nuclear Data Tables*, 38:1 – 197, 1988.

[30] Hubbell J. Review of photon interaction cross section data in the medical and biological context. *Physics in Medicine and Biology*, 44:1 – 22, 199.

[31] Berger M., Hubbell J., Seltzer S., Chang J., Coursey J., Sukumar R., Zucker D., and Olsen K. XCOM: photon cross section database (version 1.5), November 2011. `http://www.nist.gov/pml/data/xcom/index.cfm`.

[32] Shultis K. and Faw R. *Fundamentals of nuclear science and engineering.* Marcel Dekker, Inc - New York - Basel, 2002.

[33] Shultis K. and Faw R. *Radiation shielding.* American Nuclear Society, La Grange Park, IL, USA, 2002.

[34] Chilton A., Shultis K., and Faw R. *Principles of radiation shielding.* Prentice Hall, Englewood Cliffs, New York, 2002.

[35] Fernandez J. E. XRF intensity in the frame of the transport theory. *X-ray Spectrometry*, 18:271 – 279, 1989.

[36] Cullen D., Chen M., Hubbell J., Perkins S., Plechaty E., Rathkopf J., and Scofield J. Tables and graphs of photon interaction cross-sections from 10

ev to 100 gev derived from the llnl evaluated data library. Technical report, Lawrence Livermore National Laboratory Report UCRL-5400, 1989.

[37] Bearden J. and Burr F. Reevaluation of X-ray atomic energy levels. *Reviews of Modern Physics*, 39:125 – 142, 1967.

[38] McMaster W., Kerr del Grande N., Mallett J., and Hubbell J. Compilation of X-ray cross-sections. Technical report, Lawrence Livermore National Laboratory Report UCRL-50174, 1969.

[39] Kane P., Kissel L., Pratt R., and Roy S. Elastic scattering of gamma- and X-rays by atoms. *Physics Reports*, 140:75 – 159, 1986.

[40] Nelms A. and Oppenheim L. *Journal of Research - National Bureau of Standards*, 55:53 – 62, 1955.

[41] Hubbell J., Veigele W., Briggs E., Brown R., Cromer D., and Howerton R. Atomic form factors, incoherent scattering functions, and photon scattering cross-sections. *Journal of Physical Chemistry*, 4:471 – 538, 1975.

[42] Hubbell J. and OverbøI. Relativistic atomic form factors and photon coherent scattering cross-sections. *Journal of Physical and Chemical Reference Data*, 9:69 – 105, 1979.

[43] Schaupp D., Schumacher M., Smend F., Rullhusen P., and Hubbell J. Small-angle rayleigh scattering of photons at high energies: tabulation of relativistic hfs modified atomic form factors. *Journal of Physical and Chemical Reference Data*, 12:467 – 508, 1983.

[44] Kane P., Kissel L., Pratt R., and Roy S. Elastic scattering of g-rays and X-rays by atoms. *Physical Reports*, 140:75 –159, 1986.

[45] Chantler C., Olsen K., Dragoset R., Chang J., Kishore A., Kotochigova S., and Zucker D. Detailed tabulation of atomic form factors, photoelectric absorption and scattering cross section, and mass attenuation coefficients

for z =1 to z = 92 from e = 1.0-10.0 ev to e = 0.4-1.0 mev. *Journal of Physical Chemistry*, 24:597 – 1048, 1995.

[46] Compton A. A quantum theory of the scattering of X-rays by light elements. *Physical Review*, 21:483 – 502, 1923.

[47] Compton A. The spectrum of scattered X-rays. *Physical Review*, 22:409 – 413, 1923.

[48] Klein O. and Nishina Y. Über die streuung von strahlung durch freie elektronen nach der neuen relativistischen quantendynamik von dirac. *Zeitschrift für Physik A Hadrons and Nuclei*, 52:853 – 868, 1929.

[49] Waller I. and Hartree D. On the intensity of total scattering of X-rays. *Proceedings of the Royal Society of London*, A124:119 – 133, 1929.

[50] Thomas L. The calculation of atomic fields. *Proceedings of Cambridge Philosophy Society*, 23:542 – 551, 1927.

[51] Fermi E. Eine statistische methode zur bestimmung einiger eigenschaften des atoms und ihre anwendung auf die theorie des periodischen systems der elemente. *Zeitschrift für Physik A Hadrons and Nuclei*, 48:73 – 79, 1928.

[52] Veigele W., Tracy P., and Henrry E. Compton effect and electron binding. *American Journal of Physics*, 34:1116 – 1121, 1966.

[53] Smith V., Thakkar A., and Chapman D. A new analytical approximation to atomic incoherent X-ray scattering intensities. *Acta Crystallographica*, 31:391 – 392, 1975.

[54] Cooper M. Compton scattering and electron momentum determination. *Reports on Progress in Physics*, 4:415 – 419, 1985.

[55] Biggs F., Mendelsohn L., and Mann J. Hartree-fock compton profiles for the elements. *Atomic Data and Nuclear Data Tables*, 16:201 – 309, 1975.

[56] Ribberfors R. and Berggen K. Incoherent X-ray scattering functions and cross-sections by means of a pocket calculator. *Physical Review*, 29:3325 – 3333, 1982.

[57] Fernandez J. and Sumini M. Adjoint calculations for multiple scattering of compton and rayleigh effects. *Nuclear Instruments and Methods*, 71:111 – 115, 1992.

[58] Bell G. and S. Glasstone. *Nuclear Reactor Theory.* Krieger Publishing Company, 1979.

[59] Lewins J. *Importance: the adjoint function.* Pergamon Press, Oxford, 1965.

[60] Craig I. and Brown J. *Inverse problems in astronomy.* Adam Hilger, Bristol, 1986.

[61] Cuppen J. *A numerical solution of the inverse problem in electrocardiography.* PhD thesis, Department of Mathematics, University of Amsterdam, 1983.

[62] Santosa F., Pao Y.-H., Symes W., and Holland C. *Inverse problems of acoustic and elastic waves.* SIAM, Philadelphia, 1984.

[63] O'Sullivan F. A statistical perspective on ill-posed inverse problems. *Journal of Statistical Science*, 1:502 – 527, 1996.

[64] Groetsch C. *The theory of Tikhonov regularization for Fredholm equation of the first kind.* Pitman Boston, 1984.

[65] Hansen P.-C. Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6:1 – 35, 1994.

[66] Bai Z. and Demmel J. Computing the generalized singular value decomposition. *SIAM Journal on Scientific Computing*, 14:1464 – 1486, 1993.

[67] Paige C. Computing the generalized singular value decomposition. *SIAM Journal on Scientific and Statistical Computing*, 7:1126 – 1146, 1986.

[68] Anderson E. Bai Z., Bischof C., Blackford S., Demmel J., Dongarra J., Du Croz J., Greenbaum A. an Hammarling S., McKenney A., and Sorensen D. *LAPACK user guide, Third edition*, siam, society for industrial and applied mathematics edition, 1999.

[69] Sulij E. and Mayers D. *An introduction to numerical analysis*. Cambridge, 2003.

[70] Cheney E. and Kincaid D. *Numerical mathematics and computing - Sixth Edition*. Thomson Brooks / Cole, 2008.

[71] Love D. and Nelson A. Unfolding the response function of high-quality germanium detectors. *Nuclear Instruments and Methods in Physics Research A*, 274:541 – 546, 1988.

[72] Radford D., Ahmad I., Holzmann R., Janssens R., and Khoo T. A prescription for the removal of compton-scattered gamma rays from gamma-ray spectra. *Nuclear Instruments and Methods in Physics Research A*, 258:111 – 118, 1987.

[73] Knoll G. *Radiation detection and measurement*. John Wiley and Sons, Inc., third edition edition, 2000.

[74] Tikhonov A. Solution for incorrectly formulated problems and the regularization method. *Doklady Akademii Nauk*, 4:1035 – 1038, 1963.

[75] Tikhonov A. and Arsenin V. *Solution of ill-posed problems*. Winston and Sons, Washington, D. C., 1977.

[76] Hansen P. Perturbation bounds for discrete Tikhonov regularization. *Inverse Problems*, 5:41 – 44, 1989.

[77] Hansen P. The truncated SVD as a method of regularization. *BIT - Computational Mathematics*, 27:543 – 553, 1987.

[78] Hansen P. Truncated SVD solutions to discrete ill-posed problems with ill-determined numerical rank. *Journal on Scientific and Statistical Computing*, 11:503 – 518, 1990.

[79] Varah J. On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems. *Journal on Numerical Analysis*, 10:257 – 267, 1973.

[80] Miller K. Least squares methods for ill-posed problems with a prescribed bound. *Journal on Mathematical Analysis*, 1:52 – 74, 1970.

[81] Lawson C. and Hanson R. *Solving least squares problems.* Prenctice-Hall, Englewoods Cliffs, 1974.

[82] Hansen P. and O'Leary D. The use of the l-curve in the regularization of discrete ill-posed problems. *Journal on Scientific and Statistical Computing*, 14:1487 – 1503, 1993.

[83] Hansen P. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review*, 34:561 – 580, 1992.

[84] Matzke M. Unfolding methods. *Radiation Protection Dosimetry*, 107:155 – 174, 2003.

[85] McElroy W. N., Berg S., Crockett T., and Hawkins R. G. A computer-automated iterative method for neutron flux spectra determination by foil activation. Technical report, AFWL-TR-67-41, U.S. Air Force Weapons Laboratory, 1967.

[86] Matzke M. Unfolding of pulse height spectra: the HEPRO program system. Technical report, Physikalish-Technische Bundesanstalt, 1994.

[87] Matzke M. Unfolding of particle spectra. In George Vourvopoulos, editor, *International conference: neutrons in research and industry*, number SPIE 2867, pages 598 – 607, 1997.

[88] Reginatto M. *The multi-channel unfolding programs in the UMG package: MXDMC33, GRVMC33 and IQUMC33.* Physikalisch-Technische Bundesanstalt (PTB) - Braunschweig, Germany, 2004.

[89] Matzke M. Unfolding of pulse height spectra: the HEPRO program system. Technical Report PTB-N-19, Physikalisch-Technische Bundesanstalt, 1994.

[90] Matzke M. Unfolding of pulse height spectra: the HEPROW program system. Technical report, Physikalisch-Technische Bundesanstalt, 2004.

[91] Benmosbah M. *Spectrométrie des neutrons lents: étude de la réponse d'un ensemble de compteurs proportionnels.* PhD thesis, Université de Franche-Comté, Ecole Doctorale Louis Pasteur, 2007.

[92] Shannon C. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379 – 423, 1948.

[93] Jaynes E. Information theory and statistical mechanics. *Physical Review*, 106:620 – 630, 1957.

[94] Jaynes E. Prior probabilities. *IEEE Transactions on Systems, Man and Cybernetics*, 4:227 – 241, 1968.

[95] Hobson A. A new theorem of information theory. *Journal on Statistical Physics*, 1:383 – 391, 1969.

[96] Skilling J. Classic maximum entropy. In Dodrecht / Boston / London Kluwer Academic Publishers, editor, *Maximum Entropy and Bayesian Methods*, 1989.

[97] Reginatto M. Maxed, a computer code for the deconvolution of multi-sphere neutron spectrometer data using the maximum entropy method. *Nuclear Instruments and Methods in Physics Reasearch*, 476:242 – 246, 2002.

[98] Reginatto M. and Goldhagen P. Maxed, a computer code for maximum entropy deconvolution of multisphere neutron spectrometer data. *NIM*, 77:579 – 583, 1999.

[99] Wilczek R. and Drapatz S. A high accuracy algorithm for maximum entropy image restoration in the case of small data sets. *Astronomy and Astrophysics*, 142:9 – 12, 1985.

[100] R. H. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16:1190 – 1208, 1995.

[101] C. Zhu, R. H. Byrd, and J. Nocedal. L-bfgs-b: Algorithm 778: L-bfgs-b, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23:550 – 560, 1997.

[102] Reginatto M. and Goldhagen P. Maxed, a computer code for the deconvolution of multisphere neutron spectrometer data using the maximum entropy method. Technical report, Environmental Measurements Laboratory Report, EML-595, 1998.

[103] Fernandez J. and Sumini M. SHAPE: a computer simulation of energy-dispersive X-ray spectra. *X-Ray Spectrometry*, 20:315 – 319, 1991.

[104] Quarteroni A., Sacco R., and Saleri F. *Méthodes numériques*. Springer-Verlag, 2007.

[105] Axelsson O. *Iterative solutions methods*. Cambridge University Press, Cambridge, 1994.

[106] Young D. *Iterative solutions for large linear systems*. Academic Press, New York, 1971.

[107] Kahan W. *Gauss-Seidel methods of solving large systems of linear equations*. PhD thesis, University of Toronto, 1958.

[108] Hackbush W. *Iterative solutions of large sparse systems of equations*. Springer Verlag, New York, 1994.

[109] Pella P., Feng L., and Small J. An analytical algorithm for calculation of spectral distributions of X-ray tubes for quantitative X-ray fluorescence analysis. *X-ray Spectrometry*, 14:125–135, 1985.

[110] Pella P., Feng L., and Small J. Addition of m- and l-series lines to NIST algorithm for calculation of X-ray tube output spectral distributions. *X-ray Spectrometry*, 20:109–110, 1991.

[111] Ebel H. X-ray tube spectra. *X-ray spectrometry*, 28:255 – 266, 1999.

[112] Bearden J. X-ray wavelengths. *Reviews of Modern Physics*, 39:78 – 124, 1967.

[113] Deslattes R., Kessler E., Indelicato J., de Billy L., Lindroth E., and Anton J. X-ray transition energies: new approach to a comprehensive evaluation. *Reviews of Modern Physics*, 75:35 – 99, 2003.

[114] Rossi P.-L. Analisi spettrometica di un fascio di Röntgendiagnostica ricostruito dallo spettro X Compton. Tesi di Laurea in Fisica, Università degli Studi di Bologna, Facoltà di Scienze Matematiche, Fisiche e Naturali, 1997.

[115] Fernandez J. and Scott V. Simulation of the detector response function with the code mcshape. *Radiation in Physics and Chemistry*, 78:882 – 887, 2009.

[116] Salvat F., Fernandez-Varea J., Acosta E., and Sempau J. PENELOPE, a code system for Monte Carlo simulation of electron and photon transport. In *Proceedings of a Workshop / Training Course*, 2001.

[117] Sempau J., Acosta E., Baro J., Fernandez-Varea J., and Salvat F. An algorithm for Monte Carlo simulation of the coupled electron-photon transport. *Nuclear Instruments and Methods*, 132:377 – 390, 1997.

[118] Savitzky A. and Golay J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36:1627 – 1639, 1964.

[119] Steinier J., Termonia Y., and Deltour J. Comments on smoothing and differentiation of data by simplified least square procedure. *Analytical chemistry*, 44:1906 – 1909, 1972.

[120] Farebrother W. *Linear least-squares computations, First Edition.* CRC Press, 1988.

Appendices

Appendix

A Appendix: Direct numerical methods for the resolution of linear systems

Direct numerical methods form a first class of efficient methods for solving linear systems of equations. A particularity of direct methods is that they compute the solution to a linear system of equations in a finite number of steps, entirely determined by the size of the coefficient matrix. These methods would give the precise answer if they were performed in infinite precision arithmetic. In practice, however, this situation is very theoretical, because finite precision is used for the computation. Consequently, assuming numerical stability of the algorithm, the resulting vector is an approximation of the true solution.

The substitution technique

Let us define a lower triangular system of n linear equations with unknown $\overrightarrow{s}$:

$$F\overrightarrow{s} = \overrightarrow{b}, \quad F \in \Re^{n \times n} \tag{A.1}$$

where:

$$F = \begin{bmatrix} f_{11} & 0 & \cdots & 0 \\ f_{21} & f_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}, \quad \overrightarrow{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{A.2}$$

The matrix F is assumed to be nonsingular. Since F is nonsingular, all the diagonal terms are different from zero, it is possible to successively determine the values of the unknowns s_i for $i = 1, 2, \ldots, n$, by:

$$s_1 = \frac{b_1}{f_{11}} \tag{A.3}$$

$$s_i = \frac{1}{f_{ii}} \left(b_i - \sum_{j=1}^{i-1} f_{ij} s_j \right), \quad i = 2, 3, \ldots, n \tag{A.4}$$

This algorithm is usually called forward substitution, and relations A.3 and A.4 as descent formula's.

The system of equations defined by an upper triangular matrix F may be solved by a similar treatment. In this case, the algorithm is referred to as backward substitution, and is described in a general case by:

$$s_n = \frac{b_n}{f_{nn}} \tag{A.5}$$

$$s_i = \frac{1}{f_{ii}} \left(b_i - \sum_{j=i+1}^{n} f_{ij} s_j \right), \quad i = n-1, n-2, \ldots, 1 \tag{A.6}$$

Forward and backward substitution algorithms need $n(n+1)/2$ multiplications and divisions, and $n(n-1)/2$ additions and subtractions. The global arithmetic complexity of the algorithm is then n^2 floating point operations.

The GAUSS elimination technique

The GAUSS elimination algorithm aims to reduce the matrix system of equations $F\overrightarrow{s} = \overrightarrow{b}$ in an equivalent form $U\overrightarrow{s} = \overrightarrow{b^*}$, where U is an upper triangular matrix and $\overrightarrow{b^*}$ is a properly modified independent term. During the system transformation, elementary row operations (linear combinations) are used to reduce the coefficient matrix in an equivalent triangular form. The final system may be solved by using a substitution algorithm.

Let us define the nonsingular matrix $F \in \Re^{n \times n}$, whose first diagonal term a_{11} is assumed to be different from zero. Let us denote $F^{(1)} = F$ and $b^{(1)} = b$.

The values:

$$m_{i1} = \frac{f_{i1}^{(1)}}{f_{11}^{(1)}}, \quad i = 2, 3, \ldots, n \tag{A.7}$$

where $f_{ij}^{(1)}$ are the elements of $F^{(1)}$, are called multipliers. The unknown s_1 may be eliminated from rows $i = 2, 3, \ldots, n$ by subtracting from it m_{i1} times the first row, and by performing the same operation on the independent term. The following quantities may then be defined:

$$f_{ij}^{(2)} = f_{ij}^{(1)} - m_{i1} f_{1j}^{(1)}, \quad i, j = 2, 3, \ldots, n \tag{A.8}$$

$$b_i^{(2)} = b_i^{(1)} - m_{i1} b_1^{(1)}, \quad i = 2, 3, \ldots, n \tag{A.9}$$

At this stage of the elimination procedure, a second system of linear equations of the following form is obtained:

$$\begin{bmatrix} f_{11}^{(1)} & f_{12}^{(1)} & \cdots & f_{1n}^{(1)} \\ 0 & f_{22}^{(2)} & \cdots & f_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & f_{n2}^{(2)} & \cdots & f_{nn}^{(2)} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix} \tag{A.10}$$

or, in matrix form:

$$F^{(2)} \overrightarrow{s} = \overrightarrow{b}^{(2)} \tag{A.11}$$

Of course, the second system of equations is equivalent to the starting one. This second system may again be transformed in a way to eliminate the unknown s_2 from the rows 3 to n. By performing such successive transformations, it is possible to obtain a finite sequence of systems:

$$F^{(k)} \overrightarrow{x} = \overrightarrow{b}^{(k)}, \quad 1 \leq k \leq n \tag{A.12}$$

where the matrix $F^{(k)}$ has the following form:

$$F^{(k)} = \begin{bmatrix} f_{11}^{(1)} & f_{12}^{(1)} & \cdots & \cdots & \cdots & f_{1n}^{(1)} \\ 0 & f_{22}^{(2)} & & & & f_{2n}^{(2)} \\ \vdots & & \ddots & & & \vdots \\ 0 & \ldots & 0 & f_{kk}^{(k)} & \cdots & f_{kn}^{(k)} \\ \vdots & & \vdots & & & \vdots \\ 0 & \ldots & 0 & f_{nk}^{(k)} & \cdots & f_{nn}^{(k)} \end{bmatrix} \tag{A.13}$$

where it has been assumed that $f_{ii}^{(k)} \neq 0$ for $i = 1, \ldots, k-1$. Between the k-th and $(k+1)$-th system, the multipliers are given by:

$$m_{ik} = \frac{f_{ik}^{(k)}}{f_{kk}^{(k)}}, \quad i = k+1, \ldots, n \tag{A.14}$$

and the successive coefficient elements by:

$$f_{ij}^{(k+1)} = f_{ij}^{(k)} - m_{ik} f_{kj}^{(k)}, \quad i, j = k+1, \ldots, n \tag{A.15}$$

$$b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)}, \quad i = k+1, \ldots, n \tag{A.16}$$

For $k = n$, an upper triangular system of linear equations $F^{(n)}\overrightarrow{s} = \overrightarrow{b}^{(n)}$ is obtained, which has the form:

$$\begin{bmatrix} f_{11}^{(1)} & f_{12}^{(1)} & \cdots & f_{1n}^{(1)} \\ 0 & f_{22}^{(2)} & \cdots & f_{2n}^{(2)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & f_{nn}^{(n)} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{bmatrix} \tag{A.17}$$

It is usual to denote the upper triangular matrix $F^{(n)}$ by U. The terms $f_{kk}^{(k)}$ are commonly called the pivots, and must obviously be different from zero for $k = 1, 2, \ldots, n-1$. The system A.17 may easily (but not always accurately) be solved using backward substitution.

The GAUSS elimination technique applied on a $n \times n$ system of equations requires $6n(n-1)(n+1) + n(n-1)$ operations, plus n^2 operations for the backward substitution of the resulting triangular system. Approximately $(2n^3/3 + 2n^2)$ floating point operations are then necessary to solve the system. The arithmetic complexity of the GAUSS algorithm is consequently of $2n^3/3$ floating point operations. This algorithm is commonly used computed for systems with thousands of equations and unknowns, providing very often excellent results [120]. It can be demonstrated that the algorithm is numerically extremely stable for diagonally dominant and positive definite matrices [16, 104].

The GAUSS elimination method fails if at least one pivot element is equal to zero. In case of a zero pivot element, interchanging rows is necessary. This

strategy is generalized by searching in $F^{(k)}$, at each step k of the elimination, a nonzero pivot among the terms of the k-th column, for k going from line k to line n. Furthermore, in order to insure as best as possible results with the GAUSS elimination, it is generally desirable to choose a pivot element with the largest absolute value. This improves the numerical stability of the algorithm, by reducing the successive effect of round-off error propagation. Consequently, in order to insure a maximal stability of the numerical calculations, the pivot element at the k-th operation is chosen to be the greatest element in absolute value of the k-th column, for k going from line k to line n. Even if it is not strictly necessary, the permutation of the pivoting element is generally done at each step of the procedure. Of course, pivoting adds more operations to the computational cost of the algorithm. The additional cost is of the order of n^2 floating point operations. This method is usually referred to as GAUSS elimination with partial pivoting.

The LU factorization

The LU factorization is a procedure for decomposing a nonsingular square matrix F into the product of a lower triangular matrix L and an upper triangular matrix U.

Let us define a nonsingular matrix $F \in \Re^{n \times n}$. The LU decomposition of the matrix F has the following form:

$$F = LU \tag{A.18}$$

where:

$$L = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{bmatrix}, \quad U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \tag{A.19}$$

The LU decomposition is basically a modified form of the GAUSS elimination technique, with $U = F^{(n)}$. By defining the vector:

$$\overrightarrow{m}_k = [0, \ldots, 0, m_{k+1,k}, \ldots, m_{n,k}]^T \in \Re^n \tag{A.20}$$

and the matrix M_k by:

$$M_k = \begin{bmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ldots & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & -m_{k+1,k} & 1 & \ldots & 0 \\ \vdots & \ldots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & -m_{n,k} & 0 & \ldots & 1 \end{bmatrix} = I_n - \overrightarrow{m}_k \, \overrightarrow{e}_k^T \tag{A.21}$$

as the k-th GAUSS transform matrix, we have:

$$(M_k)_{ip} = \delta_{ip} - (\overrightarrow{m}_k \, \overrightarrow{e}_k^T)_{ip} = \delta_{ip} - m_{ip} \, \delta_{kp}, \quad i, p = 1, 2, \ldots, n \tag{A.22}$$

From equation A.15:

$$\begin{aligned} f_{ij}^{(k+1)} &= f_{ij}^{(k)} - m_{ik} \, \delta_{kk} \, f_{kj}^{(k)} \\ &= \sum_{p=1}^{n} (\delta_{ip} - m_{ik} \, \delta_{kp}) \, f_{pj}^{(k)}, \quad i, j = k+1, \ldots, n \end{aligned} \tag{A.23}$$

or, equivalently:

$$F^{(k+1)} = M_k \, F^{(k)} \tag{A.24}$$

Consequently, at the end of the elimination procedure, matrices M_k (for $k = 1, \ldots, n-1$) and U have been built. These matrices are such that:

$$M_{n-1} \, M_{n-2} \ldots \, M_1 F = U \tag{A.25}$$

Matrices M_k are lower triangular matrices, whose diagonal coefficients are equal to 1, and whose inverse is given by:

$$M_k^{-1} = 2I_n - M_k = I_n + \overrightarrow{m}_k \, \overrightarrow{e}_k^T \tag{A.26}$$

The products $(\overrightarrow{m}_i\ \overrightarrow{e}_i^T)(\overrightarrow{m}_j\ \overrightarrow{e}_j^T)$ are equal to zero for $i \neq j$, then:

$$F = M_1^{-1}\, M_2^{-1} \dots\, M_{n-1}^{-1} U \tag{A.27}$$

$$= (I_n + \overrightarrow{m}_1\ \overrightarrow{e}_1^T)(I_n + \overrightarrow{m}_2\ \overrightarrow{e}_2^T) \dots (I_n + \overrightarrow{m}_{n-1}\ \overrightarrow{e}_{n-1}^T) U \tag{A.28}$$

$$= \left(I_n + \sum_{i=1}^{n-1} \overrightarrow{m}_i\ \overrightarrow{e}_i^T \right) U \tag{A.29}$$

$$= \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ m_{21} & 1 & & & \vdots \\ m_{31} & m_{32} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ m_{n1} & m_{n2} & \dots & m_{n,n-1} & 1 \end{bmatrix} U \tag{A.30}$$

Denoting $L = (M_{n-1}\, M_{n-2} \dots\, M_1)^{-1} = M_1^{-1} \dots\, M_{n-2}^{-1}\, M_{n-1}^{-1}$, the previous expression becomes:

$$F = LU \tag{A.31}$$

It is worthwhile noting that all the elements under the main diagonal of the L matrix are the multipliers m_{ik} generated by the Gauss method, and the diagonal terms are all equal to one.

Given a matrix equation $F\overrightarrow{s} = \overrightarrow{b}$, the solution is calculated in two steps. Both matrices L and U are first computed, and the resolution of the system of equations is done by successively solving two triangular systems:

$$L\overrightarrow{y} = \overrightarrow{b} \tag{A.32}$$

$$U\overrightarrow{s} = \overrightarrow{y} \tag{A.33}$$

Computing the LU decomposition requires $2n^3/3$ floating point operations. Obviously, the arithmetic complexity of the LU decomposition is similar to the Gauss elimination complexity.

The Cholesky decomposition

The Cholesky decomposition is a matrix factorization of symmetric positive-definite matrices into the product of a lower triangular matrix and its conjugate

transpose. When applicable, the CHOLESKY decomposition is roughly twice as efficient as the LU decomposition for solving large linear systems of equations.

