Robust Multivariate Analysis Methods for Single Cell Raman Spectroscopy

by

Nikita Kuklev
B.Sc., University of Victoria, 2012

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Physics and Astronomy

Robust Multivariate Analysis Methods for Single Cell Raman Spectroscopy

by

Nikita Kuklev
B.Sc., University of Victoria, 2012

Supervisory Committee

Dr. Andrew Jirasek, Supervisor
(Department of Physics and Astronomy)

Dr. Alexander Brolo, Outside Member
(Department of Chemistry)

Supervisory Committee

Dr. Andrew Jirasek, Supervisor
(Department of Physics and Astronomy)

Dr. Alexander Brolo, Outside Member
(Department of Chemistry)

ABSTRACT

Usefulness of a particular clinical assay is directly correlated with its ability to extract highest possible signal from available data. This is particularly relevant for personalized radiation therapy since early plan modifications confer greater benefits to treatment outcome. Recent studies have demonstrated capability of single-cell Raman microscopy to detect cellular radiation response at clinical (below 10Gy) doses, but only in certain strongly responding cell lines and after at least two day incubation. One possible cause is rather unoptimized signal processing used. This work investigates application of several advanced multivariate methods - weighted principal component analysis (WPCA), robust PCA, probabilistic PCA, and nonlinear PCA to increase radiation response signal. Representative datasets from strongly (H460 - human lung) and weakly (LNCaP - human prostate) responding cell lines were analysed in 0-50Gy and 0-10Gy dose ranges and results quantified to determine relative and absolute algorithm performance. It was found that with careful tuning, significant improvements in sensitivity and better signal separation could be achieved as compared to conventional PCA.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank first and foremost my supervisor, Dr. Andrew Jirasek, for the extreme patience and expert guidance in performing this research. He went above and beyond to get me through this, and no words can adequately express my gratitude.

Thank you to Dr. Quinn Matthews and (soon to be Dr.) Samantha Harder for their mentorship, and for laying the groundwork for my research.

I am also very grateful to my other committee member, Dr. Alexander Brolo, as well as many of the excellent researchers at both UVic and BC Cancer Agency Vancouver Island Center for their constructive criticism and suggestions.

Finally, I would like to thank National Science and Engineering Research Council of Canada for their financial support.

Chapter 1

Introduction

This chapter introduces topics of radiation therapy and radiobiology (sections 1.1-1.2) with emphasis on modern approaches and limitations, demonstrating the need for accurate radiosensitivity assays to improve treatment outcomes. Section 1.3 presents Raman spectroscopy as a possible candidate and compares it with available alternatives. Chapter concludes with brief description of thesis scope in section 1.4.

1.1 Radiation therapy

Cancer is one of the leading causes of morbidity and mortality worldwide, accounting for approximately 13% of all deaths [1]. Biologically, it is a result of cell growth regulation failure leading to impairment of normal bodily functions. A wide variety of treatments such as chemotherapy, surgical removal, and radiotherapy (RT) are available, with efficacy varying based on site and type of tumour [2].

In general, radiotherapy is prescribed to approximately a third [3] of all patients. It has a number of advantages - high cost efficiency [4, 5] (for both patient and health care system), non-invasive nature, quick outpatient sessions, ability to use it as adjunct to other treatments, and palliative regimens, among others [6]. The main disadvantage is damage to normal tissues and associated side effects [7].

The primary goal of RT is to eliminate or slow down tumour growth while minimizing side effects. Majority of RT is done via external beam radiation therapy (EBRT), where an external ionizing source is used to deliver prescribed dose (commonly fractionated at 2Gy [8] to allow for normal tissue recovery) into a specific volume. While initially Co-60 gamma-ray sources were used with crudely made beam

shaping blocks, current designs utilize linear accelerators (LINACs) and multi-leaf collimators (MLCs), allowing for exquisite control of beam energy, intensity, and shape [9]. Modern treatment methods such as intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) are thus capable of achieving highly conformal dose distributions while near perfectly minimizing damage to surrounding normal tissues [10, 11].

1.2 Radiobiology

RT treatment planning is based largely on knowledge of cellular interactions with ionizing radiation and resulting damage. This field is known as radiobiology, introduced in this section along with a short explanation of typical interactions, cellular structure, and cell cycle.

1.2.1 Ionizing radiation

Ionization is the process of neutral atoms acquiring net charge, positive or negative. As the name suggests, ionizing radiation is a type of radiation energetic enough to create these ions by splitting neutral atoms into ion pairs. Directly ionizing particles such as electrons, protons, and α-particles are capable of ionizing via sufficiently energetic direct collisions [12]. On the other hand, photons and neutrons are called indirectly ionizing particles since they can only produce ions by first liberating directly ionizing particles.

Photon interactions

Once photons enter a medium they lose energy and are eventually stopped in a process effectively described as exponential decay. If starting with N_0 particles, the number that reaches depth x is given by

$$N = N_0 e^{-\mu x} = N_0 e^{-\left(\frac{\mu}{\rho}\right)\rho x}$$

where μ is the (total) linear attenuation coefficient. Note that it is conventional to work with μ/ρ, mass attenuation coefficient, since it does not depend on density and can be compared between materials.

For energies (< 20MeV) relevant to medical applications, attenuation is caused by four major types of interactions. These are coherent scattering, photoelectric effect, Compton effect, and pair production. By convention their attenuation coefficients are denoted as σ_{coh}, τ, σ_c and π. Total mass attenuation coefficient is then

$$\left(\frac{\mu}{\rho}\right) = \left(\frac{\sigma_{coh}}{\rho}\right) + \left(\frac{\tau}{\rho}\right) + \left(\frac{\sigma_c}{\rho}\right) + \left(\frac{\pi}{\rho}\right)$$

During coherent (elastic) scattering no energy transfer occurs. For the purposes of this work, these events can be ignored. Photoelectric effect occurs when a photon is absorbed by an atom and results in one of inner orbital electrons being ejected with almost all of the absorbed energy. Cross-section (and hence attenuation coefficient) roughly scales with photon energy E as $\tau/\rho \propto 1/E^3$ [13], such that photoelectric effect becomes negligible above ~ 100KeV in water.

Compton scattering is the process of inelastic collisions with outer shell electrons (also called 'free' due to low binding energy relative to E). During the collision, electrons absorb a portion of energy and are ejected into the medium. A photon of $E' < E$ is also emitted and continues to interact. Compton mass attenuation coefficient is near constant for all materials, and decreases slightly with energy. This process is dominant in 100KeV-10MeV range in water.

Pair production is the creation of positron-electron (e^+-e^-) pairs in the presence of strong nuclear electric field. Due to energy conservation this can only happen above 1.022MeV. Any excess energy becomes kinetic energy (split equally on average), which both particles lose by same processes. Eventually e^+ annihilates with another e^-, emitting two 511KeV γ-rays. Cross-section energy scaling of pair production is $ln(E)$. It becomes the dominant effect at energies above 10MeV in water.

Charged particle interactions

All of above processes end up in production of charged particles - namely, electrons. These are directly ionizing and interact via four processes - scattering, ionization (creation of ion pairs), excitation (raising energy level of atomic electrons), and bremsstrahlung. Of these, the latter two can produce more photons which again interact as described above. However, first three also lead to local energy deposition and consequent damage.

Eventually, all photons will either escape or produce free electrons, which in turn

will deposit energy while slowing down and finally getting captured back by local atoms. Total energy absorbed at a point, called absorbed dose, can then be calculated. It is usually expressed in units of Gray (1Gy=1J/kg).

1.2.2 Cellular interactions

To understand biological consequences of above interactions, it is necessary to review cellular structure and basic biochemistry.

Cell structure

Speaking broadly, human cells studied in this work are a set of closed environments separated by phospholipid membranes which restrict exchange of molecules between them. Primary barrier enclosing everything else is cell membrane, with inner contents called protoplasm. It consists of a multitude of vesicles, mitochondria, cytoskeleton fibres, and others along with a nucleus - a membranous structure containing chromosomes, which store deoxyribonucleic acid (DNA). DNA encodes almost all cellular proteins and through differential expression determines cell structure and function.

Molecular make-up

In terms of chemical make-up, cells are approximately 70% water [14] with remainder made up of proteins ($\sim$ 15%), lipids, nucleic acids, and inorganic solutes [15, 16]. Proportions of these change depending on cell cycle phase or environmental stresses and can be detected by Raman spectroscopy. A brief discussion of function and make-up of cell constituents is necessary to better understand Raman signals.

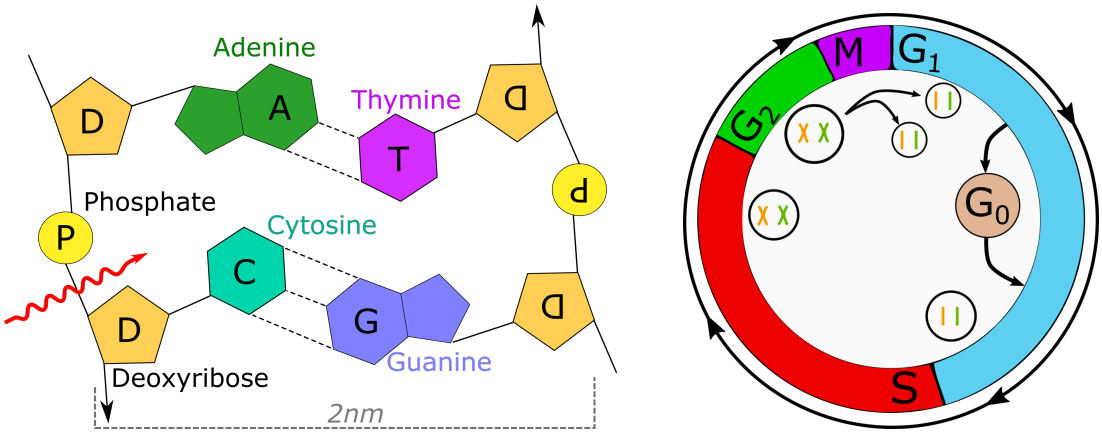
Proteins are chains of amino acids (one of 20) joined by peptide bonds. They are synthesized in cytoplasm from an RNA template which in turn is transcribed from one or more segments of DNA. Proteins can assume secondary, tertiary, and quaternary structure and perform vast majority of biochemical and signalling functions of the cell.

Lipids encompass most small hydrophobic and amphiphilic macromolecules. They fulfill a variety of functions such as membrane bilayer formation, energy storage, and cell signalling. Lipids are synthesized in the cytoplasm and endoplasmic reticulum.

Carbohydrates are one of key energy storage molecules (i.e. glycogen) and serve as precursors in a wide variety of synthetic reactions. They are, as name suggests,

characterized by long CH_2 chains as well as an aldehyde or ketone group (in open conformation).

Nucleic acids in the form of DNA and RNA are key in expression and transfer of generic material, as well as short term energy storage. RNA also performs some biochemical roles. Their basic building block, a nucleotide, has three components - phosphate group joined to a sugar molecule (that together form the backbone), and attached to each sugar another molecule called base. Four such bases are used in DNA - adenine (A), thymine (T), cytosine (C) and guanine (G). Due to geometrical arrangement of hydrogen bonds, C-G and A-T are always found in pairs (called complimentary base pairs) when two DNA strands join to form a double helix [17], which plays important role in DNA repair mechanisms. This energetically favourable conformation is shown in figure 1.1a.



(a) (b)

Figure 1.1: (a) Visualization of DNA ladder structure of complimentary pairs, with red arrow denoting a possible track of directly ionizing particle that will result in a single strand break. (b) Typical mammalian cell cycle.

Cell cycle

The most basic function of cell cycle is to ensure cellular division is possible by duplicating the vast amounts of DNA and then segregating each copy along with half of cytoplasm to each daughter cell [18]. Two major phases of cell cycle are S phase (synthesis of DNA copy) and M phase (mitosis - cell division). They are separated by two G phases (growth - increase in cytoplasm, protein, and lipid content), with G_1 occurring after mitosis and G_2 after S phase, as is visualized in figure 1.1b.

Unless special synchronization techniques are used, at any given time cells will be spread out over different phases. For a rapidly proliferating population, it is expected that 30-40% will be in S phase, 50% in G_0/G_1 and 20% in G_2 [19]. However, if conditions are unfavourable or growth regulation signals are present, cells may delay progress through G_1 or even enter resting state G_0, where only basal metabolic rate is maintained.

Cells in different cell cycle phases can be distinguished experimentally by the ratio of protein/lipid to nucleic acid content, for example via Raman spectroscopy or fluorescence staining. It must be noted that radiosensitivity also depends on phase since main radiation damage pathway involves DNA breakage, as discussed below. Thus an unsynchronized population may show natural variations in its radiation response, an important consideration for radiosensitivity assays.

DNA ionizing radiation damage

DNA double helix is a polymer built up of thousands of nucleotides, with information stored in their specific ordering. It is used to make RNA via transcription, which in turn participates in cell signalling or is translated into proteins. In an actively cycling cell, DNA breaks could be catastrophic and prevent further proliferation. For example, in mutant yeast strain with no break repair mechanism, even a single double DNA break is fatal [20]. No other molecule is as important to cell survival, and for this reason DNA is primary target of radiation therapy.

Recall that it is charged particles that deposit dose via excitations and ionizations. They can damage DNA through direct interactions, but a more prevalent mechanism involves creation of nearby free radicals - molecules with unpaired valence electrons. About 60-70% are hydroxyls [21], formed via $2H_2O \rightarrow H^+ + OH^- + OH^\bullet + H^\bullet$. Once created, they can diffuse over 20nm to react with and break DNA strands (and other nearby molecules) via a variety of mechanisms [22–24].

Such reaction can damage DNA by changing bases or cross-linking strands, but more importantly create single (SSB) and double strand breaks (DSB) in the DNA backbone (red line in figure 1.1a). Cells are capable of detecting and repairing these via several pathways depending on severity and amount of damage. Cell cycle progress is also modulated to allow more time for repair. Base changes, crosslinks, and SSBs are most often successfully repaired (by using complimentary strand), however DSB repair can result in mutations or complete chromosome failure. More detailed discussion can

be found in relevant textbooks [12, 25].

Possible cell fate

Once DNA damage occurs, healthy cells have a number of possible end states - the ideal scenario is full repair and continuation of cell cycle. If sufficient damage accumulates, programmed cell death known as apoptosis may be triggered. Alternatively, necrosis may occur whereby the cell loses ability to maintain its homeostasis and disintegrates uncontrollably. Mitotic catastrophe is possible if mitosis is attempted with damaged DNA, and also results in cell death. Cells may also enter G_0 phase indefinitely, becoming senescent. Final and least favourable outcome is that DNA gets repaired incorrectly, causing a mutation but not cell death. Depending on precise location and type of mutation such cells may gain ability to divide uncontrollably, becoming cancerous.

1.2.3 Biological response models

Quantification of radiation damage is usually done through measuring the fraction of cells that survive (i.e. are able to proliferate) after a certain dose, denoted by $S(D)$. It is measured by seeding a known number of cells and counting fraction that formed colonies after sufficient growth time. Due to the range of values involved, it is customary to plot $ln(S)$ vs D in what are called survival curves.

Multiple studies have confirmed that cell survival is correlated with rate of DNA damage by directly observing DSBs, or via fluorescence labelling and other measurements. They concluded that rate of DSB creation is linearly proportional to dose and variations in radiosensitivity with dose or cell type can be assigned to differences in repair speed [26, 27].

Linear quadratic model

Linear quadratic (LQ) model is the most basic and yet clinically used description of cell radiation response. Building on knowledge of linear dose-DSB relationship stated above, suppose that single DSB can be repaired via first order reaction with some time constant. However, if two DSBs are present at the same time a repair mismatch may occur. Moreover, certain radiation lesions are fatal through other mechanisms such as uncorrected mutations. The latter two processes correspond to terms quadratic in

dose (D) and linear in it respectively. Assigning them model parameters α and β, survival fraction is determined as

$$S = e^{-\alpha D - \beta G D^2}$$

with $G = 1$ for single fraction case [28]. Taking natural logarithm of both sides,

$$ln(S) = -\alpha D - \beta D^2$$

which makes apparent the linear and quadratic terms.

LQ model gives consistent results for single doses up to 15Gy, given that correct model parameters are known. These are usually stated as α/β ratio - the dose at which linear kill is equal to the quadratic one. Early responding normal tissues tend to have $\alpha/\beta \sim 10$, while late responding ones are lower at $\alpha/\beta \sim 3$. Most tumours have $\alpha/\beta > 10$. These values mean higher survival of normal tissues at low doses, a sparing effect used as basis for 2Gy fractionated treatments in modern radiation therapy.

High dose models

At higher doses relevant to some modern treatments such as radiosurgery, consistent deviations from LQ model are observed [29], whereby the LQ model curve continues to get steeper while experimental data suggests a linear ln(S) vs D relationship. Consequently, a variety of new models have been proposed, such as linear-quadratic-linear model (LQ-L) or linear-quadratic-cubic (LQC) model [30], which both add a third parameter to better describe high dose range. However, usage of these models in clinical treatments is so far limited.

TCP/NTCP

In order to effectively compare treatment regimens, tumour control probability (TCP) is usually used. It is defined as probability of extinction of clonogenic tumour cells at the end of treatment [31]. Conversely, normal tissue complication probability (NTCP) is defined as probability of complications (site dependent) in surrounding healthy tissues, which often receive doses close to those of the tumour. Historically, TCP/NTCP are often based on general published population values [32] combined with personal radiologist experience. Recently, there were several efforts to system-

atize and quantify three dimensional dose/volume/outcome data [33] which however still only provided general guidelines.

1.3 RT treatment planning

LQ model is used extensively in treatment planning, where its predictions are combined with other (sometimes semi-empirical) factors to arrive at a set of TCP/NTCP values for several possible regimens. Such factors can be roughly divided into three categories - tumour related (size, origin, oxygenation, genetic markers), patient related (age, medical history, other medical conditions) and treatment related (beam arrangements, treatment volume) [34, 35]. However, to a large extent planning will depend on knowing correct radiation response parameters, unique to each tumour/normal tissue.

1.3.1 Personalized radiation therapy

Patient-to-patient differences in radiation response are called patient variability [36], and within LQ framework correspond to uncertainty on α/β ratio and all corresponding results (S/TCP/NTCP). Unfortunately, current clinically used model parameters (and derived dose prescriptions) are based on averaged previous treatment data, which can lead to either unsatisfactory TCP or excess normal tissue damage.

A predictive assay with high sensitivity and specificity that could be used to reliably determine tumour radiosensitivity either pre-treatment or during early treatment stages would confer significant patient benefits. This approach is known as personalized radiation therapy, and it has seen much attention recently with a variety of proposed assays based on measuring clonogen doubling time, intrinsic radiosensitivity, hypoxic fraction, genetic/protein markers, and others [37]. Major assay types are reviewed briefly below.

Clonogenic assays

Clonogenic assays were explained above in context of measuring SF. For certain cancer types, they show correlation of *in vitro* clonogenic cell survival with *in vivo* local tumour control [38], although predictive power is usually low. Furthermore, these assays are highly labour intensive and provide little molecular information [39, 40].

Functional assays

A large class of assays based on measuring bulk tumour properties, functional assays have seen some success, mostly with local tumour oxygenation (due to oxygen-dependent nature of reactive oxygen species production) [41]. Unfortunately, direct measurements are either highly invasive or require injection of radioactive isotopes [42], while indirect protein markers show weak predictive power. Recently, methods based on diffuse optical monitoring have been proposed [43] but are not yet sufficiently accurate.

Correlations of proliferation rate assays (measured by mitotic index) have also been reported [44], but detection methods are again limited by necessity of marker injection and only apply to certain cell lines.

Genomic/proteomic assays

This type of assay quantifies expression of genes and proteins (but not corresponding cellular biochemistry), attempting to correlate it with treatment outcomes. In early single-gene studies, candidate biomarkers such as p53 gene have shown strong correlation with apoptosis [45, 46] in some cell lines, but not in others.

With remarkable advancements in cost and speed of genomic sequencing, it recently became feasible to do genome-wide assays in what is now known as study of radiogenomics [47]. Multiple recent genome-wide association studies showed significant results [48–50] for a large variety of markers, and several candidate gene polymorphisms have also been identified [51]. Unfortunately, most of this data lacks clinical validation and often yields false predictions due to limited cell line selection and lack of *in vivo* related gene expression in training samples. Nonetheless, with economies of scale, continuing improvements of sequencing technology, and introduction of bioinformatic algorithms based on machine learning (such as IBM Watson), genomic assays are expected to become commonplace in the next decade [52]. Several projects have been started to gather and systematize tissue samples and patient data on national and international levels [53, 54] in order to build up adequate reference databases, with results expected in next few years.

Imaging assays

Positron emissions tomography (PET) and magnetic resonance imaging (MRI) are the two primary methods for imaging assays, and have been used *in vivo* both sepa-

rately [55] and together [56] with above assays (such as in hypoxia radioactive marker imaging). While several promising tracers and biomarkers have been reported [57], further research and clinical studies are necessary to assess and validate their impact on radiation response [58].

1.3.2 Raman spectroscopy

It is clear that a wide variety of radiosensitivity assays exist, and yet there is no method that has achieved everything required for direct clinical application. While Raman spectroscopy (RS) does not claim to be this method, it is nonetheless a very promising option that can complement any of above assays with detailed biochemical information.

RS is based on measuring frequency and intensity of scattered light, with ability to attribute signals to particular vibrational modes of chemical bonds. It offers label-free, non-invasive, and non-destructive *in vitro* and *in vivo* approach combined with ability to capture and resolve highly complex signals from many types of molecules in a single acquisition. Detailed theory and data processing discussion is deferred to chapter 2, while here a brief summary of RS usage in biological sample analysis and specifically cancer research is provided.

RS of cells and tissues

RS application to biological samples is attractive due to above advantages as well as the ability to probe micron-scale sampling volumes (which is typically called Raman microscopy). This high resolution makes it possible to measure subcellular components such as chromosomes [59] as well as a variety of other molecules and microorganisms [60, 61]. RS is sufficiently sensitive to distinguish individual live and dead cells, as well as classify them by cell cycle phase [62]. More recently, there has been significant work on RS classification of disease signals (i.e. stomach dysplasia [63], diabetes [64]) and other abnormalities, where RS and its' more advanced modalities have shown impressive results with high specificity and sensitivity, becoming a key tool in a field of molecular fingerprinting [65]. Other biological applications include non-invasive 3D imaging (for instance ability to monitor real-time epidermic drug delivery [66]), fabricated biological sensors [67], and others too numerous to describe.

Given the wide applicability of RS, it unsurprisingly has seen extensive use in cancer research. There has been a large effort in using RS to distinguish various

malignancies from healthy tissues (prostate cancer, cervical cancer, etc.), with mostly good results [64, 68, 69]. In fact, some studies reported more fine-grained classification into several grades by severity. Moreover, several recent works concluded that Raman techniques have reached clinically relevant levels for cancer diagnosis applications [70–72].

RS in radiobiology

To make RS an effective radiosensitivity assay it is necessary to detect and quantify continuous biochemical changes related to radiation damage (in contrast with classification studies above). While direct RS observation of radiation induced damage in DNA, lipids, and other biomolecules [73, 74] is possible, so far this required doses significantly above RT ones and as such is not a viable clinical approach. Thus the only possible signal could come from radiation-induced biochemical changes, such as via stress response and DNA repair pathways. This has been observed in bulk RS measurements for cervical tissues after doses of 4Gy [75].

Due to cell cycle distribution and stochastic nature of radiation damage, response variability is always expected. For bulk RS measurements this would average out, confounding desired radiation signal. However, RS has capability for single-cell measurements, and with proper multivariate analysis techniques subsequent radiation signal separation can be achieved [64].

Extensive work by Matthews et al. [76, 77] has demonstrated that single-cell RS is capable of detecting radiation response at single fraction doses of 15Gy to 50Gy in a wide variety of cell lines. Usage of principal component analysis (PCA) has allowed for separation of dose-dependent signal from cell cycle one and others, elucidating specific biochemical changes which were in agreement with known radiation response pathways.

This was built upon by Harder et al. [78], who extended above analysis to clinically relevant dose of up to 10Gy. They found however that this analysis was possible only for certain strongly responding cell lines, while others either required higher dose data for full PCA signal separation or showed only weak RS response at all doses. Their recent publication has demonstrated that detection is possible even at doses as low as 2Gy [79].

Further improvements are possible via raising signal-to-noise ratio of Raman signal (but this is already highly optimized), increasing the amount of data (which is active

area of research using microfluidics) or improving data analysis, latter of which is the purpose of this thesis.

So far, a fairly common processing routine was used based on background subtraction and subsequent PCA dimensionality reduction. While capable of perfect signal separation on ideal data, PCA has a number of disadvantages for complex biological datasets, such as inability to work with missing values, strict Gaussian noise assumptions, and susceptibility to even a single grossly corrupted observation [80]. With current interest in 'big data', a number of more advanced multivariate dimensionality reduction techniques have been proposed such as robust PCA [81], independent component analysis (ICA) [82], self-organizing maps (SOM) [83], and others, which have already found uses in many financial and scientific fields [84, 85]. If these allow for radiation response detection at lower doses, that would lead to earlier treatment corrections and/or make RS assay suitable for more weakly responding cell lines.

1.4 Thesis Scope

The primary goal of this work is to improve current sensitivity of single cell RS to low dose radiation induced biochemical changes through application advanced multivariate analysis techniques. This is done by implementing several prospective algorithms and applying them onto previously collected high and low radiation response datasets from H460 and LNCaP cell lines respectively. Obtained signals are compared to reference results and performance improvements quantified to determine the optimal parameters.

This process is presented in chapters 2 through 8 below, starting with a more detailed overview of RS theory, apparatus, and signal processing and analysis algorithms in chapter 2. This is followed by brief description of how spectral datasets were collected in chapter 3 and a discussion of algorithm implementations in chapter 4.

Results are presented in three chapters, starting with a general discussion of data variability and outlier detection in chapter 5. High and low response data is then presented in chapters 6 and 7, with respective performance metrics and discussions.

This work concludes with a performance summary and possible future work in chapter 8. Appendices A and B show additional outlier rejection and analysis results that were for brevity omitted from main chapters.

Chapter 2

Background

This chapter presents major background topics required to understand the rest of this work. Section 2.1 introduces key aspects and theory of RS, and describes modern apparatus design and performance. Section 2.2 provides overview of spectral processing techniques and their applicability for single cell Raman microscopy (RM).

2.1 Raman Spectroscopy

When a beam of monochromatic photons such as from a laser impinges on a sample, most of the radiation is scattered via Rayleigh (elastic) scattering, meaning photons are emitted at same wavelength as incoming ones [86]. Fields such as infrared (IR) and visible light spectroscopy are devoted to measuring and interpreting absorption of this coherent scattering. However, a small portion, about 1 in 10^6 photons, undergoes inelastic scattering and is emitted at a shifted wavelength specific to the molecule it interacted with. This process forms the basis of Raman spectroscopy. First observed by Sir C. V. Raman in 1929 [87], today this technique has a multitude of research and industrial application, from analysing defects in LCD films to archaeology [88] to *in situ* tumour cell measurements [89]. This section is devoted to giving an overview of Raman - description of theoretical models, modern equipment design, and signal processing techniques.

2.1.1 Theory of Raman scattering

Raman scattering process begins when electric field of incoming beam induces polarization in the molecule, exiting it into a virtual vibrational state. These states

are inherently unstable and decay near instantaneously into one of stable ground energy levels, as is demonstrated in figure 2.1. If decay is into same state, no energy change takes place and Rayleigh scattering occurs. However, if decay is into a different state, the emitted photon has new, different energy. Those with lower energy (higher wavelength) produce Stokes radiation, while higher energy ones form anti-Stokes radiation.

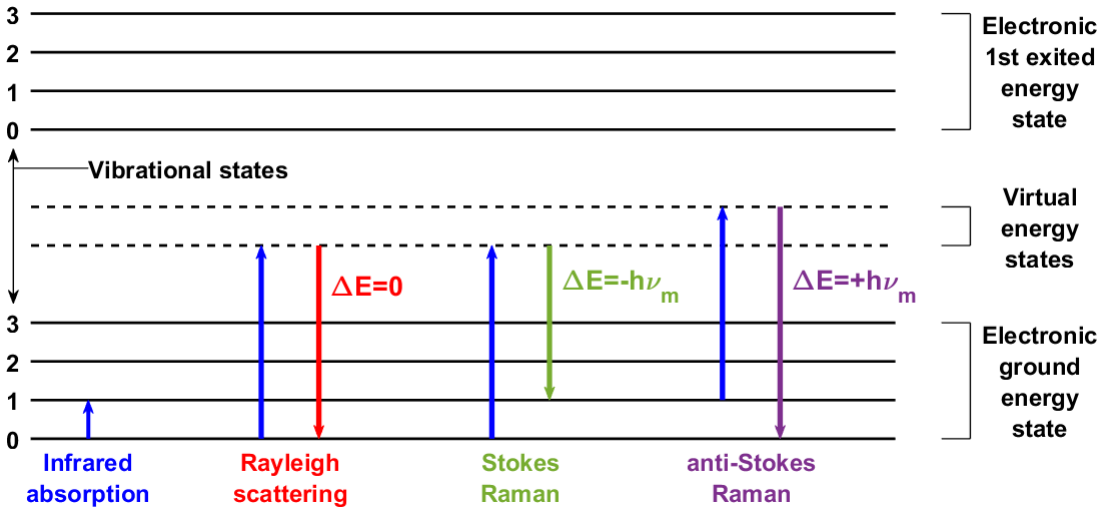


Figure 2.1: Several possible energy level transitions for a representative molecule, starting from ground state (most populated). Blue arrows denote photon absorption; emissions processes are colour matched with respective labels. Vibrational level spacing is $h\nu_m$.

Quantitative description of this process is quite involved and can be found in a variety of sources [90, 91]. Only major results are noted here to give intuition as to the scaling and order of effects involved.

Let ν denote the frequency of photons and λ the respective wavelength. A '0' subscript will be used to denote incoming photon values, and no subscript for scattered photons. Constant c is speed of light as usual. A convenient unit to express energy change is wavenumber $\bar{\nu}(\mathrm{cm}^{-1})$, defined as

$$\bar{\nu} = \frac{1}{\lambda_0} - \frac{1}{\lambda} = \frac{\nu_0}{c} - \frac{\nu}{c} \tag{2.1}$$

It is clear that $\bar{\nu}$ increases for larger changes, and is by definition positive for Stokes Raman. When expressed in (cm^{-1}) units, most signals are contained within $400 - 4000\,\mathrm{cm}^{-1}$ range.

For monochromatic light of frequency ν_0, electric field varies as

$$\mathbf{E} = \mathbf{E}_0 cos(\omega t) = \mathbf{E}_0 cos(2\pi\nu_0 t) \tag{2.2}$$

where boldface denotes vectors in Cartesian 3D space. If this field interacts with a molecule with polarizability tensor α_{ij} $(i, j = 1, 2, 3)$ then resulting dipole moment is given by

$$\mathbf{P} = \alpha\mathbf{E} \tag{2.3}$$

This change in polarizability will cause vibrations of nuclei around their equilibrium positions q_0 at normal mode frequency ν_k. For i_{th} position,

$$q_i(t) = q_{i,0} cos(2\pi\nu_k t) \tag{2.4}$$

By Taylor expanding $\alpha = \alpha(q(t))$ around q_0 as

$$\alpha = \alpha_0 + \left(\frac{\delta\alpha}{\delta q_i}\right)_{q_0} \cdot q_i + O(q_i^2) \tag{2.5}$$

and plugging into 2.3, it can be shown that

$$\mathbf{P} = \alpha_0\mathbf{E}_0 cos(2\pi\nu_0 t) + \left(\frac{\delta\alpha}{\delta q_i}\right)_{q_0} \cdot q_{i,0}\mathbf{E}_0[\underbrace{cos(2\pi(\nu_0 - \nu_k)t)}_{\text{Stokes}} + \underbrace{cos(2\pi(\nu_0 + \nu_k)t)}_{\text{anti-Stokes}}] \tag{2.6}$$

where first term is Rayleigh scattering (recall - oscillation of dipole moment produces EM radiation) and second consists of Stokes and anti-Stokes contributions as labelled. It must be noted that at room temperature, the majority of electron population occupies ground state in accordance with Boltzmann statistics, meaning actual anti-Stokes signal contributions are several orders of magnitude lower than Stokes ones unless specialized techniques are used. In this work, only Stokes Raman signals are collected.

Scattering intensity is directly related to Raman cross section, which for a transition from vibrational state n to state m is given by

$$\sigma_{n\to m} = C \cdot (\nu_0 - \nu_k)^4 \cdot \sum_{i,j} |\left(\alpha_{ij}\right)_{n\to m}|^2 \tag{2.7}$$

with C constant proportional to beam power, sum running over molecule-fixed coor-

dinates, and α determined via Kramer-Heisenberg-Diracs (KHD) dispersion theory (see [90]). Equation 2.7 indicates that Raman intensity depends on fourth power of incoming photon frequency since ν_k is molecule specific and fixed. Moreover, from 2.6 it is clear that Raman signal is not present unless $\left(\frac{\delta\alpha}{\delta q_i}\right)_{q_0} \neq 0$ (this corresponds to sum in 2.7 not being 0). In other words, Raman signal scales with fourth power of ν_0 and linearly with intensity as long as polarizability tensor changes asymmetrically with molecular displacements about equilibrium.

To determine the latter, it is necessary to consider possible molecular motions (vibrational modes) and corresponding available degrees of freedom (DOFs). An N-atom system has in most general case 3N DOFs corresponding to translations in 3D. Once atoms form a molecule, it is necessary to subtract 3 DOFs for translation of whole molecule (atoms must stay together) and 3 for rotations. Thus, molecules generally have 3N-6 DOFs. In a special case of linear molecule, one axis becomes axis of rotational symmetry - such molecules have 3N-5 DOFs.

To each DOF one can associate a normal mode vibration. For example, linear CO_2 molecule consists of three atoms and has 4 normal modes - visualization of 2 of these is given in figure 2.2.

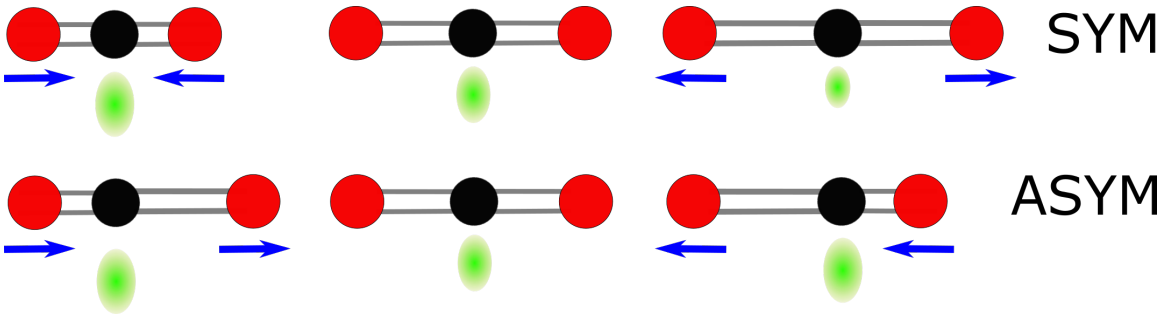


Figure 2.2: Two of four CO_2 vibrational modes. Carbons atoms are black, oxygen atoms are red, and polarizability ellipsoids are given below molecules in green. Omitted bending modes are Raman-inactive.

Below each molecule is a polarizability ellipsoid constructed via $r = 1/\sqrt{\alpha}$. Changes in its' shape correspond to changes in α and maintain same symmetry. For instance, asymmetric stretching mode ellipsoid is symmetric around equilibrium position, meaning $\left(\frac{\delta\alpha}{\delta q_i}\right)_{q_0} = 0$ and consequently this vibration is Raman inactive. Symmetric stretching however is Raman active because there is a clear asymmetry in polarizability.

For larger molecules, the situation gets complicated quickly but in general symmetric/asymmetric vibrations continue to be active/inactive respectively (for symmetric

molecules). The important conclusion is that Raman spectroscopy can only see certain vibrational modes, and thus only certain chemical bonds (which are nonetheless plentiful). By matching observed wavenumbers with lists of known signals, this allows for near unique identification of chemical makeup.

2.1.2 Raman spectroscopy and microscopy

With constant improvements in optics and semiconductor technology, Raman instrumentation has been continuously changing throughout its nearly century long history. While initial setups used mercury lamps and photographic plates, modern designs employ a wide variety of lasers, optical filters, and charge-coupled devices (CCDs) to achieve remarkable signal to noise ratio (SNR), low collection times, and high spectral resolution.

A conventional Raman system consists of many components. Light is produced with a laser source (at wavelengths from UV to IR depending on application), and steered onto the sample with appropriate optics. Scattered light is then collected and filtered to remove excitation frequency while letting Stokes radiation through. Using a spectrometer, signal is dispersed (spread out by wavelength) and projected onto CCD - a matrix of photosensitive semiconductor elements, pixels. Once sufficient signal is collected, readings along same wavelength are binned and read out to obtain Raman intensity as function of wavenumber. Modern systems can achieve spectral resolution of around $1 - 2\,\mathrm{cm}^{-1}$ [92].

To analyse small samples, it is natural to couple Raman system to a microscope, allowing for precise sample positioning and sampling volumes of just a few μm. The diagram of such a setup is shown in figure 2.3. Note that excitation beam and signal share same optical path, made possible by highly attuned 45^o mirror/filter which is reflective for Rayleigh signal but transparent to Stokes radiation.

Performance of Raman microscopy (RM) setups is highly dependent on the choice of objective. It needs to collect as much light as possible and also be made of high purity materials to eliminate stray fluorescence. The former can be quantified with numerical aperture (NA) given by NA$= n \cdot sin(\alpha)$ with α the half-angle of maximum light cone [93] and n the refractive index of the medium. NA is related to minimum sampling volume width (in plane perpendicular to beam) s by

$$s = \frac{0.61 \cdot \lambda}{\mathrm{NA}} \tag{2.8}$$

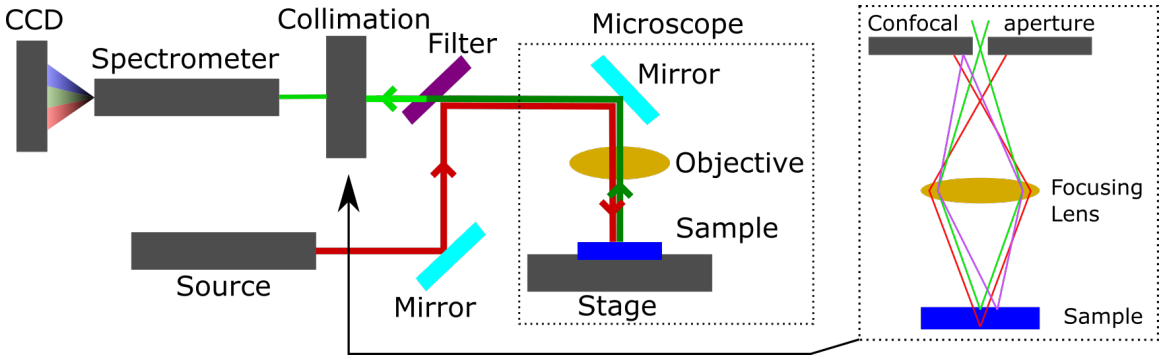


Figure 2.3: Typical Raman microscopy setup such as that used in current work. Dark red denotes excitation signal, dark green all scattered light and bright green only Stokes signal.

for ideally aligned, infinity focused case with excitation wavelength λ. While this is never achieved in practice, it gives correct inverse proportionality. Moreover, depth of focal spot (along the beam, i.e. depth of field) [94] is given by

$$DOF = \frac{n \cdot \lambda}{\mathrm{NA}^2} \tag{2.9}$$

which is even more dependent on high NA. With modern Raman systems and visible light such as used in this work, resolution of 0.5/2 μm for s/DOF respectively is possible under optimal conditions.

2.1.3 Single cell RS

Analysis of biological samples, especially live and small ones, is subject to a number of additional considerations.

Spatial resolution

In general, adherent eukaryotic cells tend to splay out along the substrate with thickness of 1-5 μm and order of magnitude higher width. However, once detached as is done in this work, they have diameters of around 10μm. It is required that spatial and confocal resolutions be of this order to analyse single cells, which as discussed above is feasible.

Spectral resolution

Distinguishing between nearby Raman signals, on the order of 5cm^{-1}, is necessary to resolve relevant biological features. Modern systems can achieve this.

Laser power and frequency

Recall that Raman intensity is higher for shorter wavelength and more energetic excitation sources. Unfortunately, biological samples exhibit strong fluorescence [95] in < 650nm range, sometimes swamping Raman signal by several orders of magnitude. Fluorescence is significantly lower in near-IR range, which is however sub-optimal in terms of signal power. Moreover, substrate fluorescence is also frequency dependent and must be tuned appropriately.

Second consideration is sample degradation due to photodamage. This effect depends on both wavelength and power density. Several studies indicate [96, 97] that at 514nm even low power density of 5mW is sufficient to induce chromosome damage (which cannot be attributed to heating due to low power density). However, at 785nm no damage was observed even at 115mW power level. In this work 785nm laser was used to avoid these issues, with more detailed description given in chapter 3.

2.2 RS spectral analysis

Usefulness of biological Raman measurements is often limited due to significant inherent background/noise [98], some sources of which were described above. The wide variety of molecules sampled, while an advantage, also causes large signal overlaps making direct interpretation hard even if SNR is kept adequately high. Consequently, RS, more than other spectroscopic techniques, depends on advanced signal processing to maximize system performance. This section provides general overview of processing pipeline and gives a summary of current and proposed analysis methods.

2.2.1 Spectral preprocessing

Preprocessing can be classified as any data manipulation that is done prior to main dimensionality reduction or another analysis algorithm. It serves to remove or reduce undesirable characteristics, such as calibration drift and background fluorescence, while raising SNR. Typical correction steps are discussed below in the most commonly

used order, although in some cases rearrangements are beneficial [99]. Some steps not relevant to current work, such as feature selection, detrending, and transformations are omitted. See [100] for applicability and details of these.

Preliminary inspection

Immediately after collection, spectra are visually assessed for poor SNR, cosmic rays or other imperfections. If these are obvious, outright rejection is reasonable [101], although care should be taken to control user bias though outlier rejection algorithms. Some defects, such as cosmic rays, can be detected and removed later in an automated fashion [102] while others are harder to correct.

SNR improvements

Depending on application, it may be necessary to raise SNR prior to main analysis. A number of methods exist such as Savitzky-Golay (SG) smoothing, wavelet denoising (WDN), two-point maximum entropy (TPMEM) smoothing or application of a dimensionality reduction algorithm (such as PCA) [103, 104]. However, these have to be applied carefully due to inevitable degradation of spectral features.

Baseline removal

The main part of preprocessing is removal of extraneous signals, primarily from sample and background fluorescence. Exact details are strongly dependent on spectral window, tissue makeup, and laser wavelength. In certain situations, such as with chlorophyll containing plant cells, fluorescence levels might be intractably high (but this is rare).

A variety of background subtraction approaches have been proposed such as simple linear estimation, more sophisticated polynomial treatments [105], wavelet-based methods [106, 107], maximum likelihood (ML) estimators [108], iterative SG signal removal [109], and a multitude of other combinations [110]. Furthermore, automated selection of best method has been very recently suggested based on genetic algorithms [111].

Normalization

Final preprocessing step is spectral normalization. Most common methods are normalization to total area under the curve [100] (corresponding to total sampling vol-

ume), to min-max range, or only to specific spectral peaks (that may be correlated linearly with amount of biological content). The latter requires a reference spectrum or window selection which can reduce sensitivity to changes in these regions.

2.2.2 Spectral analysis

Due to the wealth of chemical signals present in Raman spectra, fairly advanced analysis algorithms are common. They use two general approaches - feature construction and feature selection [101], which are distinguished by, as name suggests, either constructing new features (such as via projecting data into another basis) or extrapolating existing ones (using measured data directly). However, due to significant signal overlap, selection bias, and other disadvantages, feature selection methods are rarely used on their own [112] whereas feature construction ones have been essential in biomarker extraction and pattern recognition. In this thesis only the latter approach is considered.

Another fundamental difference is whether a method is supervised or unsupervised, although these names are somewhat of a misnomer since both are easily automated. Instead, supervision refers to input of additional data such as class labels. Unsupervised methods are utilized to extract 'natural' feature-rich variables from unknown samples. Supervised methods are often used in diagnostic analysis where training on a known dataset is performed first and then analysis applied to classify test samples.

Supervised algorithms

Examples of supervised algorithms are linear discriminant classifiers (LDCs), artificial neural networks (ANNs), support vector machines (SVMs) and partial least squares regression (PLSR). Latter of these has been used [113, 114] for building cell dose estimation models but with limited success.

Unsupervised algorithms

A wide variety of unsupervised clustering and dimensionality reduction algorithms are available. Most widely known of the latter is PCA, a powerful method in itself and often the starting point of more advanced techniques. Its goal is to find a new data representation (basis) such that most variance can be explained with just a few dimensions (as compared to dimensionality equal to pixel count in original data).

Examples of PCA usage are numerous, with applications ranging from skin drug penetration [115], to food metabolite testing [116], to tumour detection as mentioned in the introduction. For datasets that have sufficiently high SNR, PCA has been remarkably effective, although its use with noisy or otherwise corrupted data has faced some challenges. For instance, in one skin biopsy study [117], while PCA was not able to distinguish between tumour types, independent component analysis (ICA) did have limited success. Unlike PCA, ICA seeks not just uncorrelated but linearly independent components, which is a significantly harder task (more formally known as blind signal separation). There have been several other reports [118, 119] of ICA application in RS, with modifications proposed that make it more suitable for biological imaging. However while ICA is statistically more powerful, its disadvantage is the inability to assign variances to particular signals. In other words, ICA can indicate if a signal is present but not its relative strength. This complicates separation of true signals from noise. Advanced hierarchical and classifier methods have been proposed to deal with this [120].

Factor analysis (FA) and related techniques are also similar to PCA but seek to find not highest but most common sources of variance, with uses reported in spatially offset [121] RS. The assumption of underlying latent model however is difficult to incorporate into current work, although some exploratory FA estimation methods are available.

Note that while differing in detail, all of above methods were implicitly linear. If however nonlinear dimensionality reduction is considered, situation becomes extremely complicated. Techniques for dealing with this are generally called manifold learning methods, of which there are several dozen. Their most prominent biological use is in analysis of proteomics datasets [122, 123]. In terms of RS applications, while an attempt was reported (usage of self-organising maps, a type of artificial neural network (ANN), to distinguish wood types [124]), in general these methods have not seen significant adoption.

2.3 Details of analysis algorithms

This section presents the theoretical basis of relevant analysis algorithms; implementation details are left to chapter 4. For reasons that will be discussed in chapter 5, three main approaches are used to achieve desired signal improvements. First of these modifies PCA analysis to better deal with spectral outliers and noise by using either

data weighting or robust predictors (weighted PCA, robust PCA). Second category of methods attempts to remove outliers during preprocessing with formal rejection tests and then performs modified missing-data tolerant PCA analysis (ALS PCA, probabilistic PCA). Finally, a recently proposed nonlinear PCA approach based on autocorrelating ANN is attempted (NLPCA).

For the following discussion, it is useful to assume a notational convention: let n denote number of spectra in set, and let p denote number of measurements, assumed same for all spectra. Supposing that dataset is placed into $n \times p$ matrix $\mathbf{X}$, assign index i to stand for i_{th} spectrum, and index j to represent a particular WN/pixel. Hence $\mathbf{X}(i,j) \equiv \mathbf{X}_{ij}$ is a value of type double representing intensity measurement in i_{th} spectrum at j_{th} pixel. MATLAB notation will be used to denote vector selection, such as $\mathbf{X}(:,j) \equiv \mathbf{x}_j$ (note lack of capitalization) defining vector $\mathbf{x}$ that contains all n values of j_{th} pixel. In accordance with usual notation, transpose is denoted by $\mathbf{X}^{\mathrm{T}}$, mean by $\bar{\mathbf{x}}$, values along matrix diagonal by $diag()$, and rank (dimensionality of vector space spanned by matrix column vectors) by $rank()$.

2.3.1 PCA

Principal component analysis is one of the most widely used dimensionality reduction techniques, which as name suggests seeks to transform (linearly map) data into more convenient form with just a few key components. More formally, PCA seeks a linear transformation that converts a set of observations of possibly correlated variables into another set of values of linearly uncorrelated (orthogonal) variables under the condition that the first variable accounts for maximum variance, second one for most remaining variance and so forth. Mathematical details can be found in a number of sources [125, 126] and so will be discussed just briefly.

Mathematical background

PCA algorithm usually starts by centering data since offsets will induce fake variance not related to actual information in the set. Define j_{th} column mean as

$$\bar{\mathbf{x}}_j = \sum_{i=1}^{n} \mathbf{X}(i,j)/n$$

Subtracting this means vector from each row of $\mathbf{X}$ gives centered data matrix $\tilde{\mathbf{X}}$. Next step is to calculate a covariance matrix $\mathbf{C}$ with $\mathbf{C}_{i,j}$ the (co)variance between i_{th} and j_{th} spectra. From definition of variance of vector $\mathbf{x}$

$$var(\mathbf{x}) = \sigma_x^2 = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^2}{n-1}$$

covariance is similarly defined as

$$cov(\mathbf{x}, \mathbf{y}) = \sigma_{xy}^2 = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})}{n-1}$$

where n is length of both vectors. Generalizing this operation to a matrix of vectors,

$$\sigma_{\tilde{X}}^2 = \frac{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}}{n-1} = \mathbf{C}$$

where off-diagonal entries correspond to covariance while diagonal ones represent variance.

Finally, eigenvalues λ_i (represented as matrix $\mathbf{\Lambda}$ with $diag(\mathbf{\Lambda}) = \lambda_{1\ldots p}$) and eigenvectors α_i (represented as matrix $\mathbf{A}$ with $\mathbf{A}(:, i) = \alpha_i$) of $\mathbf{C}$ are obtained by solving

$$\mathbf{ACA}^{\mathrm{T}} = \mathbf{\Lambda}$$

where $\mathbf{A}$ is orthogonal. Actual iterative procedures used in libraries such as LAPACK are beyond scope of this thesis and can be found in appropriate documentation [127].

It is now possible to perform the desired PCA transformation, since explaining observed data with orthogonal components is equivalent to removing covariance between different variables - i.e. diagonalizing $\mathbf{C}$. Thus PCA can be reformulated as finding some $\mathbf{P}$ such that for $\tilde{\mathbf{Y}} = \mathbf{P}\tilde{\mathbf{X}}$

$$\mathbf{C}_Y = \frac{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathrm{T}}}{n-1}$$

is diagonalized. This however was already done above - eigenvectors of $\mathbf{C}$ turn it into

a diagonal matrix. Picking $\mathbf{P} = \mathbf{A}$, after some algebra it can be shown that

$$\begin{aligned}\mathbf{C}_Y &= \frac{(\mathbf{A}\tilde{\mathbf{X}})(\mathbf{A}\tilde{\mathbf{X}})^{\mathrm{T}}}{n-1} \\ &= \frac{(\mathbf{A}\tilde{\mathbf{X}})(\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}})}{n-1} \\ &= \mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}} \\ &= \mathbf{\Lambda}\end{aligned}$$

Thus, columns of $\mathbf{A}$ are identified as principal components (PCs) and $\mathbf{\Lambda}$ as matrix of respective variances, $\lambda_i = var(\tilde{\mathbf{X}} \cdot \alpha_i)$. Sorting columns of $\mathbf{\Lambda}$ in descending order (with appropriate $\mathbf{A}$ rearrangements) yields desired PCA results - a set of orthogonal vectors is obtained that maps data such that successively maximum amount of variance is explained.

It is convenient for analysis to define a measure of how much each new component contributes to old signals. This value is called a score, defined by

$$z_{i,j} = \tilde{\mathbf{X}}(i,:) \cdot \alpha_j \equiv \mathbf{Z}$$

Thus score is effectively a projection of j_{th} PC onto i_{th} spectrum. Using these scores, one can recover original data as

$$\mathbf{X}(i,:) = \bar{\mathbf{x}} + z_{i,1}\alpha_1 + z_{i,2}\alpha_2 + ... + z_{i,n-1}\alpha_{n-1}$$

under assumption that $n < p$. Limits change to $1...p$ otherwise, since there are at most p PCs. Moreover, sometimes only k PCs are obtained. This can happen due to data dimensionality ($k = min(p, n-1)$), machine precision issues, performance considerations, or by choice. Then, $\mathbf{A} \to \mathbf{A}(p \times k)$, $\lambda \to \lambda_{1...k}$, $\mathbf{Z} \to \mathbf{Z}(n \times k)$ and iteration for reconstruction is over $1...k_{th}$ component.

Another concern is how PCA treats missing data - in above discussion, there is no way to nicely incorporate this without losing any columns/rows where missing data is found. However, there exists another PCA algorithm called alternating least squares (ALS) which allows for such data defects [128]. It is an EM algorithm [129] that seeks to iteratively minimize square of reconstruction error on non-missing data only. However, since $\mathbf{A}$ and $\mathbf{\Lambda}$ are computed fully, all PCs are available as with PCA and complete dataset reconstruction is possible based only on known data. Detailed

description is omitted here for brevity and can be found in above references.

Interpretation

PCA should be treated as a dimensionality reduction technique that allows finding the natural basis of data in which most information is contained in fewest components (this property of PCA makes it useful for data compression). In this work, PCA and related techniques are applied to separate and quantify 'uncorrelated' changes in cellular content. It is expected that there are significantly fewer such changes than the total number of spectra collected, and that they explain more variance relative to noise and other undesired sources. Thus, relevant signals will be contained in first few PCs. By examining these along with corresponding score time/dose trends, it should then be possible to identify molecular signatures consistent with radiation-induced response.

However, since PCA is performed on centered data, PCs cannot be interpreted as 'building blocks' of original signal. Instead, they denote prevalence of certain features at a particular pixel. When combined with respective score values, these results make it possible to distinguish cells by presence of more/less positive features relative to negative ones. For instance, positive scores would correspond to a larger ratio of positive to negative features than dataset average.

There is a freedom to rescale PCs/scores by multiplying/diving with same constant respectively. In particular, flipping negative/positive features reverses scores and preserves relative trends. Such rescaling can occur during analysis and renders absolute score comparison meaningless - some method of normalization is necessary. Discussion of this is deferred to section 4.4.

2.3.2 Weighted PCA

Weighted PCA (WPCA) is the first and most obvious improvement of PCA which introduces weights to observations, variables, or even both. It has been used in several relevant applications, such as correction of systematic errors in photometric lightcurves [130] and increasing SNR of quasar redshift measurements [131].

Denote the weight of particular point by $w_{i,j}$, with meaning that higher weighted data should have more impact on results. Corresponding matrix is denoted by $\mathbf{W}$. Most general case of unique $w_{i,j}$ for every measurement does not have an analytical solution [132], although iterative approximation approaches have recently been pro-

posed [130]. However, the general case is not relevant to current work since no metric was defined to assign i weights (i.e. individual spectrum weights).

This only leaves assigning column weights w_j (forming vector $\mathbf{w}$ of size $1 \times p$). It is possible to use respective outlier or normality counts - for instance, $w_j = 1/f(\sum outliers)$, but this raises complex issues of what exact metric, significance level, and function to use. Instead, more straightforward approach is to use inverse variance, which effectively penalizes regions of high variability, namely $w_j = 1/\sigma_j^2$ where $\sigma_j = stdev(X(:, j))$. This is the same as applying eigenvalue/eigenvector decomposition not to covariance but to the *correlation* matrix.

Mathematical background

Define new centered data matrix as $\tilde{\mathbf{M}}$

$$\tilde{\mathbf{M}}(:, j) = \tilde{\mathbf{X}}(:, j) \cdot \sqrt{w_j}$$

and then continue PCA as usual, obtaining

$$\mathbf{C}_M = \frac{\tilde{\mathbf{M}}\tilde{\mathbf{M}}^{\mathrm{T}}}{n-1} = \frac{\tilde{\mathbf{X}}\mathbf{W}\tilde{\mathbf{X}}^{\mathrm{T}}}{n-1}$$

which is equivalent to finding covariance of matrix $\mathbf{W}^{1/2}\tilde{\mathbf{X}}$. Recalling that correlation is defined as $\frac{cov(\mathbf{x},\mathbf{y})}{\sigma_x \sigma_y}$,

$$\frac{\tilde{\mathbf{M}}\tilde{\mathbf{M}}^{\mathrm{T}}}{n-1} = \frac{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}}{\sigma_X^2(n-1)} = corr(\mathbf{X})$$

as claimed. PCs and scores are obtained as before, however former are not orthonormal (OR) due to the weightings. To recover proper PCs, it is necessary to multiply by the diagonal weight matrix

$$\mathbf{A}_{OR} = \mathbf{W}_d^{1/2}\mathbf{A} \qquad diag(\mathbf{W}_d^{1/2}) = \sqrt{\mathbf{w}}$$

Interpretation

WPCA algorithm seeks to reduce impact of variables with high variance, which could be indicative of large number of outliers that can strongly affect PCA results as discussed previously. Unfortunately, this may also be due to large actual biochemical differences as will be seen in chapter 5. In another sense, WPCA gives equal 'relative impact' to each pixel, unbiased by differences in absolute amplitude changes.

Once orthonormalized, WPCA components can be interpreted in same way as those of PCA. Scores can be compared in terms of relative score differences as will be described in section 4.4.

2.3.3 Robust PCA

Robust PCA is a name given to a number of related algorithms which aim to reduce PCA sensitivity to gross outliers. It has found extensive use in video processing [133], face recognition, and financial data analysis where infrequent but large outliers are expected to occur.

Mathematical background

In the current discussion of PCA, it was implicitly assumed that there exists some intrinsic low dimensional sub-space that explains majority of variance (i.e. data lies near such a sub-space). Then, it must be possible to separate $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{L}_0 + \mathbf{N}_0$$

where $\mathbf{L}_0$ is low-rank and $\mathbf{N}_0$ a small perturbation matrix. If $\mathbf{N}_0$ is small and contains only independent and identically distributed (i.i.d.) Gaussian variables, then minimization of $||\mathbf{X} - \mathbf{L}_0||$ subject to $rank(\mathbf{L}_0) \leq k$ is an equivalent reformulation of PCA (this criterion would be satisfied only if $\mathbf{L}_0$ was built from k highest variance principal components).

Robust PCA treats a more general problem of having

$$\mathbf{X} = \mathbf{L}_0 + \mathbf{S}_0$$

where $\mathbf{S}_0$ is a sparse matrix of arbitrary magnitude and unknown support (i.e. not necessarily Gaussian). With minimal assumptions ($\mathbf{L}_0$ not sparse, $\mathbf{S}_0$ pattern fairly uniform), it can be rigorously proven that a full recovery of $\mathbf{L}_0$ and $\mathbf{S}_0$ matrices is possible under condition that $rank(\mathbf{L}_0)$ is sufficiently low (see [134] for details).

However, practical implementation is not trivial due to requirement for effective robust estimates of means, eigenvectors, and eigenvalues. Some of the proposed approaches mimic PCA but compute a robust covariance matrix $\mathbf{C}$ with methods such as minimum covariance determinant (MCD). Others use robust principal component

pursuit (PCP) techniques directly [135], as well as outlier pursuit [80], LTS-subspace estimators, and several other [136].

Interpretation

Resulting PCs and scores can be interpreted similarly to PCA results.