Let us define a matrix $F \in \Re^{n \times n}$ as a symmetric positive-definite matrix. It exists a unique lower triangular matrix $H \in \Re^{n \times n}$ with strictly positive diagonal elements such that:

$$F = HH^T \tag{A.34}$$

where:

$$h = \begin{bmatrix} h_{11} & 0 & \dots & 0 \\ h_{21} & h_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix} \tag{A.35}$$

The matrix H is commonly called the CHOLESKY triangle. In practice, the elements of H are constructed directly, rather than forming the $L^{(1)}$ and $U^{(1)}$ matrices first, as in the LU decomposition. This step is done in a way similar to the LU factorization. Assuming that $i \leq j$, the CHOLESKY decomposition then requires that:

$$f_{ij} = \sum_{k=1}^{i} h_{ik}\, h_{jk}, \quad 1 \leq i \leq j \leq n \tag{A.36}$$

In equation A.36, the fact that $(H^T)_{kj} = h_{jk}$ has directly been used. The sum only extends up to $k = i$ since H is lower triangular. Of couse, the same equation will also hold for $i > j$, since F is a symmetric matrix. For $i = j$, i.e. for the main diagonal matrix coefficients, equation A.36 gives:

$$h_{11} = \sqrt{f_{11}} \tag{A.37}$$

$$h_{ii} = \left(f_{ii} - \sum_{k=1}^{i-1} h_{ik}^2 \right)^{1/2}, \quad 1 < i \leq n \tag{A.38}$$

It is here interesting to note that the calculations are done column by column, starting with the first diagonal element, then continuing in the same column,

passing after to the second diagonal element, etc. As F is a positive-definite matrix, f_{11} (and therefore h_{11}) is a positive real number (all the principal minors have to be positive). Further, all the diagonal coefficient h_{ii} are also positive. The non diagonal coefficients are given by:

$$h_{ji} = \frac{1}{h_{ii}} \left(f_{ij} - \sum_{k=1}^{i-1} h_{ik} h_{jk} \right), \quad 1 \leq i < j \leq n \tag{A.39}$$

For large n the number of the operations required by the CHOLESKY algorithm is approximately $n^3/3$ floating point operations which, as might be expected, is half the number given for the LU factorization of a non symmetric matrix. This algorithm is very stable with respect to the round-off error propagation [104].

Appendix

B Appendix: Carbon scattering system

Unpreconditioned carbon system

In this part of the appendixes, the figures of the reconstructed vectors obtained with the different numerical methods are given for the unpreconditioned carbon matrix system. The figures correspond to the vectors reported in Table 8.4, p. 101.

Vector	Material	Page
$\overrightarrow{s}_{\mathrm{rec,\,SVD}}$	carbon	p. 179
$\overrightarrow{s}_{\mathrm{rec,\,LU}}$	carbon	p. 180
$\overrightarrow{s}_{\mathrm{rec,\,Sub}}$	carbon	p. 181
$\overrightarrow{s}_{\mathrm{rec,\,BE}}$	carbon	p. 182
$\overrightarrow{s}_{\mathrm{rec,\,G}}$	carbon	p. 183
$\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$	carbon	p. 184
$\overrightarrow{s}_{\mathrm{rec,\,J}}$	carbon	p. 185
$\overrightarrow{s}_{\mathrm{rec,\,SOR}}$	carbon	p. 186

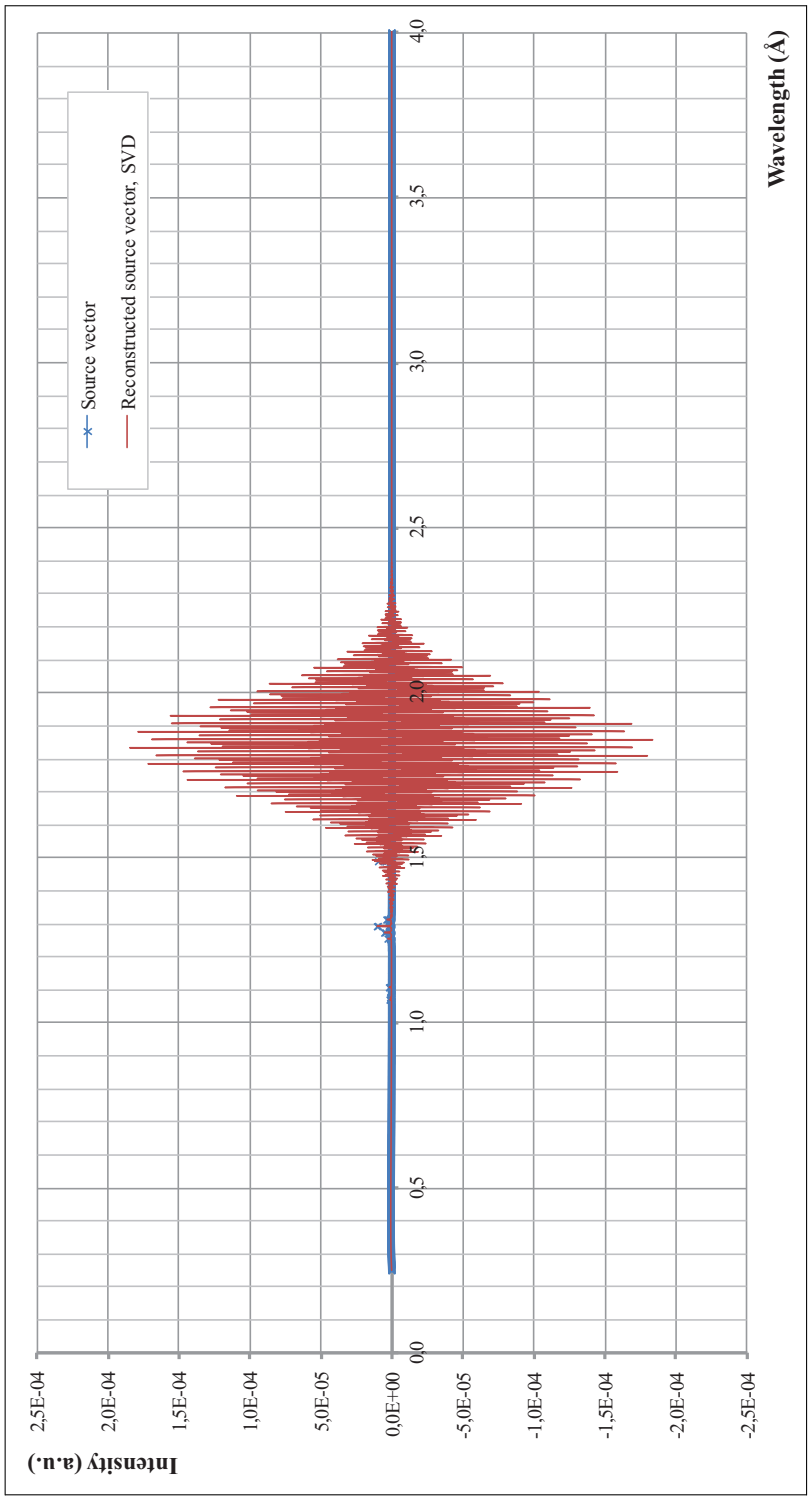

Figure B.1: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SVD}}$. The system has not been preconditioned. Carbon sample.

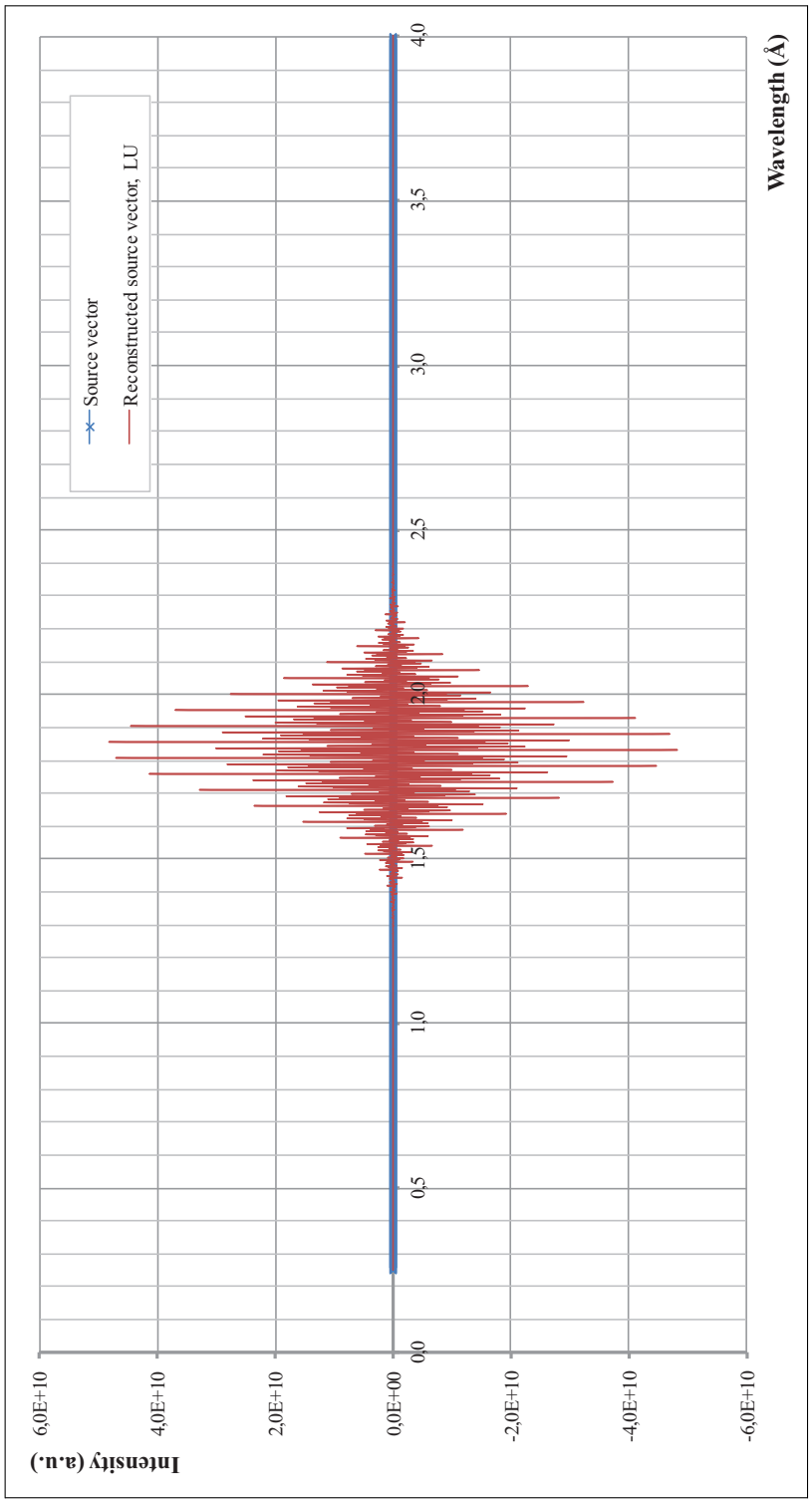

Figure B.2: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, LU}}$. The system has not been preconditioned. Carbon sample.

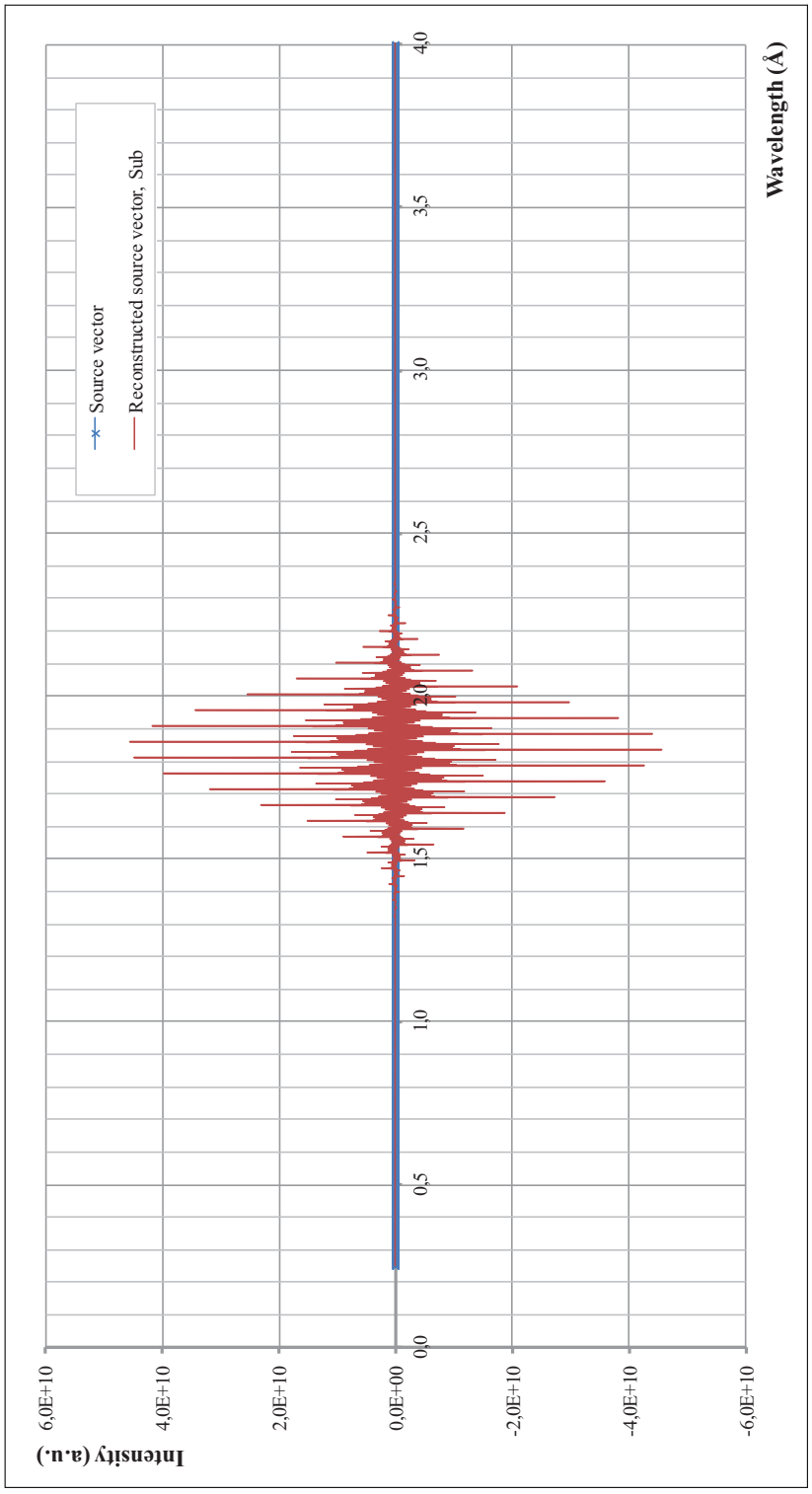

Figure B.3: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. The system has not been preconditioned. Carbon sample.

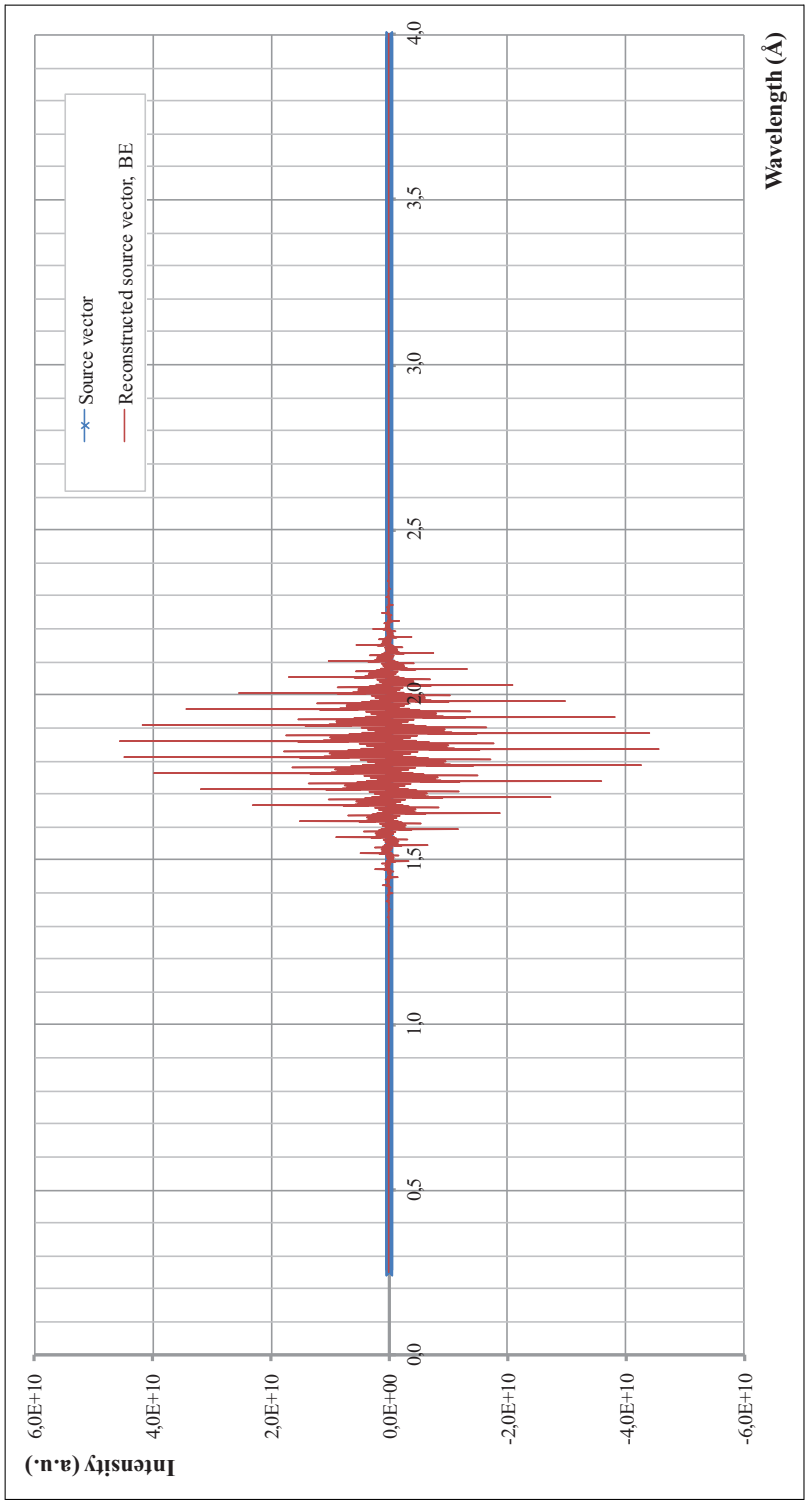

Figure B.4: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, BE}}$. The system has not been preconditioned. Carbon sample.

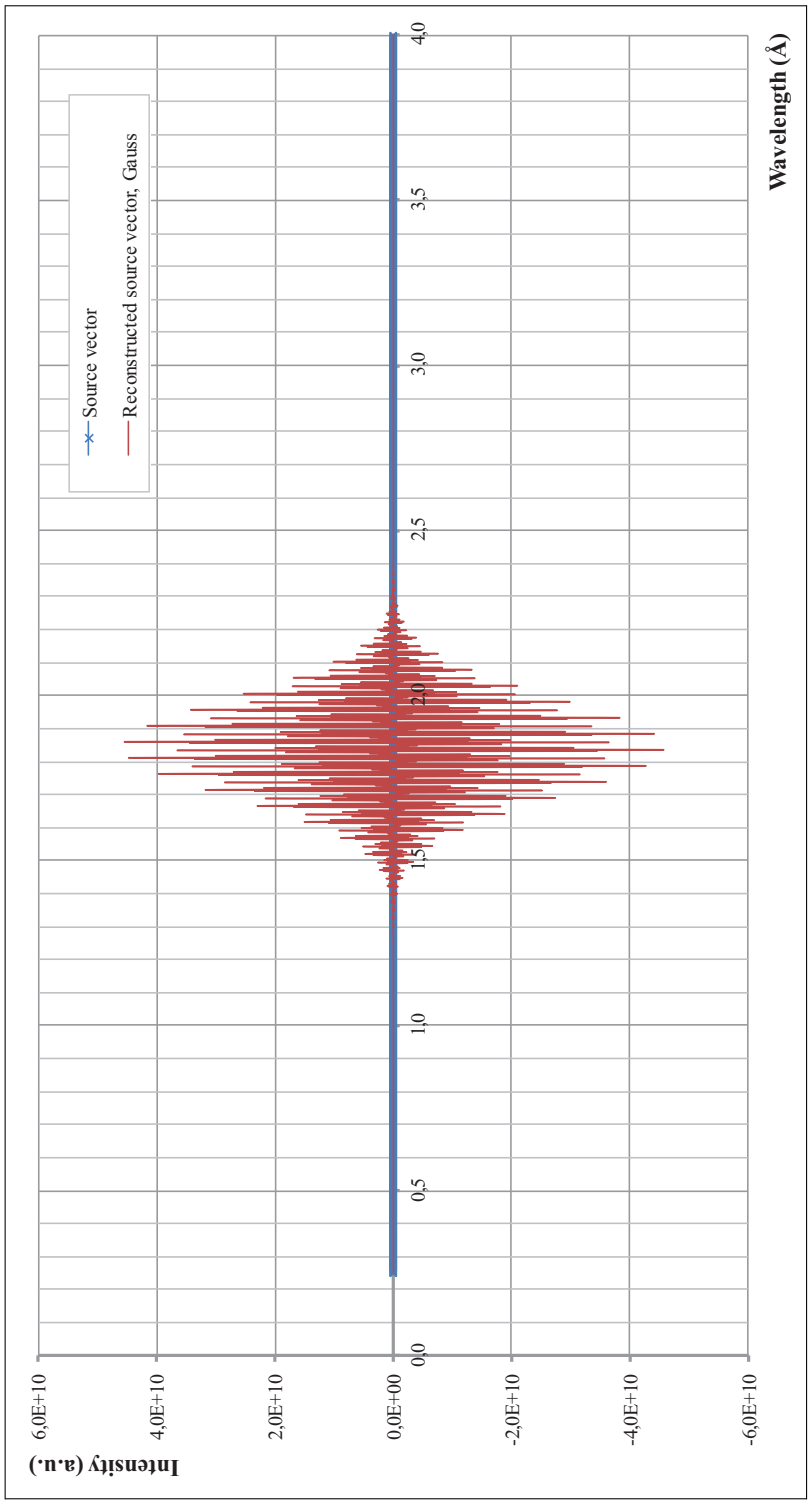

Figure B.5: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, G}}$. The system has not been preconditioned. Carbon sample.

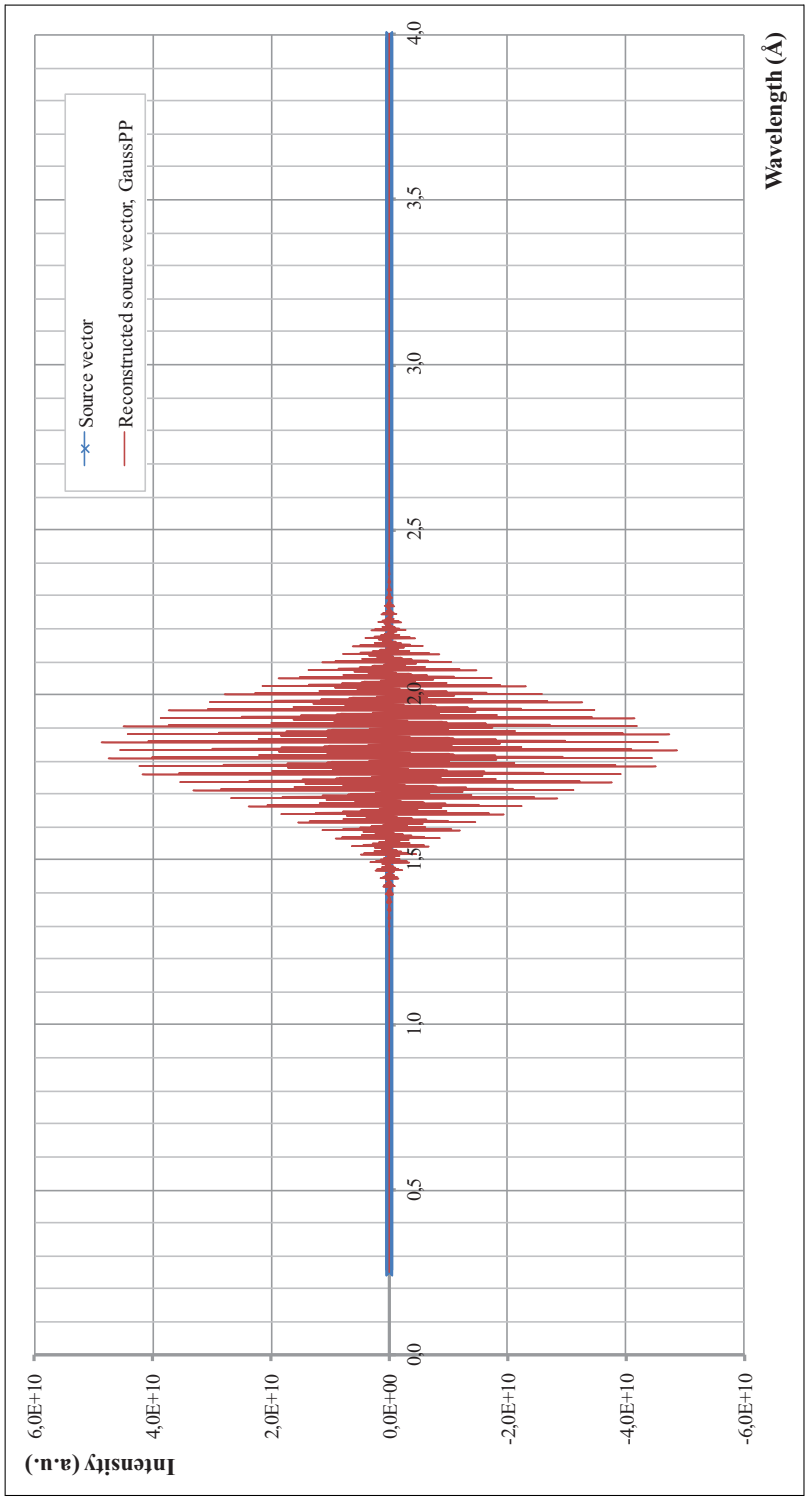

Figure B.6: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Gpp}}$. The system has not been preconditioned. Carbon sample.