2.3.4 Probabilistic PCA

Probabilistic principal component analysis (PPCA) is a reformulation of PCA under formal statistical treatment (i.e. as a latent variable model) first proposed by Tipping and Bishop in 1999 [137]. As all such models, PPCA has an associated likelihood function and it can be shown that there exists only one likelihood stable local maximum which also is the global maximum. Thus, results of maximum likelihood estimation will uniquely converge to PCA.

Some of PPCA advantages include aforementioned missing data tolerance with possibly better reconstruction as compared to ALS. It is also capable of component by component processing, which can mitigate numerical precision errors. Finally, it is possible to add a prior (i.e. initial belief) for PCs which is then taken into account during ML extremization to produce a posterior distribution (in which case the process is called variational Bayesian PCA).

Mathematical background

PPCA models observed vectors $\mathbf{x}_j$ based on an isotropic error model through the equation

$$\mathbf{x}_j = \mathbf{W}\mathbf{y}_j + \mu + \epsilon_j$$

where $\mathbf{y}_j$ $(k \times 1)$ is the vector of latent variables, ϵ the error term, and μ the mean, which serves as an offset. The respective probability distributions are

$$p(\epsilon_j) = \mathcal{N}(0, v_x\mathbf{I})$$

$$p(y_j) = \mathcal{N}(0, \mathbf{I})$$

with $\mathbf{I}$ identity matrix of rank k (same as size of $\mathbf{y}$), $v_x > 0$ the residual variance, and $\mathcal{N}(\mu, \Sigma)$ denotes normal distribution of mean μ and covariance Σ. As such

$$\mathbf{y}_j \sim \mathcal{N}(0, \mathbf{W}\mathbf{W}^T + v * \mathbf{I})$$

To determine latent variables, it is necessary to solve for $\mathbf{W}$, μ and v_x simultaneously. However, no analytical solutions exist and EM iterative processes must be used. As for PPCA convergence to plain PCA results, the proof is quite arduous and does little for a reader not well versed in statistical modelling. Moreover, the requirement for fast-converging EM algorithms further complicates things. This discussion is left to a number of relevant publications [128, 129, 137, 138] and omitted here for brevity.

Interpretation

Resulting PCs and scores can be interpreted similarly to PCA results. Furthermore, one can reconstruct data with estimated missing values and compare accuracy to initial dataset. Outliers can then be determined as those points with largest estimation errors.

2.3.5 Nonlinear PCA

A natural evolution of above linear methods is to perform a nonlinear PCA analysis, in which orthogonality is only preserved locally but principal component planes become curved manifolds. There are a number of prosed algorithms but most lack serious validation. A somewhat popular approach is based on auto-associative neural networks, also known as bottleneck networks [139]. It has been shown to excel in some bioinformatics applications, and moreover allegedly gives better missing data reconstruction than linear methods [140].

Mathematical background

Arduous detail [141] are again left to respective publications, with only general description given here. Artificial neural networks are in essence a crude model of biological neurons (nodes) and are comprised of an input node layer connected to several hidden layers which in turn connect to an output layer. Nodes can 'talk' with those in the next layer by propagating their values at a certain ratio. For instance, if input node A1 connects to three nodes B1/B2/B3 at weights of 1/2/3, then if a '1' is

supplied to A1 and signals propagated, B nodes take on values of 1/2/3 respectively. They may in turn connect to more layers, eventually reaching output nodes that can be read out. Such networks are trained by comparing output to desired one (i.e. using a training set with known results) and tuning weights iteratively using gradient descent or other algorithms.

Bottleneck networks are a special ANN case with inputs and outputs of same size but with hidden (inner) layers containing fewer nodes. For instance, $500 \rightarrow 10 \rightarrow 3 \rightarrow 10 \rightarrow 500$ is a bottleneck ANN with 500 inputs/outputs, 2 nonlinear layers of 10 nodes and a central hidden layer of only 3 nodes. During training, output target is set equal to the input, and weights tuned by square error on the reconstruction. This in effect achieves dimensionality reduction, since smaller hidden layers must account for as much variance as possible to minimize errors (going by above example, only 3 central node values must be used to reconstruct all 500 outputs). The application of this process to NLPCA is simply done by changing the tuning criterion, such that ANN now seeks to project data by the shortest distance onto a manifold. While this extracts k principal 'curves', there is no information as to their ordering. However, fairly effective modifications have been proposed [142] based on hierarchical learning, and are used in this work.

Interpretation

Nonlinearity of principal subspace means it is not possible obtain PCs and associate them with chemical signals as with methods above. However, scores are still computable via shortest distance projection of spectra, and can be directly compared with linear results.

Chapter 3

Materials and Methods I - Raman Data Collection

This chapter gives a brief description of protocols used to collect the datasets analysed in this work. Section 3.1 describes cell lines used and their properties, and protocols used for culturing, maintaining, and irradiating them. It also provides a rationale for choosing these cell lines for analysis. Section 3.2 outlines RS setup and procedures used in data collection.

Note that all work described below was performed by Samantha Harder as part of her MSc thesis, and for brevity only a summary is provided here. Refer to [78] and [143] for a more thorough description.

3.1 Cell line properties and protocols

For this work, data was available from four cell lines (each with duplicate experiments A and B) - H460(lung), MCF-7(breast), MDA-MB-231(breast), and LNCaP(prostate). Previous studies [78][144] have shown that these can be divided roughly into two categories in accordance with intensity of radiation response, with former two having high signal and latter two a weaker signal. Table 3.1 shows pertinent properties of these cell lines.

Because of time constraints and low likelihood to obtain new information in analysing all available data, it was decided to work with one representative cell line per group. For high signal category, H460 was chosen as it reproducibly demonstrated the strongest response. A low signal group was represented by LNCaP cell line, as it

Cell line	Response[143]	Source tissue	SF_2[76]	p53 status[144]
H460	Strong	Lung	0.64	wt
MCF-7	Strong	Breast	0.64	wt
LNCaP	Weak	Breast	0.71	mt
MDA-MD-231	Medium	Breast	0.27	wt

Table 3.1: Comparison of cell line properties. Chosen ones are in bold. Radiation response strength is evaluated by magnitude of corresponding principal component (explained in chapter 4). Abbreviations: SF_2 - surviving fraction after 2Gy dose, wt - wild type, mt - mutated type (nonfunctional).

demonstrated the most challenging behaviour and therefore was considered a reasonable worst case scenario. For brevity, following protocols omit minor detail as well as treatment of unused cell lines, but these can be accessed in respective references [78, 143].

3.1.1 Storage and maintenance

All cells were maintained in facilities of Deeley Research Center (DRC), part of British Columbia Cancer Agency's (BCCA) Vancouver Island Centre (VIC). Culturing protocols used were specific to each cell line but in general followed accepted standards [145, 146]. Cell stocks were obtained from American type culture collection (ATCC) and cryopreserved in liquid nitrogen until required. Prior to experiments, cells were thawed and resuspended in culturing medium, which was the same for both cell lines studied (see table 3.2). They were then seeded (with appropriate aliquot) into T-75 flasks suitable for surface attached cell lines and placed into an incubator at 37^oC and 5% CO_2. Cultures were maintained in exponential growth phase until required by reseeding (splitting population at ratios indicated below) at 50-60% confluency, generally every 3-4 days.

Cells were prepared for irradiation by seeding 24 T-75 flasks via same process as above and then left to incubate for 96hrs to reach approximately 50% confluency.

Cell line	Media Composition	Resilience	Splitting Ratio
H460	RPMI-1640 base + 10% FBS + 25mM HEPES + L-Glutamine	High	1:30
LNCaP		Low	1:3

Table 3.2: Cell maintenance parameters. Abbreviations: RMPI - Roswell Park Memorial Institute, FBS - fetal bovine serum, HEPES - 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid (buffering agent). Resilience refers to subjective sensitivity of cells to stresses such as late reseeding or mechanical vibration.

3.1.2 Irradiation and collection

Growth medium was replaced one hour prior to irradiation. All flasks were then simultaneously removed from the incubator and transferred carefully to the LINAC, with care taken to minimize time spent outside the incubator. Irradiation was performed in BCCA-VIC facilities, using Varian 21EX LINAC in 6MV photon mode. Geometry was programmed to deliver 6 Gy/min to 30.0cm x 30.0cm field at isocentre, where flask surfaces containing cells were positioned. Appropriately sized 5cm water equivalent material was placed on both sides to ensure sufficient dose buildup and backscatter. Single fractions were delivered to three flasks at a time, with monitor units required calculated on VIC MU Calculator version 5.05 (which contained clinically used calibration tables). Target doses were chosen as 0Gy (control), 2Gy, 4Gy, 6Gy, 8Gy, 10Gy, 30Gy, and 50Gy, representing a reasonable clinical range and two high dose fractions in case of insufficient response to low doses. Graphic representation of this procedure, as well as collection timeline, are given in figure 3.1.

Flasks were harvested each day for every dose, with resulting post irradiation times of 18h, 42h and 66h. For each flask, cells were detached and collected, with 35% transferred into PCR tube with phosphate-buffered saline + 3% FBS solution. This tube was centrifuged to form a pellet and placed on ice for same day Raman analysis. Remaining cells were used for cell cycle and viability counting as well as western blots, but that data was not quantitatively used in this work (beyond ensuring sufficient cell viability).

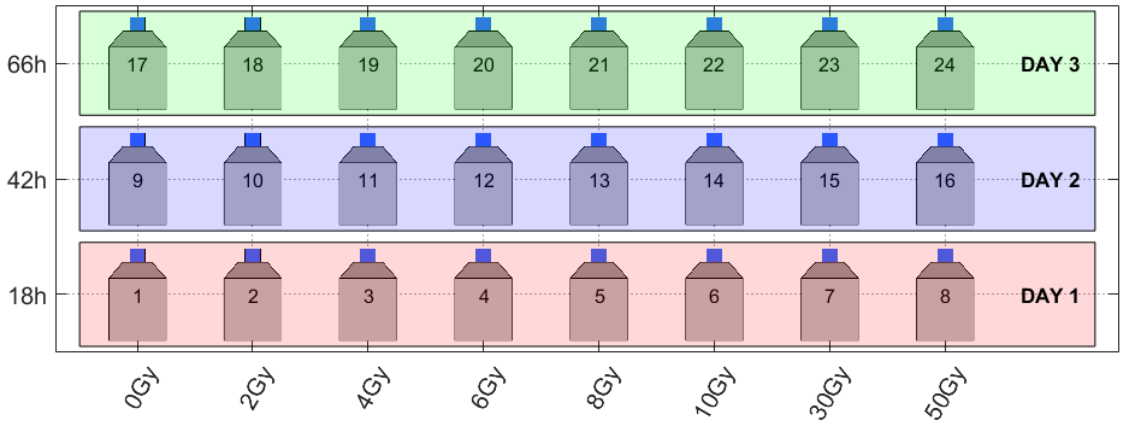


Figure 3.1: Visualization of flask doses and collection times. Numbering represents nomenclature to be used during data analysis.

3.2 Raman spectroscopy

3.2.1 Experimental setup

Raman data was collected with a Renishaw inVia confocal Raman microscope (Renishaw plc, Wotton-under-Edge, Gloucestershire, UK) located at UVic. This system consisted of 785nm continuous wave diode laser coupled to a computer controlled grating/filter/detector assembly and microscope (Leica Microsystems, Wetzlar, Germany) with motorized stage and 100x dry objective with NA of 0.9 (Leica). All components were controlled with WiRE 3.0 software package (Renishaw), which also performed calibration and data collection routines.

System parameters

Laser power was measured to be $55 \pm 5\,\mathrm{mW}$ and sampling volume was ($1.2 \pm 0.3\,\mu$m, $5.8 \pm 0.3\,\mu$m, $14.8 \pm 0.3\,\mu$m) in (x,y,z) dimensions respectively. Resulting power density in the sample was $0.53\,\frac{\mathrm{mW}}{\mu\mathrm{m}^3}$. Data collection was performed over several months and no noticeable power degradation occurred during that time.

Samples were held in place on a linearly encoded motorized stage (Prior Scientific Inc., Rockland, MA, USA) with precision of $0.1\,\mu$m in x, y and z directions. Reflected light dispersal was done with 600 line/mm diffraction grating, and collected onto Andor iDus DU-401A-BR-DD deep depletion temperature controlled CCD detector (Andor Technology Ltd., Belfast, Northern Ireland). Detection region had 1024×127 pixels (X $\times$ Y) of $26 \times 26\ \mu\mathrm{m}^2$ size, with dispersal along X and binning along Y

directions. Of those, 600 rows were used in collection for final WN region of 423.493 cm^{-1} to 1819.470 cm^{-1}.

3.2.2 Data collection

Prior to each daily run, a reference silicon spectrum was collected and peak position/intensity recorded, with results of 19000 counts $\pm 9\%$ for intensity and $520.252 \pm 0.119\,cm^{-1}$ for wavenumber indicating fairly little system drift throughout the experiment.

Cell pellets (obtained via protocol of section 3.1.2) were transported to UVic and kept on ice until acquisition, from 1 to 4 hours. For each set, a pellet was extracted from PCR tube, spread evenly on 5mm thick MgF_2 disk, and placed on the motorized stage. A representative selection of cells were located under visible light that matched several criteria - regular and live appearance, no obstructions by debris, and top layer location. Laser was aimed at such cells, focused on surface, and then stage moved 10 μm upwards (+z) to shift the focal point to expected position of the cell interior. After 10s of photobleaching at full power, a 10s acquisition (also at full power) was performed and spectrum examined for obvious corruption such as cosmic rays, spurious fluorescence or signs of dead cell prior to storing it. This process was repeated until 20 acceptable spectra were collected. Over 3 days, 24 cell samples were analysed in each trial such that final datasets consisted of 480 spectra. Processing of these is described in the next chapter.

Chapter 4

Materials and Methods II - Spectral Processing

This chapter explains in detail the preprocessing steps and subsequent multivariate algorithms applied to the RS data, as well as comparison metrics and underlying software architecture. Section 4.1 describes the general data flow, software modularity, and numerical precision considerations. Section 4.2 explains preprocessing routines and selection of parameters to maximize signal-to-noise ratio. Section 4.3 provides implementation details of multivariate analysis methods. Chapter concludes with section 4.4, which gives an overview of available comparison and performance metrics.

4.1 Software stack

4.1.1 General architecture

In writing code for this work it was quickly realized that an unfortunately common coding paradigm of 'all-in-one' scripts would quickly become burdensome and difficult to maintain and expand. Hence a framework was developed that allowed for modular data processing with a strong suite of support functions such as checkpoints, logging, and performance profiling. General structure is shown in figure 4.1.

Majority of code was written in MATLAB language with extensive use of object-oriented programming (OOP) functionality. Certain routines also called external code in C (.mex files) and R (via R-DCOM interface), for reasons of performance and access to additional statistical routines respectively.

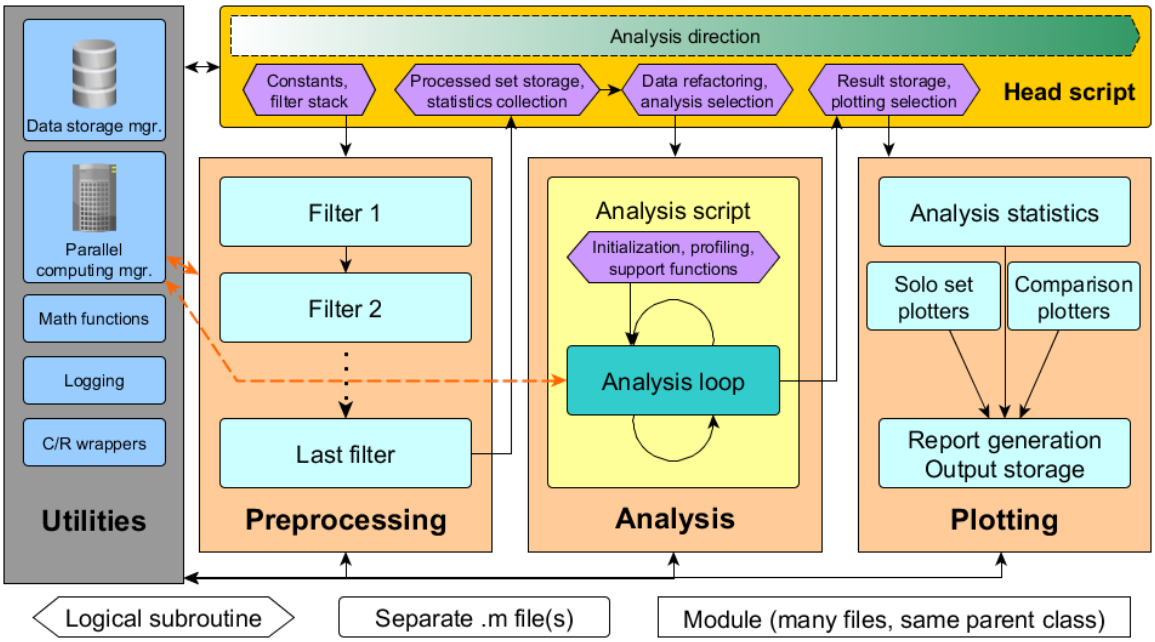


Figure 4.1: Software stack used to process RS data. Modules, m-files, and functions are denoted with different shapes as given on the bottom edge of the figure. Functionally similar components are of same colour.

Final analysis code was comprised of over 41000 lines of code (LOC) in 145 files (excluding any libraries). It was run with MATLAB 2015a (8.5.0.197613) runtime (The MathWorks Inc., Natick, MA, USA) under Windows® 8.1 64-bit on Lenovo ThinkServer TS140 (Intel® Xeon® 1225v3 - 4c@3.2GHz, 16GB ECC RAM).

Due to the variety of libraries used in this work, attributions will be given in respective sections. Complete code is made available upon request to author under GNU General Public License version 3 (GPLv3).

4.1.2 Dealing with machine precision

MATLAB environment is capable of performing most operations with either 31 (single) or 63 (double) bits of precision, and complies with IEEE 754 (standard for floating-point arithmetic) [147]. This counting system is almost ubiquitously used in modern programming languages but a detailed discussion is beyond the scope of this work. Important practical conclusion is that this limits the ability to represent an arbitrary number to the closest value available in floating point (fp) arithmetic system - a concept of floating point precision. As a practical demonstration of possible issues, consider the following computation:

Listing 4.1: Floating point precision example

```
>> 1 - 3*(4/3 - 1)
ans =
2.220446049250313e-16
>> eps(1)
ans =
2.220446049250313e-16
```

This shows how there is a loss of precision equal to distance (called ϵ) between 1.0 and next larger number. While rarely a problem for simple computations, such errors can compound for iterative algorithms resulting in significant accumulated errors, especially when linear algebra operations such as single-value decomposition (SVD) are involved (in that case, any dimension with eigenvalue $\approx \epsilon$ is considered to be ≈ 0, and does not count towards rank of the corresponding matrix). Several approaches can be used to manage this - most obvious solution is to utilize variable-precision arithmetic (VPA, part of symbolic math toolbox) or equivalent 3rd party library. Unfortunately most of MATLAB algorithms do not function well under vpa system, and those that do incur performance losses of several orders of magnitude [148]. Although commercial solutions are available that claim to have near 300x(vs VPA) speed increase in common linear algebra operations [149], it was not considered worthwhile to purchase them due to necessity of converting existing code and availability of easier mitigation strategies.

Instead, in this work data was rescaled (by multiplicative factors that did not affect any conclusions) and algorithm parameters (stopping criterion, iteration limit, data dimensionality, etc.) tuned to minimize precision loss. Where possible, analysis was run with a variety of initial random number generator (rng) seeds, and parameters tuned to obtain consistent output. Specific settings are given in appropriate sections.

4.2 Preprocessing

This section describes components of the pre-processing pipeline in the order they were applied. Multivariate analysis algorithms are presented in the next section.

4.2.1 Background subtraction

Removal of fluorescence and other background signals is a key aspect of Raman spectroscopy that is crucial to obtaining analysable spectral sets. While a constant baseline would have been easy to remove, in this work datasets contained varying backgrounds due to changes in amounts of biological material and MgF_2 substrate within sampling volume. Based on comparative studies [150] and experience of previous works, iterative SG [109] signal removal method was chosen for this work and is described below.

Let S_i denote i_{th} original spectrum, B_i^k the k_{th} iteration of baseline estimate for that spectrum, $SG()$ the filter function, and S_i^{br} the final result. Algorithm starts by obtaining first baseline estimate $B_i^1 = SG(S_i, \text{ws})$ from S_i, and then computes intermediate signal estimate as $S_i^1 = min(S_i, B_i^1)$. This is used to get next baseline estimate $B_i^2 = SG(S_i^1, \text{ws})$, and cycle is continued for further 19 times to obtain final baseline estimate $B_i^{20} = SG(S_i^{19}, \text{ws})$. Baseline removed signal is then given by $S_i^{br} = S_i - B_i^{20}$.

Output of this process depends only on the SG filter, which can be treated as a function that returns a vector of same length as its input, $\text{Data}_{\text{smoothed}} = SG(\text{Data}, \text{window size})$. While full discussion about SG algorithm (and its nice properties such as peak shape preservation) is beyond scope of this thesis, it is important to understand the single tunable parameter, window size. It controls, as name suggests, the size of sliding spectral window that is fit with a polynomial using LSQ regression. Expressing this parameter as fraction of whole range, it can take values within $[0, 1]$. Larger settings correspond to wider window and less conformal smoothing, while smaller values result in less smoothing. For this work, value of 0.07 was chosen based on previous studies [143] and visual inspection of known signal regions. A more detailed discussion is given in section 5.1.

4.2.2 Normalization

While background removal accounted for variability of extraneous material in the sampling volume, there were also differences in amounts of relevant cellular contents. This was due to inevitable variability in parameters such as cell size, shape, laser positioning, and others. Without normalizing for this, data analysis would be skewed by false sources of intrasample variability.

Previous work [144] has shown that peak normalization is not suitable for single-

cell Raman, since one of highly prominent features (glycogen, a carbohydrate) changes dramatically between irradiated and control populations. Furthermore, specific peak normalization reduces sensitivity to selected regions, which is an undesired effect. Of the other two techniques available, total area normalization was used due to its more robust behaviour and good prior results.

4.2.3 Shifting

As discussed previously, by collecting spectrum of pure silicon prior to each collection run it was established that system drift at all times was significantly smaller than 1 pixel and had no systematic trends. Nonetheless, it was important to correct for since this error was known [144] to contaminate principal components.

Shifting was done by first fitting a single peak Gaussian function to a region primarily containing the phenylalanine peak (990cm^{-1} - 1013cm^{-1}), which was chosen as it was by far the sharpest and one of strongest features, permitting reliable calibration for all datasets. The fitting algorithm used was nonlinear least squares regression via trust-region method, part of standard MATLAB library. All options were left on default settings except TolX, which was set to 1.0×10^{-7} to improve fit accuracy.

For each spectrum $S_{i=\{1-n\}}$ in the set, fitting parameter corresponding to distribution mean was extracted (b_i). First spectrum mean, b_1, was then selected as target and other spectra shifted by respective offsets, $\Delta_i = b_1 - b_i$. This was performed through linear interpolation between available points (using 'interp1' MATLAB routine). Where necessary, endpoints were extended with neighbouring values to preserve dimensionality. This data modification did not have any effect on analysis since there were no interesting signals near endpoint regions.

Shifting algorithm was repeated a second time on modified spectra $S'_{i=\{2-n\}}$ (keeping b_1 same) and resulting set $\{S_1, S''_{i=\{2-n\}}\}$ along with X axis of first spectrum (replicated n times) was taken as final preprocessed dataset. Justification for double alignment will be presented in section 5.1.

4.2.4 Outlier, significance, and normality tests

Significance tests

As will be shown below, numerous analysis results in this study gave distributions of scores or other quantitative values. It was necessary to determine whether these point

distributions were statistically significant, in other words whether they were drawn from different probability density functions. This was accomplished with two-sample t-test [151]. Null hypothesis is that two vectors are independent random samples of two normal distributions with same mean and same (unknown) variance. Alternative hypothesis is that these vectors come from distributions with different means. Test output was p-value (p_{val}) - the probability to obtain a result as extreme. For a test at significance level α, null hypothesis was rejected if $p_{val} \leq \alpha$. Specific α values used are given in respective sections.

Normality tests

Assumption of normality is key for many parametric analysis and outlier rejection methods, necessitating preliminary normality tests to determine whether given measurements came from a normal distribution.

A variety of methods are reported in literature, such as Kolmogorov-Smirnov, Lilliefors, Anderson-Darling (AD), Shapiro-Wilk (SW) [152], as well as newer, less established ones [153]. Several recent papers [154, 155] seeking to compare their statistical power (loosely speaking, accuracy) concluded that SW and its derivatives, Shapiro-Francia (SF) and Chen-Shapiro (CS) are most suited to small sample sizes (~ 30) and situations where type of non-normality and alternative distribution are unknown.

In this work, SW normality test is used for platykurtic(kurtosis < 3) and SF for leptokurtic(kurtosis > 3) data, at various significance levels. Null hypothesis is that values are a random sample from normal distribution of unknown mean and variance. Alternative hypothesis is that values do not come from a normal distribution. As before, if p-value returned by test was less than significance level specified, null hypothesis was rejected.

For completeness, it must be noted that some authors [156] dispute validity of using such tests to determine whether normality-reliant algorithms can be applied. This is because of swamping, whereby outliers themselves cause otherwise well behaved distribution to fail [157] normality tests. Instead, some suggest graphic approaches such as Q-Q plots. However, in this work it was considered infeasible to implement any supervised technique, and so most reliable normality tests were judiciously used.

Outlier detection

Outlier detection and rejection are fairly large and admittedly contentious topics, with many sources suggesting usage of outlier-resistant (robust) techniques instead. In this work, while robust methods such as RPCA were applied, other approaches required detection and removal of spectral defects so as to increase analysis sensitivity.

As with normality tests above, a wide variety of approaches exist, such as graphical (normal probability plot), hybrid (box plots), or those based on quantitative models. It was again considered impractical to use any supervised solution and so only latter category was applicable to current work.

Multiple outlier detection algorithms have been proposed such as Dixon's Q test [158], generalized extreme studentized deviate test (gESD) [159], Grubbs' test [160], modified Thompson Tau (TT) test, and others [161]. These tests each have specific application conditions, and remove varying amounts of outliers, although all are parametric and demand data normality.

In this work two different methods are used to reach a consensus vote on whether a certain point is an outlier - gESD and TT with robust estimators. These methods were chosen based on their flexibility (TT removes outliers iteratively until a preset limit, gESD dynamically determines number of outliers) and recommendations in several handbooks [157]. (In those comparative studies explored, some indicated that no method has definitive advantage [162] while others advocated for more advanced techniques [163]).

4.3 Multivariate analysis algorithms

This section provides details on algorithm implementations and corresponding parameters, where applicable.

4.3.1 PCA

To establish PCA baseline, MATLAB function pca() was used to perform the analysis. By default, it runs single value decomposition (SVD) to compute eigenvalues/eigenvectors. This is implemented via LAPACK routine and considered very reliable. An ALS algorithm option was also available, with implementation based on work of Roweis [129]. It was used only in missing data cases. Number of computed PCA

components (k) was always set to match other methods, and relative variances recalculated accordingly. Where required, a flip was performed to align PCs.

4.3.2 Weighted PCA

WPCA was implemented with MATLAB pca() function by specifying column weights as inverse of column variances. Orthonormalization was done manually as described above. For reference, both original and OR components were given. As with PCA, paremeter k was reduced to match other datasets and PC flips performed if necessary.

4.3.3 Robust PCA

In this work two methods were attempted. First method was based on augmented Lagrange multiplier (ALM) technique [164]. It has shown good performance in video signal separation [133], and can also been used for matrix completion (i.e. data reconstruction). An essentially unmodified MATLAB implementation provided by Perception and Decision Laboratory at University of Illinois Urbana-Champaign was used, which returned separated $\mathbf{L}_0$ and $\mathbf{S}_0$, former of which could then be processed using regular PCA. The major tunable parameter was λ, which scaled the sparse error term of cost function. Unfortunately, during preliminary testing it was found that dependence on this value was extreme. With default setting ($1/\sqrt{n}$), virtually no noise was detected, while changing it by only a factor of two moved the whole signal into outlier class. Automated tuning of λ to maximize score distances (see below) was attempted but found to be problematic due to algorithm instability. As such, this method was not explored further.

Second method used was a modified version of well known code robpca() from LIBRA (Library for Robust Analysis) by ROBUST group at KU Leuven. It is a hybrid method using PCP optimization with MCD estimator that returned PCs/scores directly. Several comparative studies have found it to be one of the best RPCA methods for high dimensionality ($p > n$) cases [165]. The fraction of outliers this algorithm should resist is given by $1 - \alpha$, where α is the actual input to the function (corresponding to fraction of 'good' data). Default α of 0.75 was changed to 0.8 to match the 4/20 outliers that are checked in formal rejection tests. Since in this method number of components has to be specified, k was varied within 2-5 range and best radiation response results presented.

4.3.4 Probabilistic PCA

First candidate was built-in MATLAB ppca() function, however it was found inadequate due to poor EM algorithm which caused significant numerical errors. Moreover, reliability was questionable since author found a bug in termination conditions. This was reported to Mathworks under case number #01029162, acknowledged, and fixed in R2015a release.

As such, an alternative algorithm was implemented following work of Ilin and Raiko [128], based on their MATLAB package for missing value PCA methods. Testing on a sample dataset was carried out to ensure expected results, namely that same PCs/scores are returned without missing data and that for random defects the departure from known data is reasonably small and smooth (near perfect correlation to actual PCs, insignificant score differences, low RMS error of reconstruction). During actual analysis, outlier % (via significance level) was limited to stay within range where random reconstruction performed adequately.

4.3.5 Nonlinear PCA

NLPCA MATLAB library based on work of Matthias Scholz was used [139]. Neural network options were left at their defaults (5 layers, hierarchical learning, weight decay) and first 3 components computed. In place of usual PC plots, a projection of obtained manifolds onto PCA score 3D space (i.e. X/Y/Z - PC1/PC2/PC3 score) was given to better visualize data non-linearity.