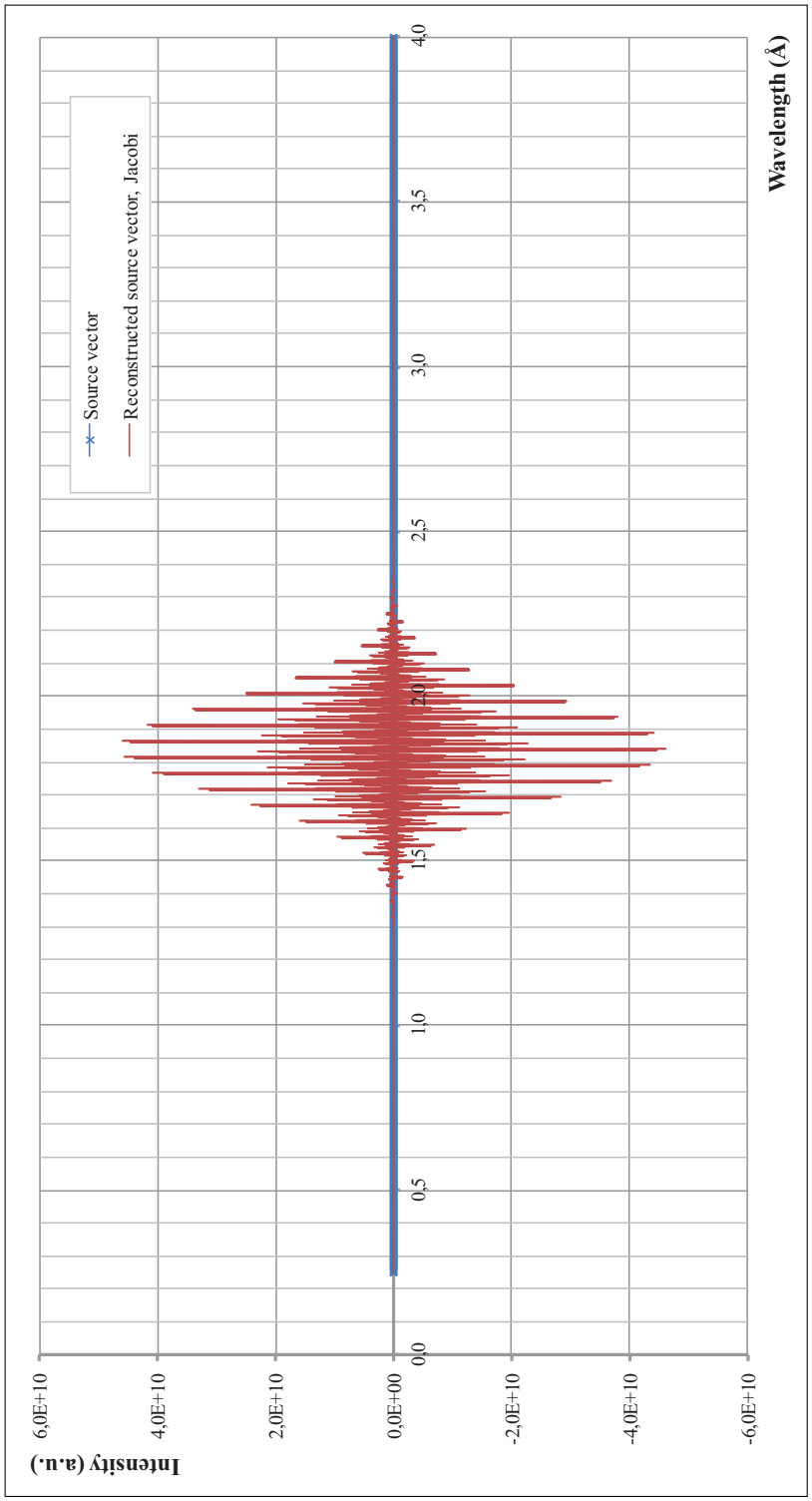

Figure B.7: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, J}}$. The system has not been preconditioned. Carbon sample.

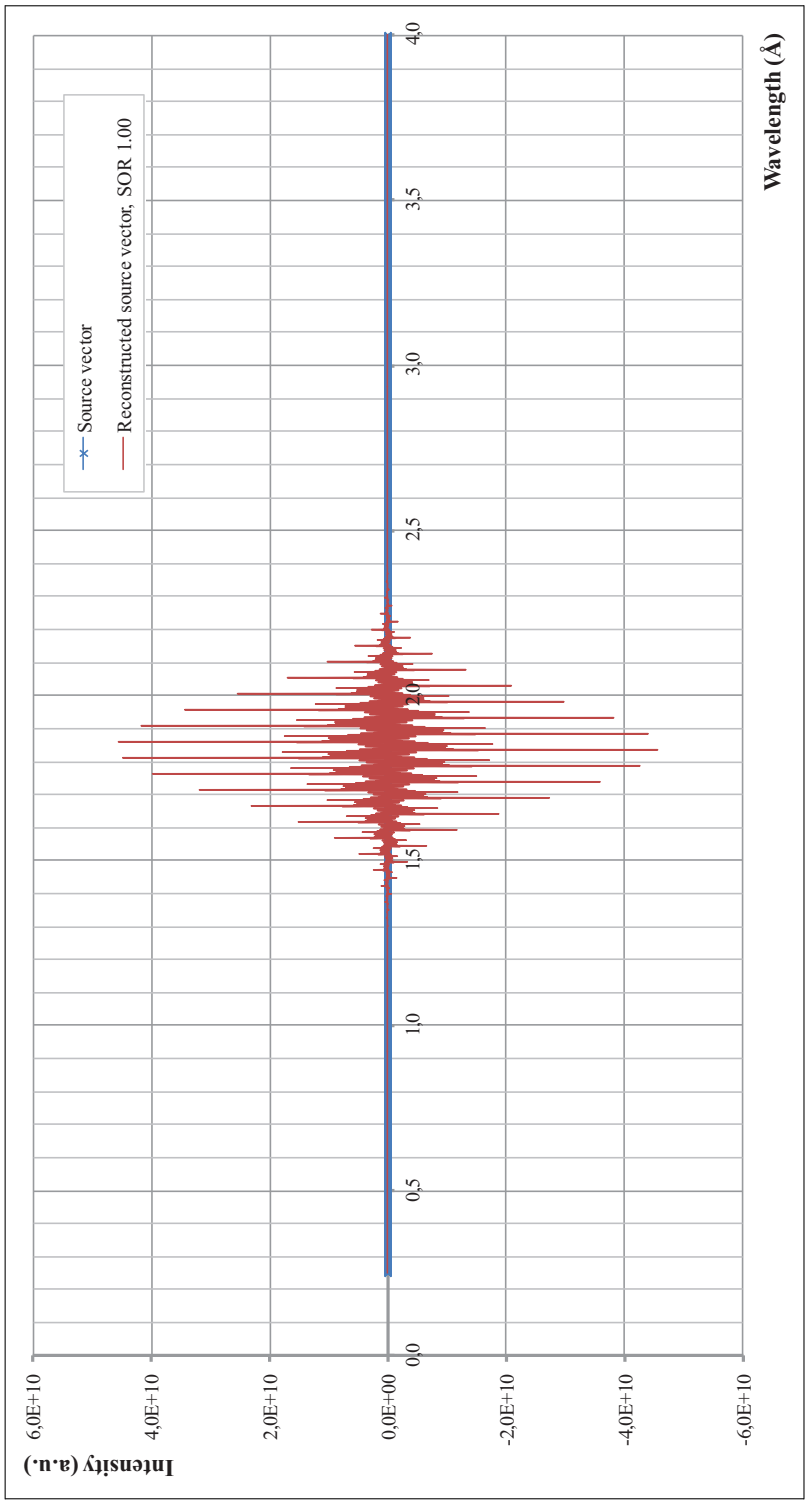

Figure B.8: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SOR}}$. The system has not been preconditioned. Carbon sample.

Right preconditioned carbon system

In this part of the appendixes, the figures of the reconstructed vectors obtained with the different numerical methods are given for the right preconditioned carbon matrix system. The figures correspond to the vectors reported in Table 8.5, p. 102.

Vector	Material	Page
$\overrightarrow{s}_{\mathrm{rec,\,SVD}}$	carbon	p. 188
$\overrightarrow{s}_{\mathrm{rec,\,Chol}}$	carbon	p. 189
$\overrightarrow{s}_{\mathrm{rec,\,LU}}$	carbon	p. 190
$\overrightarrow{s}_{\mathrm{rec,\,Sub}}$	carbon	p. 191
$\overrightarrow{s}_{\mathrm{rec,\,G}}$	carbon	p. 192
$\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$	carbon	p. 193
$\overrightarrow{s}_{\mathrm{rec,\,J}}$	carbon	p. 194
$\overrightarrow{s}_{\mathrm{rec,\,SOR}}$	carbon	p. 195

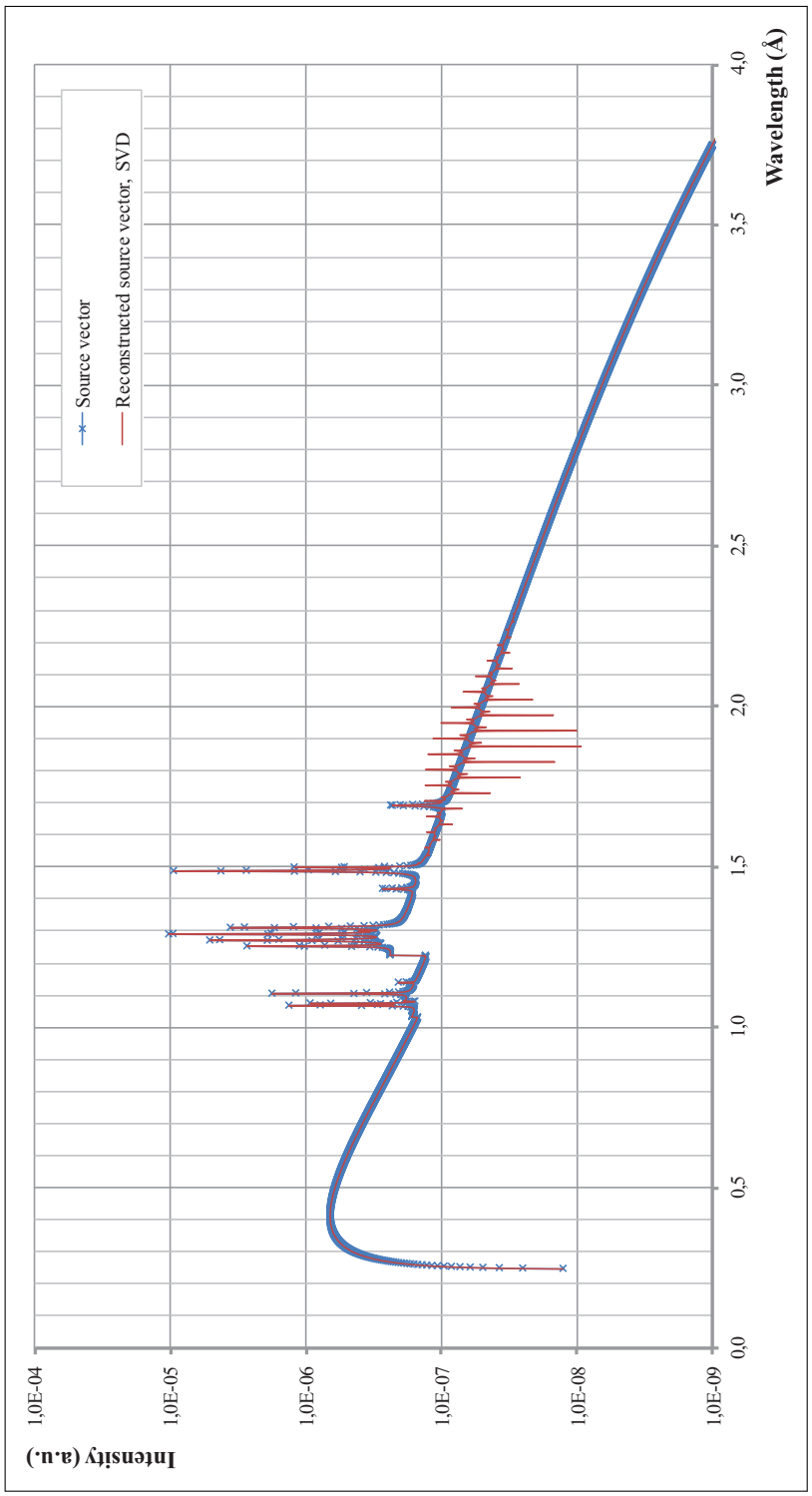

Figure B.9: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SVD}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

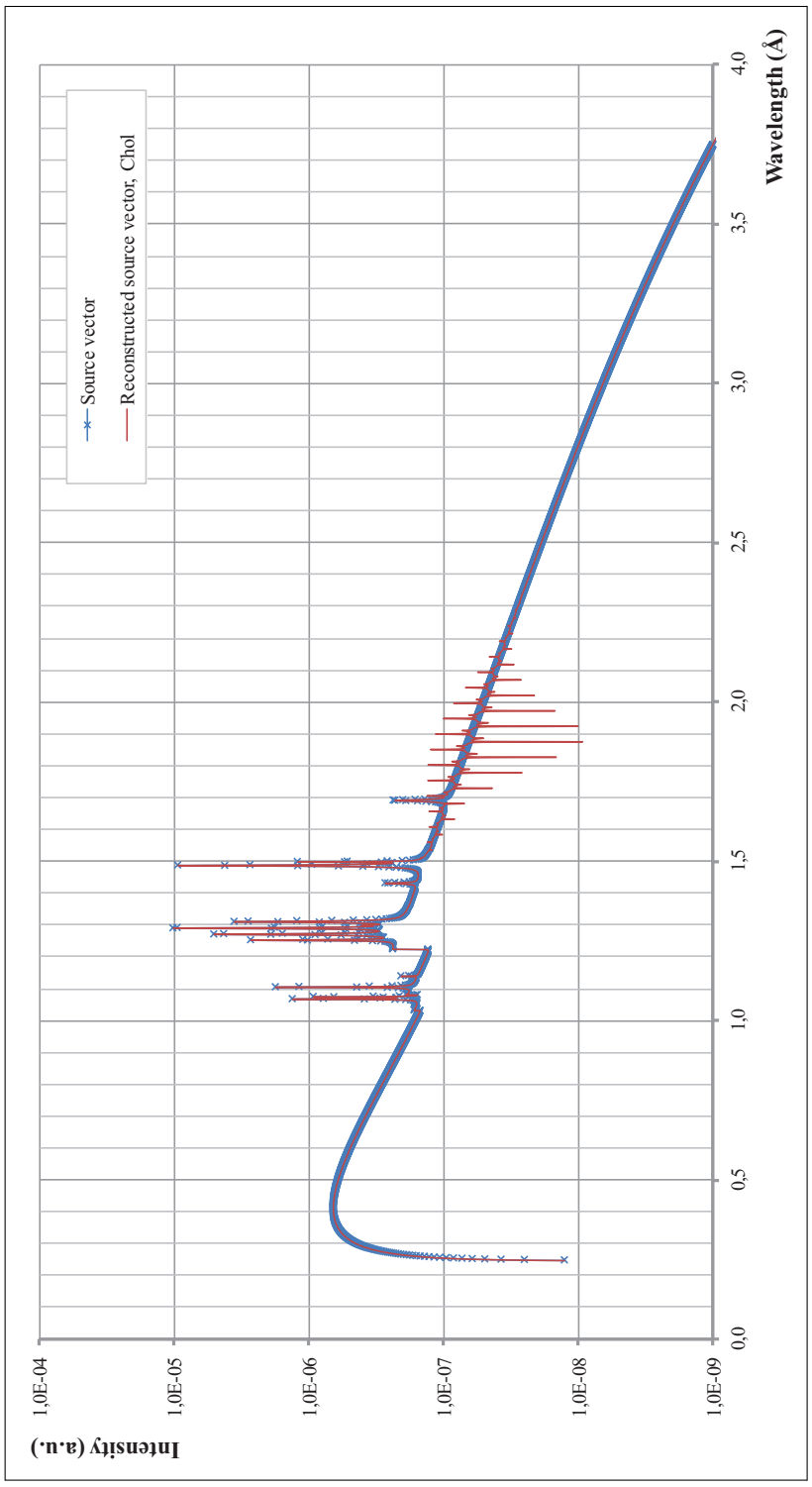

Figure B.10: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Chol}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

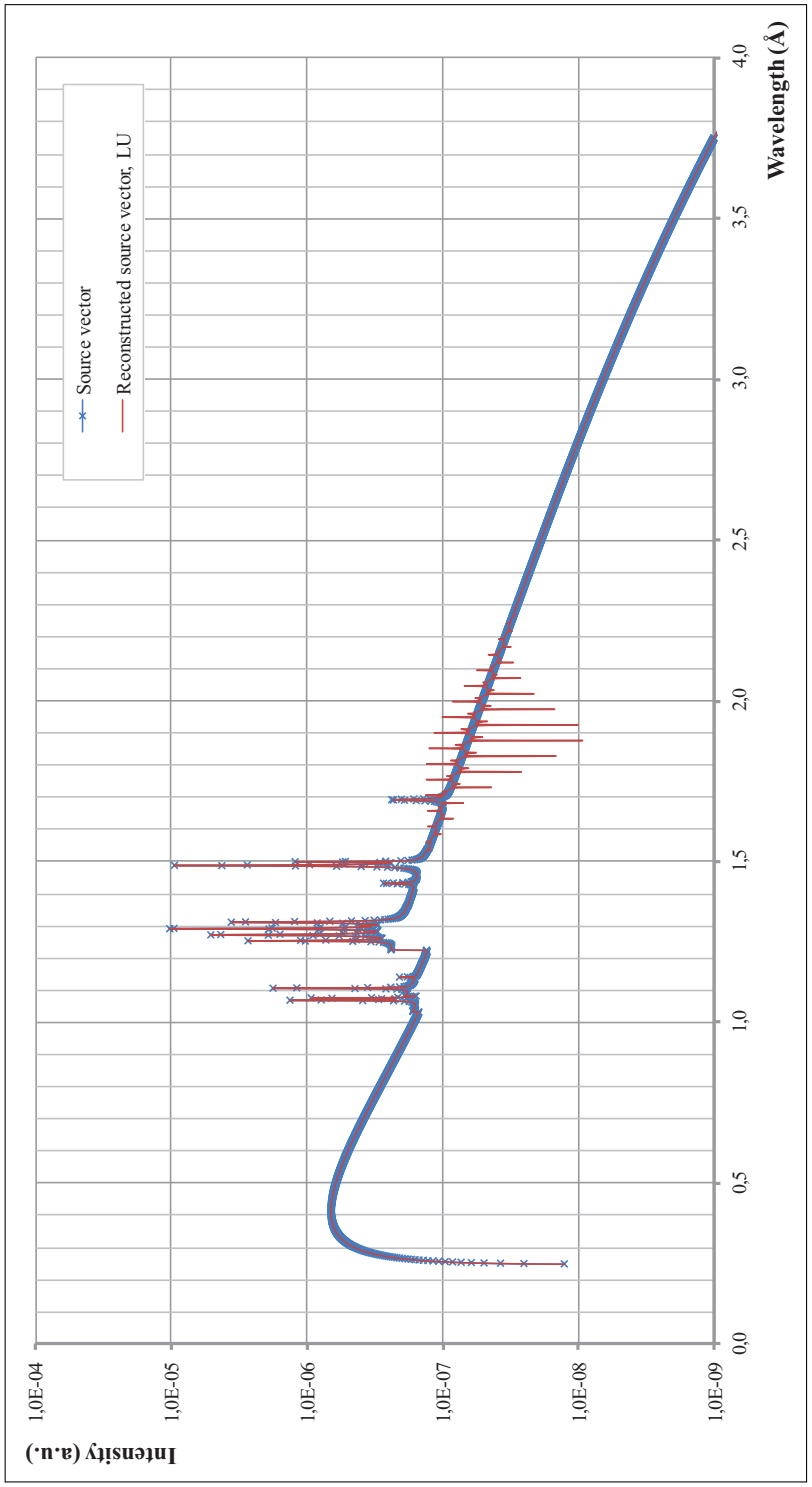

Figure B.11: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, LU}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

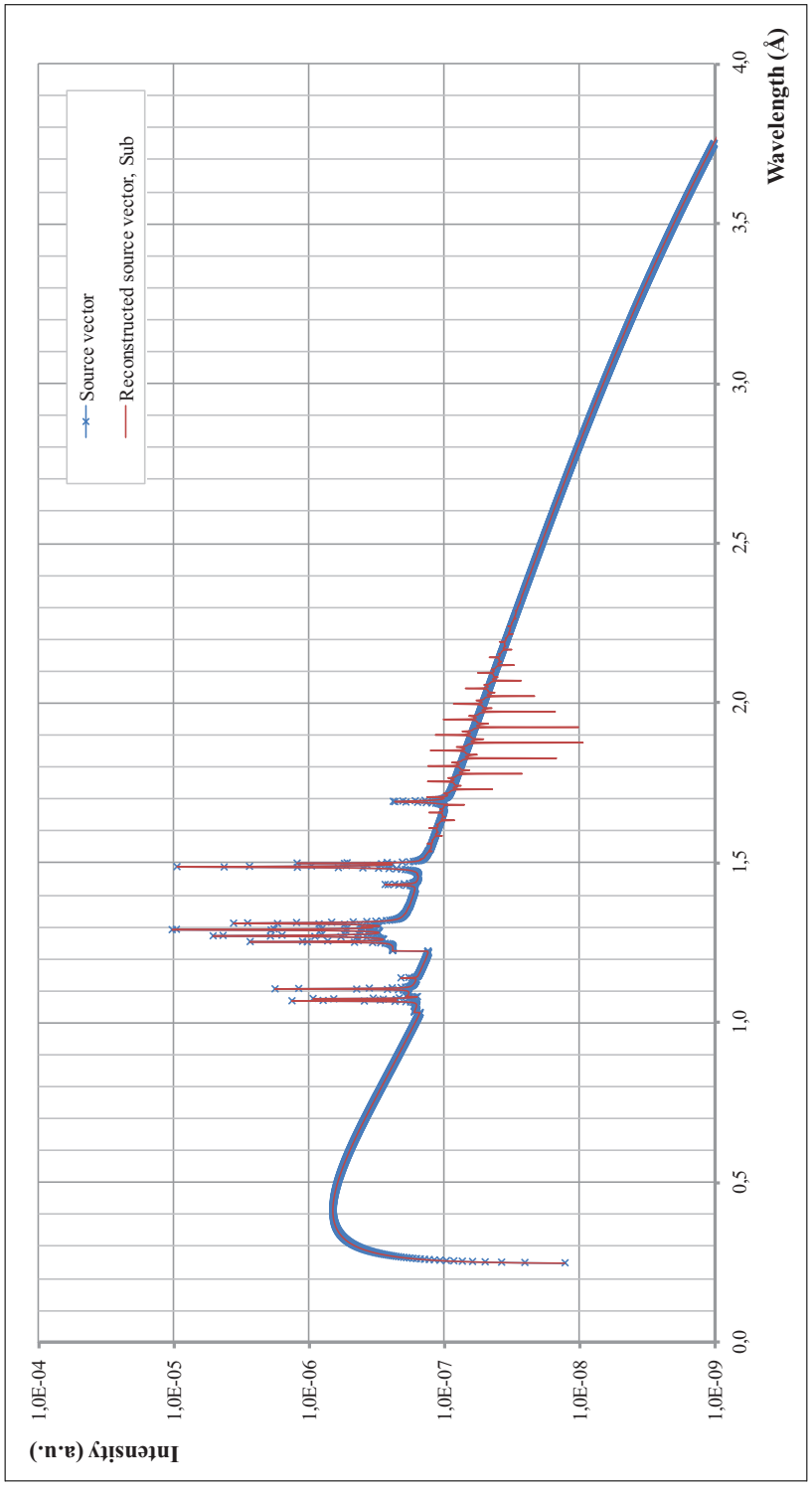

Figure B.12: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

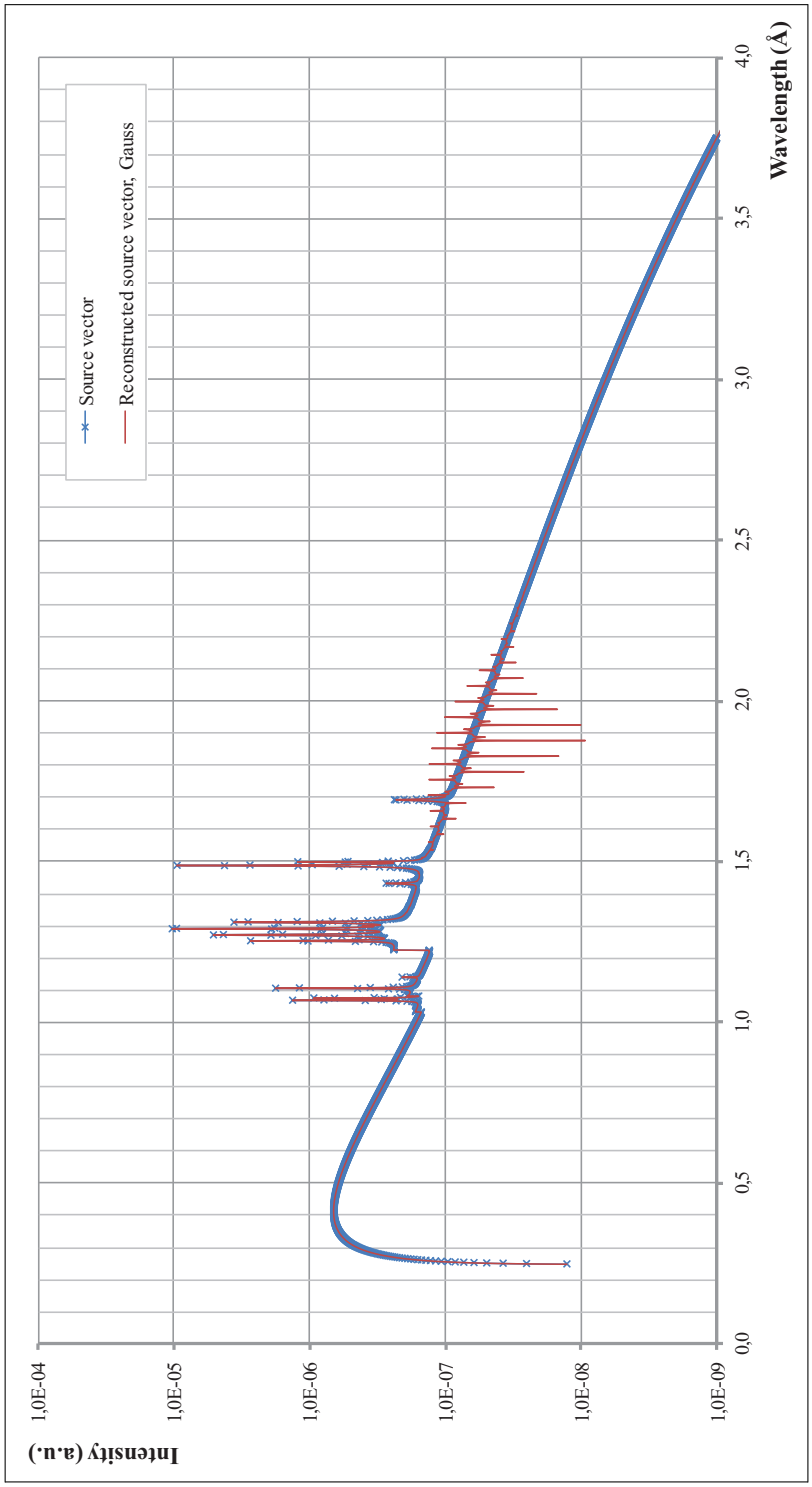

Figure B.13: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, G}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