4.4 Performance evaluation

In this work, *absolute* performance is defined as quantitative measure of ability to detect radiation-induced response - a stronger signal at lower doses would indicate better results. Correspondingly, *relative* performance is defined by how well the method fares compared to the baseline PCA signal. Note that in both cases certain rescaling/renormalization may occur and does not affect conclusions.

4.4.1 Explained variability

Direct comparison of λ values is not useful due to rescaling freedom of most methods. However, % variability is quite useful as long as number of components it is computed

against is kept same. Lower relative values for radiation-related components can be indicative of worse performance since more weight is kept in other PCs (either assigned to different biochemical signals or lost in noise-filled PCs). However, for robust methods such behaviour can also occur due to outlier suppression and/or better ability to resolve separate signals, and so examination of respective scores is necessary. Variability ratios (i.e. PC2/PC1) can be compared between methods to determine which biochemical processes are favoured, and possibly tune for highest radiation response signal.

4.4.2 Principal components

PC peaks

Peaks in principal components can be directly attributed to particular vibrational modes of chemical bonds, which allows assigning components to respective biochemical processes, such as radiation induced accumulation of glycogen or cell cycle variation of nucleotides. By noting changes in relative intensities of these, it is possible to infer which signals become more or less important relative to PCA baseline, and as above determine algorithm affinity to particular processes.

4.4.3 Principal component scores

Score significance (intrasample)

Comparison of score distributions (to determine whether statistically significant changes were detected) was done via two-sample t-test at 5% significance. Day 1, 0Gy population served as reference for day 2/3 0Gy data. All other doses were compared to 0Gy populations at respective times. In this manner, both drifts of control unirradiated population and dose dependent trends were tested. Better absolute performance corresponds with first significant result at lower dose/time.

Score significance (intersample)

Above significance results were also used for relative performance by comparing detection thresholds. Moreover, a t-test was applied at 10% between matching (day/dose) distributions, although it was not expected that results would change so drastically so as to satisfy it. Finally, a comparison metric was defined based on average of

raw p-values in certain dose range, with lower results corresponding to 'more significance'. Specifically, high dose set results would be computed as average of 2Gy-50Gy p-values (at specific time) whereas low dose computation would only use 2Gy-10Gy values. While crude in that p-values are not a linear measure of signal, this calculation still gave a good idea of detection confidence differences between methods.

Relative score distance (intersample)

Beyond significance results, a useful measure of discriminatory capability is score distance from reference distribution, defined as difference of means of respective batches. To account for possibly different score scaling, a method is required to appropriately renormalize score distributions.

There are three reasonable candidates - via average standard deviation, via minimum-maximum (MM) distance, and by reference (D1-0Gy) point. However, single point scaling does not account for possible score shifts and hence was not used. Remaining two methods each have certain advantages - standard deviation is a good measure of confidence on the means, and was already used in previous works to compare PCA results of various preprocessing methods. Moreover, it is not affected by score shifting or scaling. For instance, consider identical score distributions with just a multiplicative factor difference - since mean and deviation will both change linearly, relative score distance will vanish as desired. Min-max normalization is more useful to determine how fast a certain detection level is reached relative to extrema, most of which are expected at D1-0Gy reference point and for high-dose long-time batches. Assuming that methods at least agree about these extrema, min-max scaling then allows comparison of their sensitivity relative to those measurement. As above, same nice properties of offset and scaling invariance apply. For plotting, standard deviation scaling method was used since it provided best visualization.

Chapter 5

Results and Discussion I - Data Selection and Preprocessing

This chapter gives an overview of data preprocessing and outlier detection. Section 5.1 validates the preprocessing routines with demonstration of intermediate results while section 5.2 explores outliers and demonstrates necessity of robust analysis techniques. For brevity, just H460B set is considered with any deviation of LNCaP data specified where needed. Respective plots for H460A, LNA and LNB are given in appendix A.

5.1 Preprocessing validation

Preprocessing Raman data is a crucial step in ensuring consistency between different spectra prior to attempting any further multivariate analysis. Given the large number of tunable parameters, it is important to confirm that the procedures yield reasonable and consistent results.

Baseline removal

As explained previously, chosen baseline removal technique is based on iteration of SG filter. Two tunable parameters are available - window size and number of iterations. The former was set at 0.07 (i.e. 7%) based on comprehensive analysis by Matthews et al. [144], where it was found that for regions with possible contamination of signal, highly conformal 0.03 window is optimal while 0.07 gives best results for clean, broad featured regions. Final recommendations were to combine the 0.03 window for lower half of spectrum and 0.07 for upper half. However, since that study several

improvements were made to Raman system (better microscope slides, new camera with higher SNR, and more refined collection technique) that allowed for uniform usage of 0.07 window size. A comparison of 0.03 and 0.07 window sizes is shown in figure 5.1.

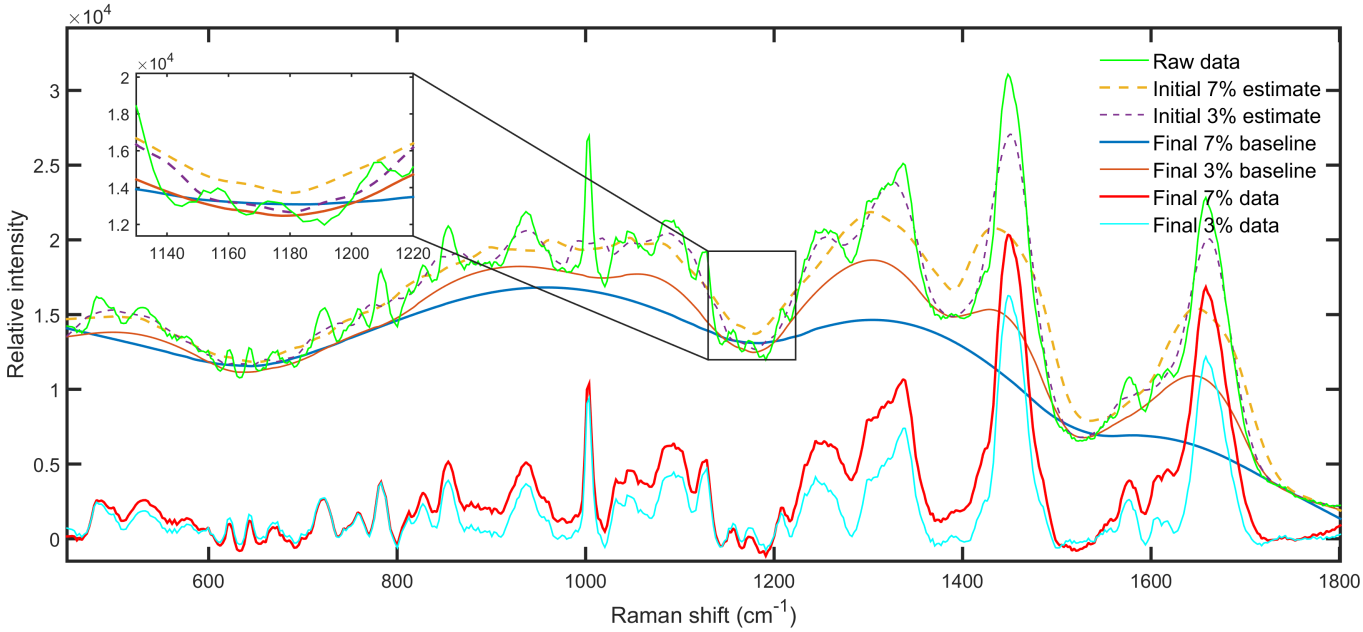


Figure 5.1: Example of baseline removal with two different window sizes.

It is important to emphasize that while the above study only optimized window size for PCA analysis, this same value was used for all methods in present work since performance comparisons (section 4.4) are most meaningful when analysis is done on identical datasets. Furthermore, due to similarity it was expected that PCA settings would be near optimal for most methods.

The second tunable parameter, number of iterations, was set at 20 not due to performance constraints but because further iterations did not contribute significantly - 20 was sufficient to obtain reliable baseline estimate for all datasets studied. To demonstrate this, 'sum of changes' per each step was calculated as sum of absolute differences between new and old baselines at each pixel, normalized to the first value. It was found that after 20 iterations relative changes never exceeded 10% for any dataset, which was deemed a sufficient criterion for convergence. A plot of this data for H460B-1 spectrum is given in figure 5.2.

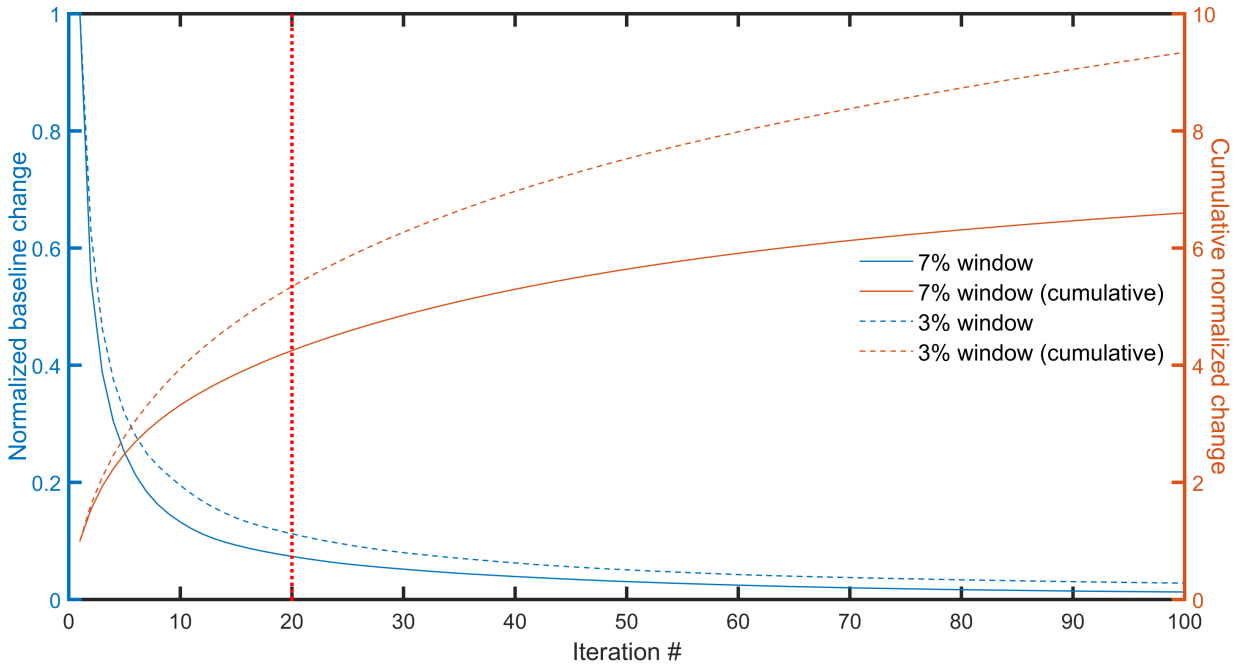


Figure 5.2: Details of SG SRM convergence. Vertical red line denotes chosen iteration cutoff.

Normalization

Total area normalization is, as discussed above, a reasonable approach and was already validated in previous studies [144]. Its necessity is clearly demonstrated by the large spread of non-normalized spectra, as is shown in section 5.2.

Shifting

Accuracy of shifting algorithm was verified by calculating Gaussian fitted curve offsets after single and double application of procedure described in section 4.2, and is shown in figure 5.3. Note that all major peak shifts happened in between trials, with largest between days consistent with expectations. Average absolute peak deviation was $(0.368 \pm 0.2)\,\mathrm{cm}^{-1}$, within 0.5 pixels. It was reduced to $(0.0211 \pm 0.01)\,\mathrm{cm}^{-1}$ after first iteration and $(0.00384 \pm 0.007)\,\mathrm{cm}^{-1}$ after second one. This validates necessity of two iterations to fully align spectra. It is likely that given more calibration peaks (and hence better fit quality over more points), shifting process could be reduced to just a single iteration but this was not possible for our studied spectral window. LNCaP dataset corrections had similar characteristics.

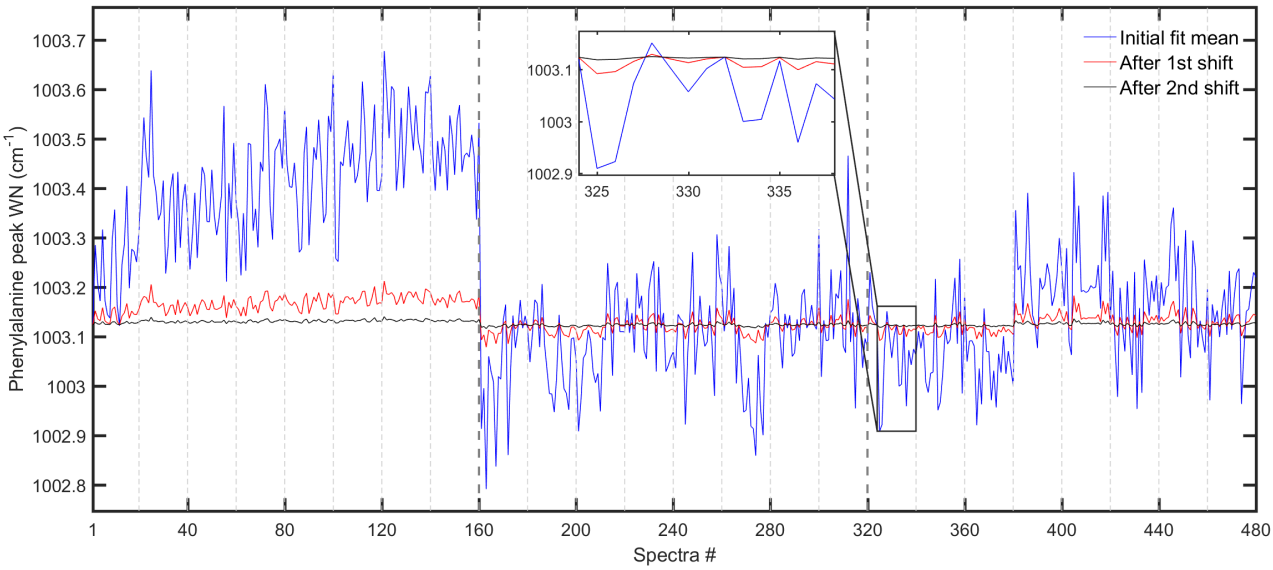


Figure 5.3: Phenylalanine peak offset values before shifting, and after 1 or 2 iterations of correction algorithm. Gray dashed lines denote different 20-sample batches with bolded ones separating days 1, 2, 3.

5.2 Data defects

This section briefly demonstrates that collected data has many imperfections that need to be dealt with in automated and robust manner.

5.2.1 Spectral variability

As is discussed in chapter 2, there exist many possible sources of variability, such as low energy cosmic rays, spurious debris within sampling volume, or other unexplained sources. It is illustrative to visualize spectral distributions at various stages of preprocessing. Figure 5.4 shows percentile areas of raw, background corrected and corrected+normalized spectra respectively for batch #1 of H460B dataset. From average standard deviations, it is clear that each step reduced variability as desired. However, it is also apparent that certain regions exhibited suspiciously high spread even for final data.

Another issue was that variability did not remain constant in time even for samples of same dose. Visualization of best case scenario, 0Gy control samples, is provided in figure 5.5. Note how the middle 1σ band remains near constant while outer regions spread out significantly with time, which is corroborated by increasing average deviation. Comparing problematic regions to figure 5.4 indicated that the majority of them

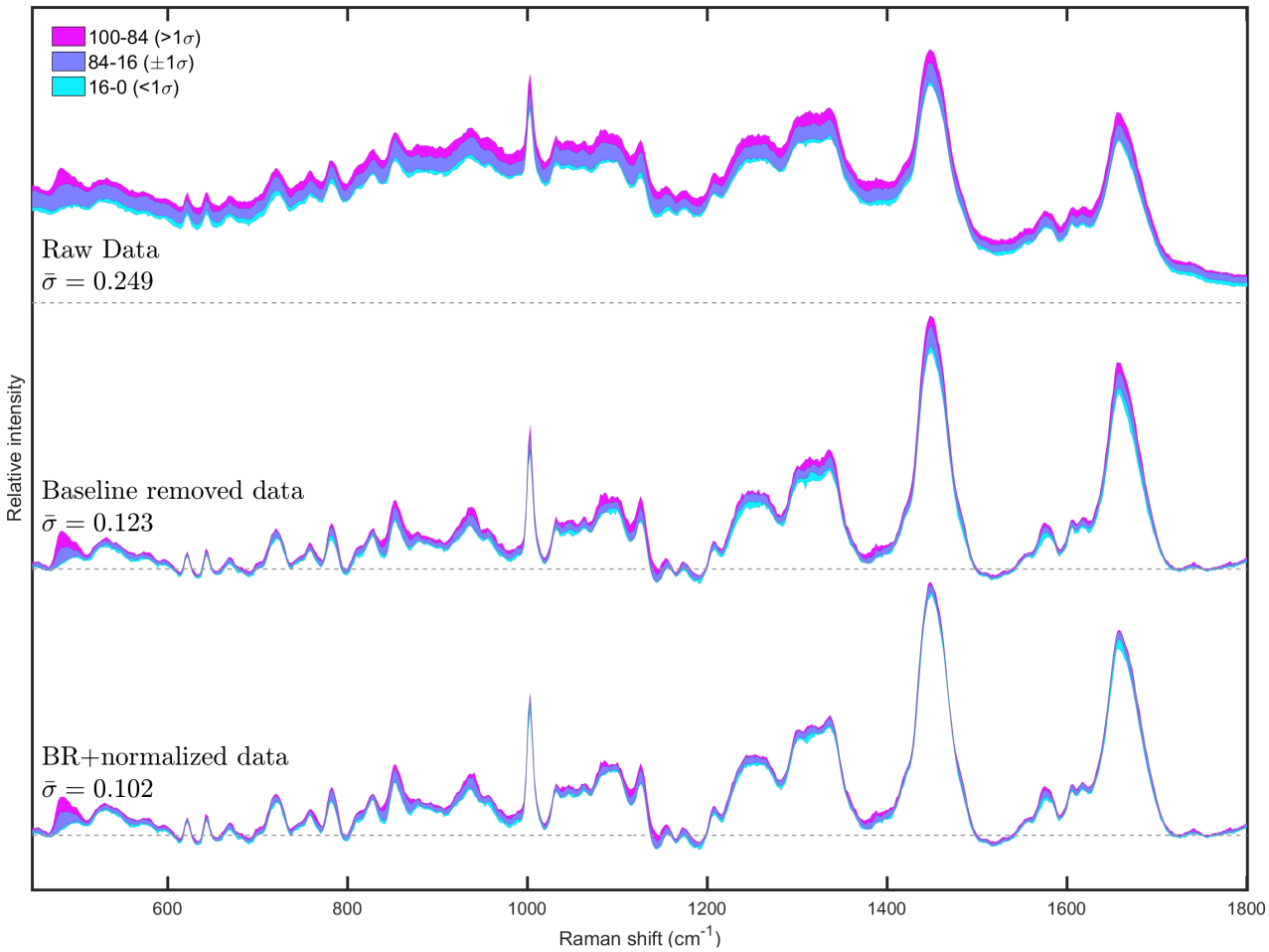


Figure 5.4: Variability of D1-0Gy data batch at various stages of preprocessing. Data was rescaled for easier visualization. Note that average standard deviation corresponds to average σ of 20 measurement at each pixel, and has arbitrary units. It is only meaningful for relative comparison purposes.

match (such as at $480\text{cm}^{-1}, 1335\text{cm}^{-1}$), suggesting possible differentiator signals in those areas.

5.2.2 Spectral outliers

The above discussion has demonstrated visually that spectral variability is a significant concern - it is highly non-uniform over the spectral window of interest and varies depending on collection time and dose. Results of quantitative outlier detection and removal that was applied to deal with this are presented below.

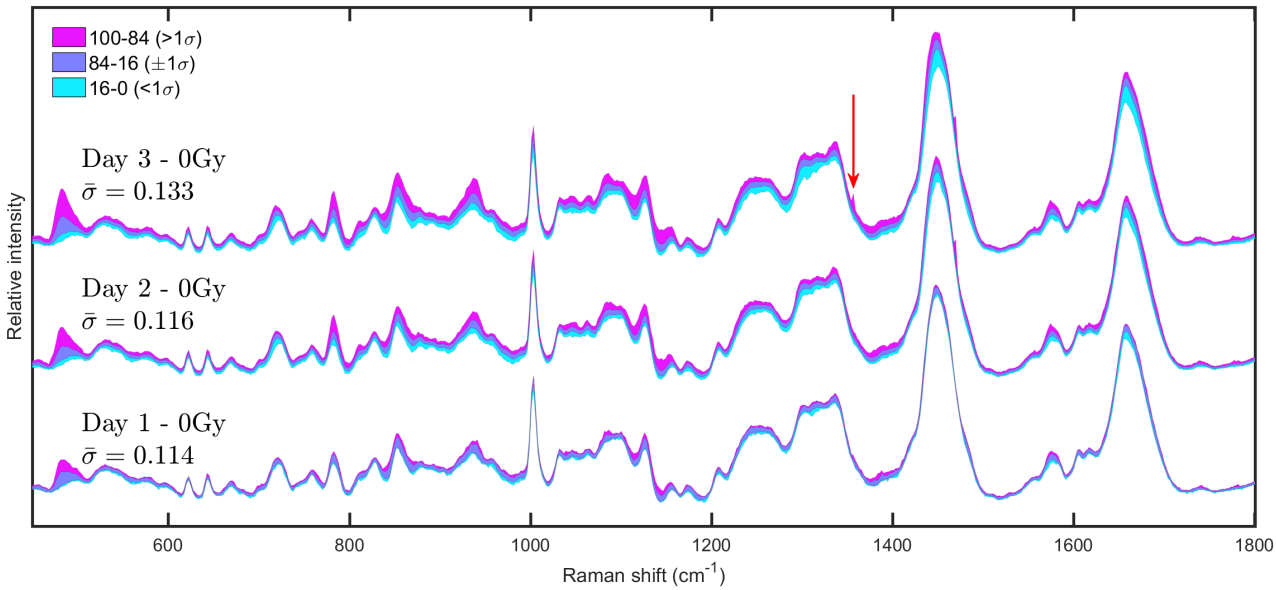


Figure 5.5: Variability of 0Gy batches collected at different days. Average standard deviation corresponds to average σ of 20 measurements at each pixel, and has arbitrary units. It is only meaningful for relative comparison purposes. Note that red arrow denotes a clear outlier that was missed during manual rejection.

Normality

Results are best displayed as 2D colour map, where each region represents the test result of specific batch at a particular wavenumber (i.e. result of testing 20 points), with colour indicating the outcome. This is shown for normality tests in figure 5.6.

Note that chosen significance level of 5% is for demonstration purposes only and was varied in actual analysis. For convenience this map was aligned with a representative percentile spectrum plot, and not surprisingly the same problematic regions noted above have yielded highest test failure counts. Especially worrying is the large non-uniformity between batches, such as D1-50Gy vs D3-50Gy. In general, both higher dose and later time batches tended to fare worse but just slightly.

Outlier removal

Results of normality tests were used as a logical mask to determine where to apply formal outlier removal tests. Again, results are best visualized as a 2D map, shown in figure 5.7a.

Of the two methods used, TT was significantly more aggressive with 25950 (9.29% of data) points flagged compared to 3605 (1.29%) for gESD. Combination rejected

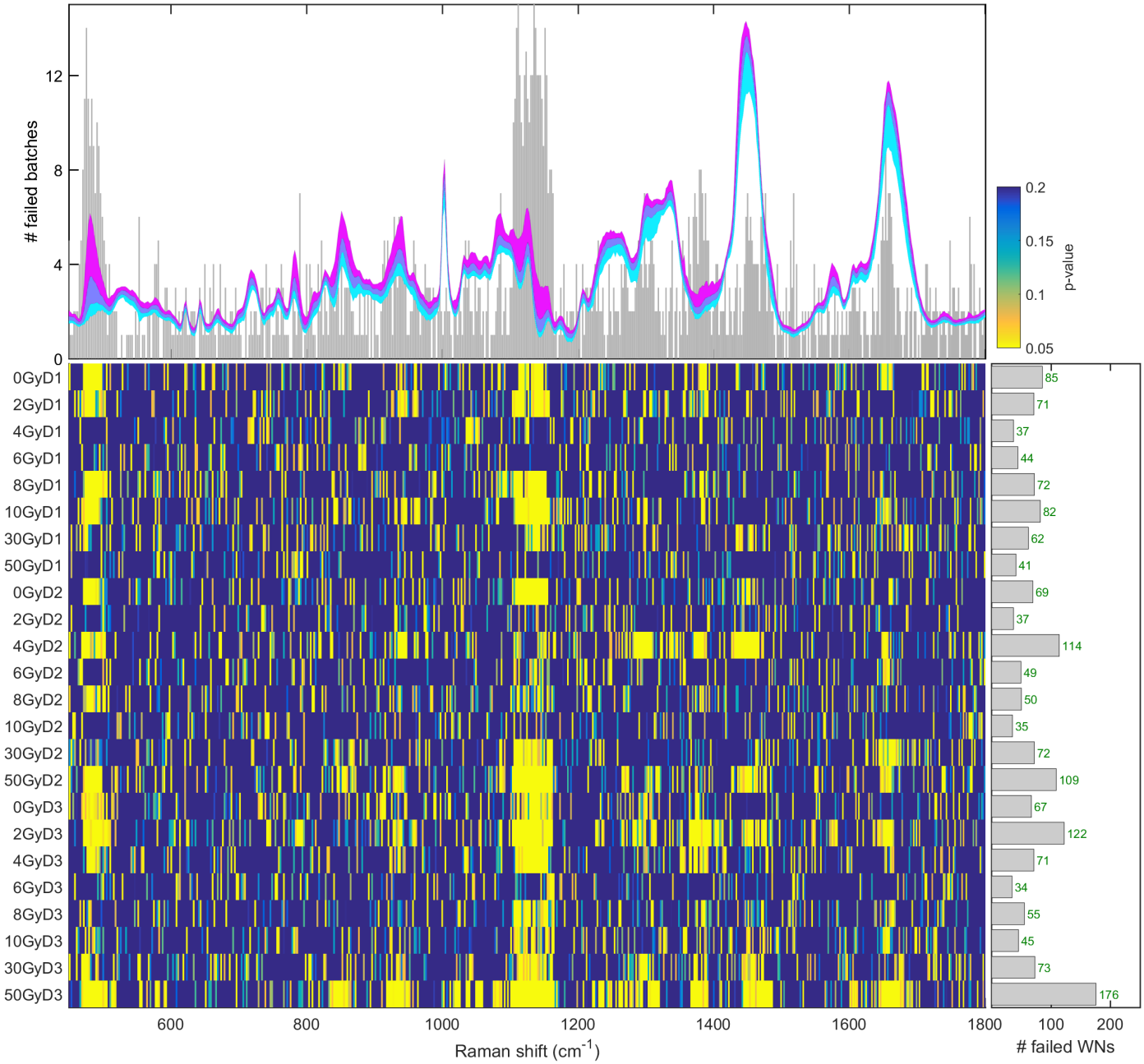
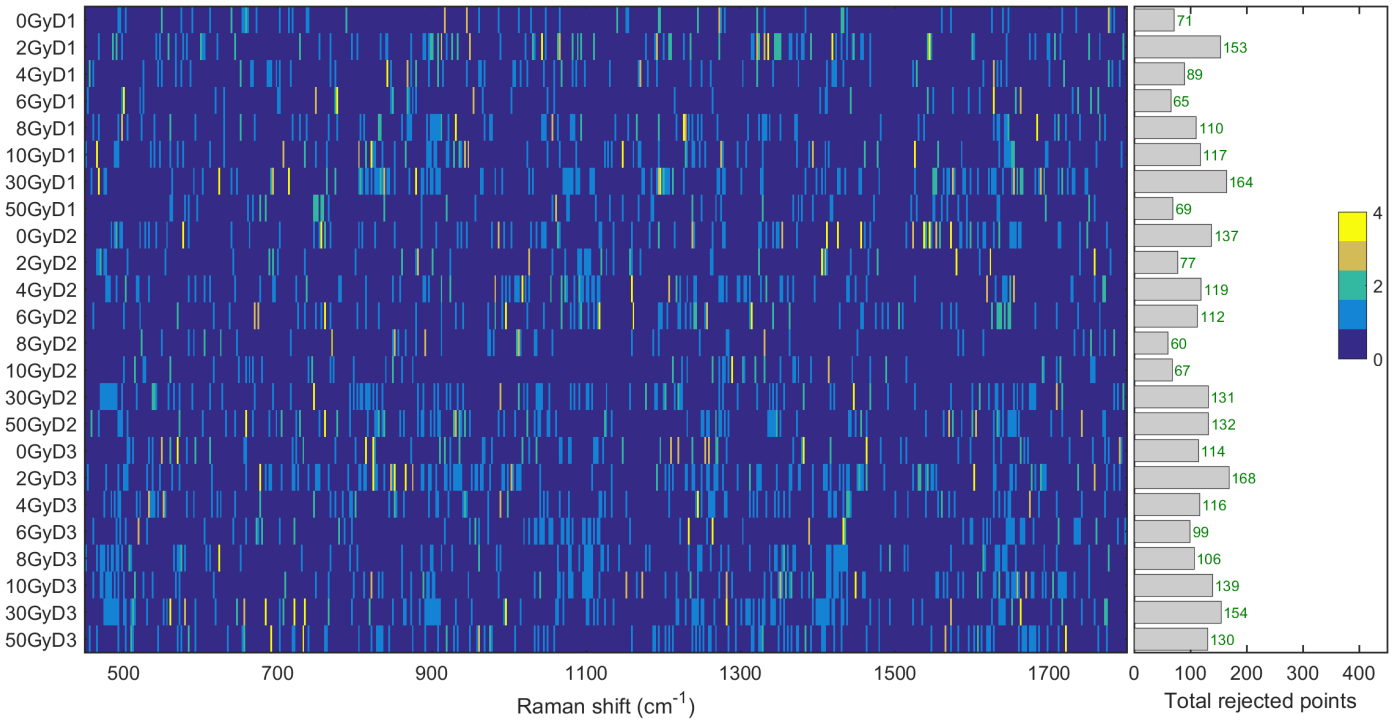
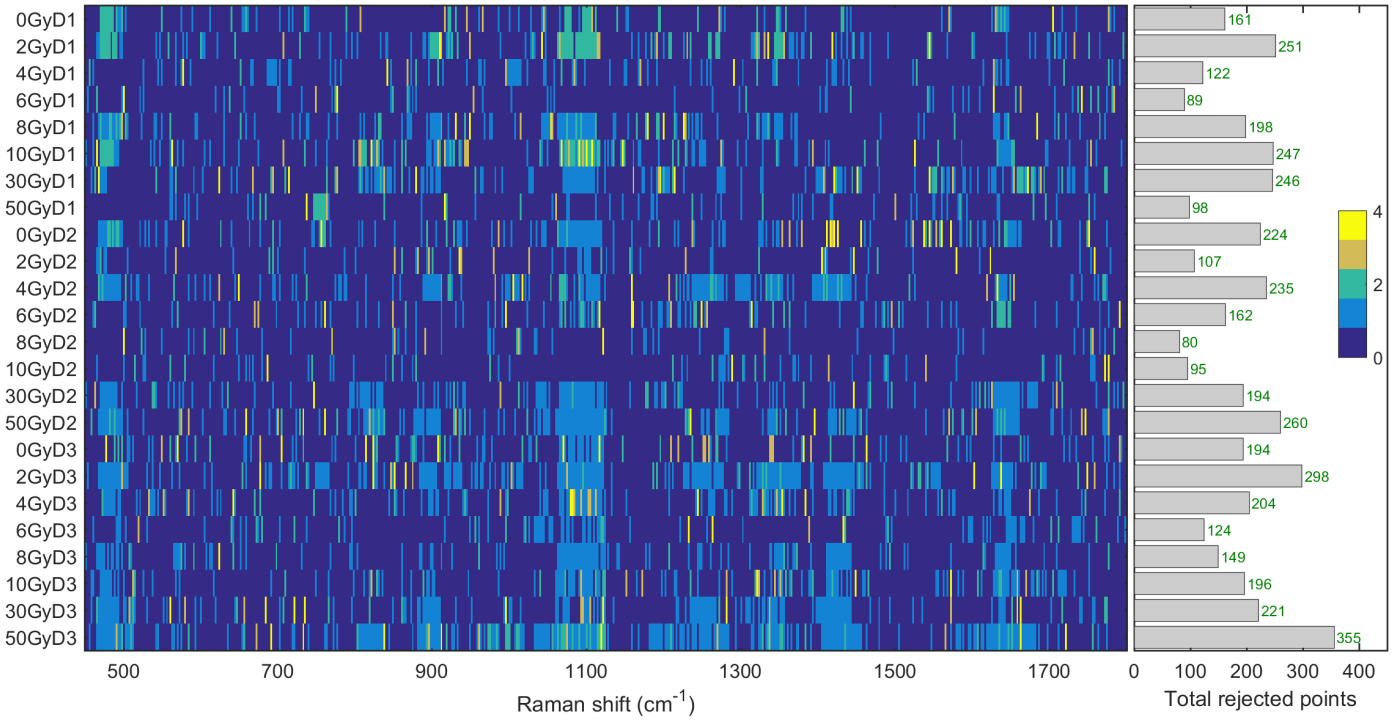


Figure 5.6: Results of H460B SW normality test at $\alpha = 5\%$ mapped into 2D colormap by batch (y), pixel number (x) and resulting p-value (z, color). Yellow results are those rejecting null hypothesis. For easier visualization, p-values of > 0.2 are all shown in blue. Along top edge is a bar plot of total failed batches, overlaid with percentile plot of D3-50Gy batch. Both are aligned with map x-axis. On right edge is bar plot of total number of failed WNs, aligned with map y-axis.



(a) Rejection of normally distributed data only.



(b) Rejection of all data.

Figure 5.7: TT + gESD outlier removal in H460B dataset at $\alpha = 5\%$ mapped into 2D colormap by batch (y), pixel number (x) and number of outliers removed (z, color). Bar graph on right is a count of all points rejected in each batch.

3576 (1.28%) points, indicating that TT encompassed gESD results almost completely. To study possibility of normality masking as discussed above, rejection was also performed on all data (regardless of normality) with results for TT, gESD, and combined counts of 28820, 5041 and 5012 (10.3%,1.80%,1.79%) respectively (figure 5.7b). This indicated a disproportionate amount of outliers was being detected in problematic regions, especially for gESD, consistent with masking. Furthermore, a closer manual examination revealed that majority of problematic 3-4 point rejection decisions were correct.

These results indicated that chosen outlier rejection tests worked as expected, but that their applicability to most crucial regions is questionable due to possible masking effects. However, it is important to remember that the goal is to optimize parameters for best signal extraction. As such, both normality test masking and significance parameters were dynamically varied during actual analysis to maximize performance metrics of section 4.4.

5.3 Summary

This chapter has demonstrated that chosen preprocessing routines yield expected results and remove extraneous sources of variability such as background signal and total sampling volume. However, it was also demonstrated that even with careful collection and processing there are significant imperfections in the data such as missed small cosmic ray events, possible spectral outliers, as well as generally increasing variability with dose and with time. This caused data in certain problematic wavenumber region to often fail normality tests, while having large numbers of proper outliers.

Clearly, robust automated techniques are required for best signal extraction from such data, and it was shown that formal outlier rejection was a reasonable option but its application questionable due to problems with normality test masking. Of the two methods used, gESD was often limiting factor in consensus rejection process and at 5% significance removed approximately 1-2% of data. The following chapters explore signal changes after outlier rejection as well as results of other robust techniques.

Chapter 6

Results and Discussion II - H460

This chapter presents results of applying previously described analysis techniques to H460B dataset. Section 6.1 discusses dataset quality, with main results presented in section 6.2. A comprehensive performance comparison is given in section 6.3.

6.1 Dataset quality

In general, H460 cells were resilient to environmental stresses. They were also the fastest growing with correspondingly highest splitting ratio. Since two independent trials A and B were available, all analysis was performed on both and only results that matched (correlation exceeding ± 0.9 where possible to evaluate) are reported. This was true for all relevant H460 results. For PCA and related techniques, only most important few components (explained variability $> 3\%$) are given. Moreover, discussion is focused on components that could be attributed to specific variability sources such as radiation response or cell cycle, usually #1 and #2.

6.2 High and low dose analysis results

Analysis algorithms were applied to both 0-50Gy and 0-10Gy datasets, with latter given on the right side of all figures. A brief description of results is given, but performance conclusions are reserved for section 6.3.

6.2.1 PCA

Component 1

Component 1 accounted for 59.7% of total variance in high dose set and 55.0% in low dose set. It is shown in figure 6.1 along with corresponding scores and p-values. Extensive molecular identification of this component was done previously [78] and so only key chemical constituents are described here. Majority of positive features were due to protein (1004, 1239 cm^{-1}) and nucleic acid (783, 1377, 1483-1577 cm^{-1}) vibrations, while negative ones were almost exclusively due to carbohydrates. More detailed comparison revealed that negative peaks matched closely to those of pure glycogen. This work does not attempt to determine specific mechanisms leading to such response, however other recent publications have elucidated one of contributing pathways [79].

Trends in PC1 scores were fairly clear, with more negative values at both later times and increasing doses. For control populations, this shift was not significant. However, starting at D2-4Gy and D3-2Gy shifts were significant and continued to be more so with increasing doses. More negative scores corresponded to higher glycogen content consistent with cell survival response (see [79] for detailed discussion). Interestingly, largest score shifts happened in low dose region, with 30Gy and 50Gy less separated from 10Gy than 10Gy from control. This suggested that component 1 represents an intense radiation response that is approaching saturation at higher doses.

Low dose results gave near identical PCs, with correlation of 1.0. Same significance levels were observed, but with somewhat higher p-values as would be expected from loss of strong signal in 30Gy and 50Gy batches (this trend is a good validation of p-score metric). Nonetheless, low dose set was clearly sufficient to fully quantify radiation response.

Component 2

Component 2 accounted for 16.1% of total variance in high dose set and 18.4% in low dose set. It is shown in figure 6.2. Based on prior work this component was identified as cell cycle Raman component [144]. It exhibited characteristic separation of nucleic acid/protein and lipid features into positive and negative contributions respectively, in agreement with expected relative variations of these constituents at different cell cycle phases (as described in chapter 1).

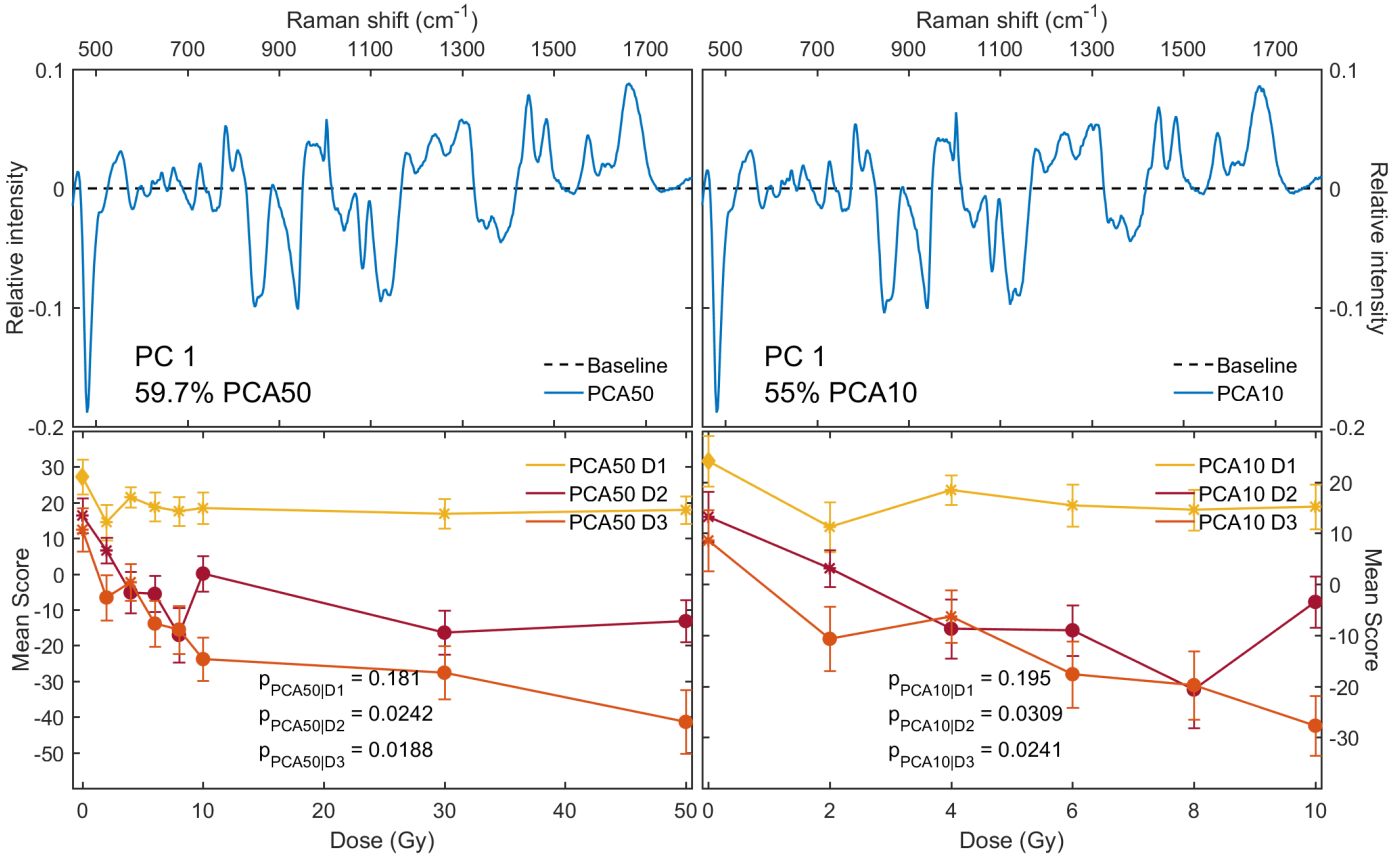


Figure 6.1: PCA component 1 and scores for H460B 50Gy/10Gy datasets. Solid circles denote significant results, asterisks non-significant ones, and diamond the D1-0Gy reference point.

In terms of time dependence, PC2 scores decreased, indicating larger ratio of lipids to other cellular contents. This was consistent with expected increase in G_1 population due to culture reaching high confluence and nutrient exhaustion, a conclusion supported by viability and FACS measurements (not shown). Dose dependence was not observed with exception of D2 (likely due to bad 0Gy batch). Low dose results were again very close, with PC2 correlation of 1.0, and a slight increase in p-values.

Component 3

Components 3 had explained variability of 4.9% and no molecular origin could be determined. It had no useful dose variation, with all samples deviating significantly but uniformly from control population. See appendix B for respective plots.

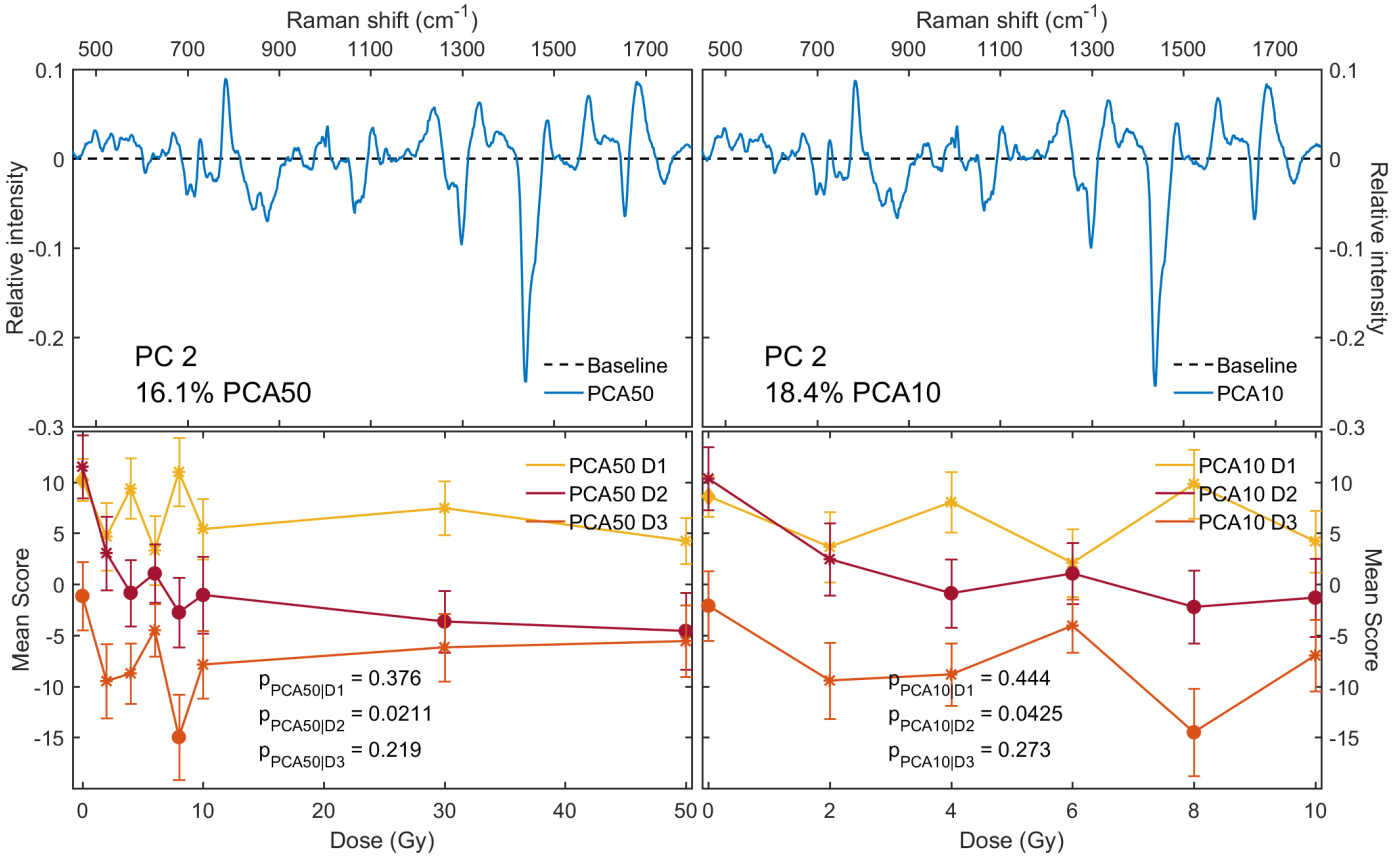


Figure 6.2: PCA component 2 and scores for H460B 50Gy/10Gy datasets. Symbol interpretation same as in fig 6.1.

6.2.2 Weighted PCA

WPCA analysis was performed with $1/\sigma$ weighting as outlined above. Obtained PC variances for both datasets were significantly lower than PCA values, as can be seen in figure 6.3, with crossover only occurring at component 3. Variances relative to first component have slightly increased, as is shown by respective ratios in the lower right corner of the figures. Note that since the full ($k = 479$) WPCA set was computed, it was meaningful to compare absolute cumulatives. They indicated that WPCA redistributed around 20% of explained variability into higher ($k > 5$) components, which are essentially noise-filled and not analytically useful. For low dose set, less variance was explained in PC1 consistent with some loss of radiation response signal.

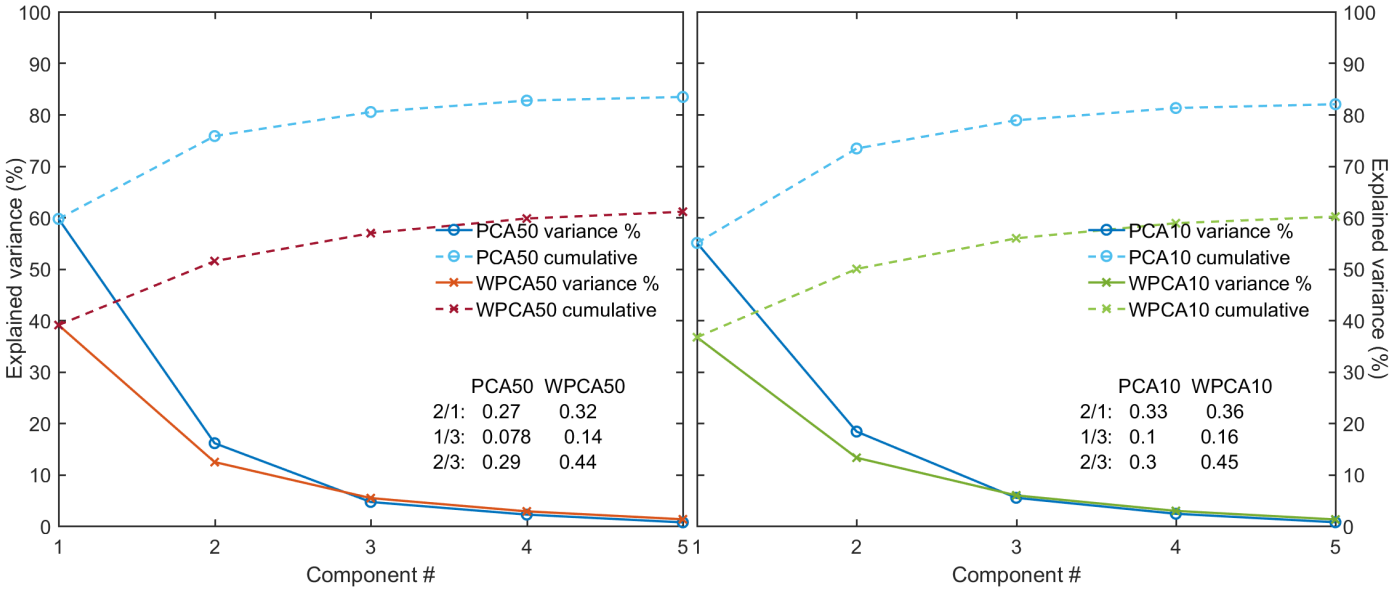


Figure 6.3: WPCA/PCA variances and respective cumulatives calculated with maximum k. Only 5 top PCs are shown. The numbers in lower right corner correspond to ratios of explained variance between indicated components.

Component 1

Component 1 and corresponding scores are shown in figure 6.4. Major differences are obvious for orthonormal WPCA (labelled as ORWPCA) component, with general trend of attenuating peak PCA regions while boosting others as can be seen near 550, 850, 1130, and 1660 cm^{-1}. However, behaviour is clearly more complex than just inverse scaling - note how the double peaks in 1400-1500 cm^{-1} region get transformed, with first one losing half its amplitude while second one remains largely intact.

Score significance differences were small. Nonetheless two results, D1-2Gy and D2-4Gy, have become statistically significant if just barely. WPCA p-values were all better, with minor differences at D1 but major improvements at D2 and D3. Low dose dataset showed similar trends - no changes in significance were observed, but all p-values have increased somewhat as expected.

To quantify similarity of various PCs, several correlation values were used. First of these was HL (high-low) correlation between ORWPCA50 and ORWPCA10, denoted by ρ_{HL}. It was found to be 1.0, confirming close low dose set results. The other two calculated correlations were between PCA and ORWPCA (called interanalysis correlations, denoted by ρ_{IA50} and ρ_{IA10}). They were 0.88/0.89 for high/low dose sets respectively, in agreement with likely larger number of outliers in high dose set.

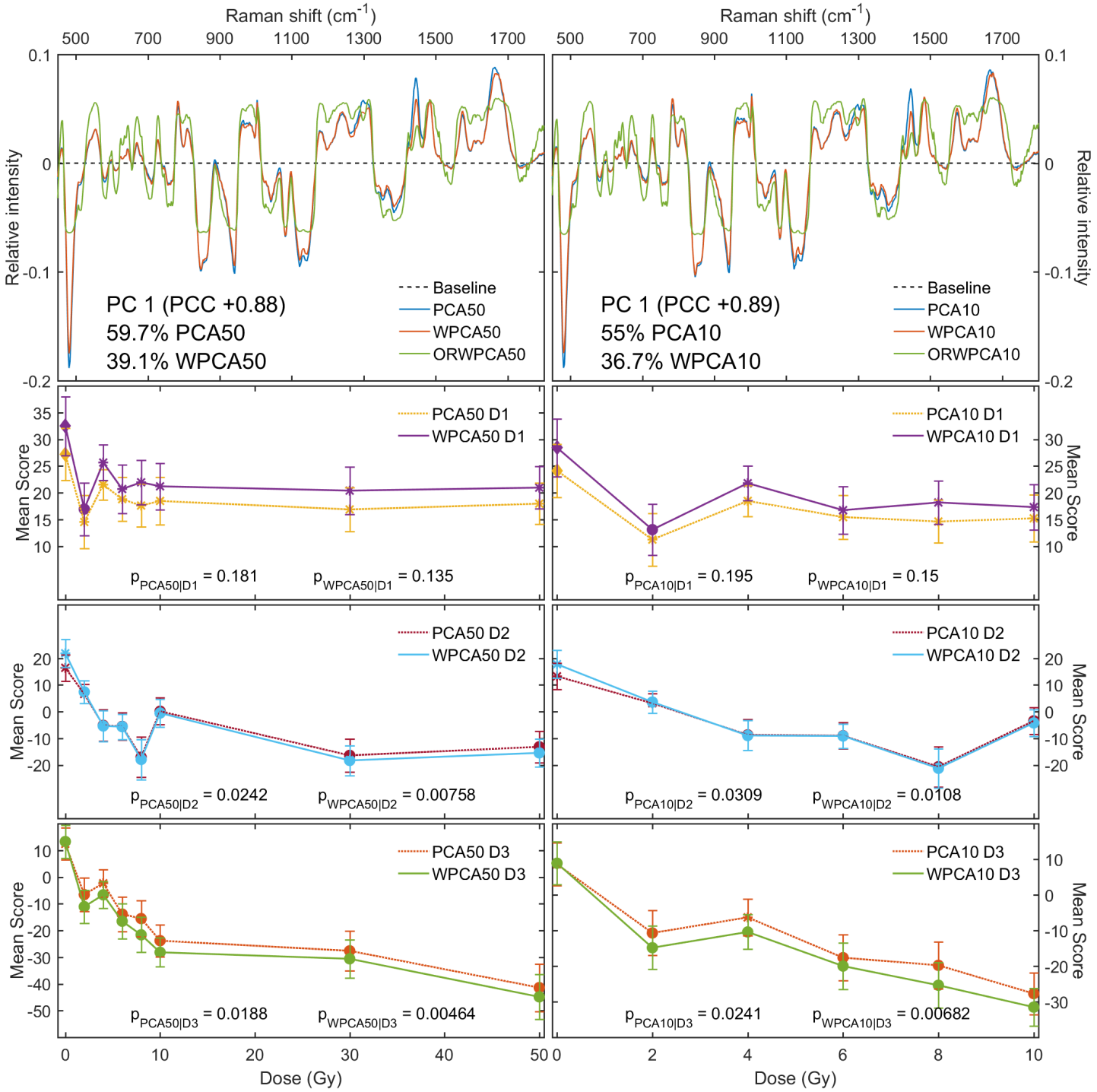


Figure 6.4: WPCA/PCA component 1 and scores for H460B 50Gy/10Gy datasets. Symbols same as in figure 6.1. Relative explained variability is given in lower left corner of PC plots along with PCC (Pearson's correlation coefficient) between PCA and ORWPCA. Score plots show respective average p-values.

Component 1 score distances

Score distances normalized by standard deviation method (σ-distances) showed good separation that rewarded lower uncertainty of WPCA results, as is shown in figure 6.5. A definite time dependent trend was observed (by inspection and via score averages), with day 1 having positive WPCA σ-distances but days 2 and 3 negative ones. This indicated that WPCA has achieved higher signal separation. For min-max method, all distances were negative which for generally decreasing scores corresponds to achieving faster response (vs max). Overall, better score distances, additional significant results, and improved p-values indicated stronger WPCA performance on component 1.

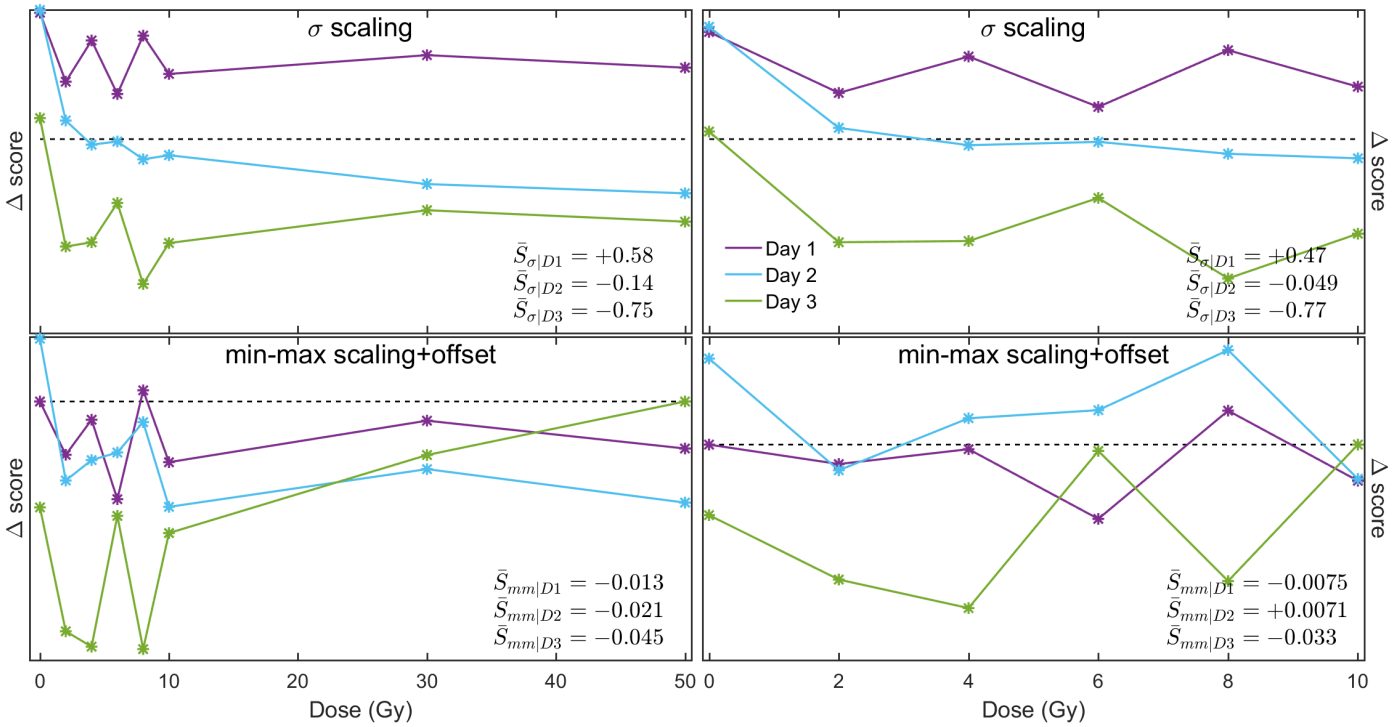


Figure 6.5: WPCA/PCA PC1 score distances between same time/dose batches, scaled with σ and min-max methods. High/low dose sets are shown on left/right respectively. Asterisks indicate insignificant results, while circles denote t-test rejection at 10%. Average score distances of non-0Gy populations for each day are given in lower right.

Component 2

Component 2 and respective scores are shown in figure 6.6. WPCA trend of attenuating peak regions while boosting others was noted again, although this was not

as obvious due domination of sharp but small features. This matched behaviour seen in component 1. Low dose results were nearly identical with $\rho_{HL} = 1.0$ and corresponding ρ_{IA50}/ρ_{IA10} of 0.83.