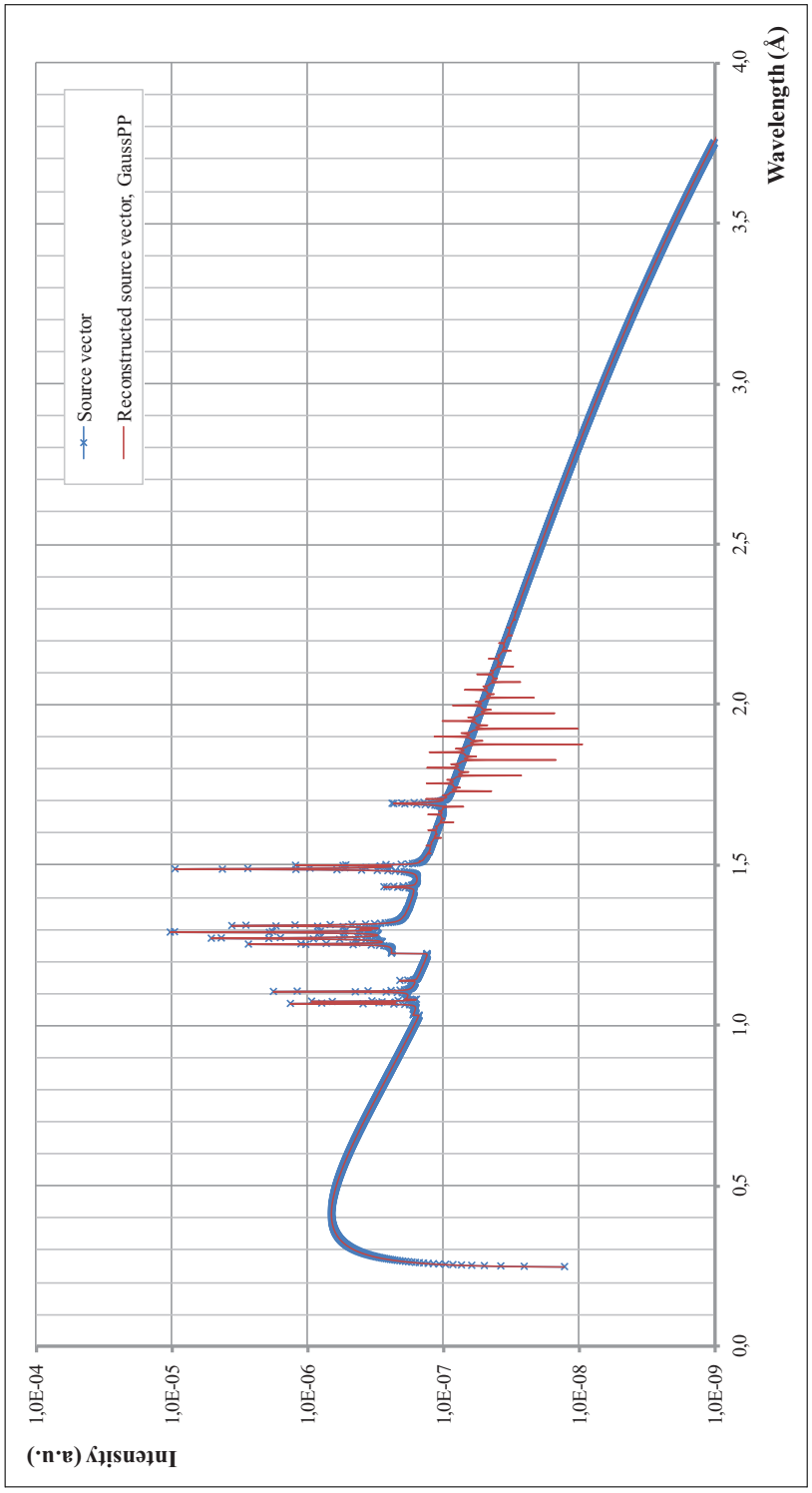

Figure B.14: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Gpp}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

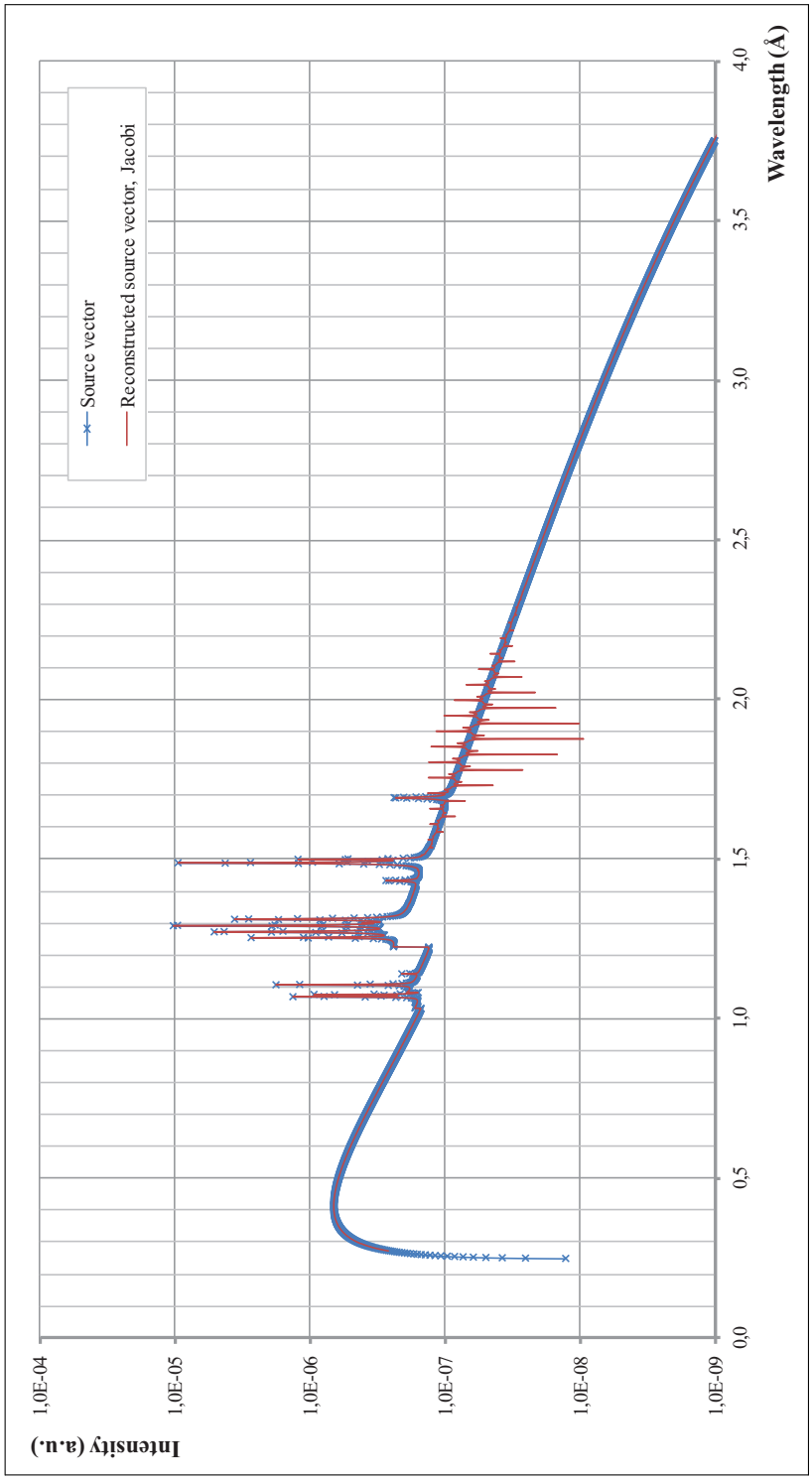

Figure B.15: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, J}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

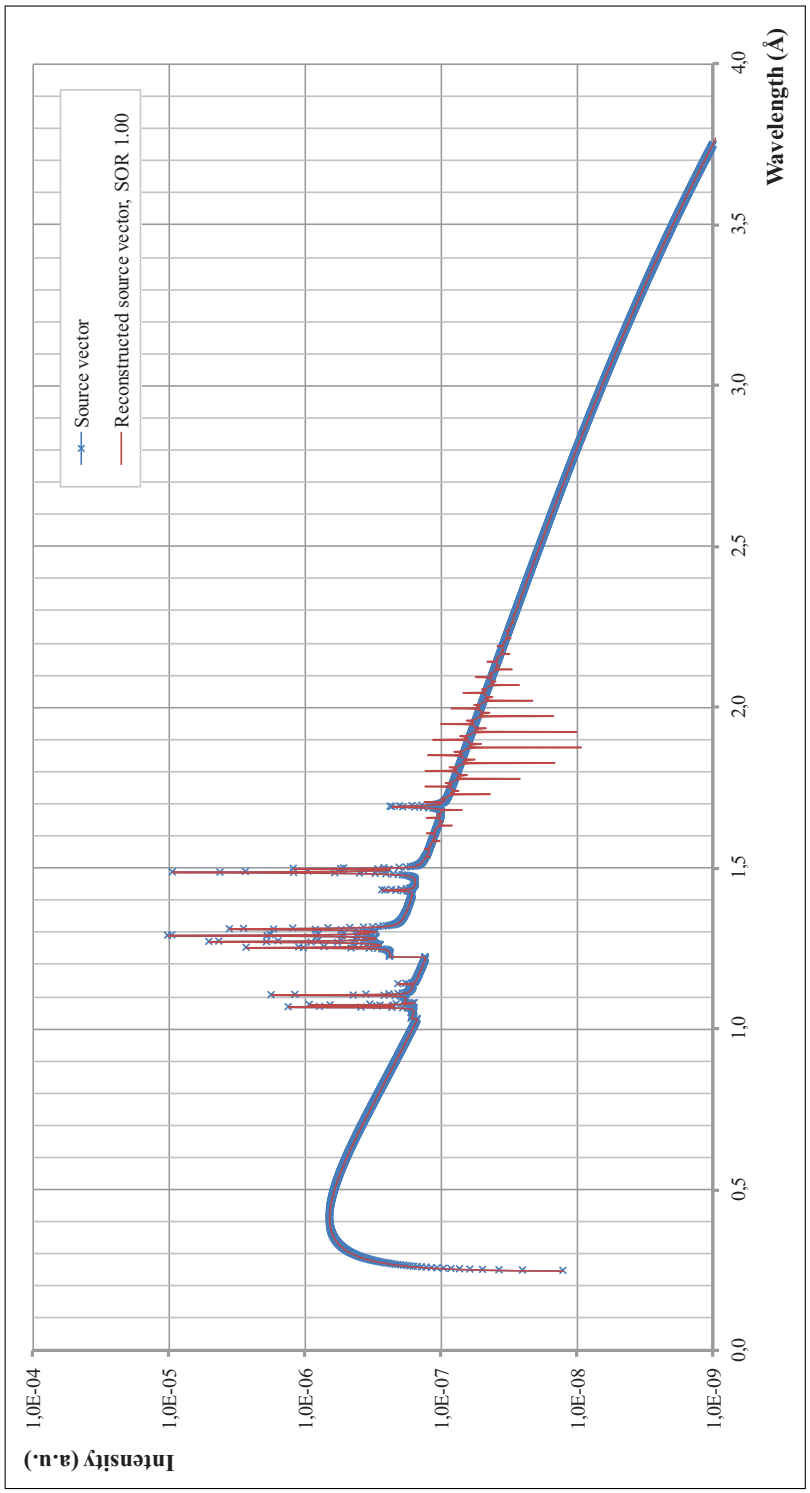

Figure B.16: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SOR}}$. The system is right preconditioned by the adjoint matrix. Carbon sample.

Left preconditioned carbon system

In this part of the appendixes, the figures of the reconstructed vectors obtained with the different numerical methods are given for the left preconditioned carbon matrix system. The figures correspond to the vectors reported in Table 8.6, p. 103.

Vector	Material	Page
$\overrightarrow{s}_{\text{rec, SVD}}$	carbon	p. 197
$\overrightarrow{s}_{\text{rec, Chol}}$	carbon	p. 198
$\overrightarrow{s}_{\text{rec, LU}}$	carbon	p. 199
$\overrightarrow{s}_{\text{rec, Sub}}$	carbon	p. 200
$\overrightarrow{s}_{\text{rec, G}}$	carbon	p. 201
$\overrightarrow{s}_{\text{rec, Gpp}}$	carbon	p. 202
$\overrightarrow{s}_{\text{rec, J}}$	carbon	p. 203
$\overrightarrow{s}_{\text{rec, SOR}}$	carbon	p. 204

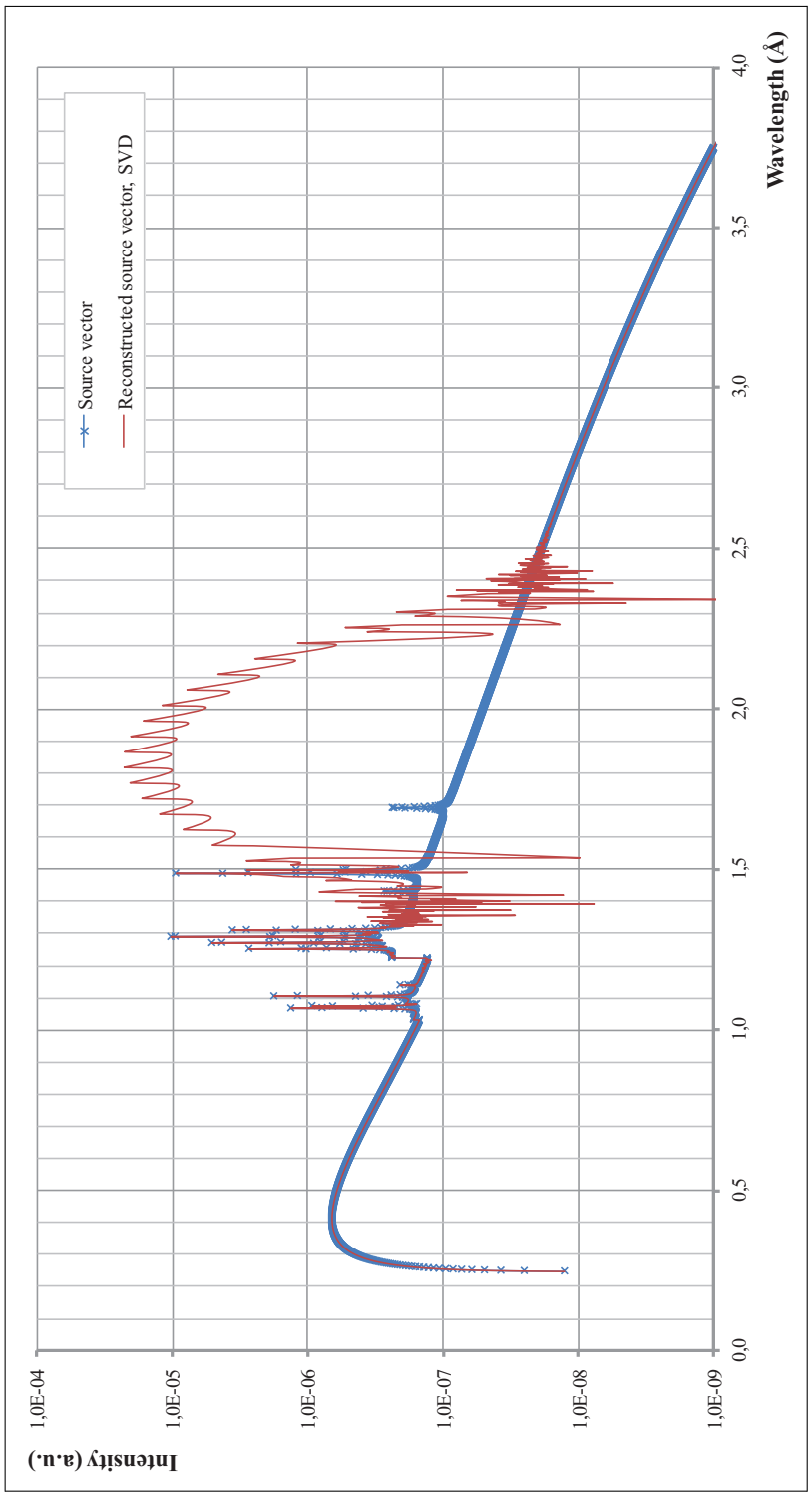

Figure B.17: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SVD}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

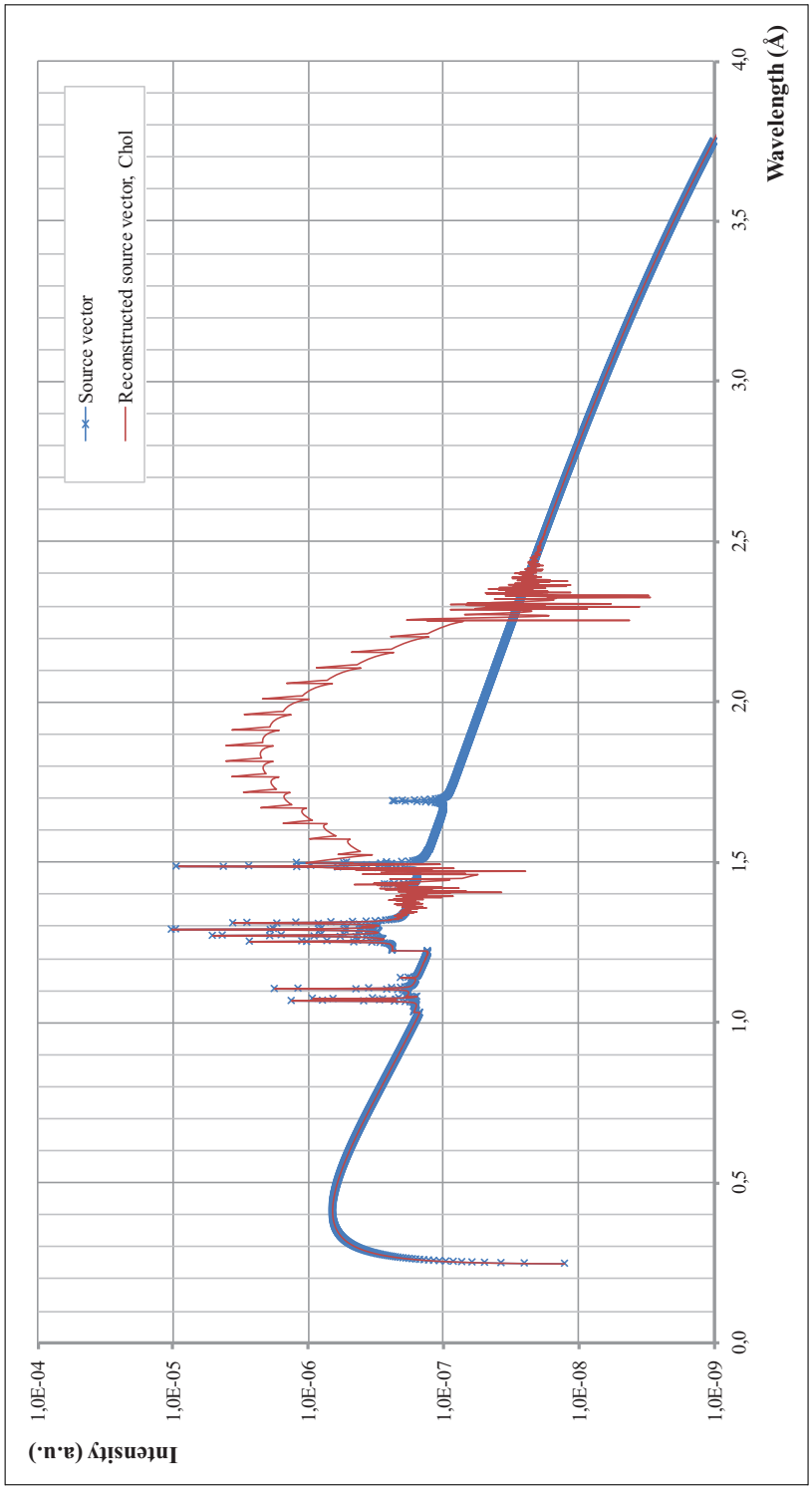

Figure B.18: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Chol}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

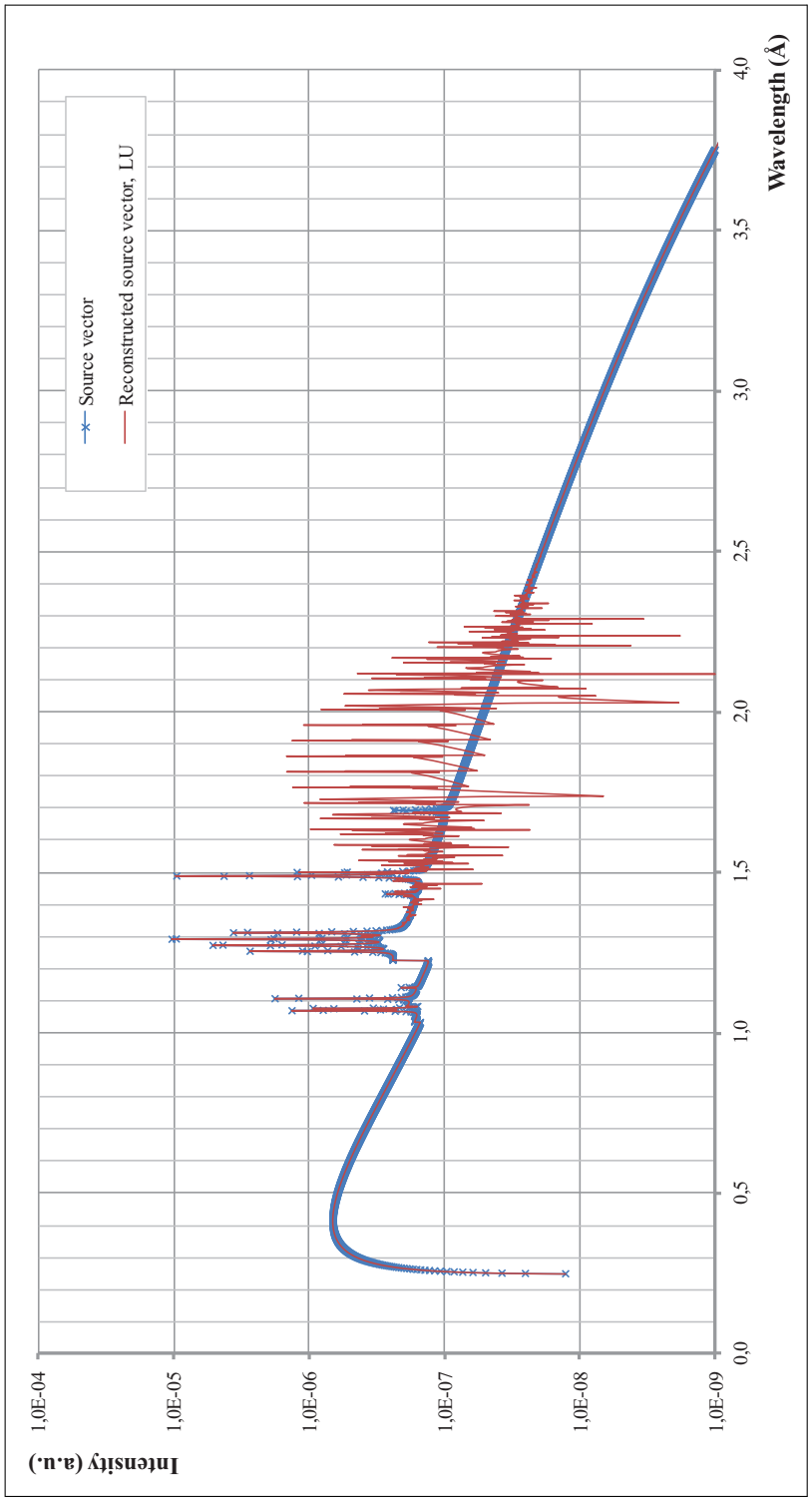

Figure B.19: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, LU}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

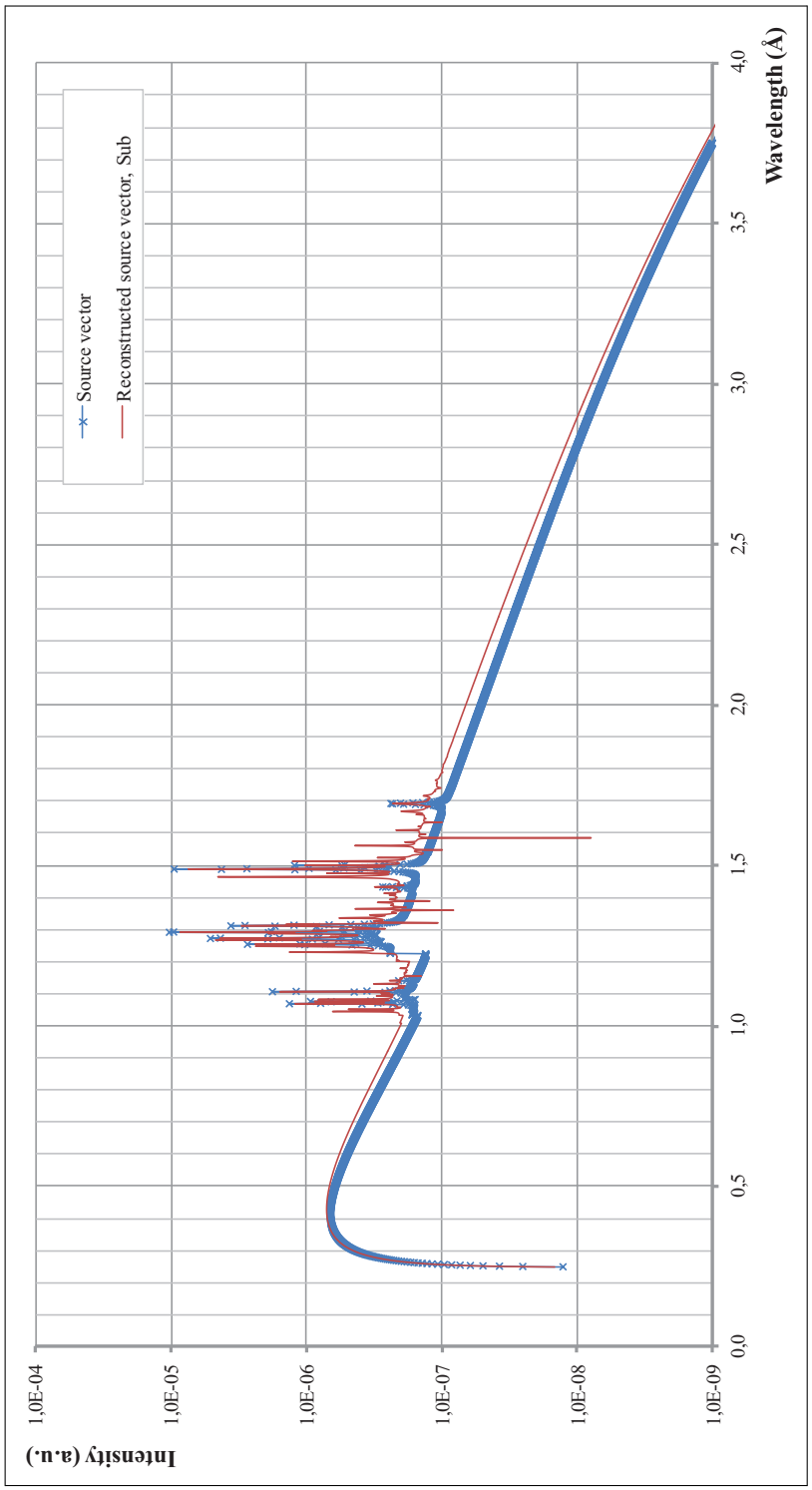

Figure B.20: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

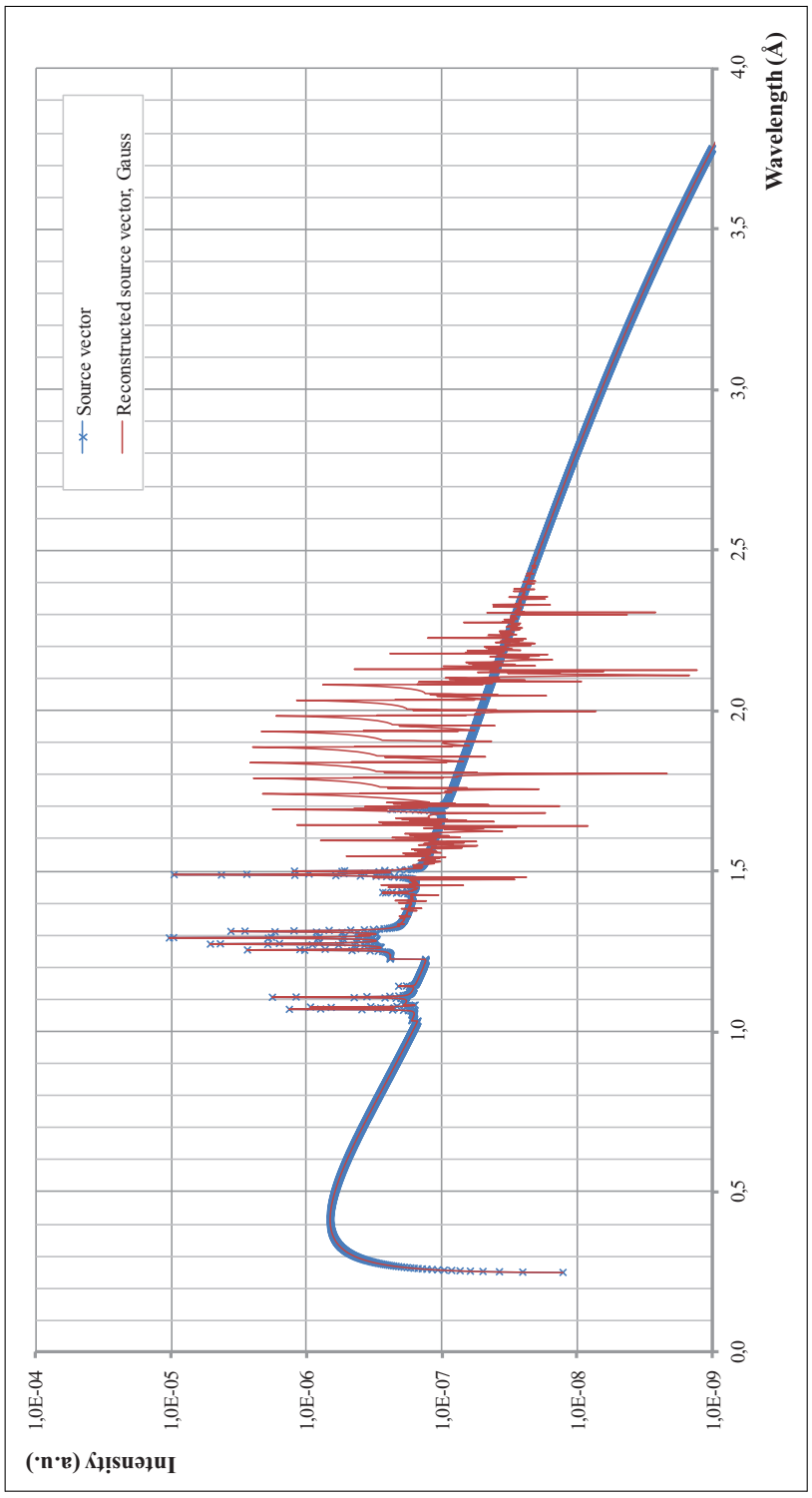

Figure B.21: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, G}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

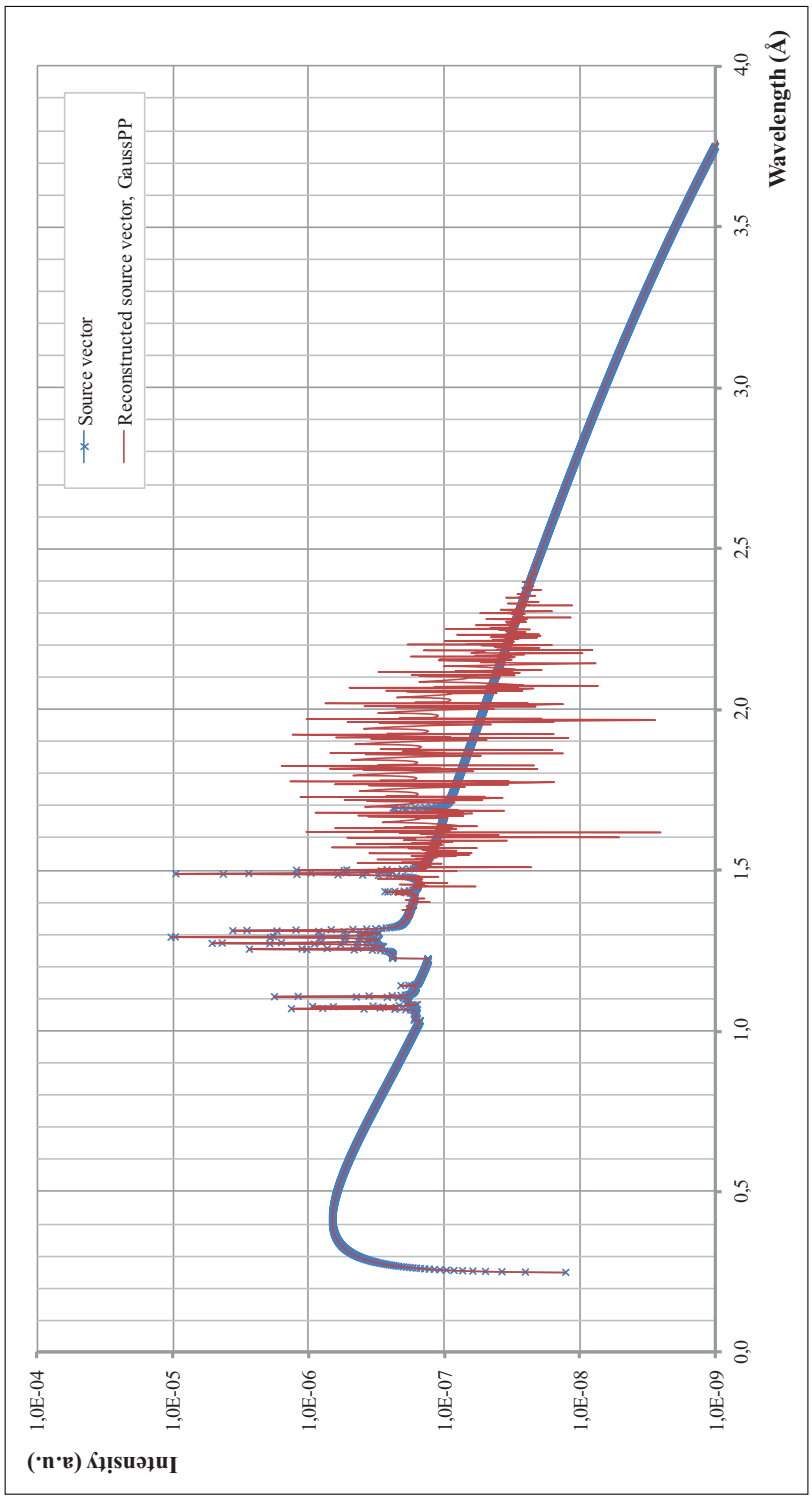

Figure B.22: Comparison between the source spectrum $\vec{s}$ and the reconstructed source vector $\vec{s}_{\text{rec, Gpp}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

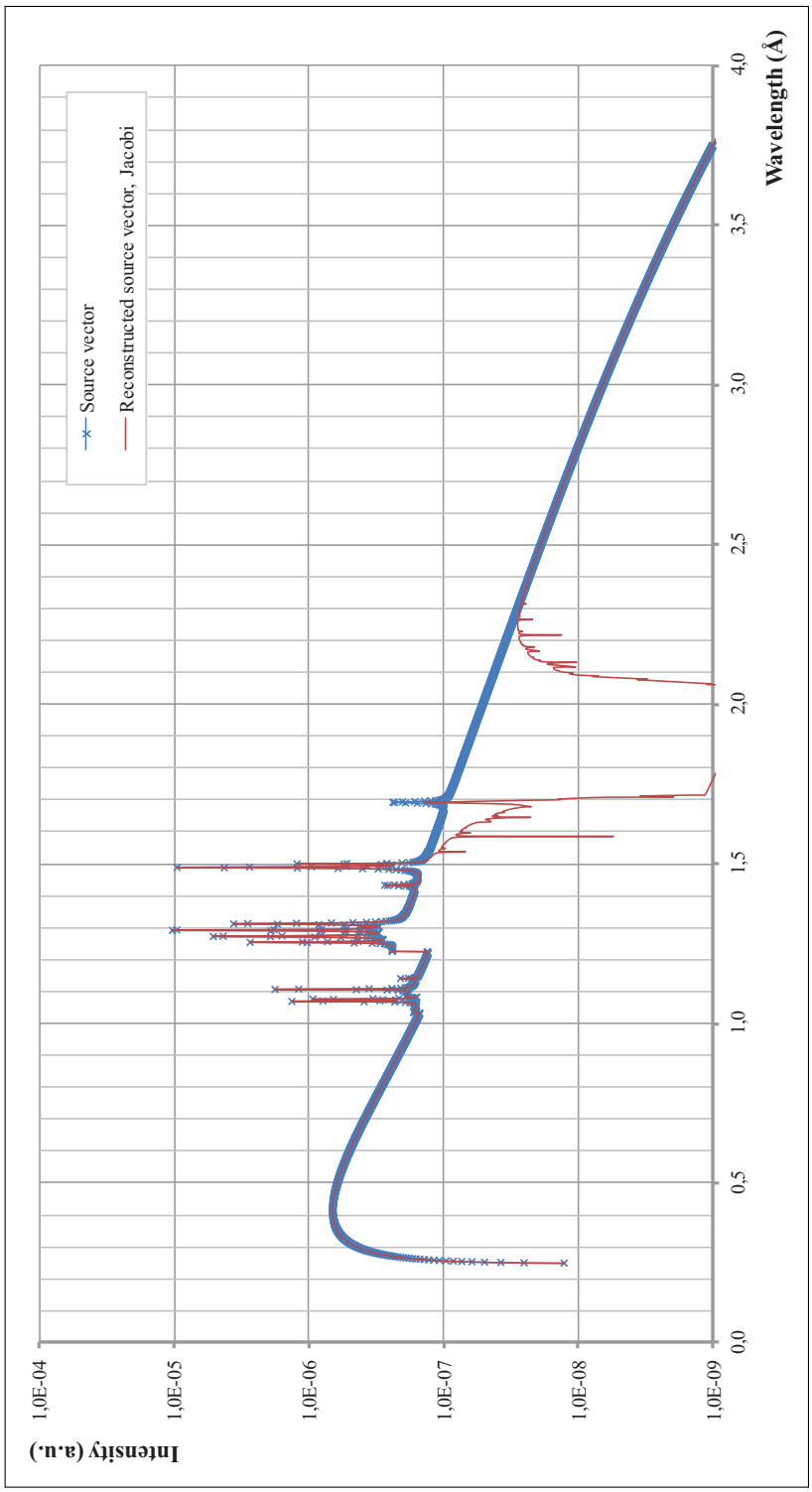

Figure B.23: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, J}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

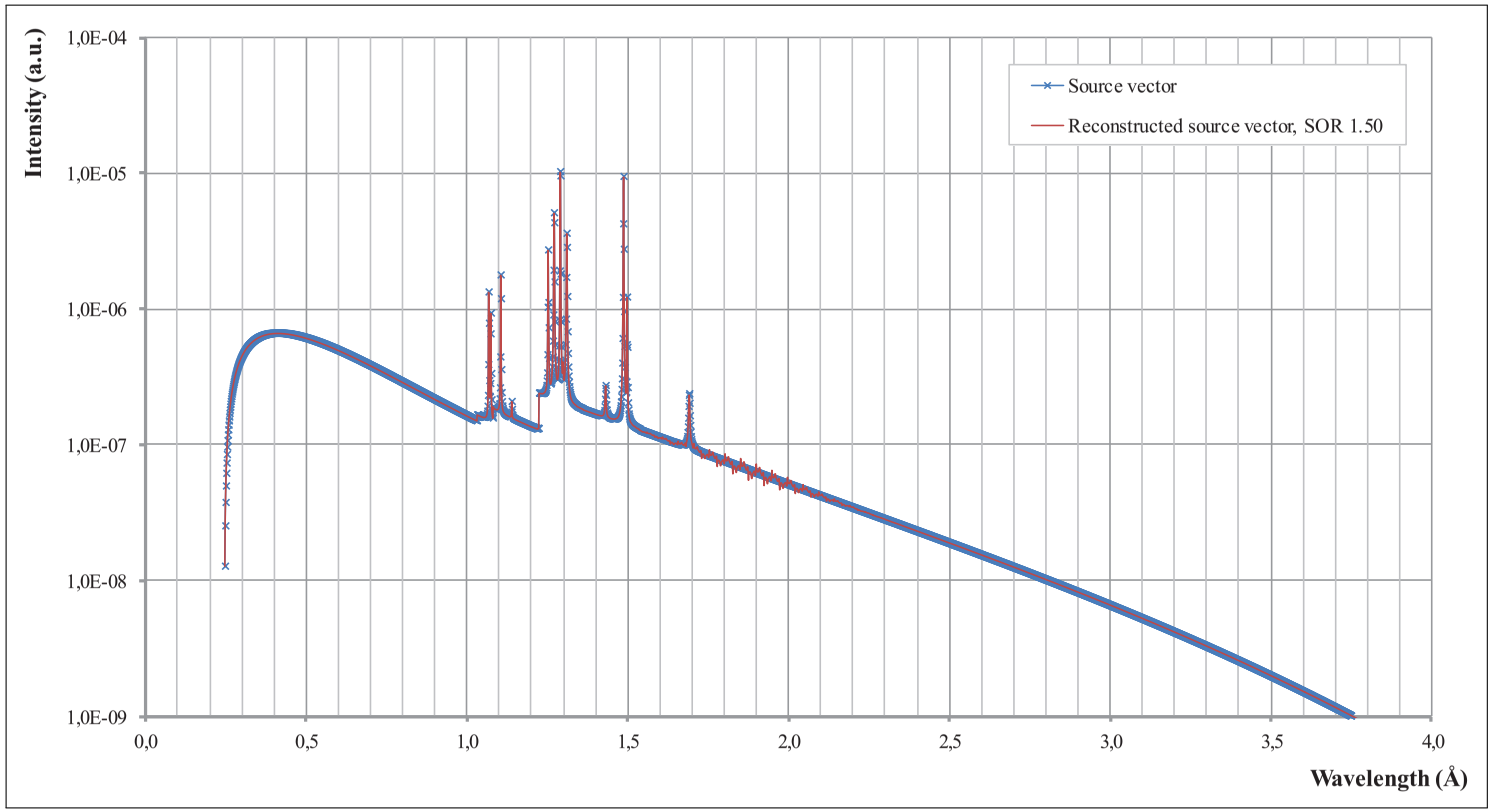

Figure B.24: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SOR}}$. The system is left preconditioned by the adjoint matrix. Carbon sample.

Appendix

C Appendix: Aluminium scattering system

Unpreconditioned aluminium system

In this part of the appendixes, the figures of the reconstructed vectors obtained with the different numerical methods are given for the unpreconditioned aluminium matrix system. The figures correspond to the vectors reported in Table 8.8, p. 110.

Vector	Material	Page
$\overrightarrow{s}_{\mathrm{rec,\,SVD}}$	aluminium	p. 206
$\overrightarrow{s}_{\mathrm{rec,\,LU}}$	aluminium	p. 207
$\overrightarrow{s}_{\mathrm{rec,\,Sub}}$	aluminium	p. 208
$\overrightarrow{s}_{\mathrm{rec,\,BE}}$	aluminium	p. 209
$\overrightarrow{s}_{\mathrm{rec,\,G}}$	aluminium	p. 210
$\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$	aluminium	p. 211
$\overrightarrow{s}_{\mathrm{rec,\,J}}$	aluminium	p. 212
$\overrightarrow{s}_{\mathrm{rec,\,SOR}}$	aluminium	p. 213

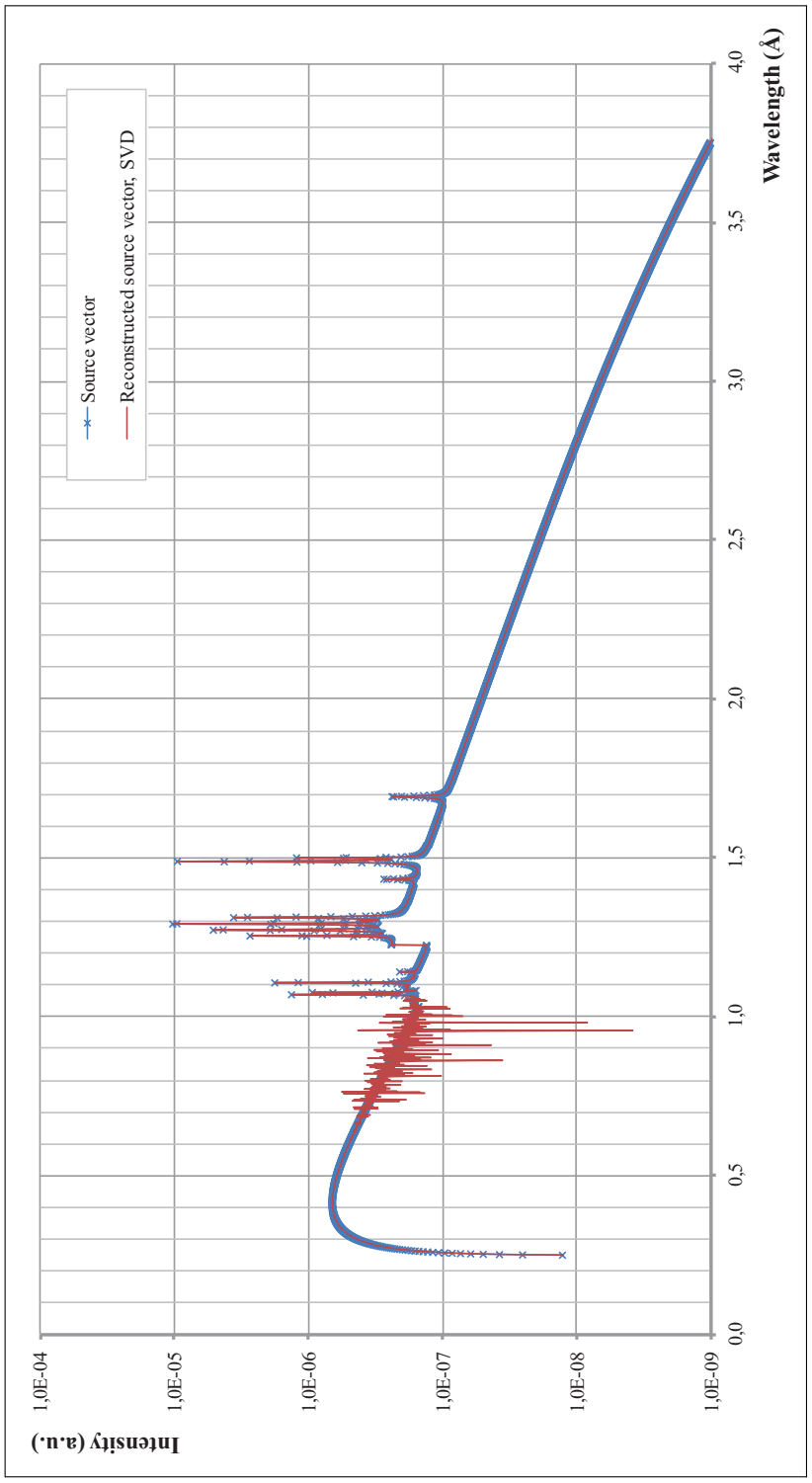

Figure C.1: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SVD}}$. The system has not been preconditioned. Aluminium sample.

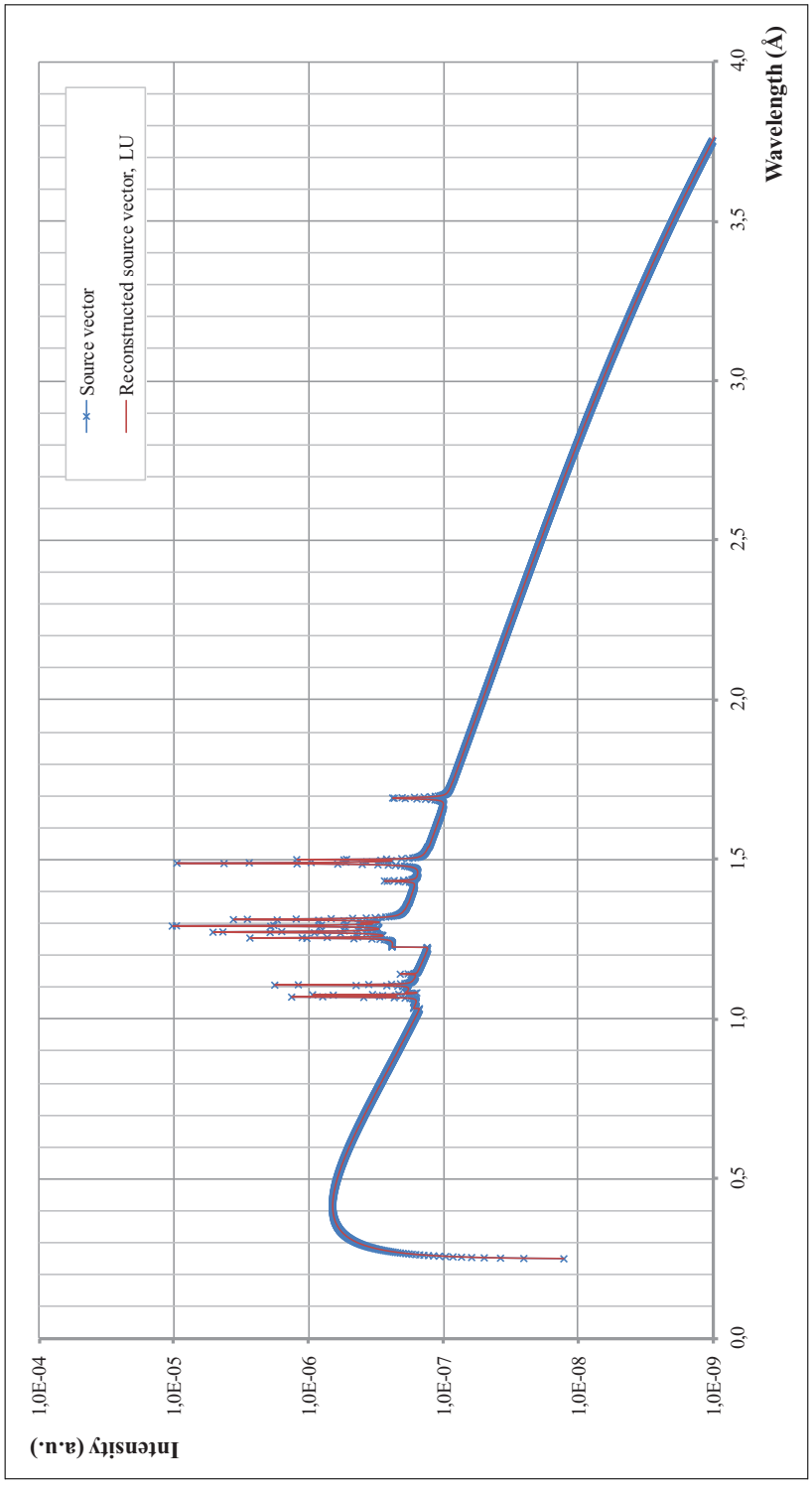

Figure C.2: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, LU}}$. The system has not been preconditioned. Aluminium sample.

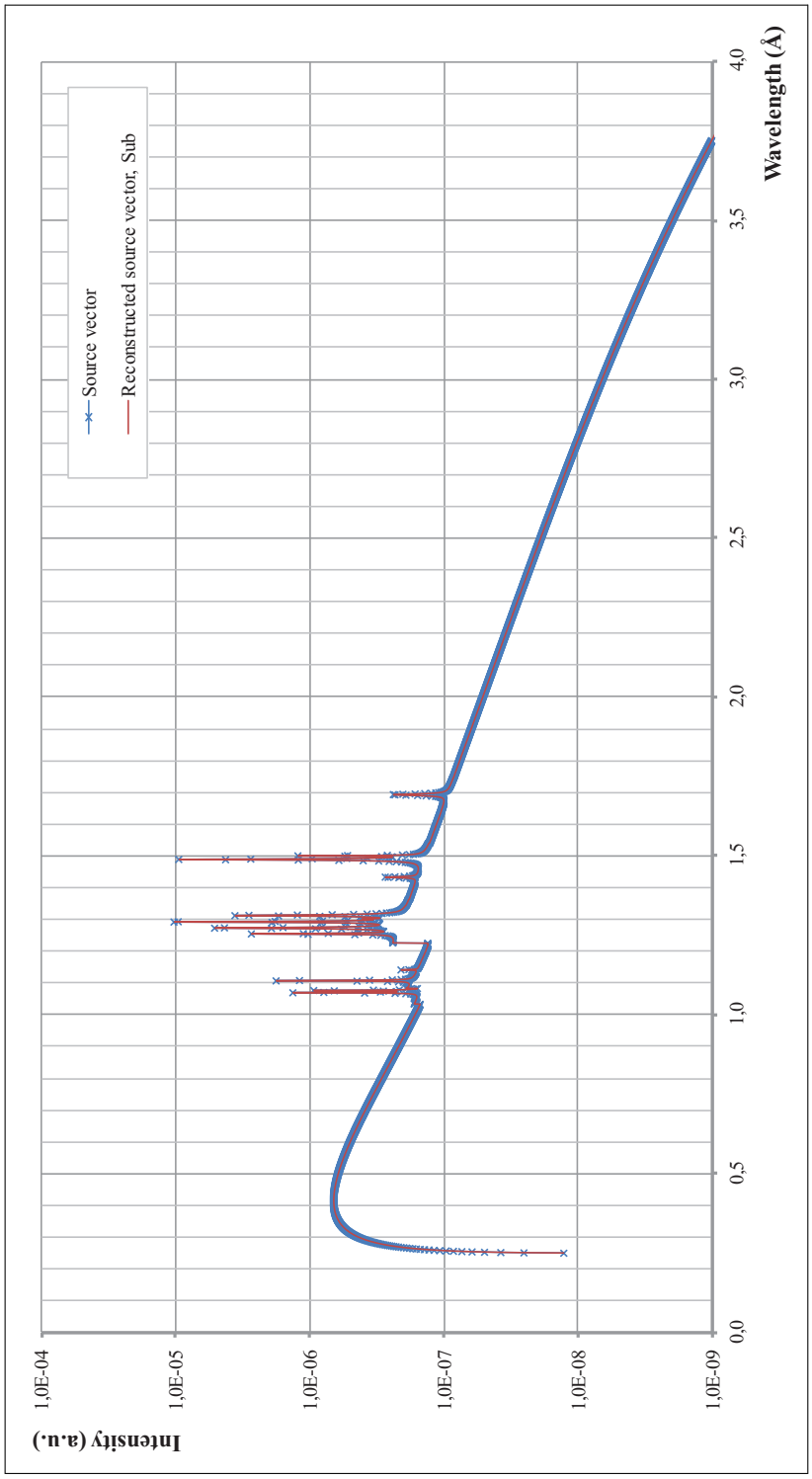

Figure C.3: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. The system has not been preconditioned. Aluminium sample.