Score-wise, WPCA had noticeably larger deviations, especially in day 1. There appeared to be a trend of lower initial WPCA scores with increase to above PCA levels by day 3. Points at D2-2Gy as well as D3-4Gy and D3-10Gy have become significant, but again just marginally. All WPCA p-values were lower as in component 1 and all but one increased in low dose dataset.

For σ-distances (figure B.3), there was a small time dependent change towards positive values but it was not significant. However, MM-scores showed a larger upward trend corresponding to lower WPCA sensitivity. Overall, 3 additional significant results and improved p-values suggested slightly stronger WPCA performance on component 2, but score distance examination was inconclusive.

Component 3

Due to similarity with PCA signal, this and subsequent components are again omitted from consideration. Results are provided in appendix B.

6.2.3 Robust PCA

Robust PCA has displayed decreased first component variance while maintaining higher one for all subsequent components in both dose ranges, as can be seen in figure 6.7. The magnitude of this shift was however significantly lower than with WPCA and crossover occurred at component 2. Note that PCA percentages have changed because of renormalization to $k = 3$. However, PC ratios remained the same as expected. For the low dose set, less variance was explained in PC1 consistent with some loss of radiation response signal.

Component 1

Results for high and low dose datasets were closer to PCA than in previous analysis, but each had clear differences resulting in ρ_{HL} of 0.99. Several shifts were observed, mostly at previously identified problematic regions near 1150 and 1450 cm^{-1}, as is shown in figure 6.8. This was consistent with the expectation of rejecting grossly corrupted observations that were detected there with statistical tests. Moreover, rejection increased for the low dose set, where 1420-1450 cm^{-1} region nearly vanished

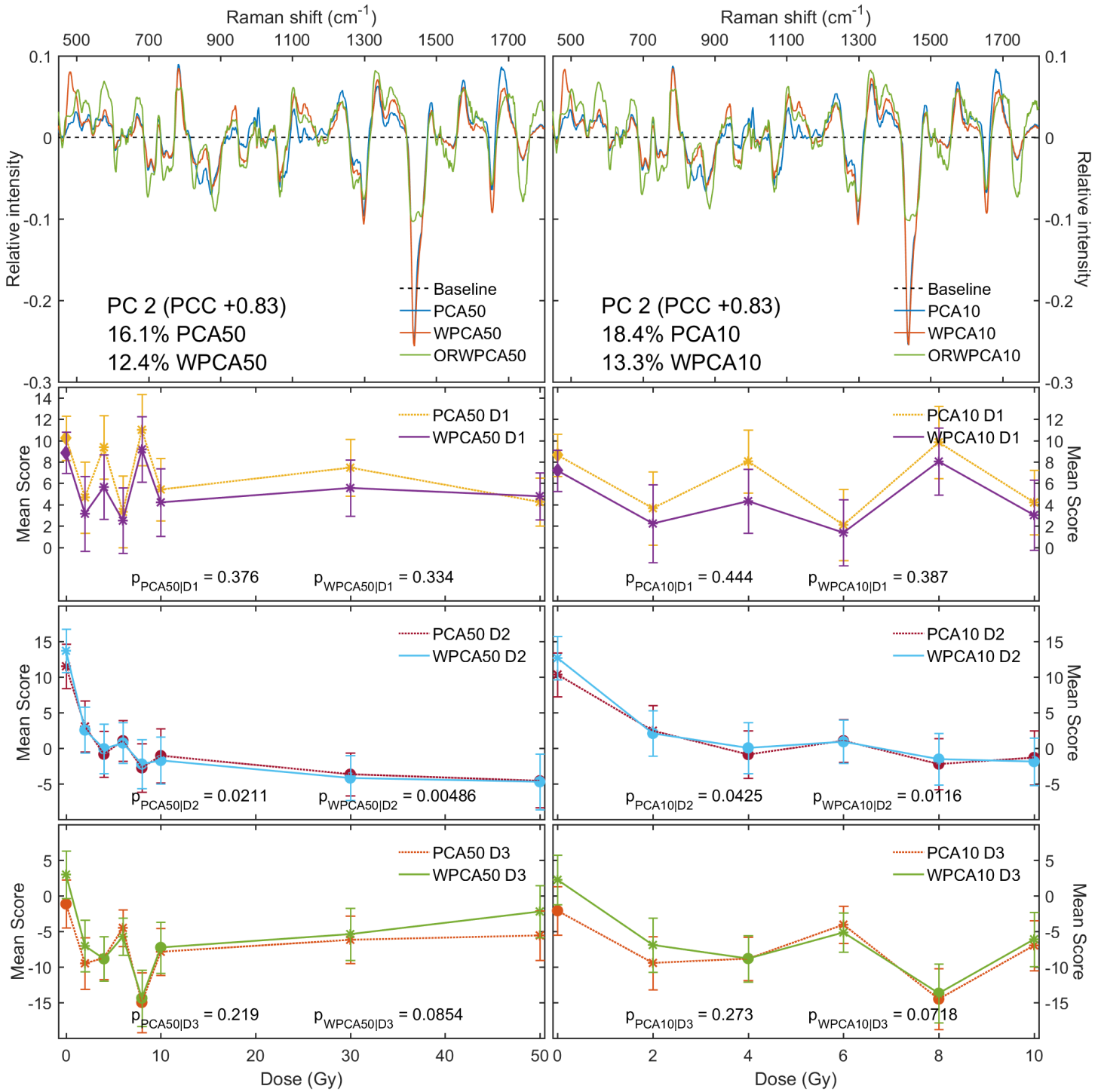


Figure 6.6: WPCA/PCA component 2 and scores for H460B 50Gy/10Gy datasets. Notation same as in figure 6.4.

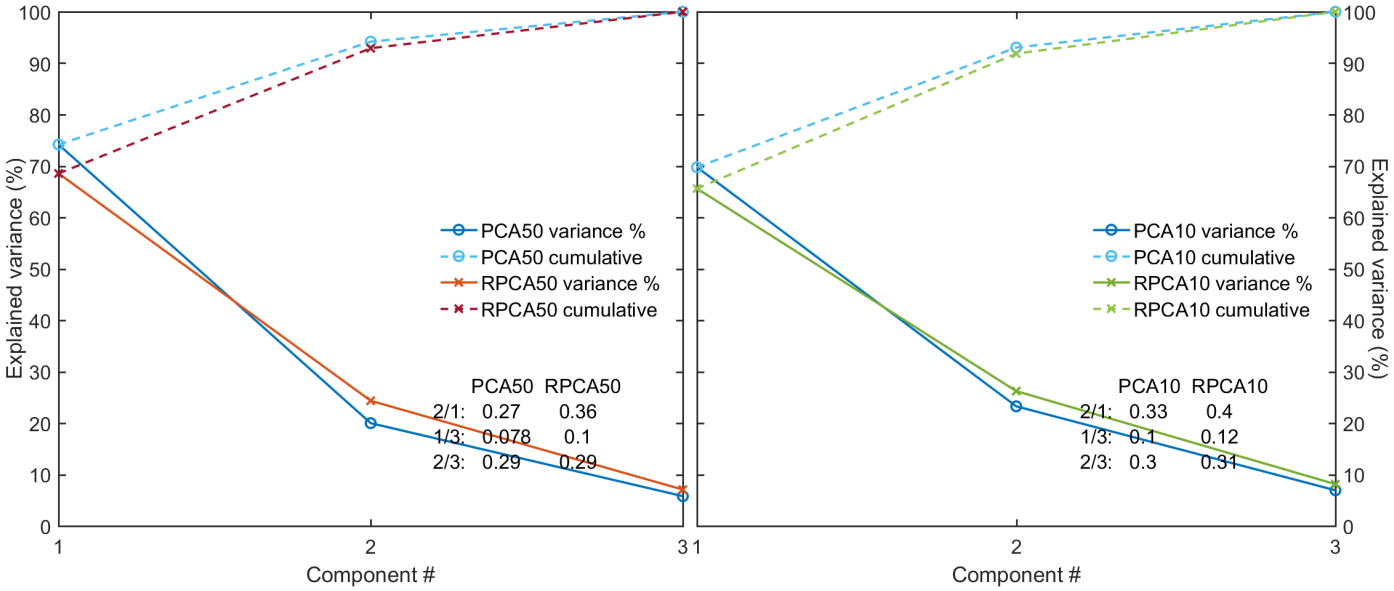


Figure 6.7: RPCA/PCA variances and respective cumulatives calculated with $k = 3$.

and stronger deviations in 1300 cm^{-1} and 1420 cm^{-1} areas appeared. Interanalysis correlation correspondingly decreased from 0.98 to 0.96. Whether this was due to high dose outliers no longer forcing unrelated signals into radiation response component, or a reverse situation of high dose radiation biochemical response being detected as outlier due to rarity is unknown.

In terms of score significances, the only new result observed was at D3-4Gy (omitting 0Gy ones) due to a relatively minor shift. Nonetheless, average p-values were all lower for RPCA indicating higher significance. Interestingly, p-values showed only minor changes for low dose analysis, increasing approximately three times less than PCA ones. This could indicate better RPCA noise rejection.

Component 1 score distances

Score distances (figure 6.9) displayed trends close to those of WPCA, with increasing time dependent negative offset of σ-scores indicating stronger signal. Min-max scores were negative at low doses and approached 0 at higher ones, consistent with earlier signal detection. Overall, RPCA performed better on component 1.

Component 2

Component 2 (figure 6.10) displayed very interesting trends with several regions increasing dramatically such as at 480 cm^{-1} and 930 cm^{-1}. These deviations grew

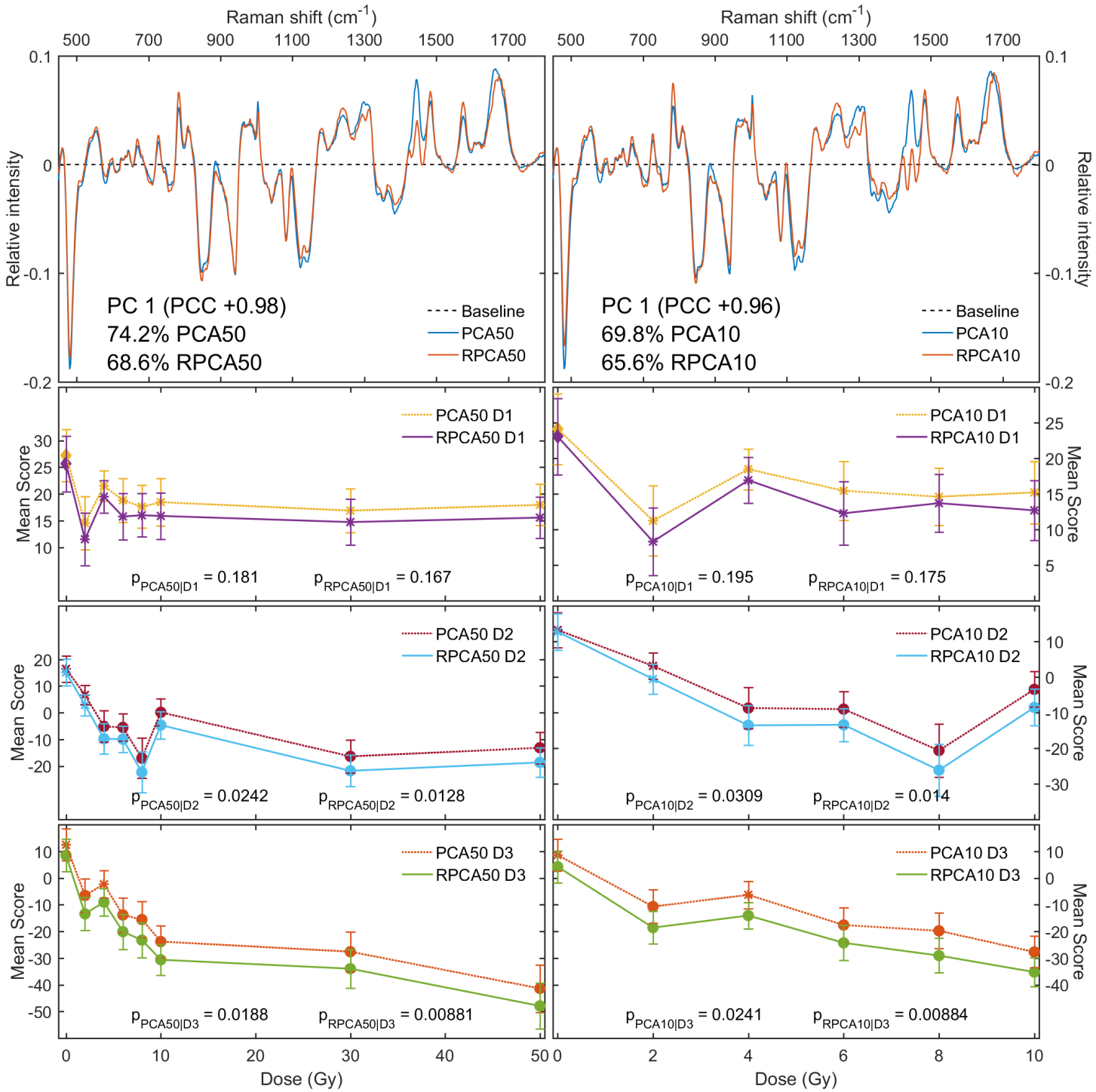


Figure 6.8: RPCA/PCA component 1 and scores of H460B 50Gy/10Gy datasets. Notation same as in figure 6.4, except that PCC is between PCA and RPCA.

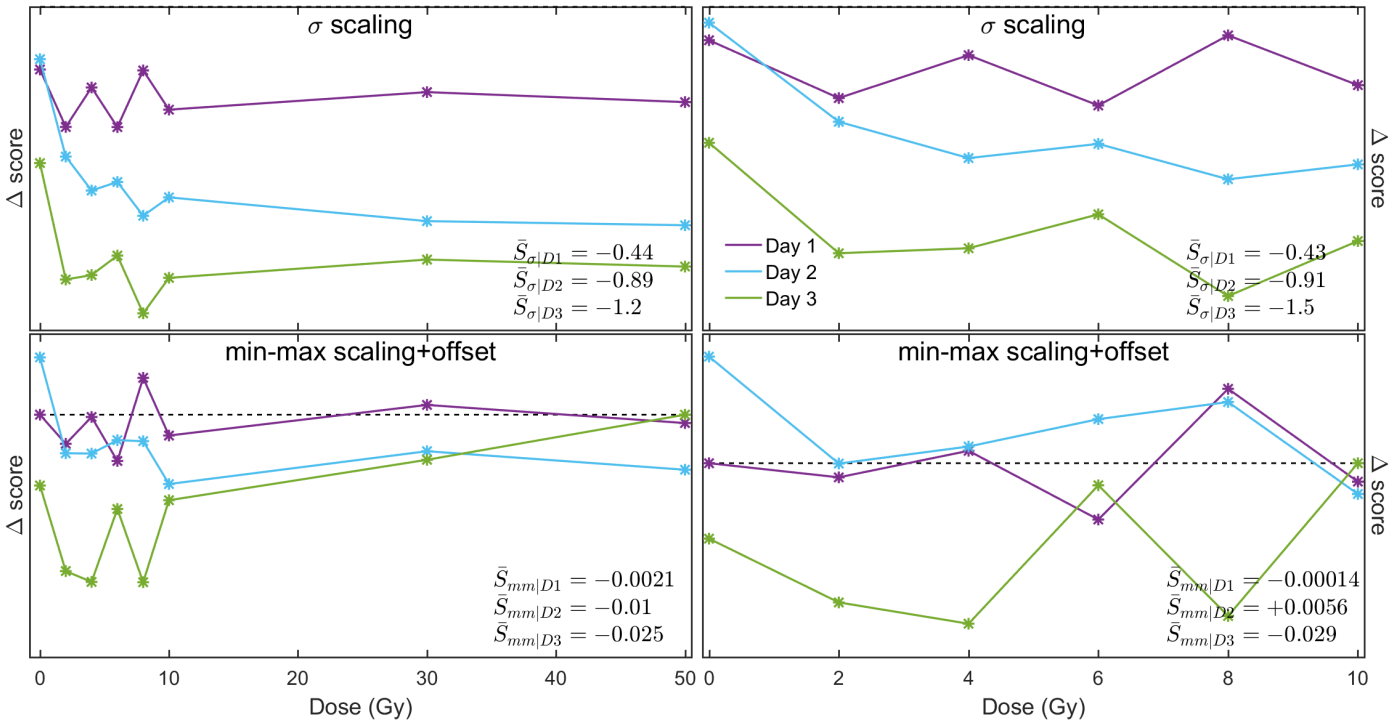


Figure 6.9: RPCA/PCA PC1 score distances for σ and min-max scaling methods. Notation same as in figure 6.5.

slightly in low dose dataset, leading to ρ_{HL} of 0.99. In general (consistent with variance % changes) RPCA removed certain peaks in PC1 and added others in PC2. Moreover, these changes did not overlap, indicating that signals must have been transferred into/from other components, possibly since they were erroneously initially placed by PCA.

Score distributions showed a clear trend of lower RPCA signal, with a change of 7 significant results in the high dose set. Correspondingly, p-values were notably higher and increased even more for low dose analysis. Note that reference (0Gy) distributions were relatively the same, and most differences occurred at higher doses and later times. This was indicative of lost or changed biological signals that RPCA was measuring, consistent with PC2 changes.

Score distances (figure B.6) indicated increasingly large loss of signal, with later σ-scores all positive. Min-max distances supported this, since they were uniformly positive as well. Several high-dose points have shifted sufficiently to satisfy the 10% significance criterion, such as at D3-50Gy in high dose set and D3-6/8/10Gy in low dose set. Overall, RPCA performance was worse in terms of distinguishing PCA-like cellular cycle signals, but due to changes in PC2 it was hard to determine whether

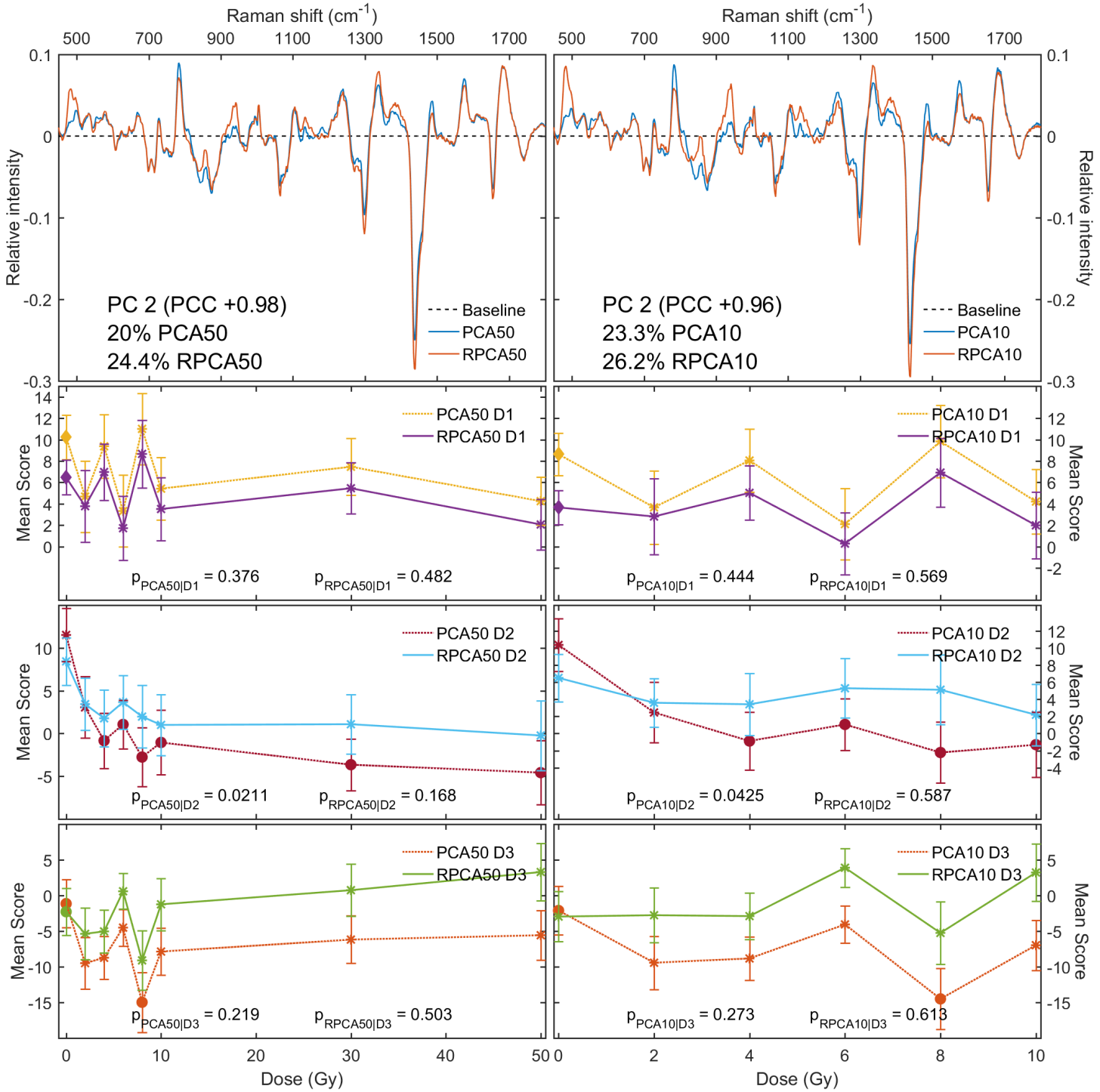


Figure 6.10: RPCA/PCA component 2 and scores of H460B 50Gy/10Gy datasets. Notation same as in figure 6.8.

this was beneficial for overall signal separation. In other words, while there was a clear loss of significance, it is unknown whether this indicates worse performance and loss of cell cycle component or instead a better elucidation of correct cell cycle signal that has smaller score changes than PCA one.

Component 3

Somewhat surprisingly component 3 had fairly high agreement with PCA, but again displayed no new information. It is given in Appendix B.

6.2.4 Probabilistic PCA

Algorithm validation

Prior to showing analysis results, validation data is presented. Using full H460B 50Gy dataset, PPCA results were virtually identical to PCA (not shown), with essentially perfect (to 7 significant digits) correlation of first four components and relative score differences near machine precision limits $\sim 10^{-11}$.

Results of data removal testing were also exemplary - randomly (via pseudo-rng roll for each value) removing up to 80% of points in the dataset still gave nearly the same PCs. To quantify the errors, deviations were monitored via four metrics - p-value and score changes, correlation of principal components, RMS error of reconstructed data, and PC subspace angle (all with reference to PCA). As such, both changes in score significance and in PCs were measured. In terms of the former, removal of up to 40% of the data did not result in p-value changes of more than 1%, with negligible score differences. As for PCs, RMS error and subspace angle gave very similar behaviour and so only RMS error and correlation are shown in figure 6.11. Note that $k = 4$ was used (sufficient for over 95% of variance) since lower PCs were never considered in this work and so their recovery was not as important. ALS results are not shown since they were found slightly inferior to PPCA, with progressively larger divergences at high alphas. Combined with worse computational performance this resulted in rejection of ALS algorithm. All further analysis was done solely with PPCA.

Results indicated that even with losses of 50%, PCs remained nearly identical. Correlation loss increased with component number, consistent with these PCs having less information. RMS errors were insignificant relative to reference (α=0) case,

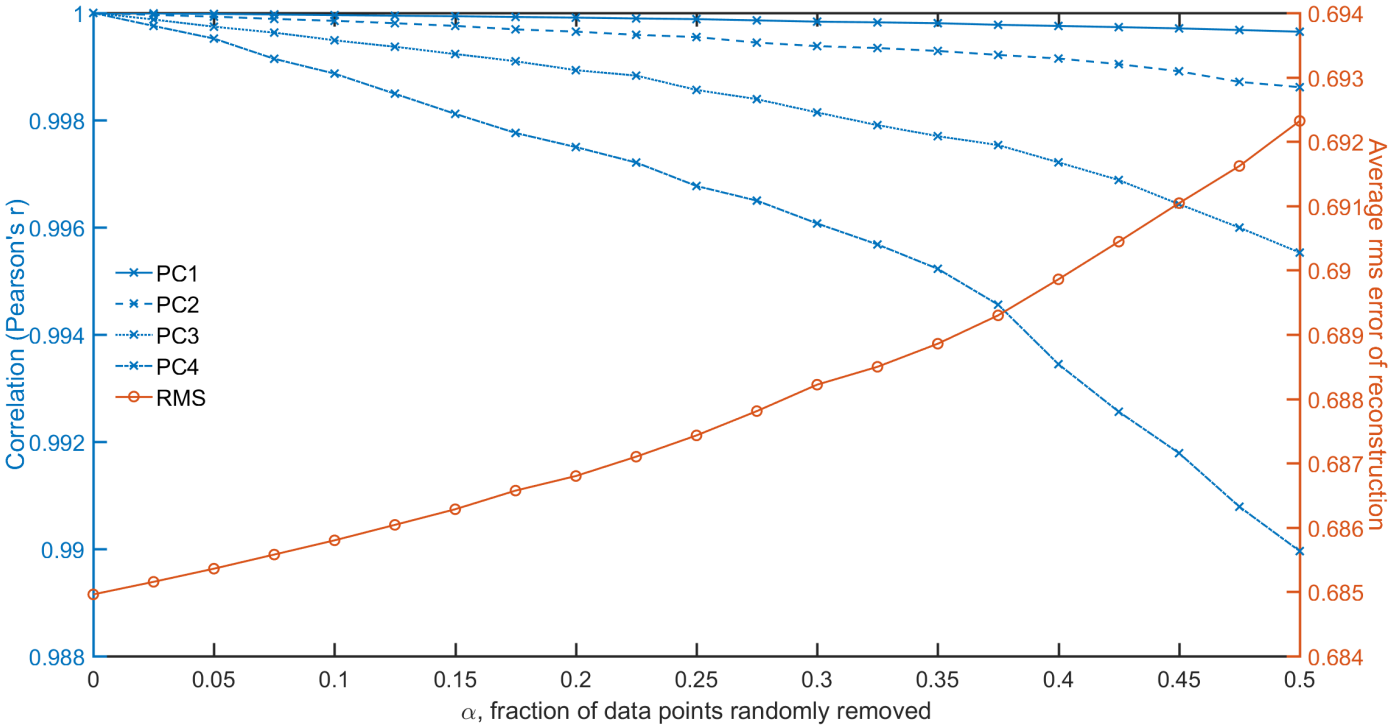


Figure 6.11: PPCA accuracy with randomly missing data. Correlation was determined individually for first four PCs of PCA and PPCA applied to H460B 50Gy dataset. RMS error was calculated on the full dataset.

with changes of under 1%. This resilience validated use of aggressive outlier removal strategies. Hence maximum possible removal range of 0-20% was chosen for actual analysis. After several optimizations rounds, final data removal algorithm was applied with significance setting of 0.4 and normality masking off to remove 15.1% of the data from high dose set and 15.0% from low dose one. This slight discrepancy was attributed to prevalence of suspicious outliers in higher dose batches in H460B set. Note that while chosen α appears very high, this was due to peculiarities in gESD test which was limiting consensus vote at all points. TT alone achieved same rejection rate at $\alpha = 0.07$, a significantly more reasonable value.

Explained variance

Explained variances remained nearly the same, as is shown in figure 6.12, indicating that little actual information was lost from top PCs.

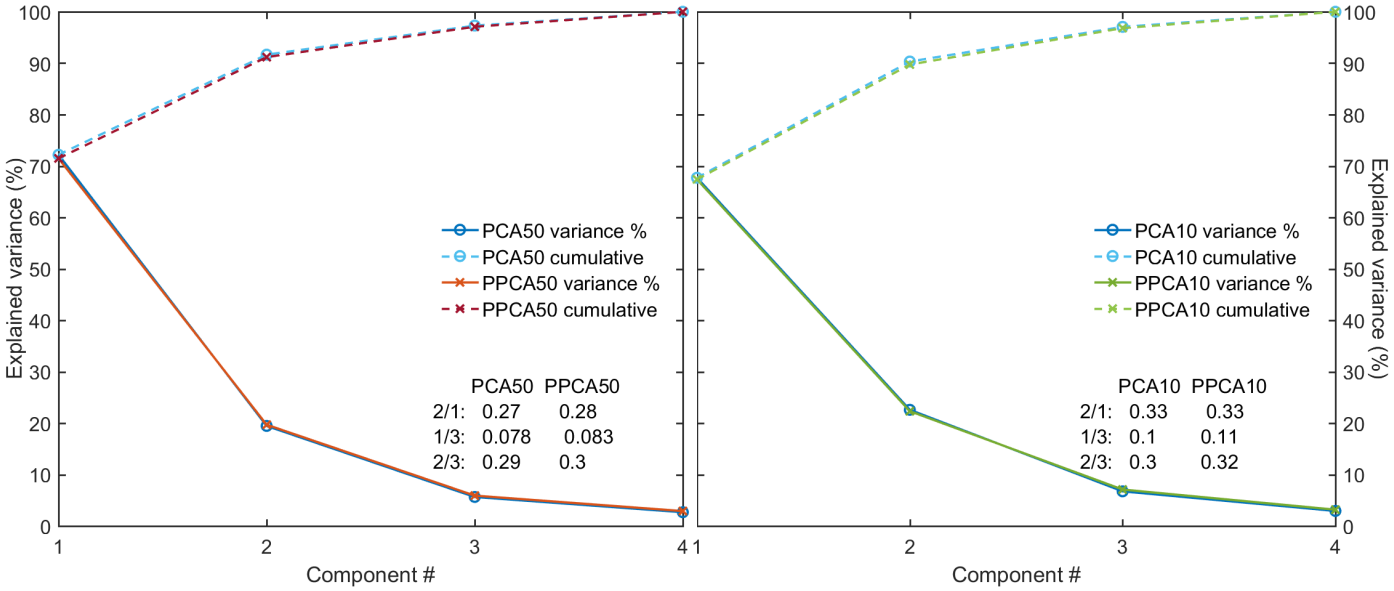


Figure 6.12: PPCA/PCA variances and respective cumulatives calculated with $k = 4$.