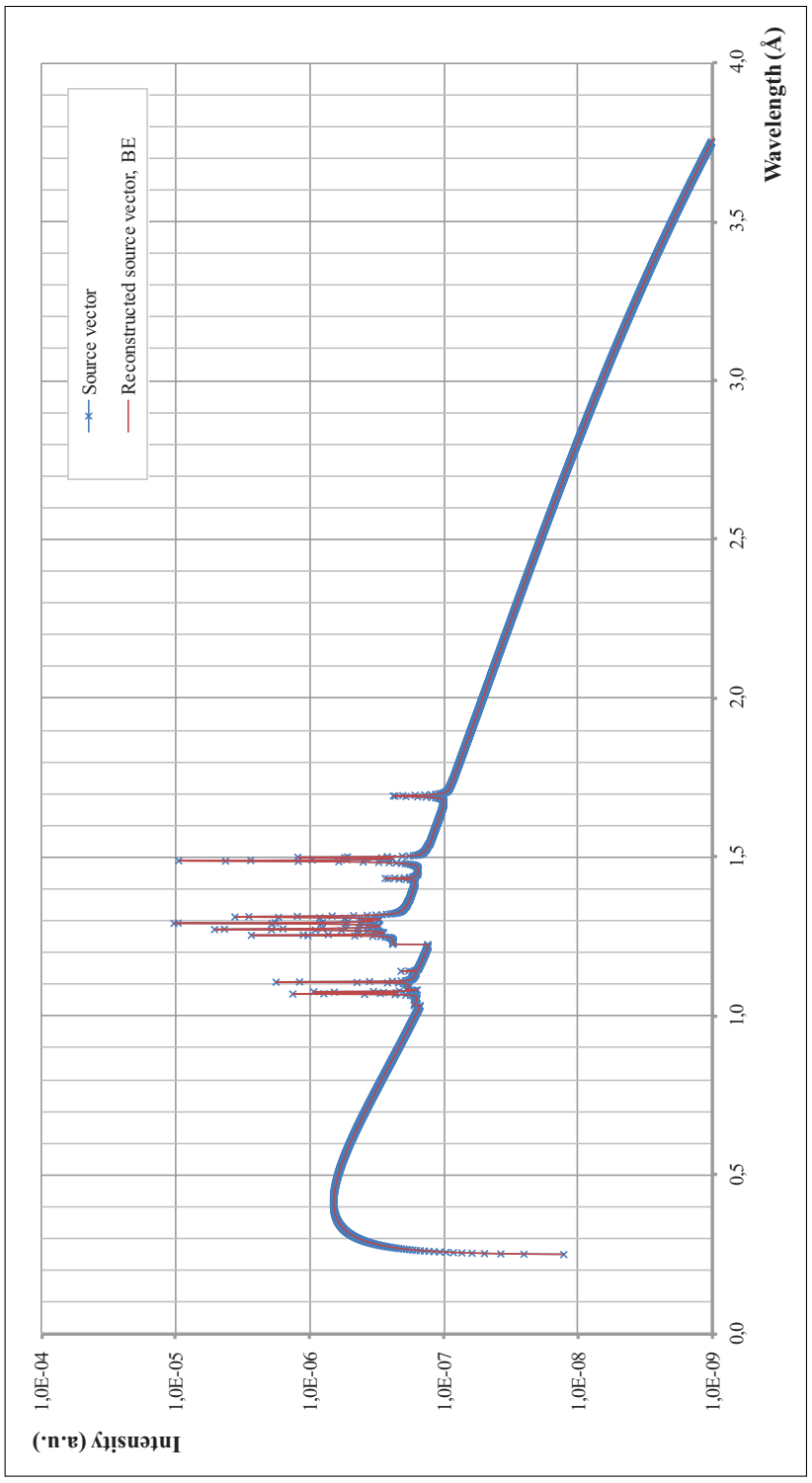

Figure C.4: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, BE}}$. The system has not been preconditioned. Aluminium sample.

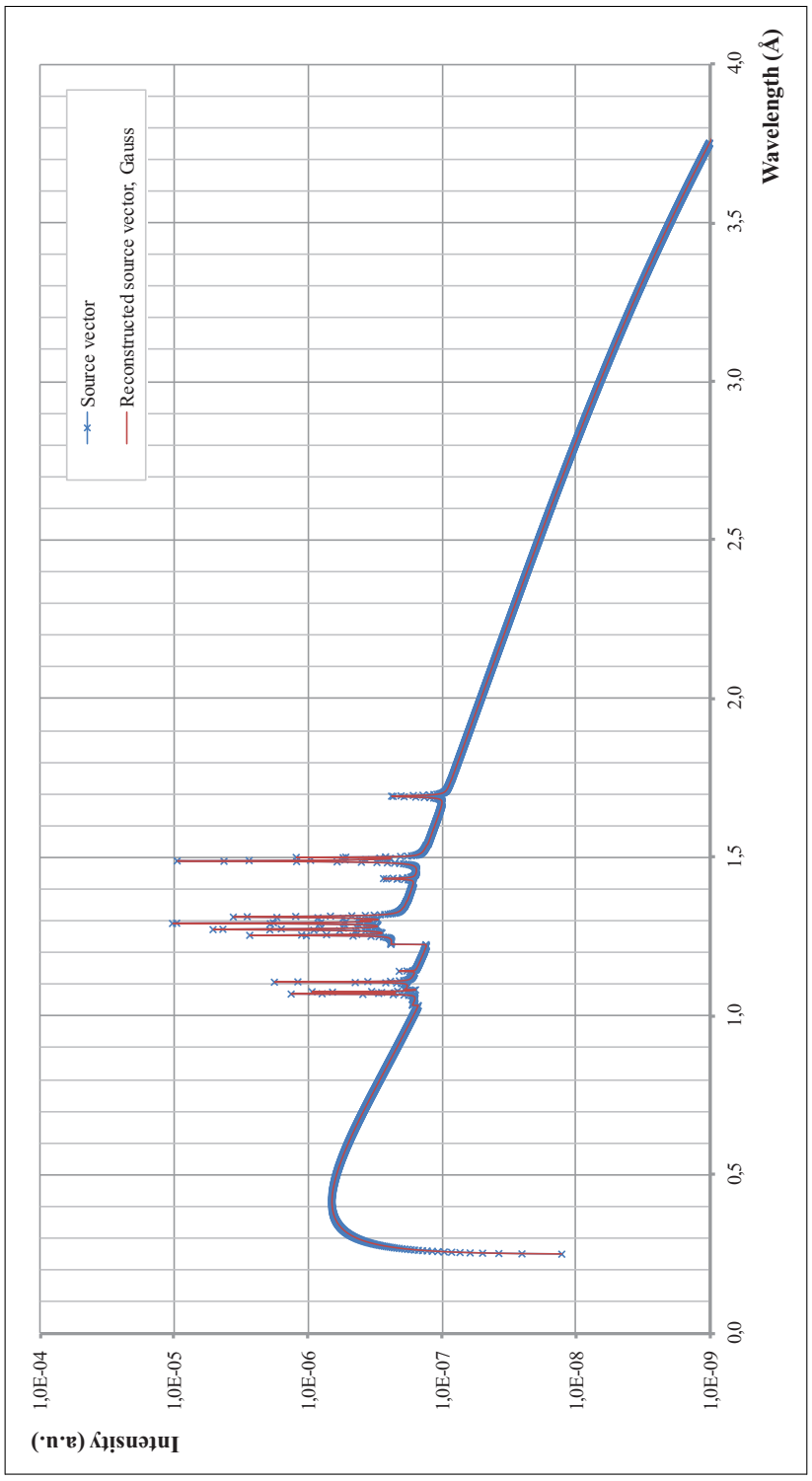

Figure C.5: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, G}}$. The system has not been preconditioned. Aluminium sample.

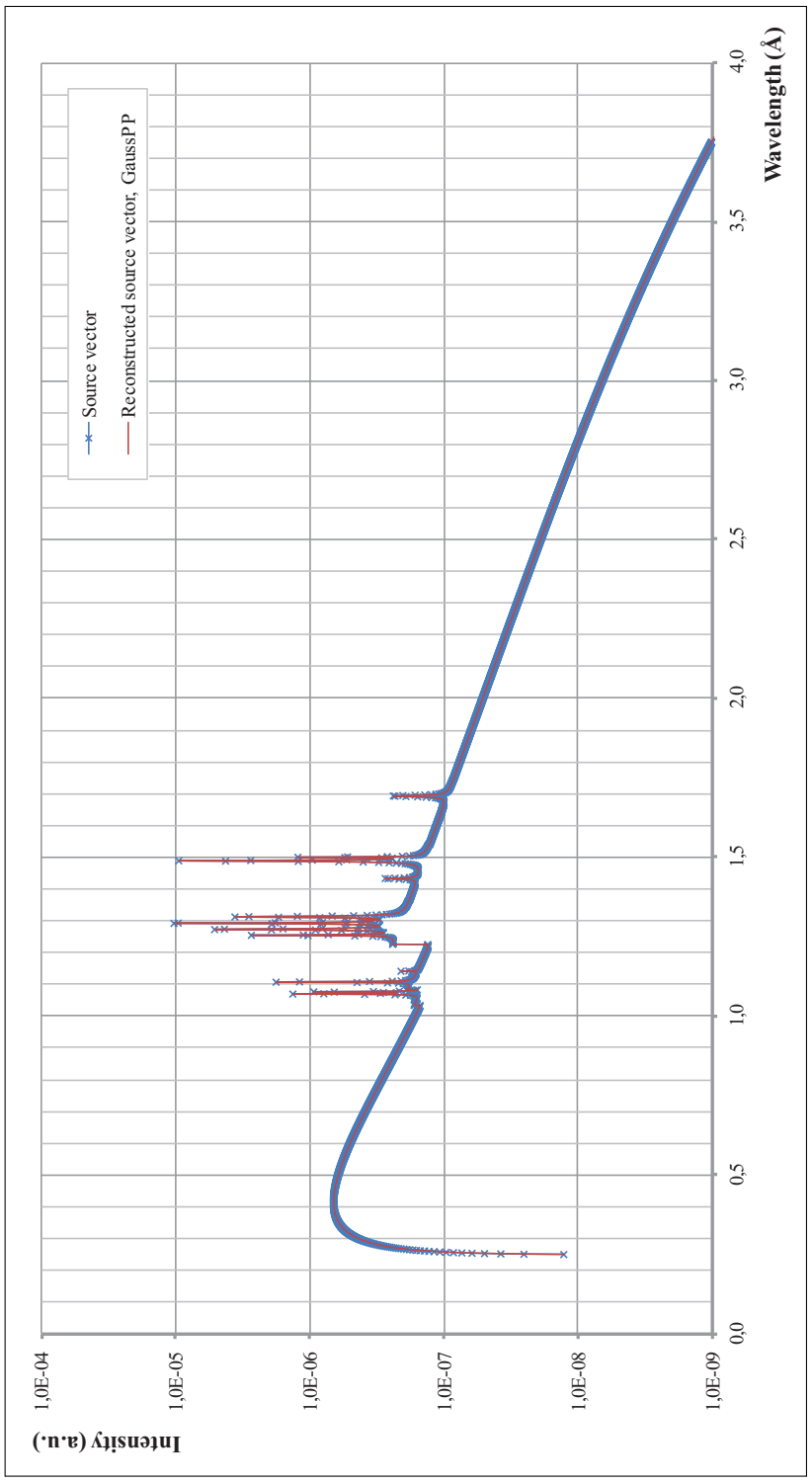

Figure C.6: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Gpp}}$. The system has not been preconditioned. Aluminium sample.

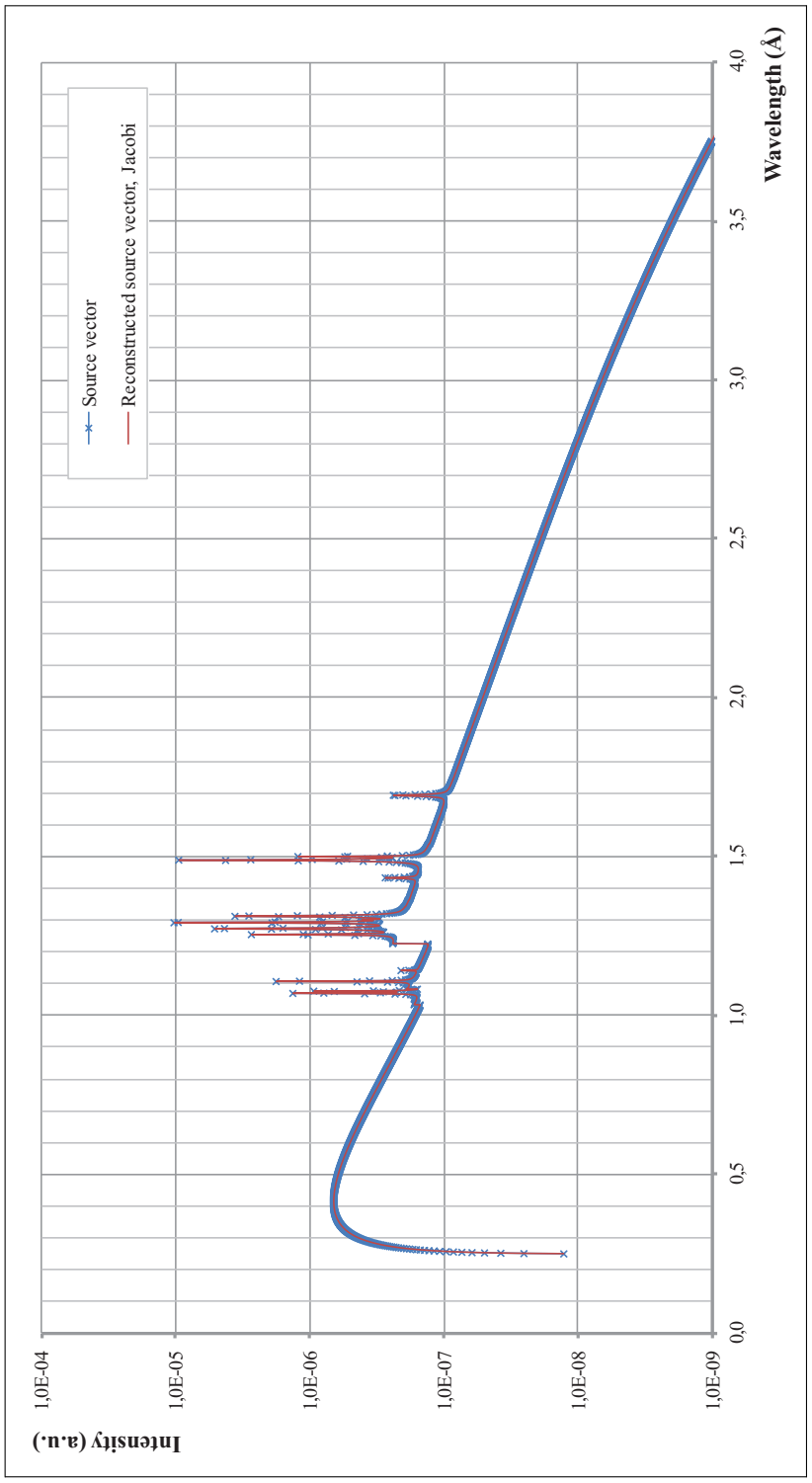

Figure C.7: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, J}}$. The system has not been preconditioned. Aluminium sample.

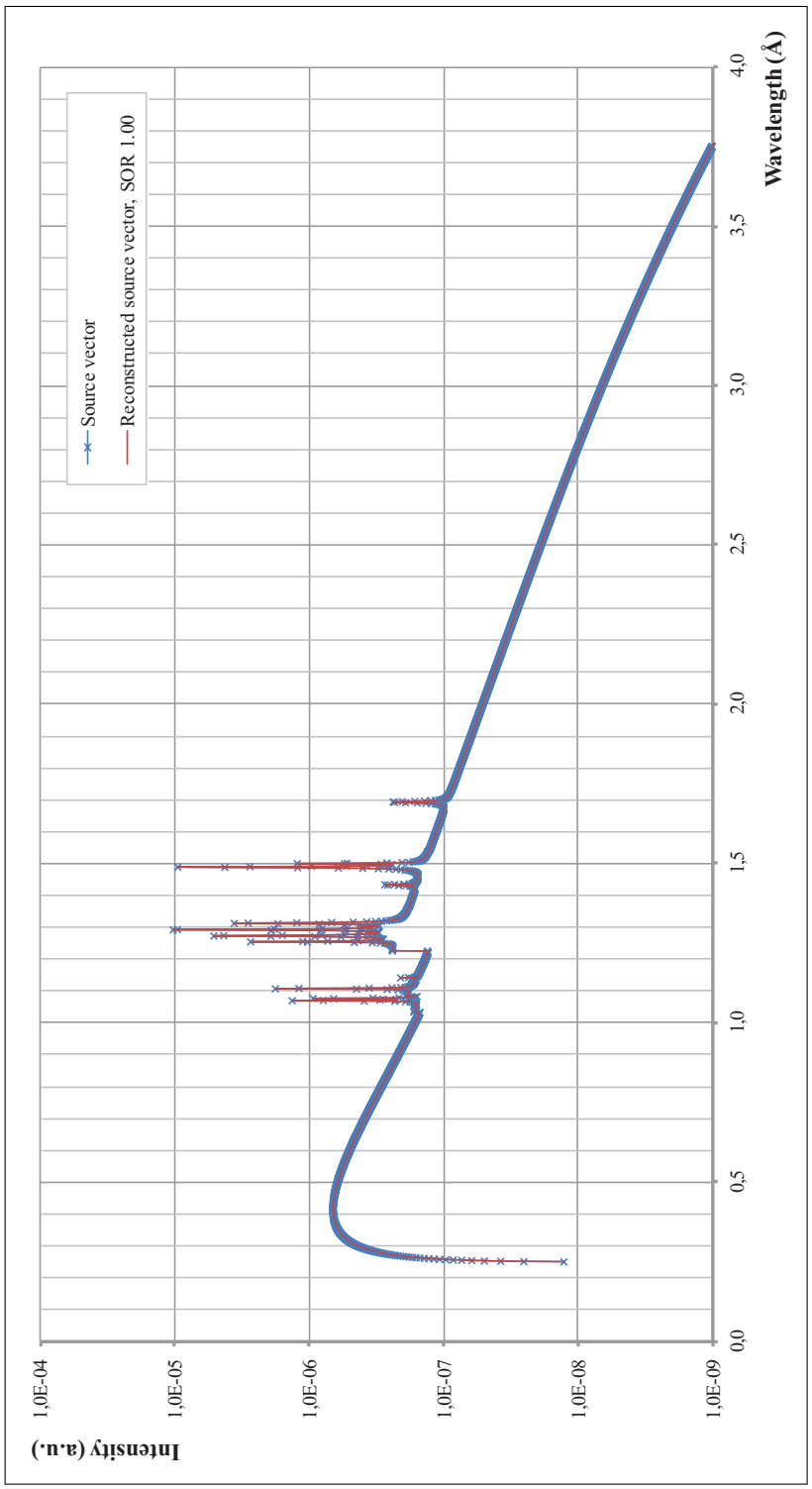

Figure C.8: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SOR}}$. The system has not been preconditioned. Aluminium sample.

Right preconditioned aluminium system

In this part of the appendixes, the figures of the reconstructed vectors obtained with the different numerical methods are given for the right preconditioned aluminium matrix system. The figures correspond to the vectors reported in Table 8.9, p. 111.

Vector	Material	Page
$\overrightarrow{s}_{\mathrm{rec,\,SVD}}$	aluminium	p. 215
$\overrightarrow{s}_{\mathrm{rec,\,Chol}}$	aluminium	p. 216
$\overrightarrow{s}_{\mathrm{rec,\,LU}}$	aluminium	p. 217
$\overrightarrow{s}_{\mathrm{rec,\,Sub}}$	aluminium	p. 218
$\overrightarrow{s}_{\mathrm{rec,\,G}}$	aluminium	p. 219
$\overrightarrow{s}_{\mathrm{rec,\,Gpp}}$	aluminium	p. 220
$\overrightarrow{s}_{\mathrm{rec,\,J}}$	aluminium	p. 221
$\overrightarrow{s}_{\mathrm{rec,\,SOR}}$	aluminium	p. 222

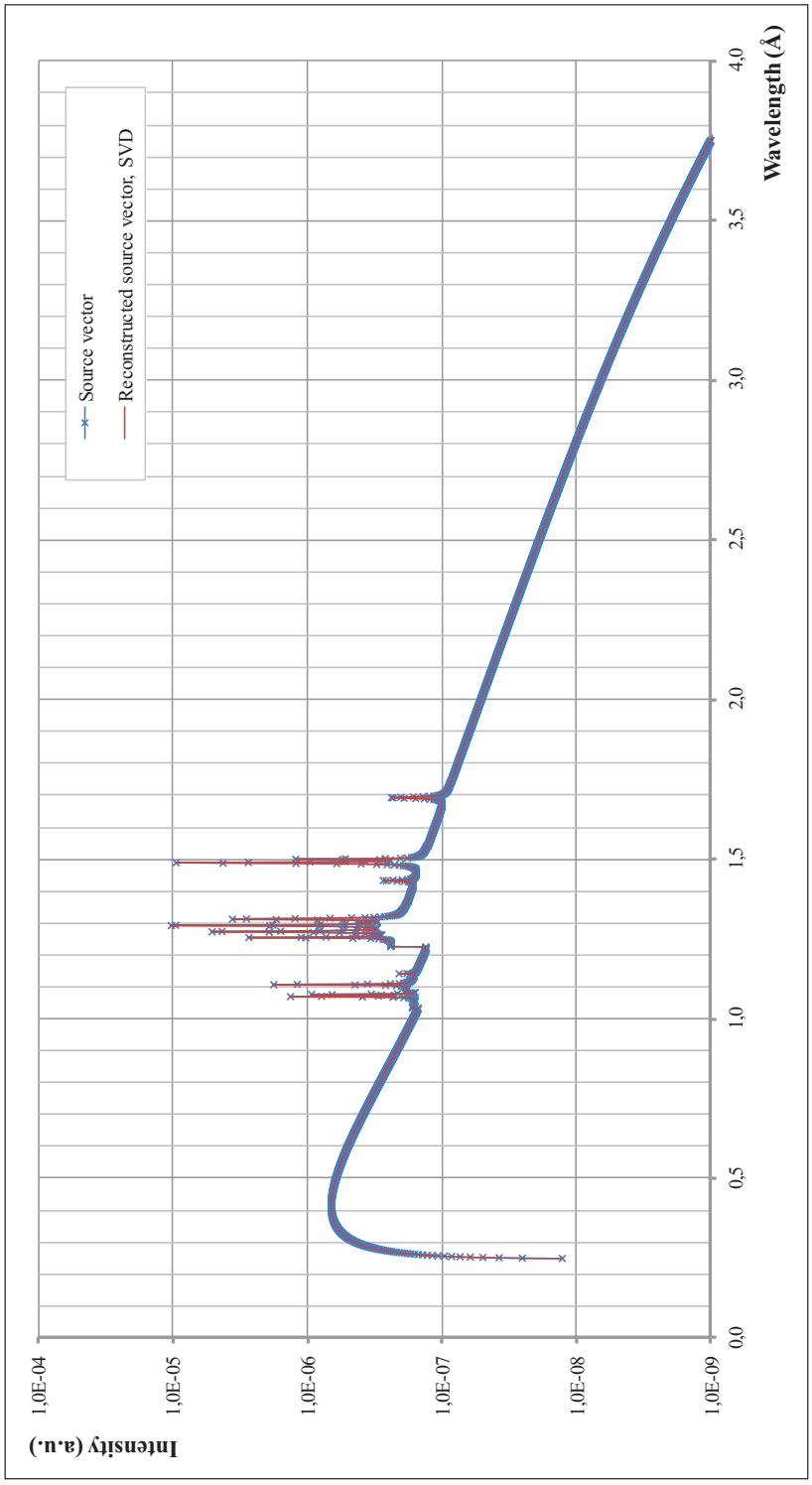

Figure C.9: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SVD}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

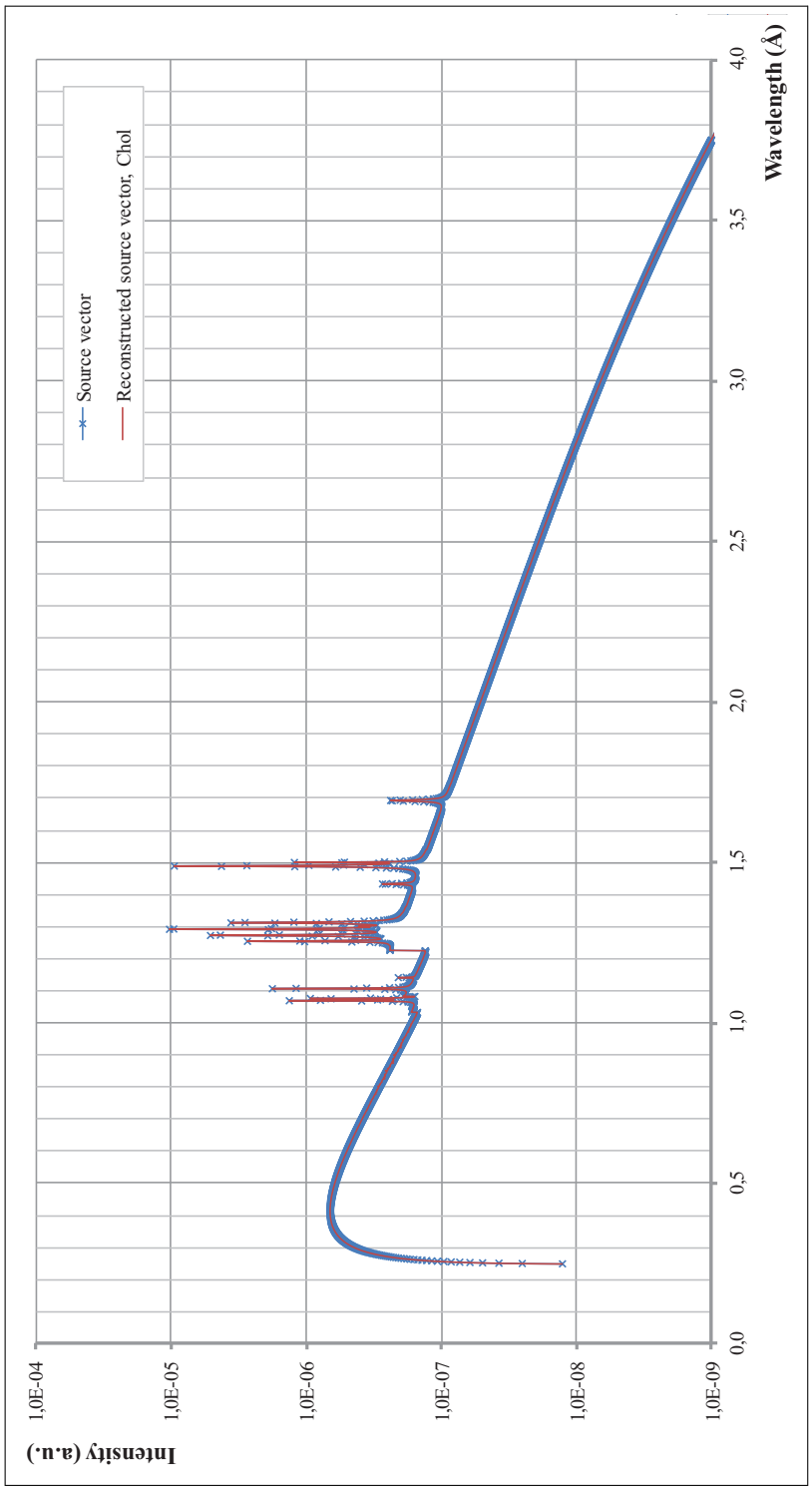

Figure C.10: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Chol}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

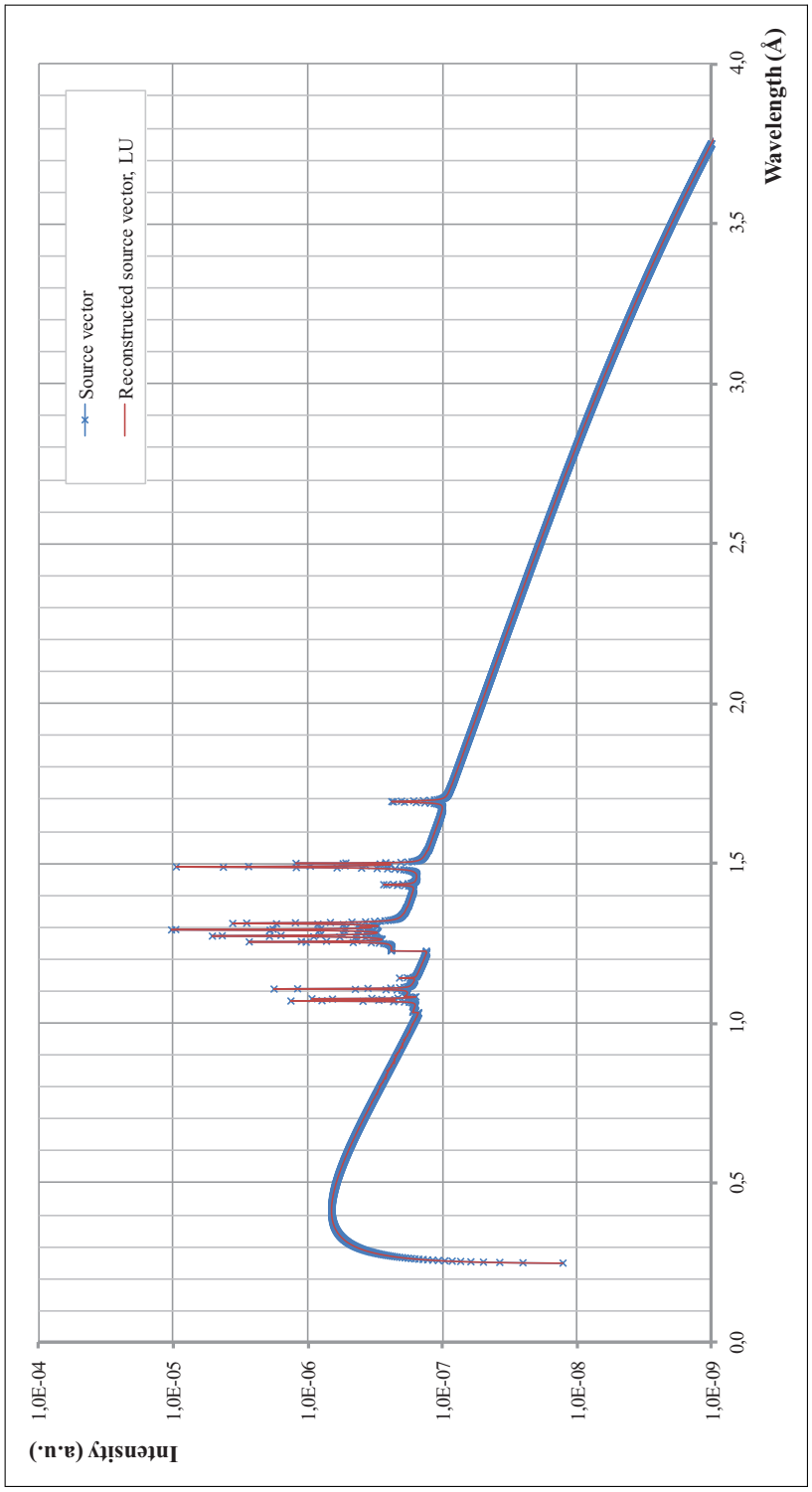

Figure C.11: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, LU}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

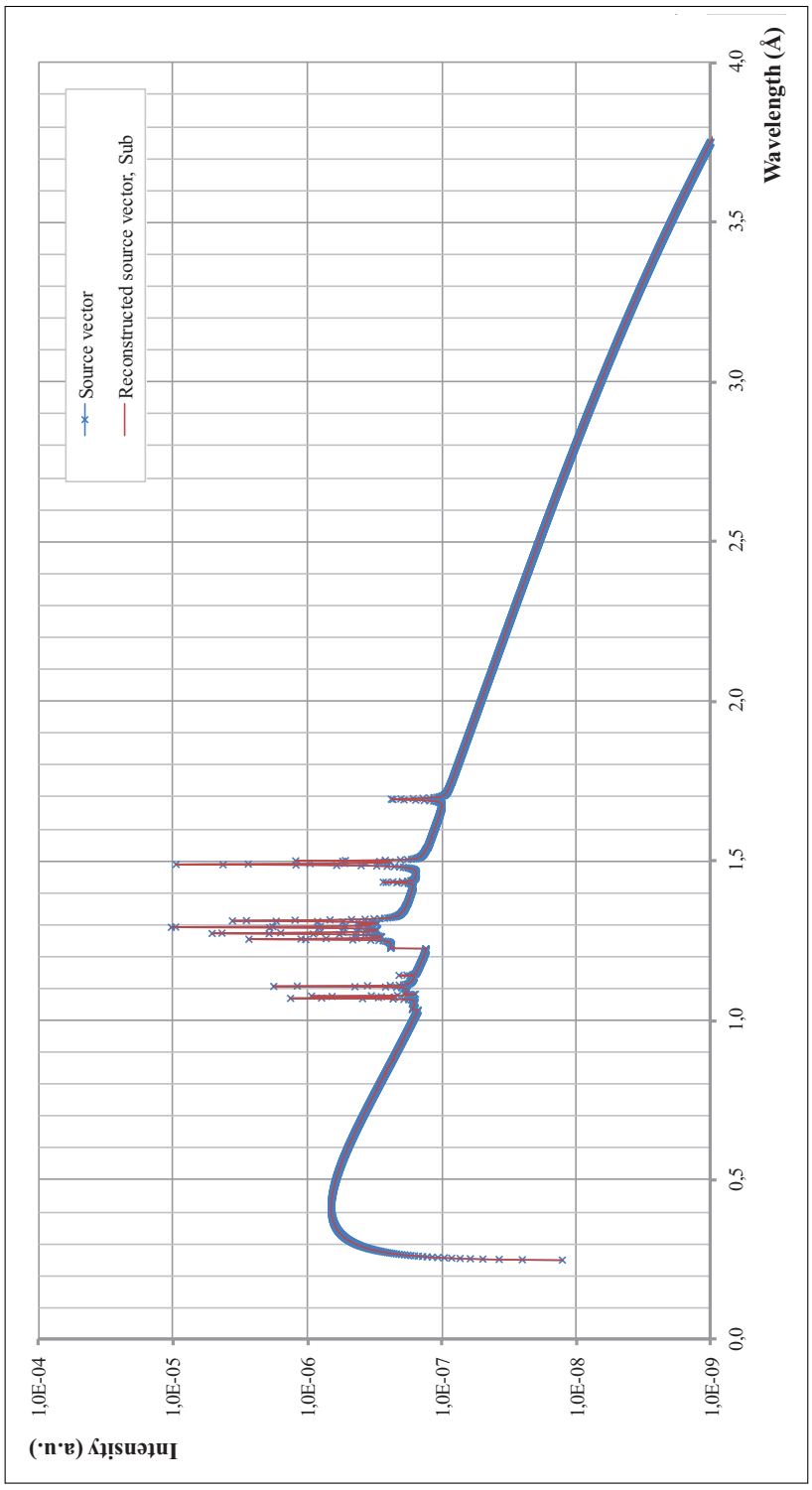

Figure C.12: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

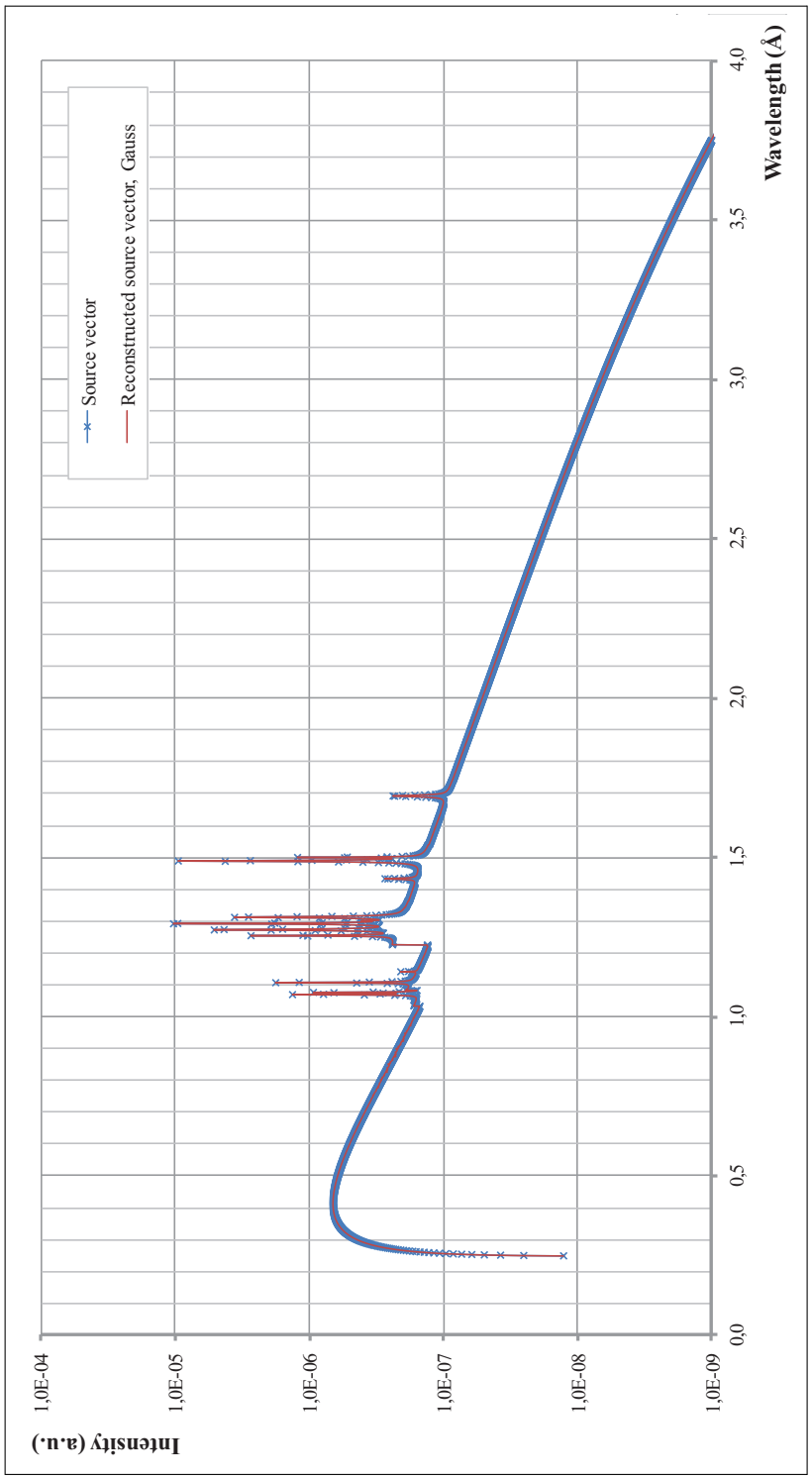

Figure C.13: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, G}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

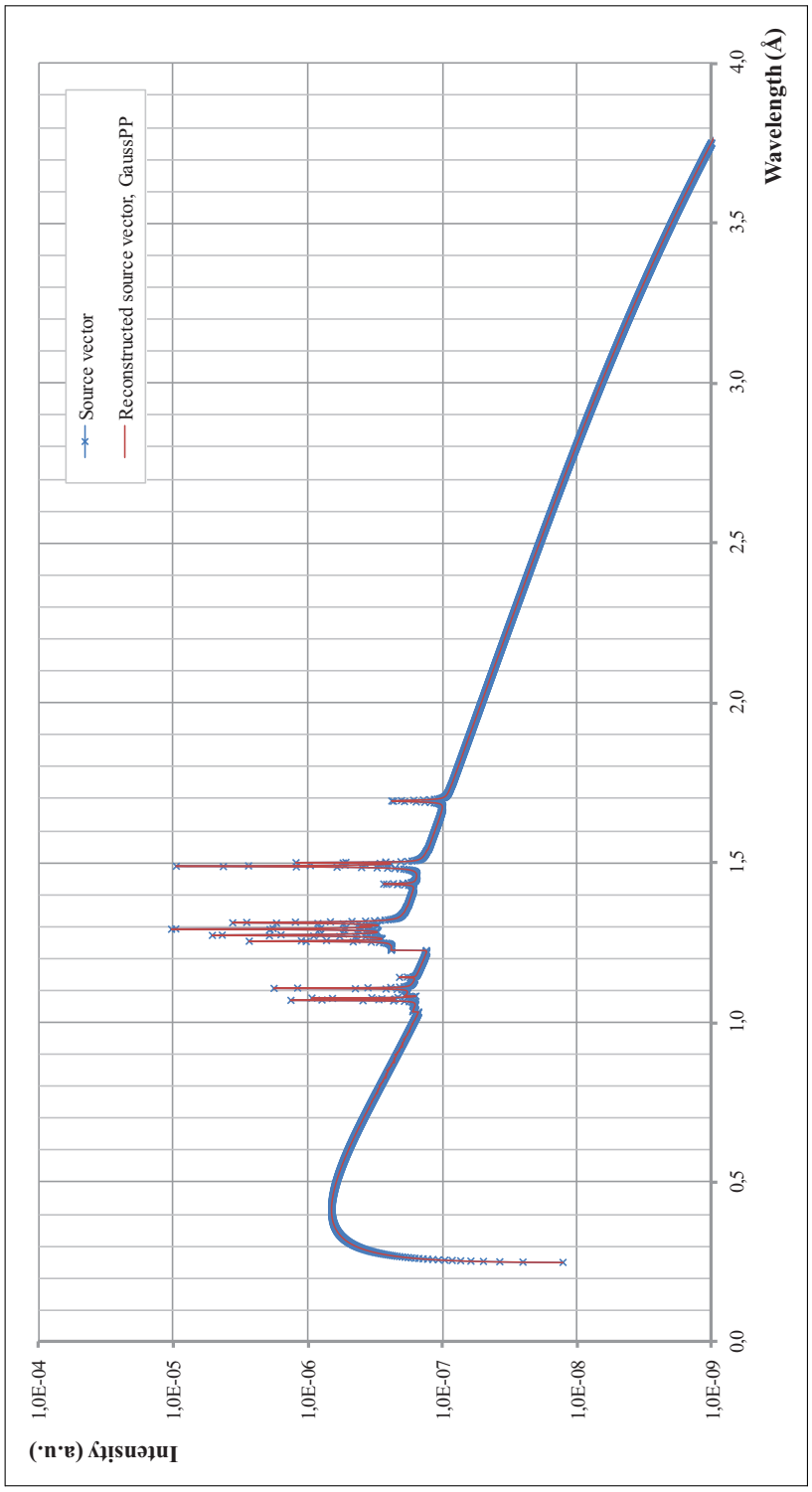

Figure C.14: Comparison between the source spectrum $\vec{s}$ and the reconstructed source vector $\vec{s}_{\text{rec, Gpp}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

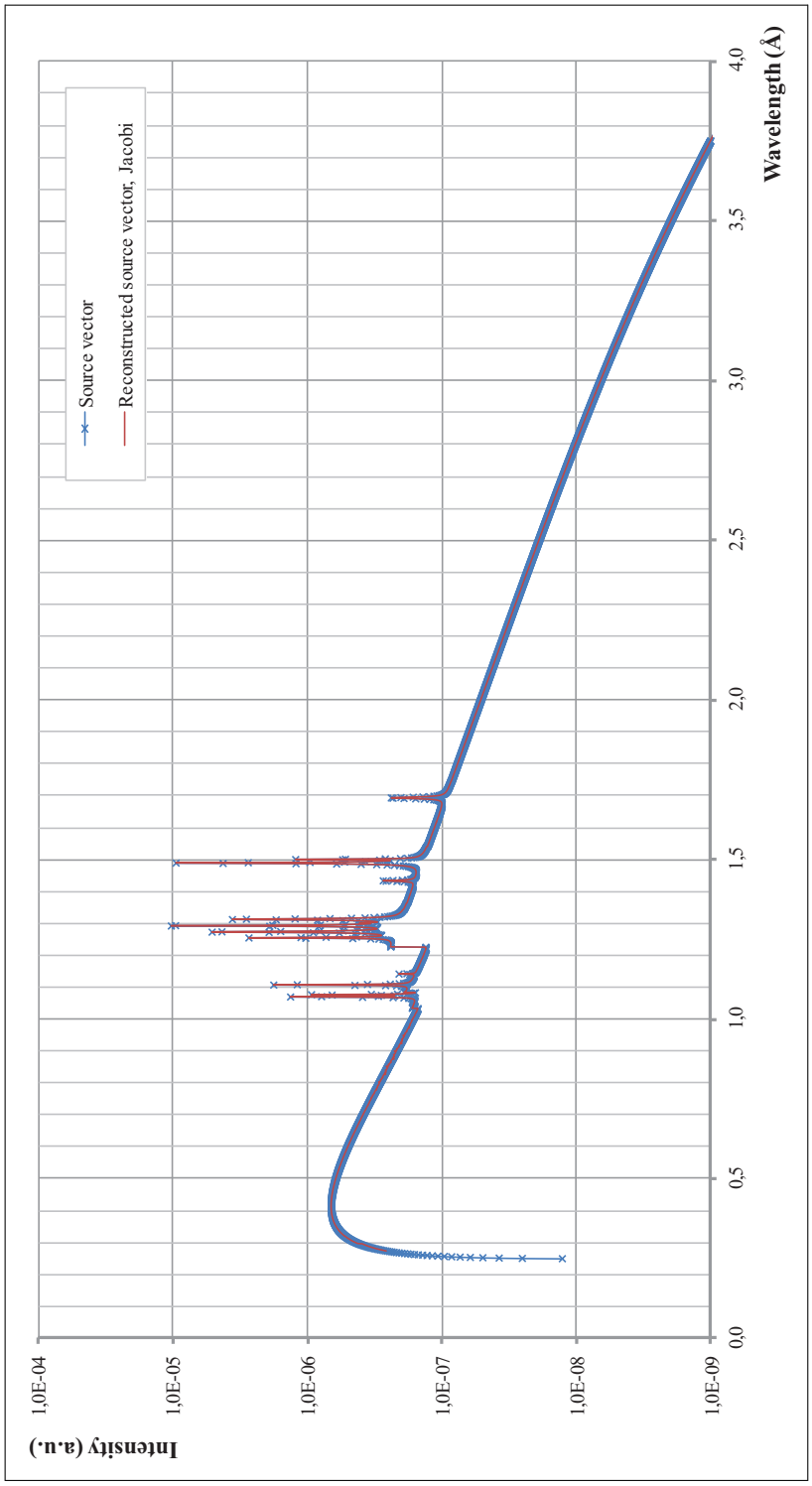

Figure C.15: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, J}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

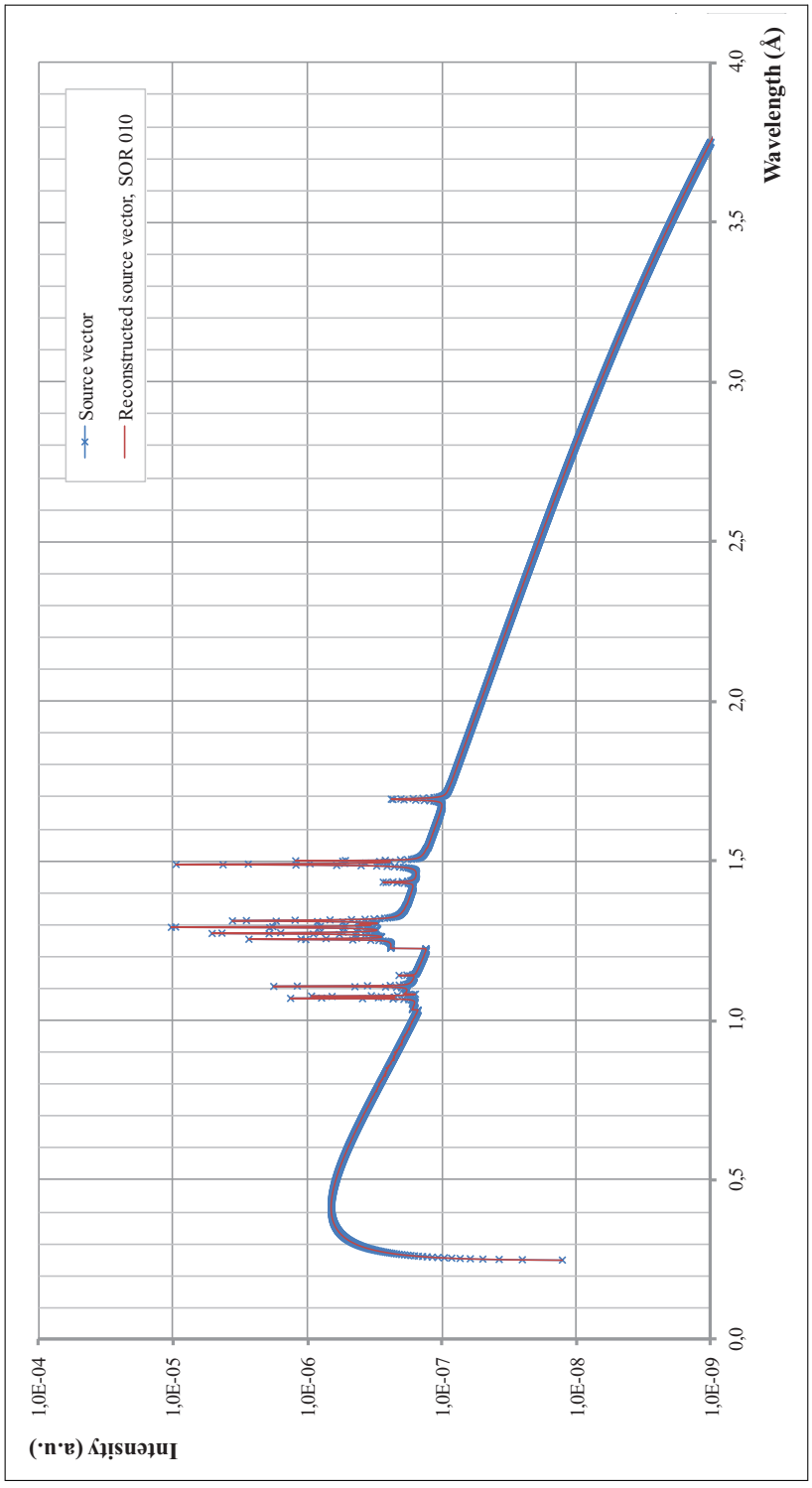

Figure C.16: Comparison between the source spectrum $\vec{s}$ and the reconstructed source vector $\vec{s}_{\text{rec, SOR}}$. The system is right preconditioned by the adjoint matrix. Aluminium sample.

Left preconditioned aluminium system

In this part of the appendixes, the figures of the reconstructed vectors obtained with the different numerical methods are given for the left preconditioned aluminium matrix system. The figures correspond to the vectors reported in Table 8.10, p. 113.