Component 1

By comparing the changes in PC1 to regions of high variability that were discussed in chapter 5, it was fairly clear that those were dampened by PPCA (figure 6.13). This included notable features at 840, 930 and 1130 cm^{-1} as well as the peaks first noticed in WPCA analysis at 1400-1500 cm^{-1}. It was therefore concluded that PPCA achieved its purpose of removing influence of spectral outliers. Low dose PC1 showed somewhat noisier attenuation, but in general features matched very well, with ρ_{HL} of 1.00 and identical 0.99 results for ρ_{IA50}/ρ_{IA10}.

In terms of scores PPCA performed better, gaining 8 significant results. Interestingly, it seems to have pushed D1 series past edge of detectability, with 5/7 results becoming significant. The low dose set generally followed similar trends. Corresponding p-values reflected the obvious PPCA improvements and were close to an order of magnitude lower than PCA ones. They also followed the usual trends of decreasing with time and increasing for low dose set.

Component 1 score distances

As with WPCA, score distances confirmed the obvious PPCA improvements (figure 6.14). Namely, σ-scores decreased with time, crossing over into negatives between D1 and D2, indicative of better signal resolution. MM-scores were uniformly negative and remained strongly so for almost full dose range, demonstrating highly improved

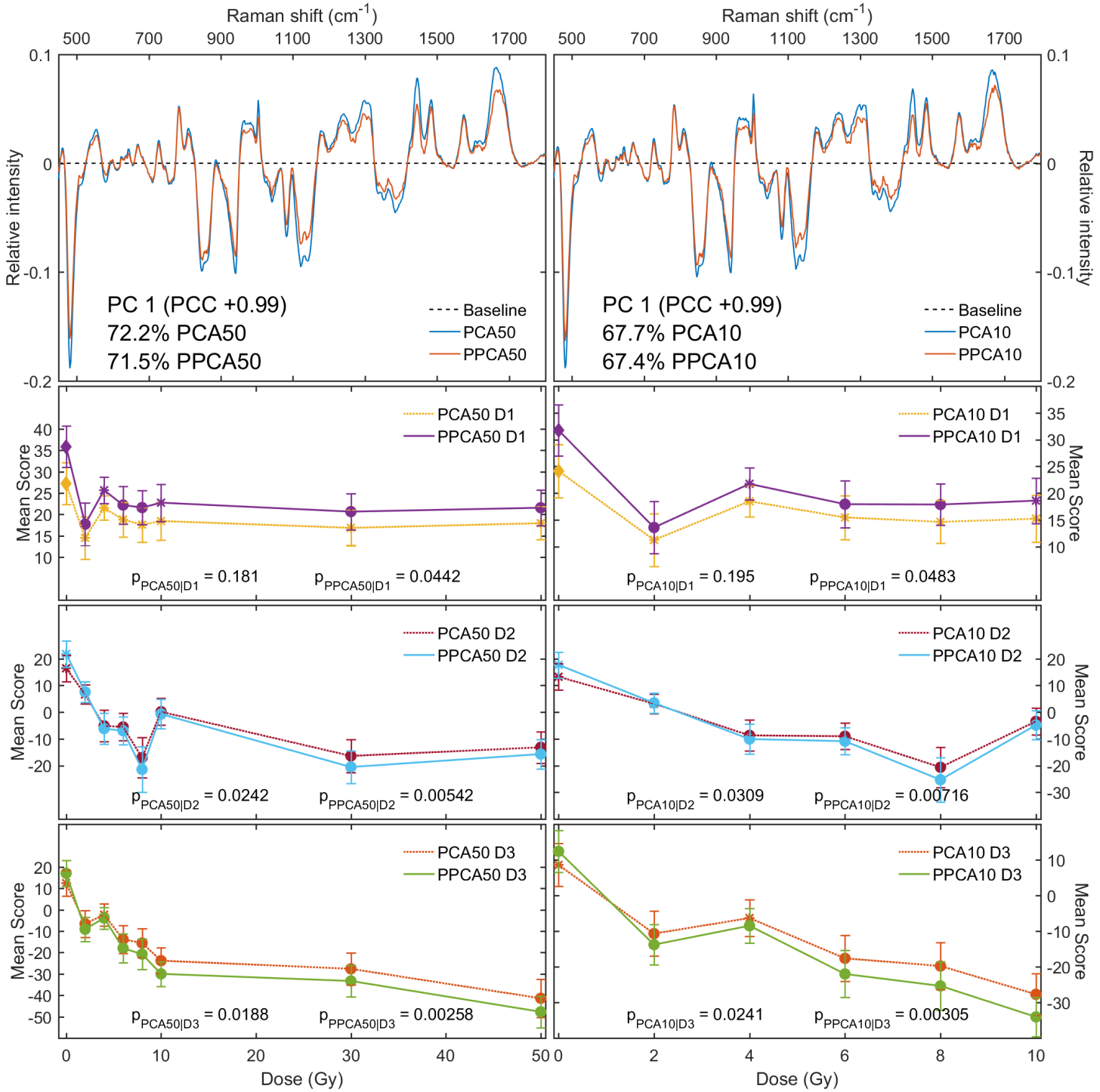


Figure 6.13: PPCA/PCA component 1 and scores of H460B 50Gy/10Gy datasets. Notation same as in figure 6.8. Note that 15.1%/15.0% of data was removed from 50Gy/10Gy sets respectively.

PPCA sensitivity. Low dose data matched above trends. As such, based on additional significant results and lower p-values, PPCA performance on component 1 was found to be better.

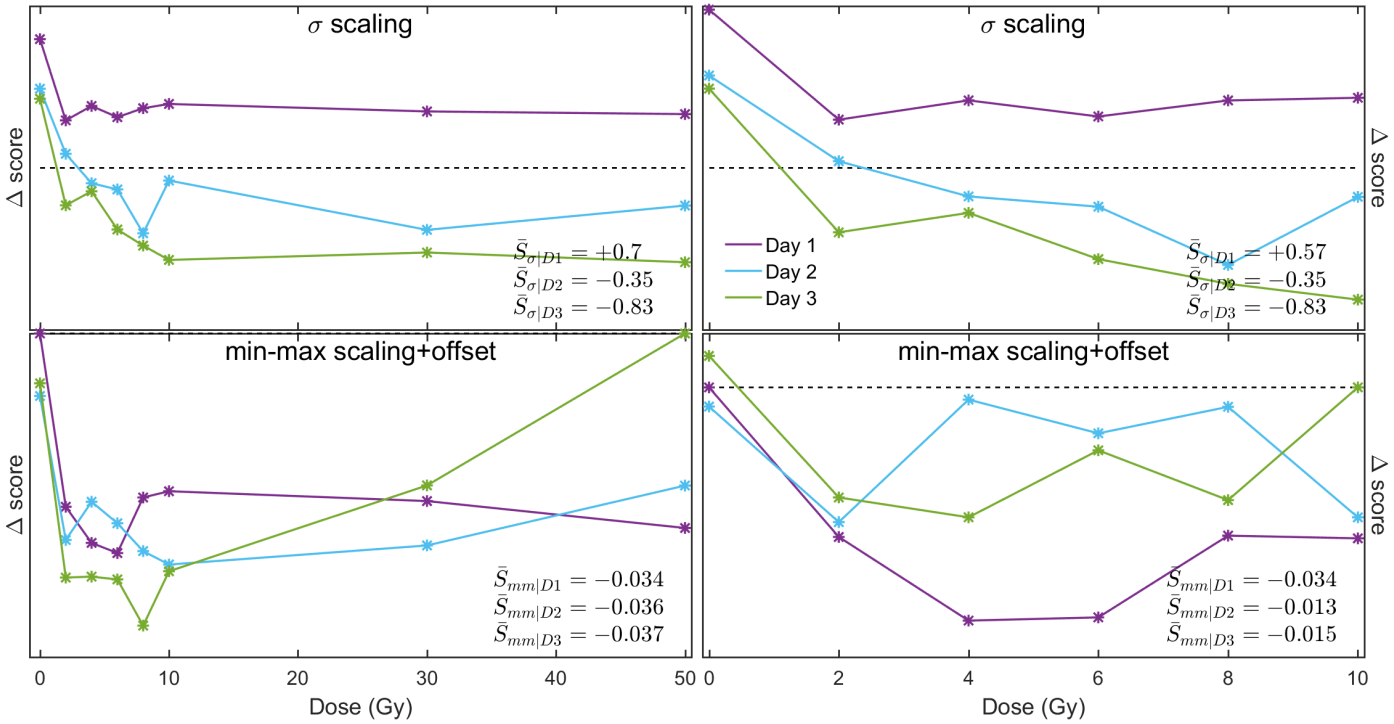


Figure 6.14: PPCA/PCA PC1 score distances for σ and min-max scaling methods. Notation same as in figure 6.5.

Component 2

Differences in cell cycle component (figure 6.15) were comparatively minor, with just slight peak attenuation. Relative to PC1, significantly fewer modification have occured, possibly related to lack of strong PC2 features in the outlier regions. HL correlation was good at 1.0 with correspondingly high IA one of 0.99 in both sets.

No significance changes were observed in high dose set, and only 1 very marginal switch at D3-8Gy in low dose set. P-values indicated that at D1 results were almost identical, but PPCA lost some signal at D2/D3. As usual, p-values increased for the low dose set. Score distances showed no discernible trends, with near equal number of positive and negative values (figure B.9), all with fairly low magnitudes. Overall, PPCA performance on component 2 was found to be slightly worse.

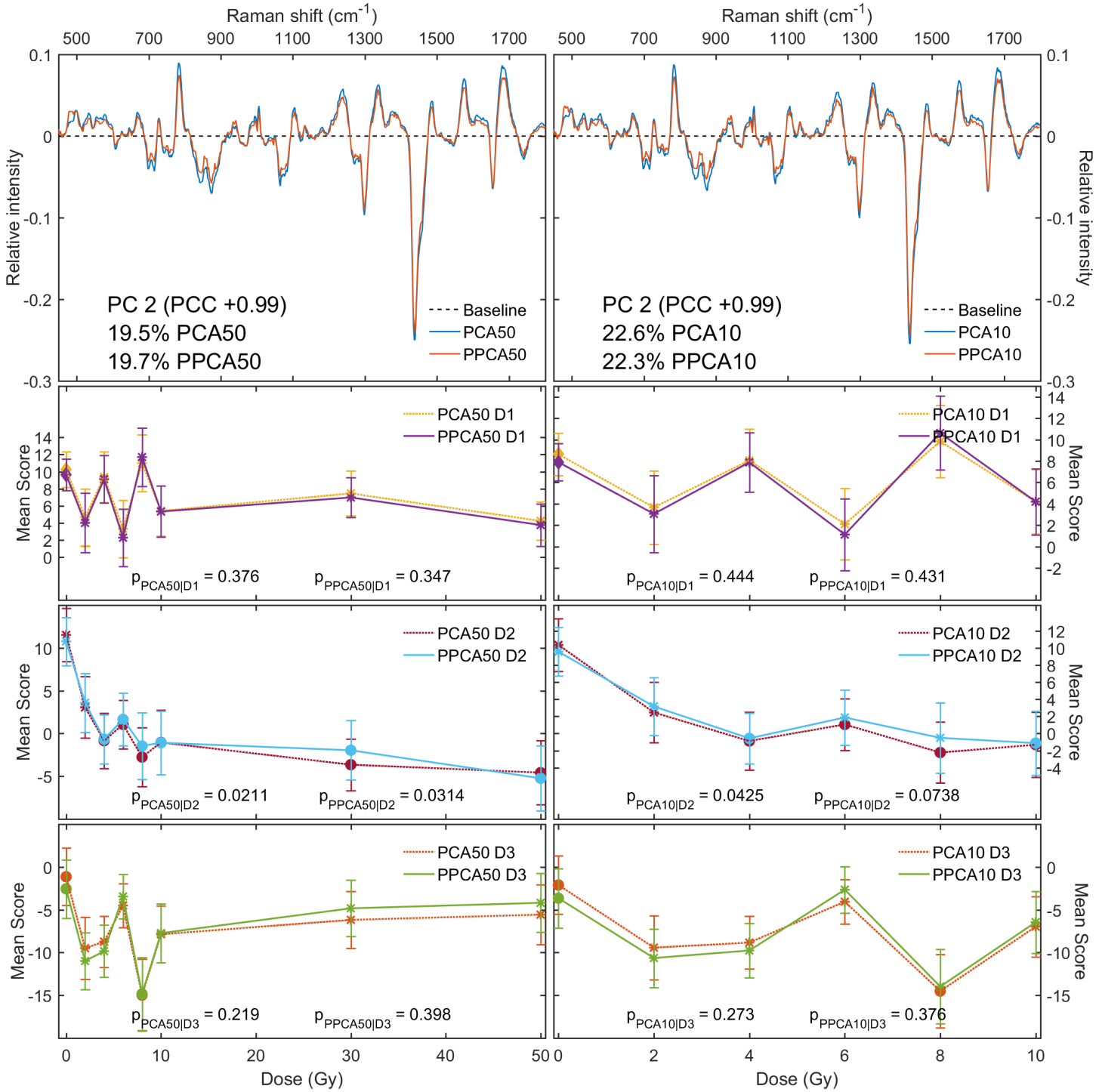


Figure 6.15: PPCA/PCA component 2 and scores of H460B 50Gy/10Gy datasets. Notation same as in figure 6.13.

Component 3

Results of component 3 matched closely with PCA, and are left for appendix B.

6.2.5 Nonlinear PCA

NLPCA was run with 2, 3, and 4 components and resulting significances analysed. Best results were obtained with $k = 3$ and are used for plots below. It must be noted that this method was extremely computationally intensive, with single run taking upwards of 10 minutes. Moreover, each optimization yielded slightly different results due to random initial seeding. Combined with a vast array of tunable parameters this was a definite disadvantage which should be taken into consideration before bulk automated usage.

Component 1

As was discussed previously, there is no method of assigning variances to curved principal components. It is however possible to view these curves as 3D projections in the PCA component space, and this is provided in place of usual Raman plots in figure 6.16. Qualitatively speaking, PC1 and PC2 formed the long backbone of the subspace, which was fairly straight majority of the time, especially in the sparse 'tail' region, indicating that linear methods are not a bad approximation. PC3 was in plane perpendicular to the turning point of the curve (best seen as the vertical line in the middle of X-Y projection). Via comparison with low dose set (right side of 6.16) it was found that 30Gy and 50Gy data formed the bulk of tail points, consistent with them having most signal. This demonstrated how a relatively small amount of high signal data can determine structure for a large portion of the subspace. Note that X-Z view is provided in PC1, X-Y in PC2, and Y-Z in PC3 plots for easier visualization.

In terms of scores, some improvements were observed with new significant results at D1-2Gy and D3-4Gy, as well as significant reference shift for D3-0Gy. P-values were correspondingly lower, with increasing differences at later times. Low dose dataset has as usual shown slightly worse results, possibly due to above mentioned changes in PCs.

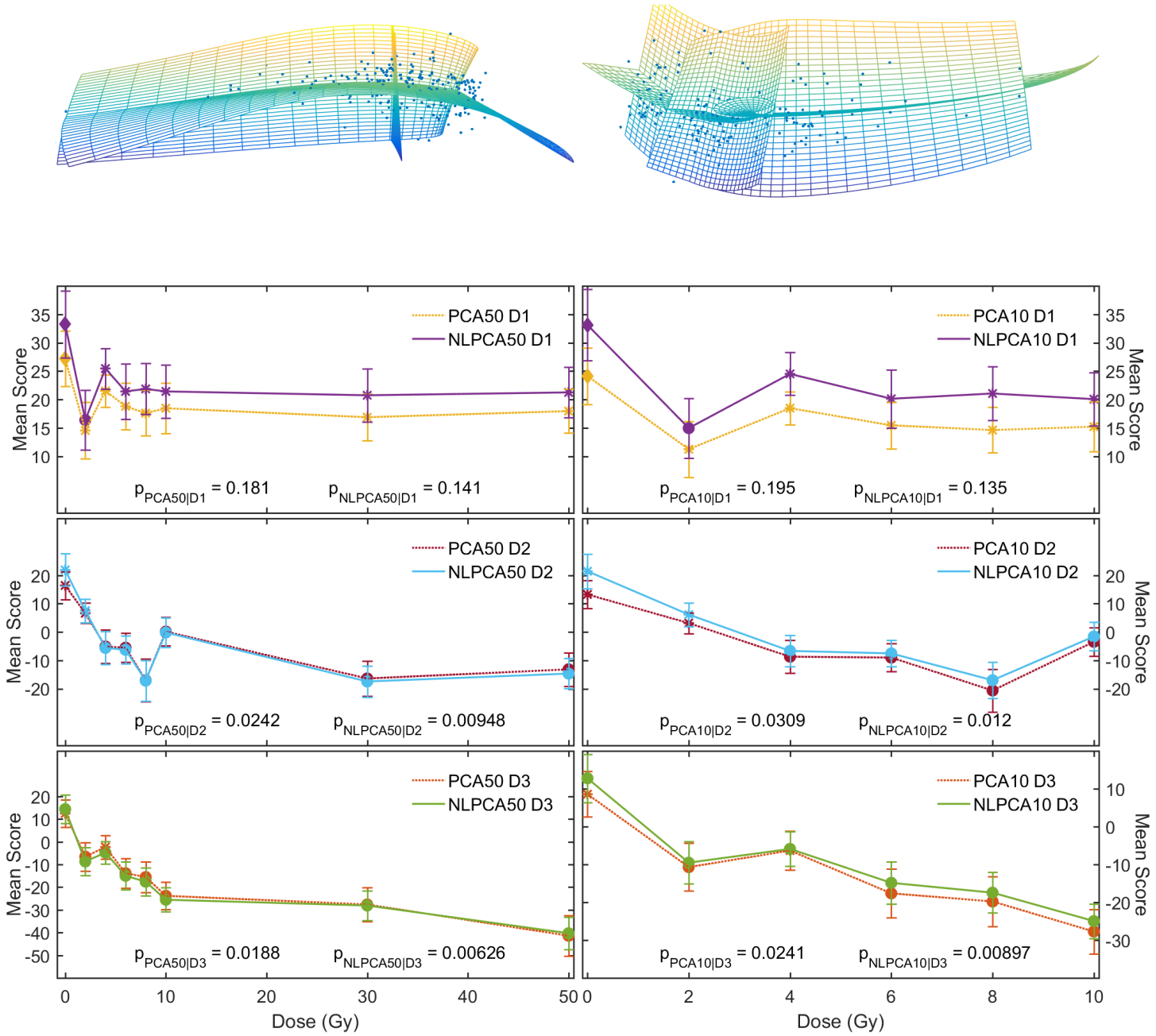


Figure 6.16: NLPCA/PCA projection and PC1 scores for H460B 50Gy/10Gy datasets. PC curves are projected into 3D principal subspace, and X-Z perspective used. Solid circles denote significant results, asterisks non-significant ones, and diamond the D1-0Gy reference point.

Component 1 score distances

Score distances (figure 6.17) were interpreted similarly to previous cases, with σ-scores showing slight NLPCA advantage while MM-scores displayed corresponding improvement in relative sensitivity. However, due to rather low magnitude of absolute score shifts these trends were not particularly important. Overall, NLPCA performed better on component 1.

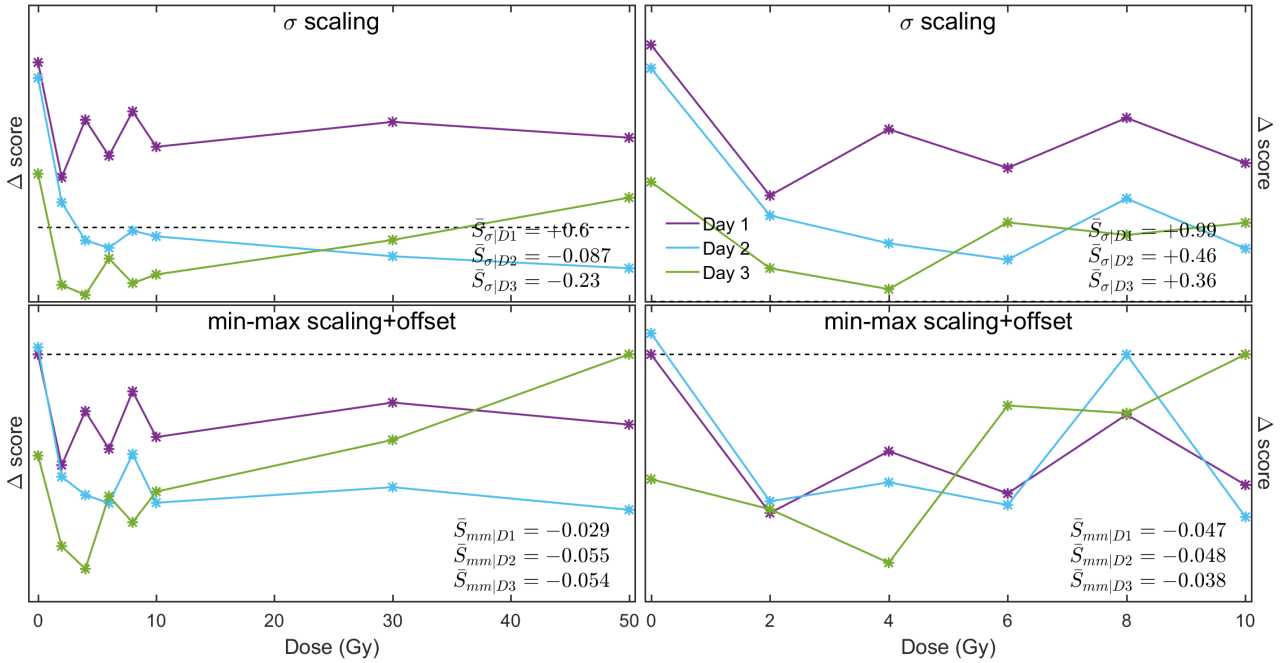


Figure 6.17: NLPCA/PCA PC1 score distances for σ and min-max scaling methods. Notation same as in figure 6.5.

Component 2

A behaviour similar to that of RPCA was observed in component 2, with a very large reduction in scores. In fact, only a single significant result remained, for total loss of 6 points in both datasets. As expected, p-values were also notably higher. Score distances (figure B.12) showed the obvious loss of significance, with σ-distances increasing with time. MM-scores were highly erratic in low dose region, separating out with time at 30Gy and 50Gy. Their positive values were interpreted as loss of sensitivity, consistent with other observations. As such, NLPCA performed worse on component 2.

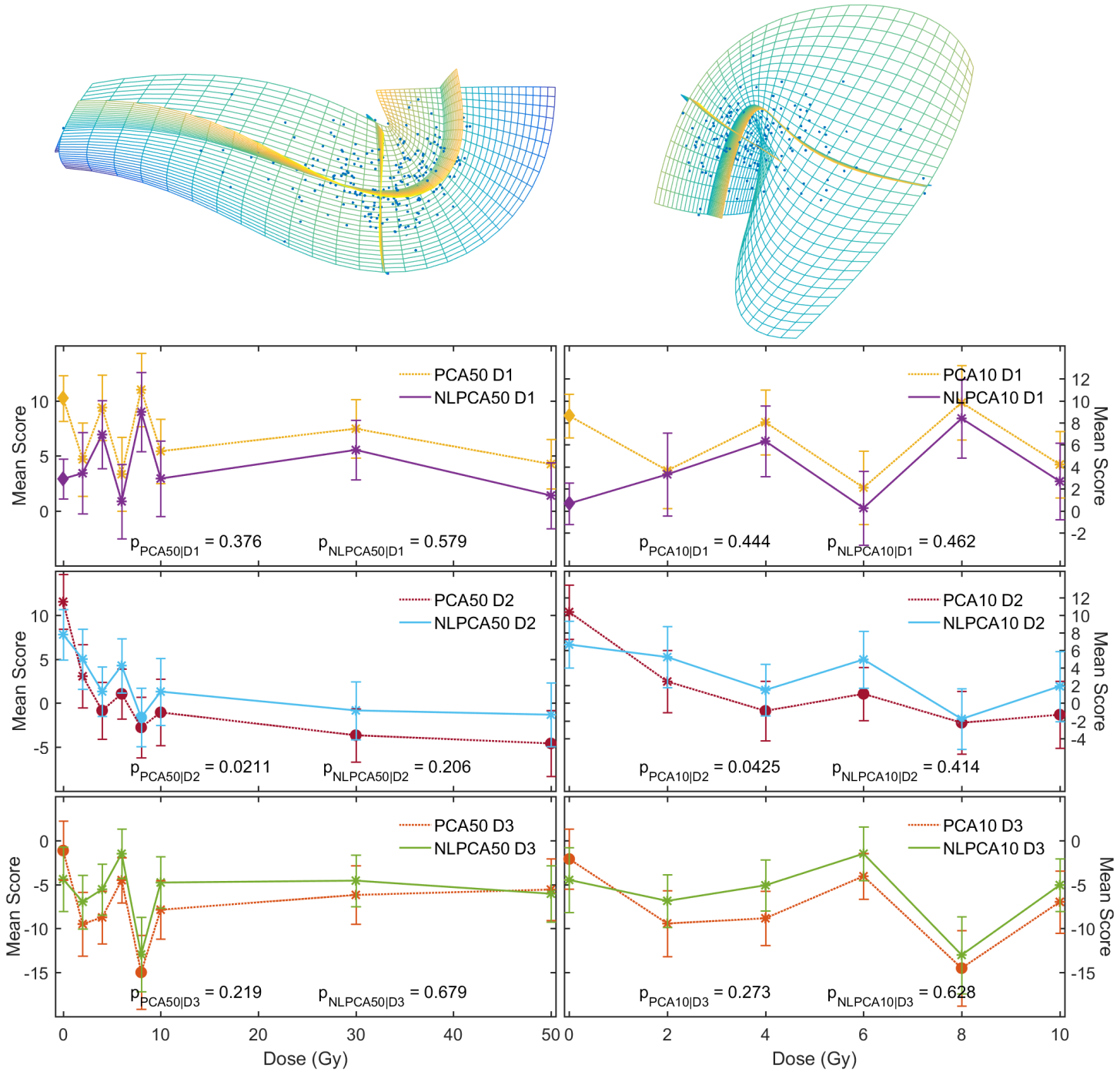


Figure 6.18: NLPCA/PCA projection and PC2 scores for H460B 50Gy/10Gy datasets. PC curves are projected into 3D principal subspace, and X-Y perspective used. Notation same as in figure 6.16.

Component 3

As with RPCA, good correspondence to PCA was noted for component 3, with slightly lower p-values.

6.3 Discussion of results

Given the vast variety of approaches tested, it is difficult to immediately determine the optimal one. This section discusses key properties and performance metrics of studied algorithms, and ranks them by component and dose set.

6.3.1 Component 2 performance

Since PC2 was associated with cell cycle, its results were not as important - they are briefly outlined first. WPCA showed interesting PC2 changes with some unattenuated strong features. It had slightly better significances than PCA but inconclusive score distances. RPCA gave less significant results by a large margin, with clear loss of cell cycle information. However, it did gain several new PC peaks in high variability regions, possibly indicative of new dose independent signals being assigned to PC2. PPCA was not as bad, but still had slightly lower significances than reference results. Also, only minor PC peak changes were observed. Finally, NLPCA showed near total loss of significant results due to major unexplained shifts in 0Gy populations. Given above results, it was concluded that PCA already extracted near maximal amounts of signal for cell cycle component, resulting in insignificant improvements from robust analysis.

6.3.2 Component 1 performance

It was known from the start that H460 dataset fared well even under usual PCA analysis due to dominating radiation response component. From the performance metrics (tables 6.1 and 6.2) however it is apparent that WPCA, PPCA, and NLPCA performed notably better even under these conditions. The other method, RPCA, also did achieve better results but had fewer significant points and its p-value improvements were not as impressive. It is interesting to note how WPCA/PPCA are such opposites in terms of complexity and yet it seems that regardless of how spectral outliers were dealt with (removed or scaled down), these solutions converged to

Metric	PCA	WPCA	RPCA	PPCA	NLPCA
First detection D1	n/a	**2Gy***	n/a	**2Gy**	**2Gy***
First detection D2	4Gy	**2Gy**	4Gy	**2Gy**	4Gy
First detection D3	2Gy	2Gy	2Gy	2Gy	2Gy
Significant result Δ D1	n/a	**+1**	0	**+5**	**+1**
Significant result Δ D2	n/a	**+1**	0	**+1**	0
Significant result Δ D3	n/a	**+2****	**+2****	**+2****	**+1**
Average p-value D1	0.18	**0.14**	**0.17**	**0.044**	**0.14**
Average p-value D2	0.024	**0.0076**	**0.013**	**0.0054**	**0.0095**
Average p-value D3	0.019	**0.0046**	**0.0088**	**0.0026**	**0.0063**
Subjective examination	n/a	+++	+	++++	++

Table 6.1: H460B 50Gy PC1 performance summary. Bolded values indicate results better than PCA, with green emphasis on best one. * indicates likely spurious detection due to insignificant higher doses. ** indicates reference distribution also became significant relative to D1-0Gy. Subjective score (+/- meaning better/worse) refers to judgement based on listed metrics, score distance trends, and PC correlations.