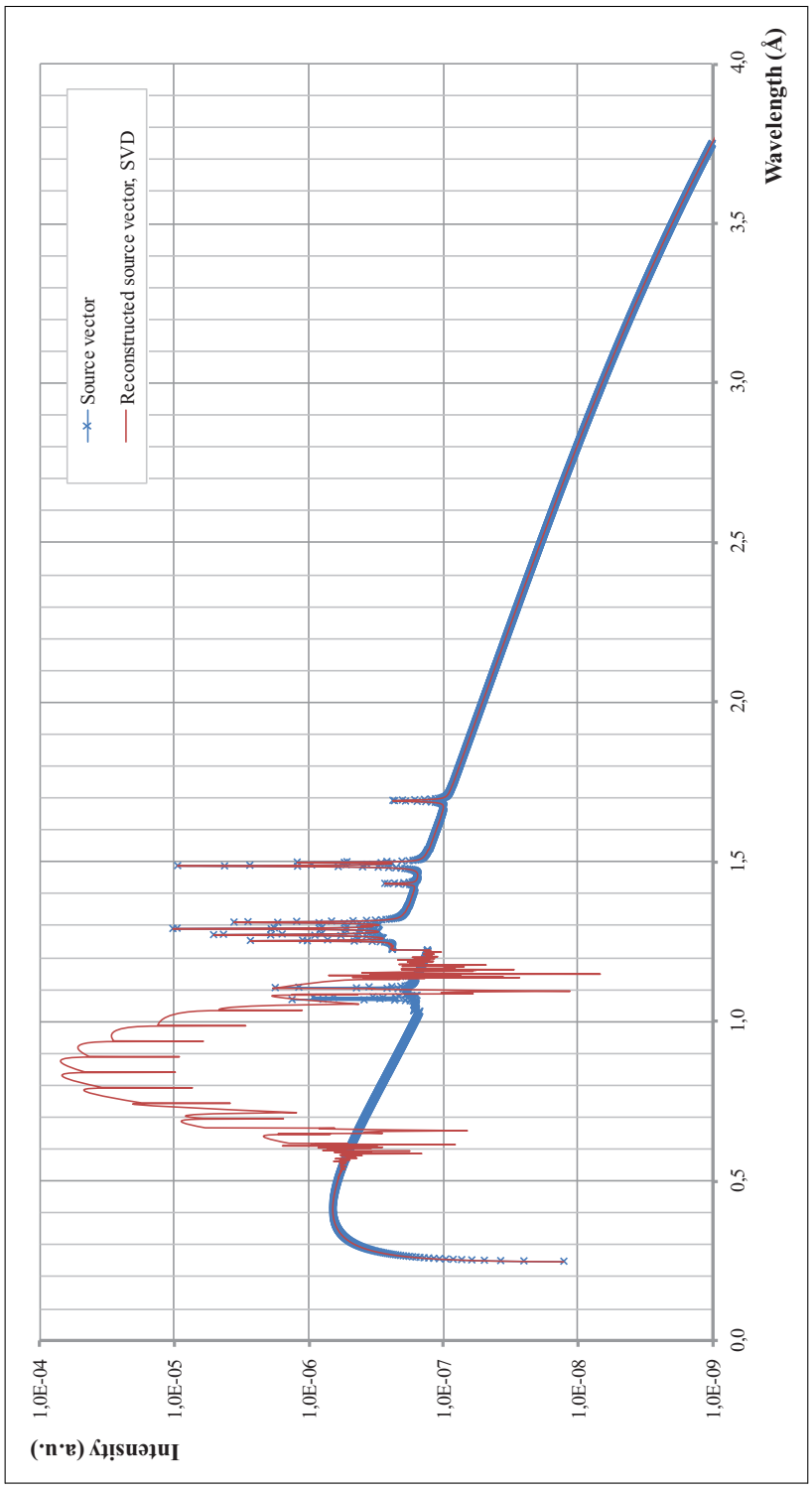

Figure C.17: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SVD}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

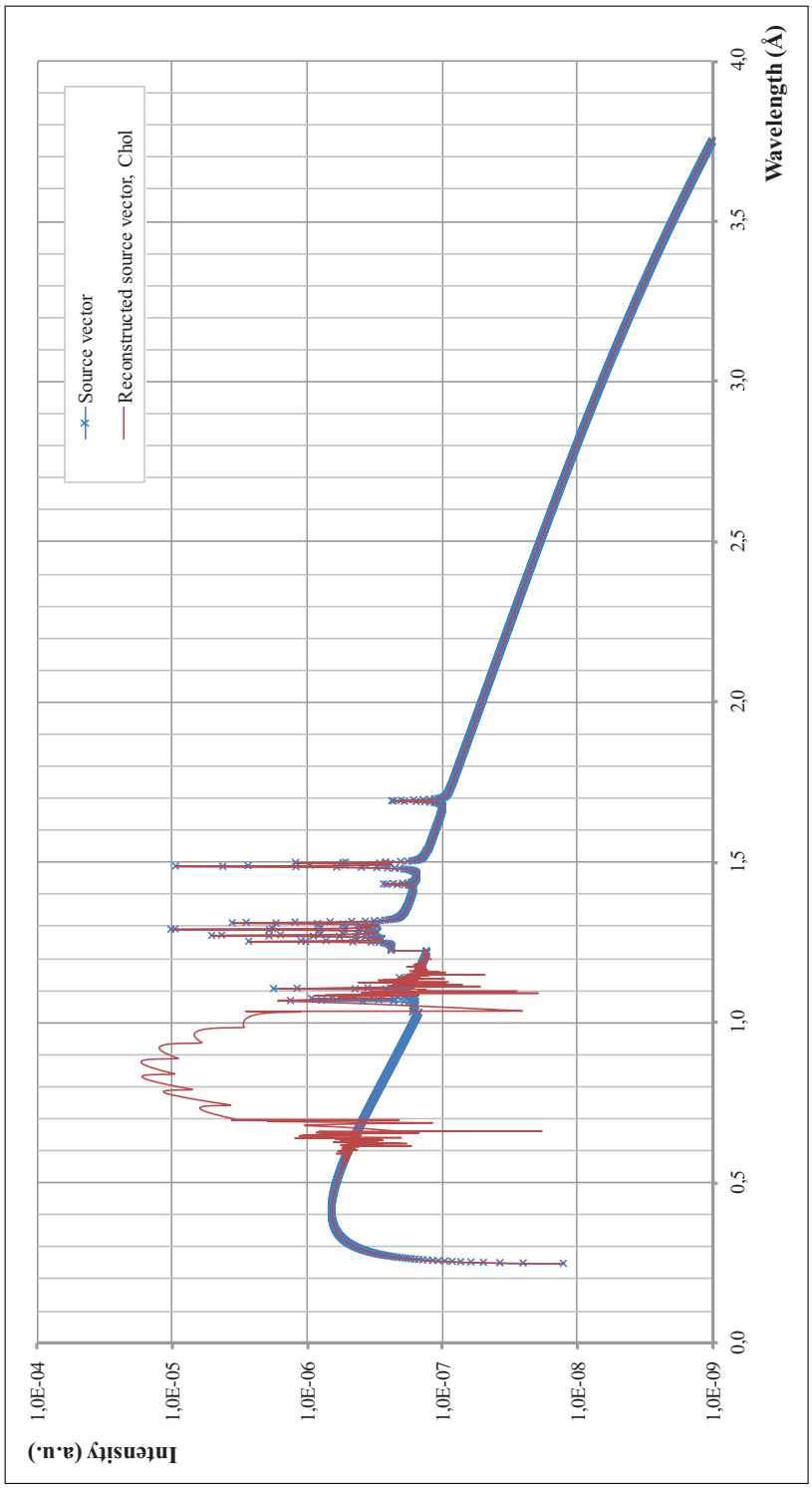

Figure C.18: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Chol}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

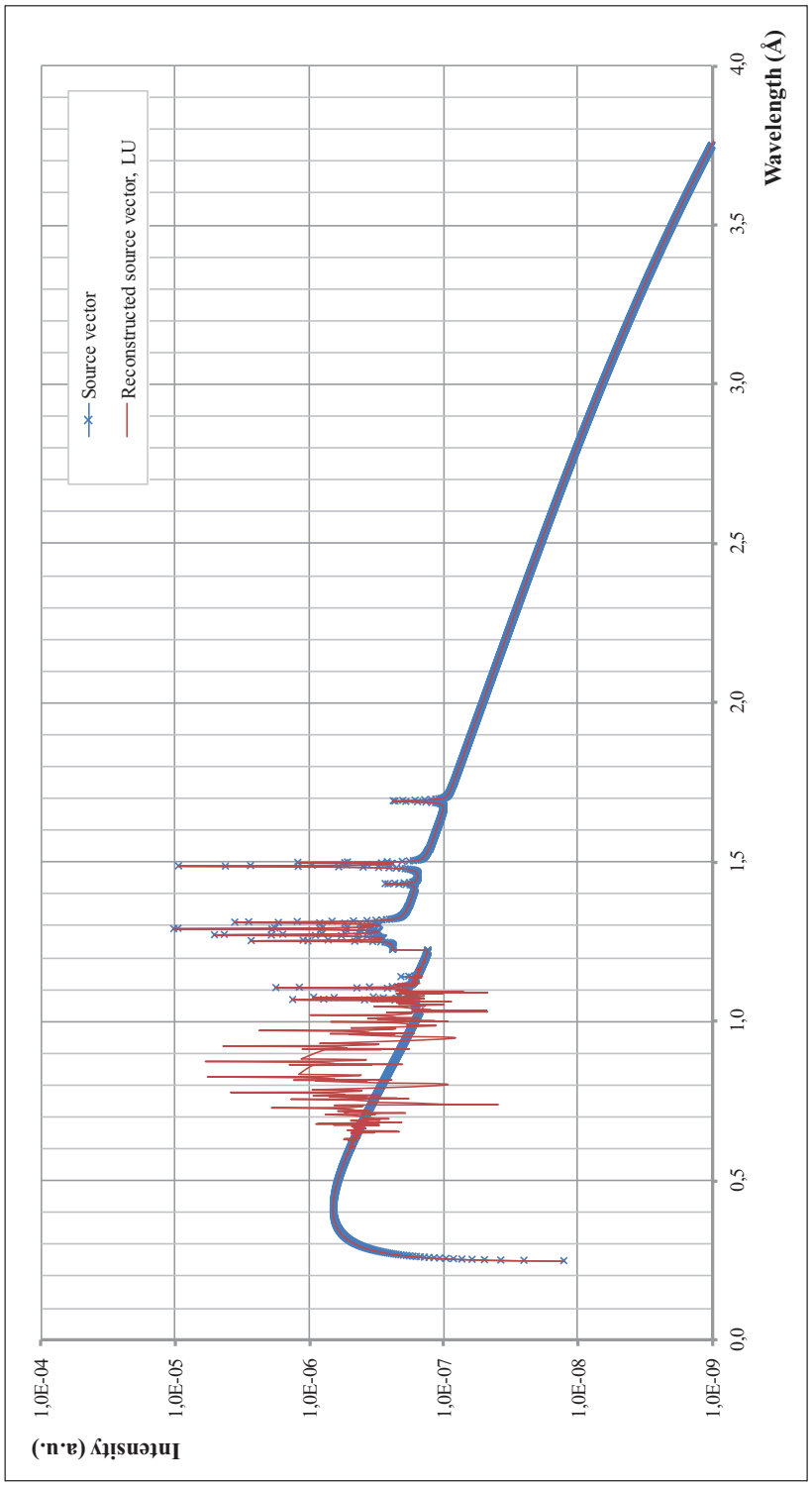

Figure C.19: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, LU}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

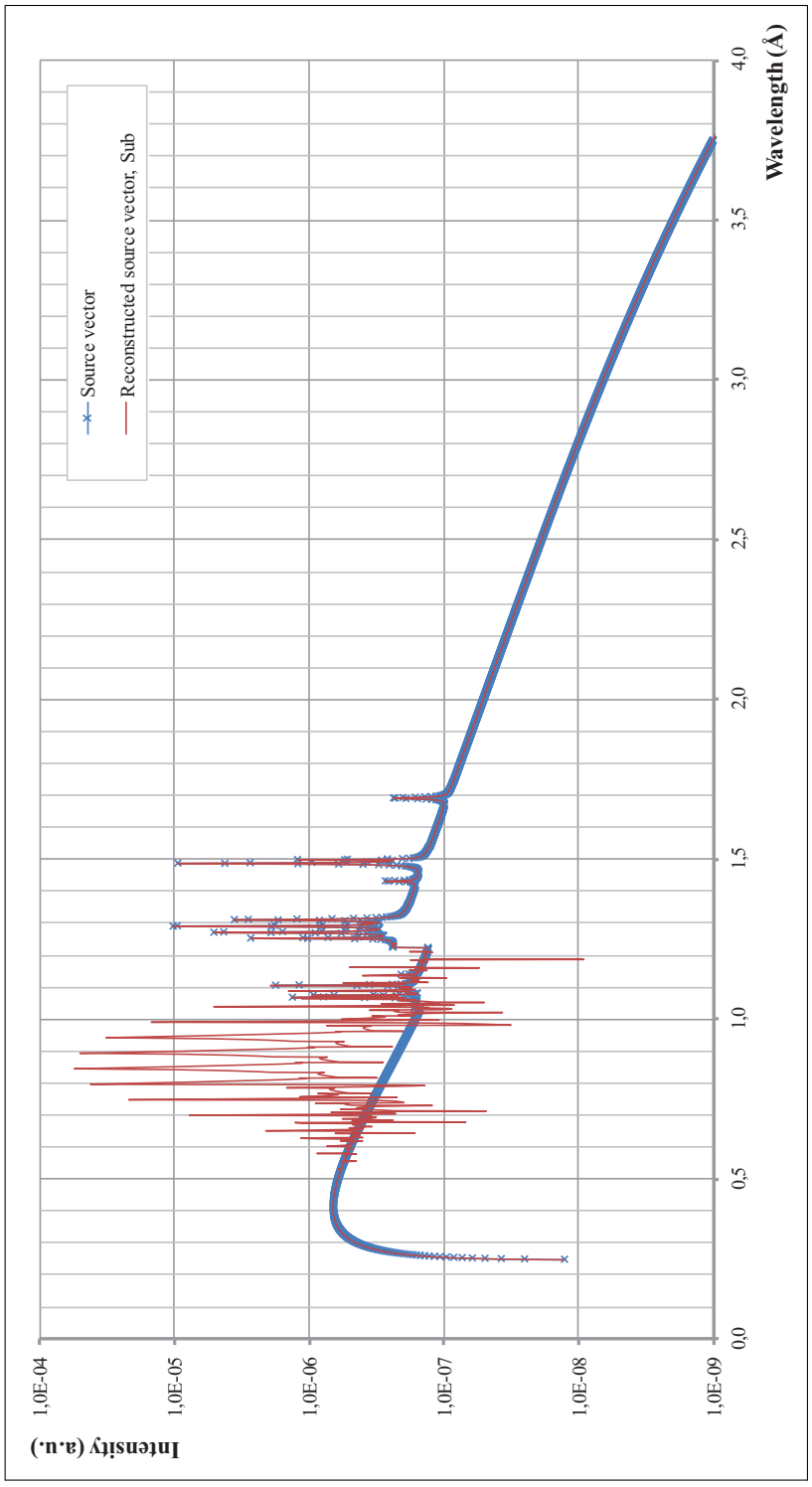

Figure C.20: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Sub}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

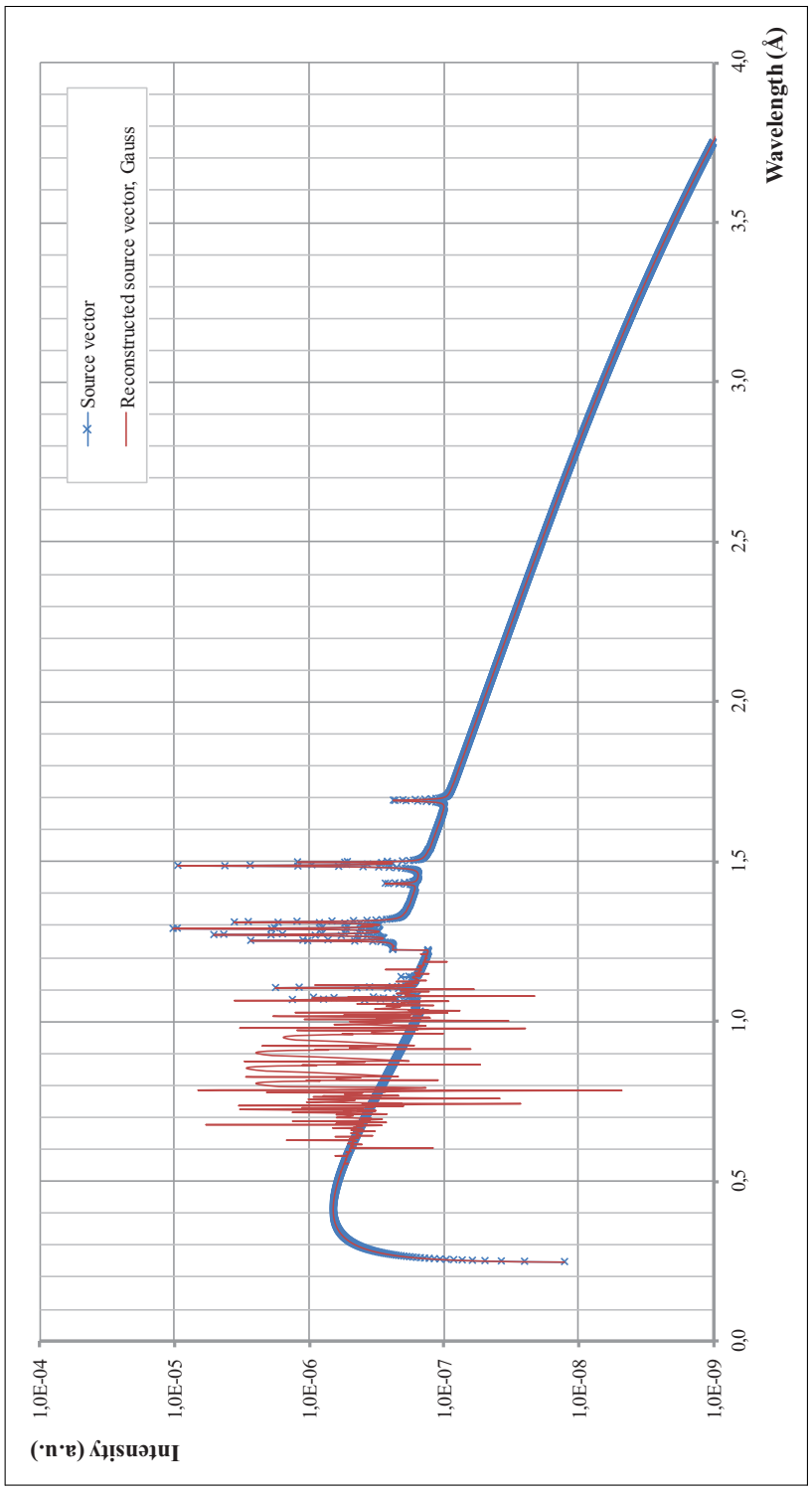

Figure C.21: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, G}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

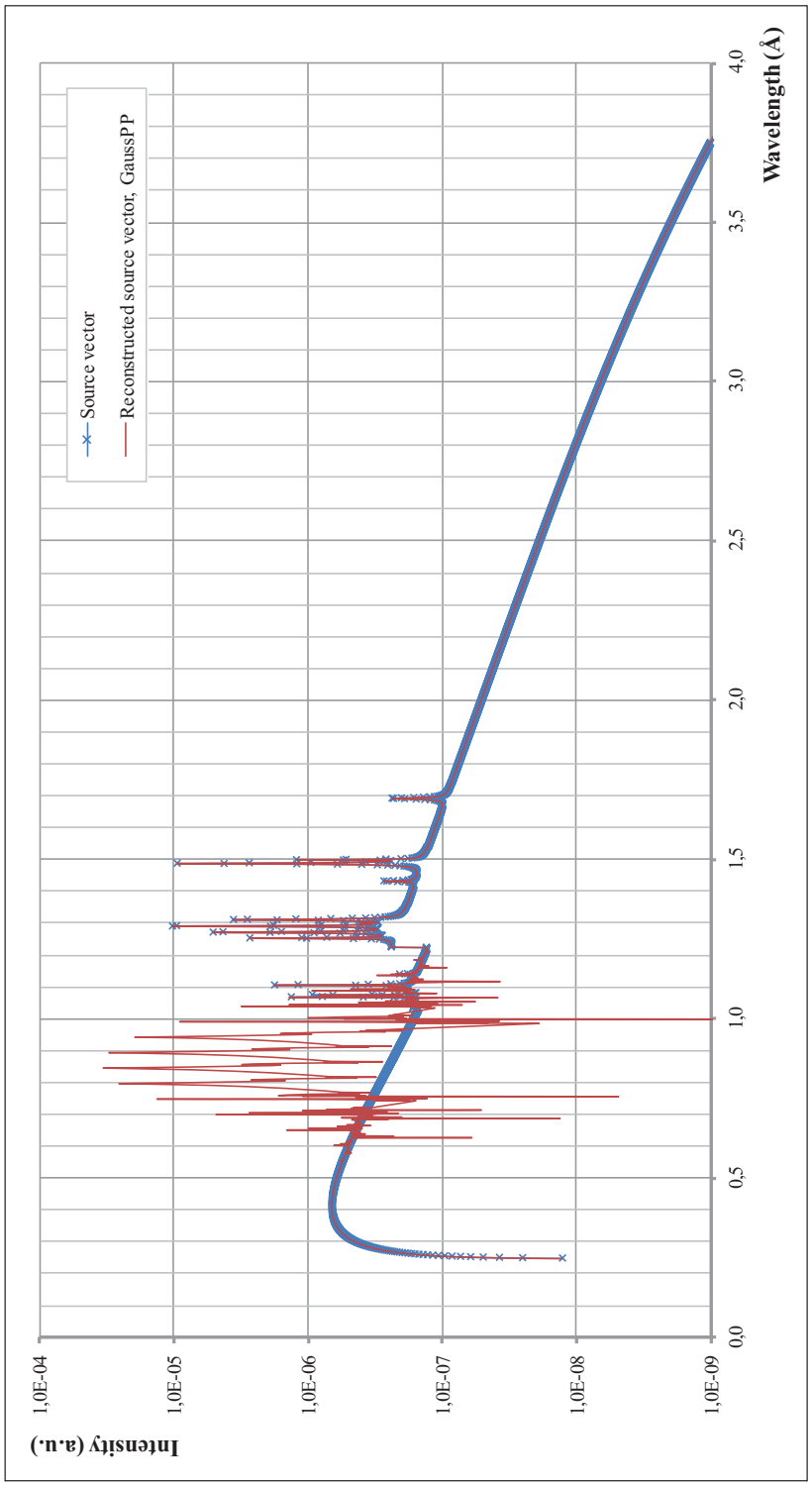

Figure C.22: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, Gpp}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

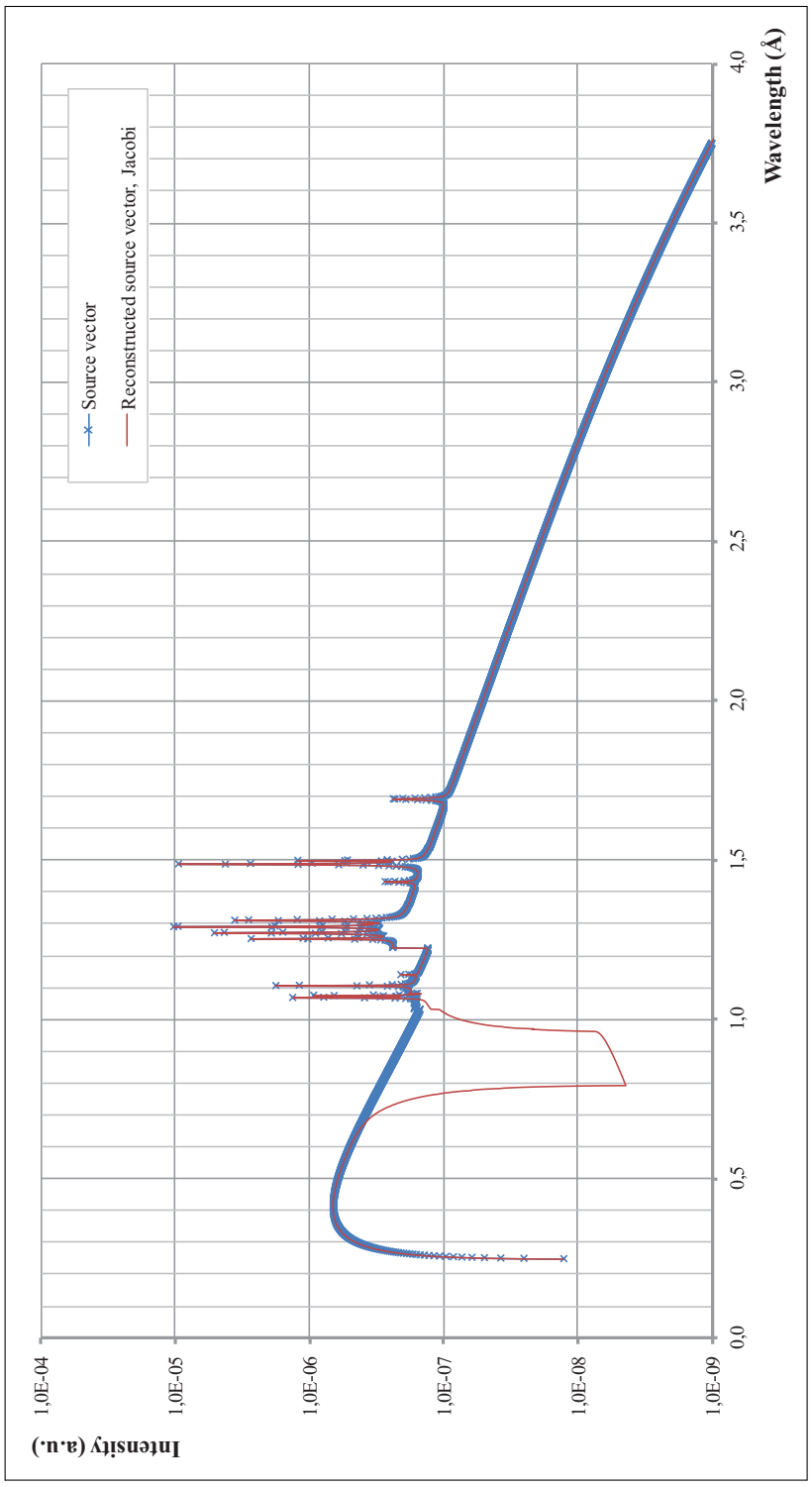

Figure C.23: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, J}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.

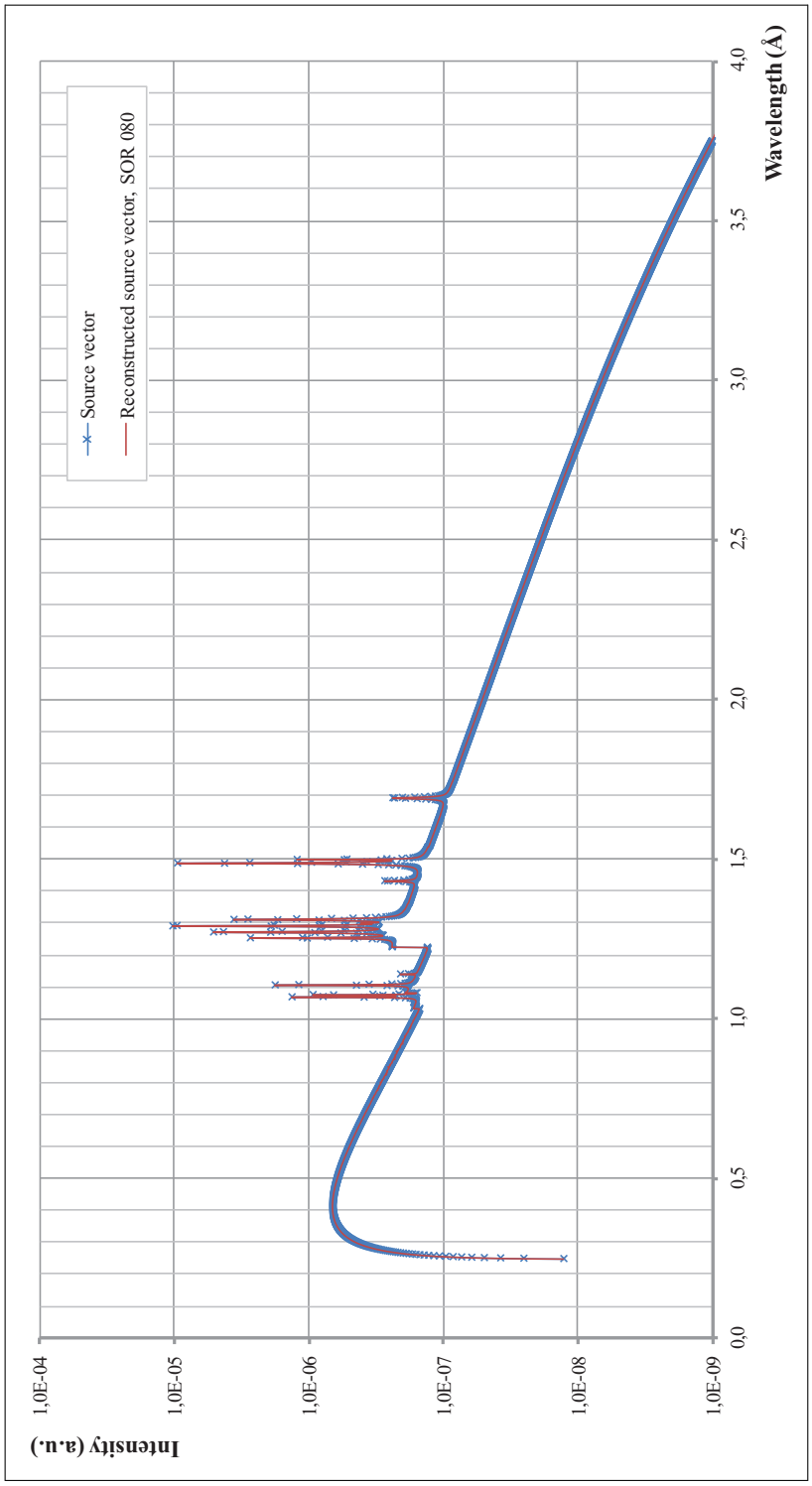

Figure C.24: Comparison between the source spectrum $\overrightarrow{s}$ and the reconstructed source vector $\overrightarrow{s}_{\text{rec, SOR}}$. The system is left preconditioned by the adjoint matrix. Aluminium sample.