Metric	PCA	WPCA	RPCA	PPCA	NLPCA
First detection D1	n/a	**2Gy***	n/a	**2Gy**	**2Gy***
First detection D2	4Gy	**2Gy**	4Gy	**2Gy**	4Gy
First detection D3	2Gy	2Gy	2Gy	2Gy	2Gy
Significant result Δ D1	n/a	**+1**	0	**+3**	**+1**
Significant result Δ D2	n/a	**+1**	0	**+1**	0
Significant result Δ D3	n/a	**+2****	**+1****	**+2****	**+2****
Average p-value D1	0.20	**0.15**	**0.18**	**0.048**	**0.14**
Average p-value D2	0.031	**0.011**	**0.014**	**0.0072**	**0.012**
Average p-value D3	0.024	**0.0068**	**0.0088**	**0.0031**	**0.0090**
Subjective examination	n/a	+++	+	++++	++

Table 6.2: H460B 10Gy PC1 performance summary. See 6.1 for notation.

strongest radiation response signals (although they did nonetheless have fairly different PCs). While somewhat disappointing, NLPCA results can be attributed to it being experimental and possibly not optimally tuned, as well as near-linearity of PC subspace. Reasons for poor RPCA performance were not clear since its estimators should have been at least more robust than simple variance weighting.

High dose set

By all available criteria - number of significant points, p-values, and detection levels, PPCA was found to be the best method. Moreover, its behaviour could be exquisitely tuned by changing outlier detection algorithms, masking, or significance levels. Due to fairly good (∼15s per set) performance, it was also deemed suitable for bulk automated processing usage.

Low dose set

For the same reasons as above, PPCA continued to dominate in the low dose set, although its advantages were smaller.

6.3.3 Comments

Interpreting p-values

Recall that p-value is the probability to obtain a test result as extreme or more so, in current case under assumption of Gaussian density function. Its non-linearity makes direct interpretation difficult. However, p-values can be converted to 'units' of standard deviation, which are more intuitive, via 1σ=0.32, 2σ=0.046, 3σ=0.0027, and 4σ=0.00006. Thus, above PPCA results are around 2.9σ, which is quite a bit 'stronger' in comparison to the $\sim 2.3\sigma$ result of PCA, a fact not made clear from just score distribution plots.

High variability of PC2 results

It was noticed that nearly all methods produced strongest deviations from PCA in component 2, which was assigned to cell cycle. In comparison, components 1 and 3 were close matches in all cases, even though the latter is just 30% of PC2 variance and $\sim 5\%$ of total variance. This behaviour was somewhat unexpected and may suggest a deeper interpretation for PC3. One possible explanation is that while cell cycle

changes have a wide variety of intermediate signals (i.e. cell can be measured at any point in the cell cycle, providing a continuous distribution of possible spectra), PC3 corresponds to some unknown but very sharply triggered mechanism, since most of its score changes happen in the 0-2Gy region. Once more low dose data is available, PC3 re-examination may be warranted.

6.3.4 Summary

For high radiation response signal dataset, it was found that at all doses the proposed algorithms have performed better than PCA in radiation component (PC1) but none achieved notable improvements in cell cycle component (PC2). By the metrics discussed above, PPCA was found to be the overall best method with average improvement of 0.63σ in result significance.

Chapter 7

Results and Discussion III - LNCaP low signal dataset

This chapter presents results of applying previously described analysis techniques to LNCaP cell line datasets. Analogous to previous chapter, section 7.1 discusses dataset quality, with main results presented in section 7.2. Performance comparison is given in section 7.3.

7.1 Dataset quality

LNCaP cell line was significantly more sensitive to environmental stresses and showed slower growth rate. As with H460, since two independent trials A and B were available, all analysis was performed on both and only results that matched were reported. Normality and outlier rejection tests (given in appendix A) showed that LNB set had a fairly uniform rejection rate (no anomalous batches or spectral windows) that was lower than that of H460A/B. LNA set had similar results with exception of one significantly outlying batch, D3-2Gy. As such, LNB set was used for analysis and plotting. Note that algorithm validation figures were omitted since they contained essentially same results as in previous chapter.

7.2 High and low dose analysis results

7.2.1 PCA

Component 1

Component 1 accounted for 54.2% of total variance in the high dose set and 54.5% in low dose set. It is shown in figure 7.1 along with corresponding scores and p-values. Its shape is very similar to PC2 of H460B, which was identified as cell cycle component. In terms of score trends, increasingly negative results were observed at higher doses. However, shifts were not significant until doses above 10Gy at D2 and D3. These observations were consistent with fairly few changes in cell cycle distribution, as confirmed by FACS measurements (not shown). Low dose results gave near identical PC with ρ_{HL} of 1.0, as well as same significant points, but with higher p-values.

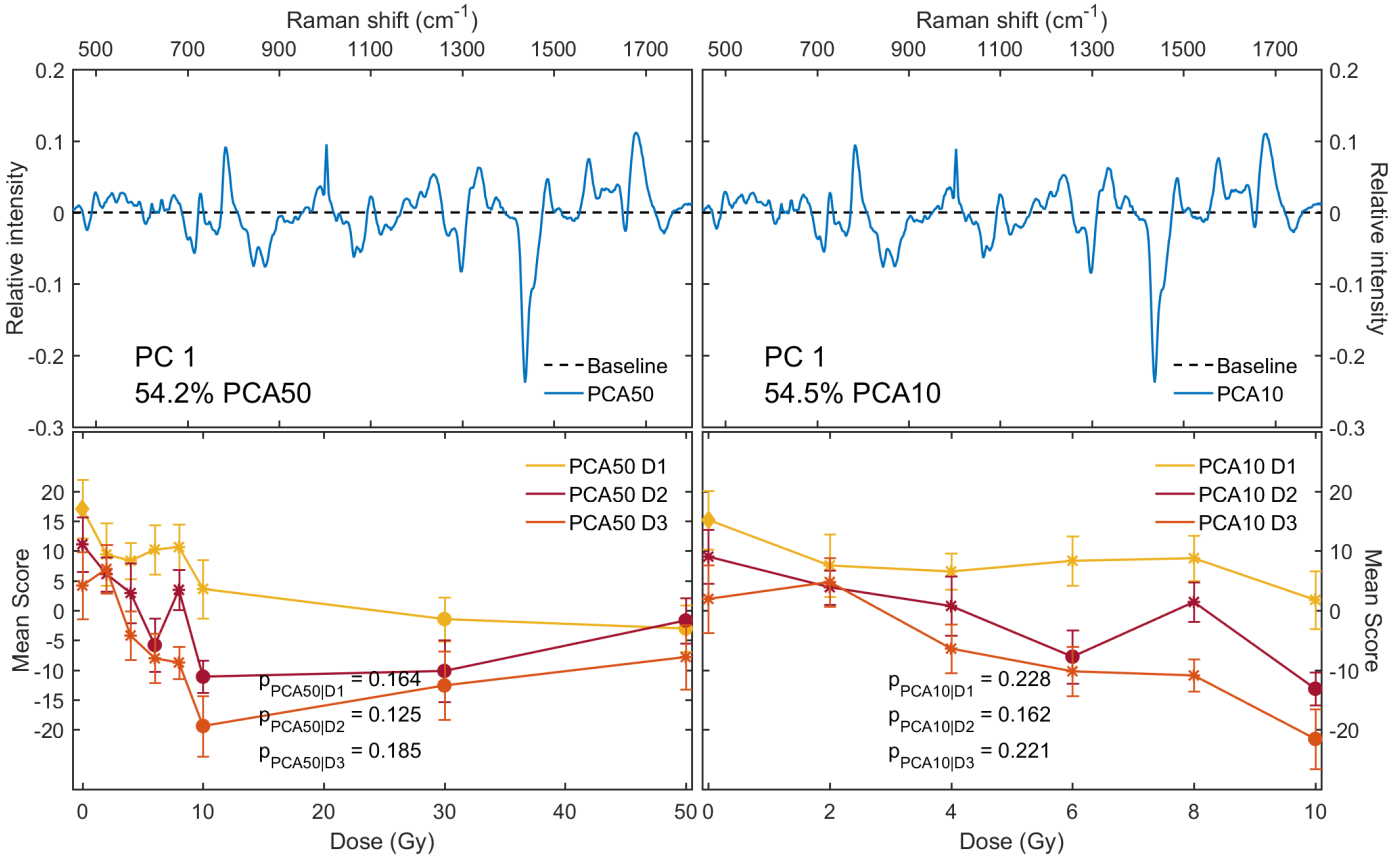


Figure 7.1: PCA component 1 and respective scores of LNB 50Gy/10Gy datasets. Solid circles denote significant results, asterisks non-significant ones, and diamond the D1-0Gy reference point.

Component 2

Component 2 accounted for 5.83% of total variance in high dose set and 6.06% in low dose set. It is shown in figure 7.2. Note how little variance PC2 accounted for compared to H460. Such fast falloff indicated more signal was being kept in low, noise filled components, consistent with generally lower LNCaP response. The main features observed were attributed to nucleic acids (+ve at 783, 1093 cm^{-1}, -ve at 733, 1136 cm^{-1}), proteins (-ve 849, 941, 1400 cm^{-1}), and lipids (+ve 1298 cm^{-1}, -ve 718 cm^{-1}). They differed markedly from those seen in H460 radiation response component. Previous works have identified some of the negative features of PC2 with choline, an important constituent of membrane phospholipids.

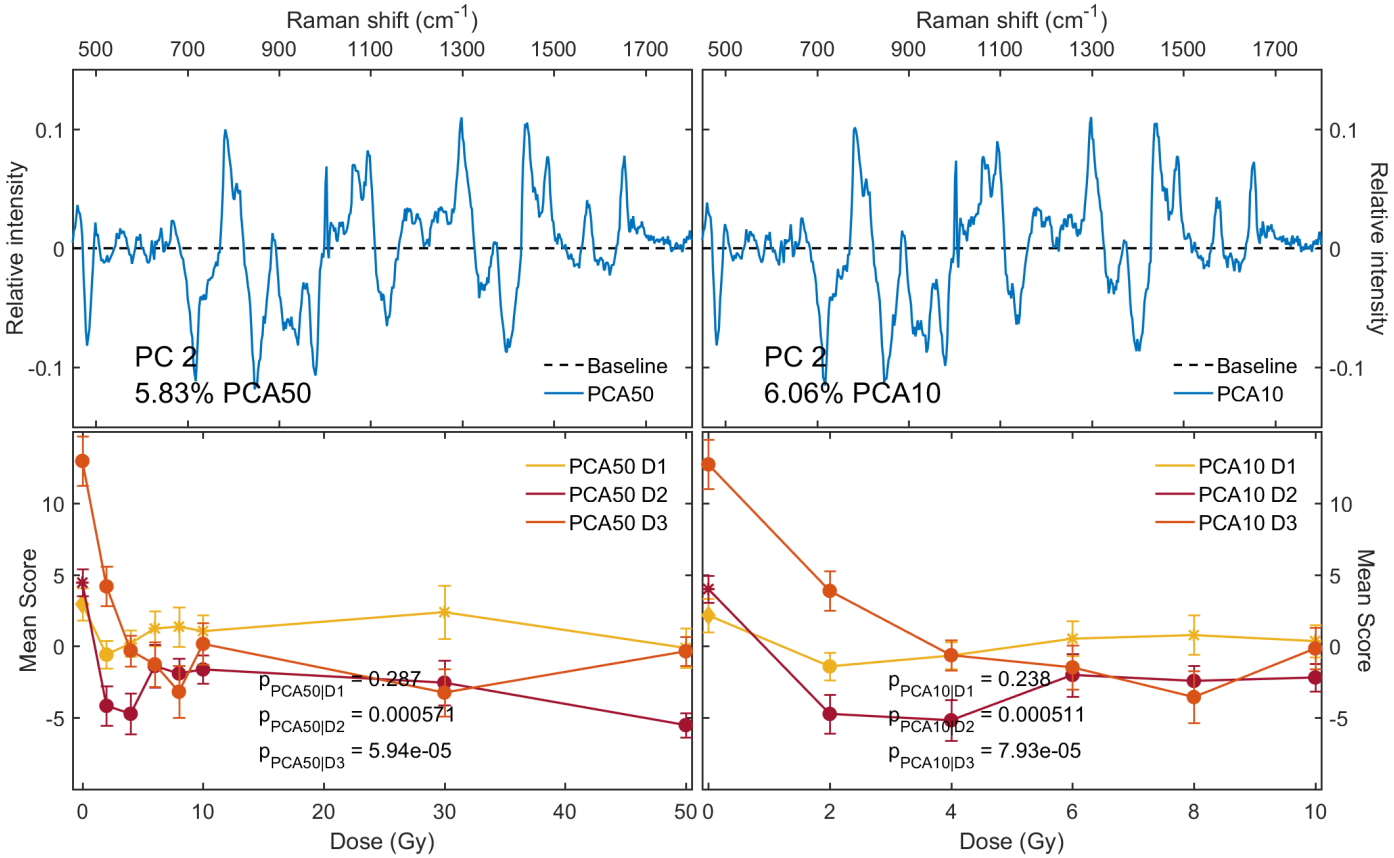


Figure 7.2: PCA component 2 and respective scores of LNB 50Gy/10Gy datasets. Symbol interpretation same as in figure 7.1.

Scores of PC2 appeared to increase with time in unirradiated population, but were reduced to near zero levels at doses above 4Gy at D1 and 2Gy at D2/D3. These trends indicated dose-dependent attenuation of choline reduction. In other words, radiation response was slowing down or even reversing the growth of positive PC2

features associated with relatively lower choline content. Low dose results were very similar, with ρ_{HL} of 1.0. Interestingly, 2 out of 3 p-values became slightly lower, possibly due to above mentioned attenuation behaviour being less prominent in low dose set and thus giving PCA a better source of variability. This will be discussed in more detail below.

Component 3

Components 3 (figure B.14) had explained variability of 4.3% in both sets and no molecular origin could be determined. It had no apparent dose variation, but showed uniform increase with time. As such, PC3 discussion is omitted for the rest of this chapter.

7.2.2 Weighted PCA

Explained variance trends were quite different from H460B case, with only first PC losing significant signal while others were approximately the same as can be seen in figure 7.3. PC2/PC1 ratio correspondingly almost doubled. Around 23% of variance was again moved into $k > 5$ noise components.

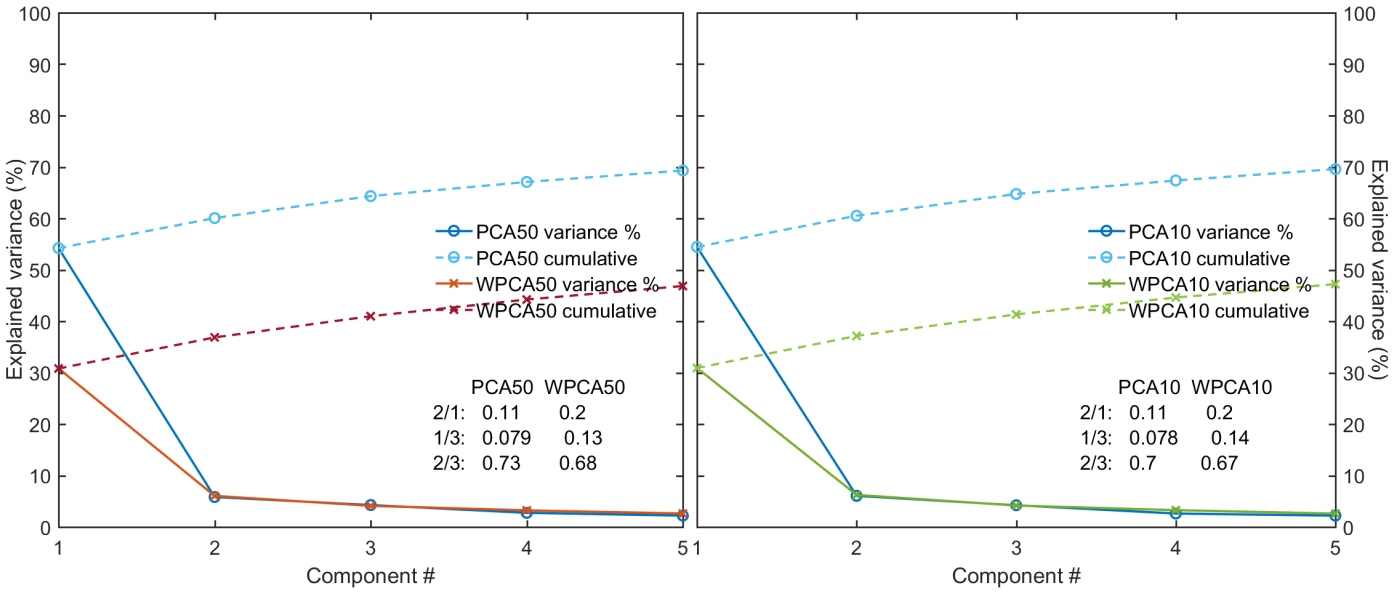


Figure 7.3: WPCA/PCA variances and respective cumulatives with $k = 479$.

Component 1

Component 1 and corresponding scores are shown in figure 7.4. Observed changes were highly similar to those of H460 PC2, with attenuation of several peaks and respective increases of smaller features. In general however there were fewer shifts as indicated by ρ_{IA50} of 0.86 (vs 0.83 in H460B). This can be attributed to more uniform outliers (and hence uniform variance) in LNB dataset. Significance differences were fairly minor with only additional results at D3-6Gy/8Gy and a loss at D3-50Gy, all of which were marginal changes. WPCA p-values were better for all days, but again just slightly. The low dose dataset showed extremely good ρ_{HL} of 1.0. No changes in significance were observed but p-values have slightly increased. Score distances (fig B.16) indicated virtually equal performance, with near zero magnitudes for both σ-scores and MM-scores. Note that all score distance plots related to this chapter were placed into appendix B due to the generally small differences observed. Overall, WPCA performance on component 1 was approximately equivalent to PCA.

Component 2

Component 2 and respective scores are shown in figure 7.5, and demonstrated interesting behaviour with certain peaks (490, 1090 cm^{-1}) remaining untouched and some even becoming stronger (1350-1400 cm^{-1}) while others were attenuated by almost a factor of two (680, 1430 cm^{-1}). Overall, results conformed more closely to reference than those of H460B with ρ_{IA50}/ρ_{IA10} of 0.89/0.90 respectively. In terms of scores, a minor but consistent trend of worse significance (and higher p-values) was observed at all times, with shifts in 0Gy distributions most notable. However, no significant results were lost. Low dose set had ρ_{HL} of 0.99, approximately equal p-values, and generally followed above trends.

In terms of score distances (fig B.17), a slight loss of signal was obvious due to predominantly positive σ-distances. Meanwhile, MM-scores consistently stayed very small indicating equal relative sensitivity. Overall, WPCA performance on component 2 was just slightly below PCA.

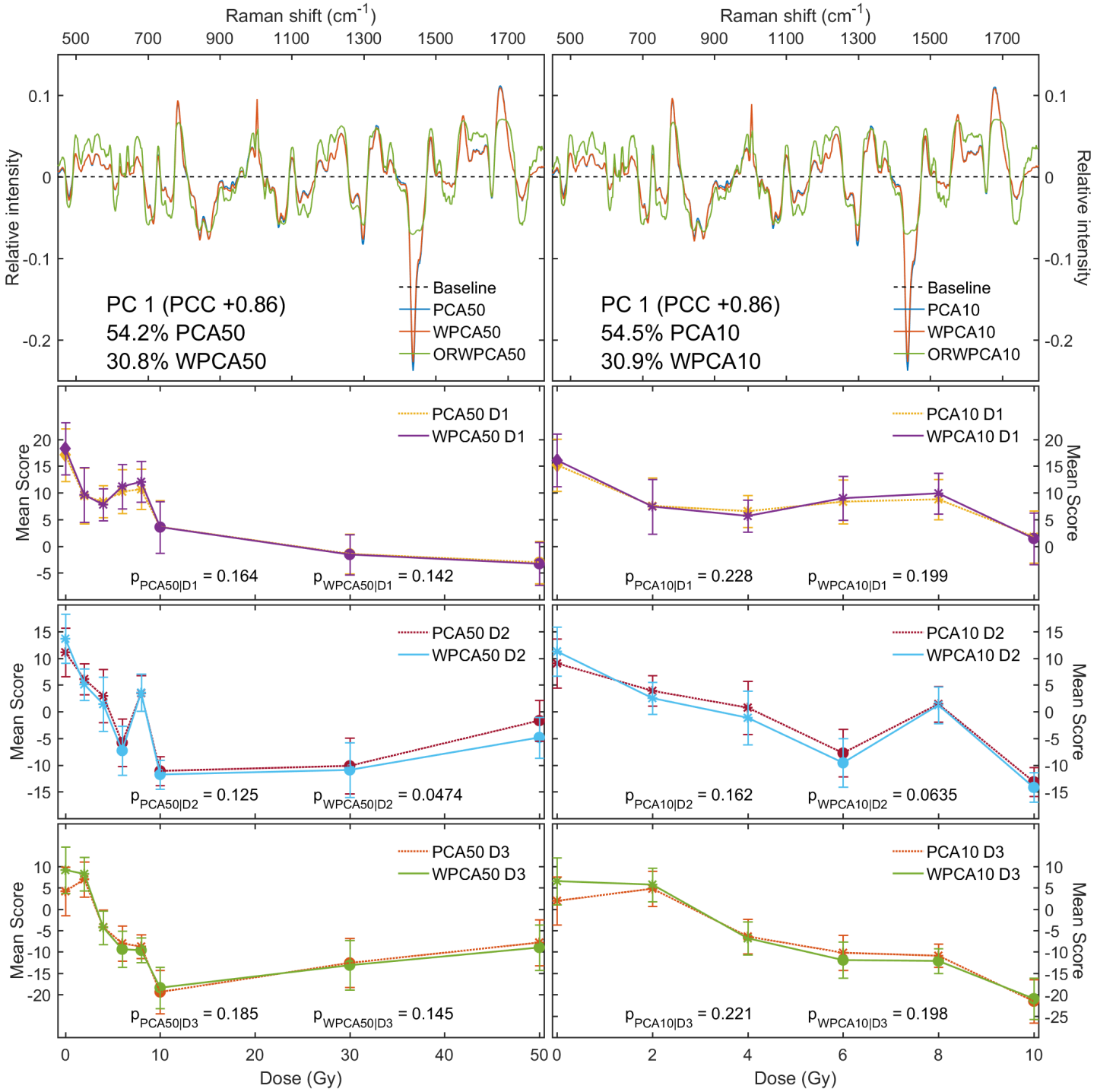


Figure 7.4: WPCA/PCA component 1 and respective scores for LNB 50Gy/10Gy datasets. Symbols same as in figure 7.1. Relative explained variability is given in lower left corner of PC plots along with PPC between PCA and ORWPCA. Score plots show respective average p-values.

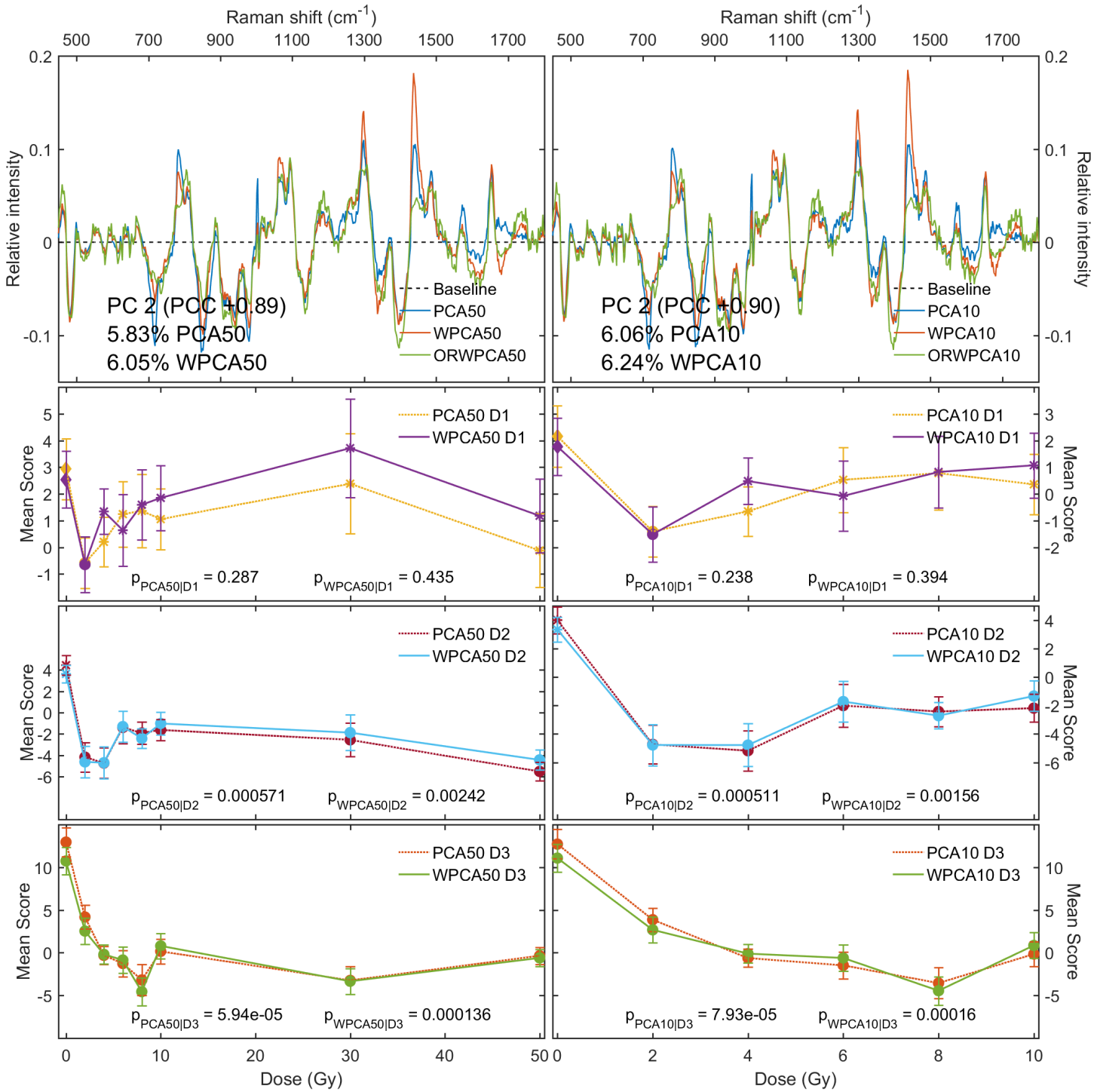


Figure 7.5: WPCA/PCA component 2 and respective scores for LNB 50Gy/10Gy datasets. Notation same as in figure 7.4.

7.2.3 Robust PCA

Variances renormalized to $k = 3$ are plotted in figure 7.6, and show first observed increase in PC1 variance. However, this was balanced by a bigger opposite difference on PC2 for a net loss of variance in top two components.

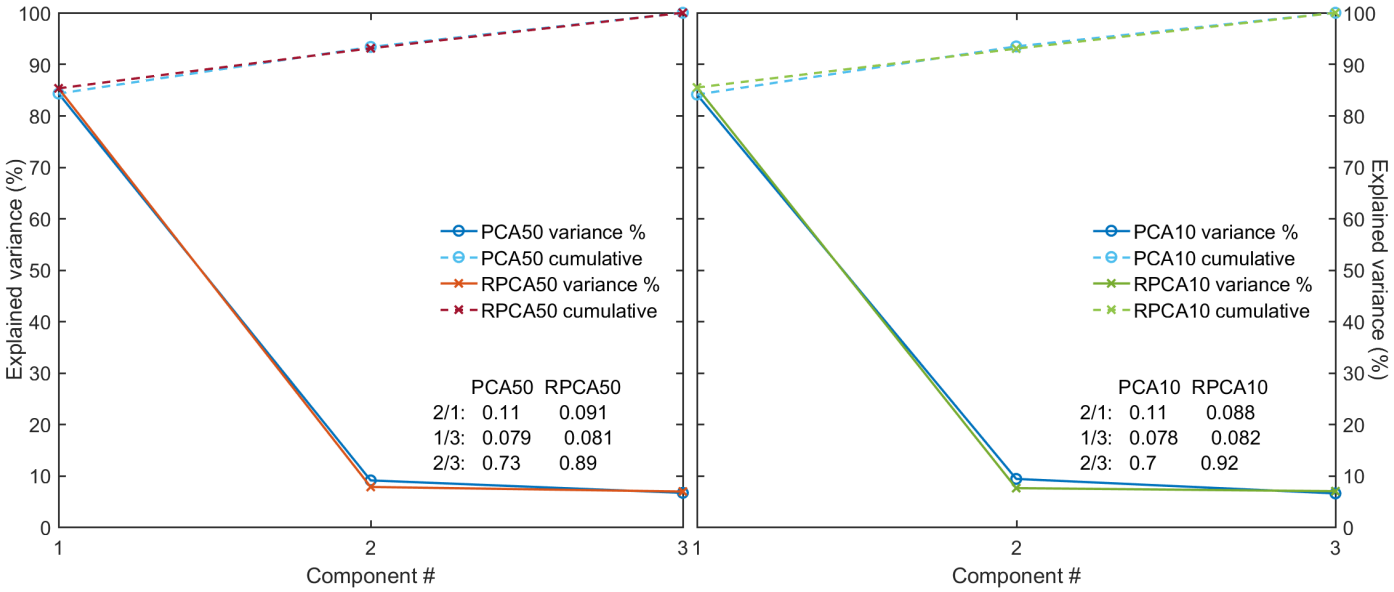


Figure 7.6: RPCA/PCA variances and respective cumulatives calculated with $k = 3$.

Component 1

Virtually no differences were observed in PC1, with ρ_{IA50} and ρ_{IA10} of 1.0, as is shown in figure 7.7. This remarkable result corresponded to barely noticeable score shifts and no significance changes, likely explained by relatively strong PC1 signal combined with few outlier in LNB set. Low dose results were correlated perfectly at ρ_{HL} of 1.0, with no significance differences observed. Corresponding score distances (figure B.20) showed no trends and very small magnitudes. Overall, RPCA performance on component 1 was approximately equivalent to PCA.

Component 2

Significantly larger changes were observed in PC2, with attenuation in 710 and 890 cm^{-1} regions and notably stronger signals around 1250 and 1660 cm^{-1}. Corresponding scores have also changed, with a loss of D1-2Gy significant result and generally higher p-values. Score distances (figure B.21) confirmed this loss of sensitivity, with almost uniformly positive σ-distances.

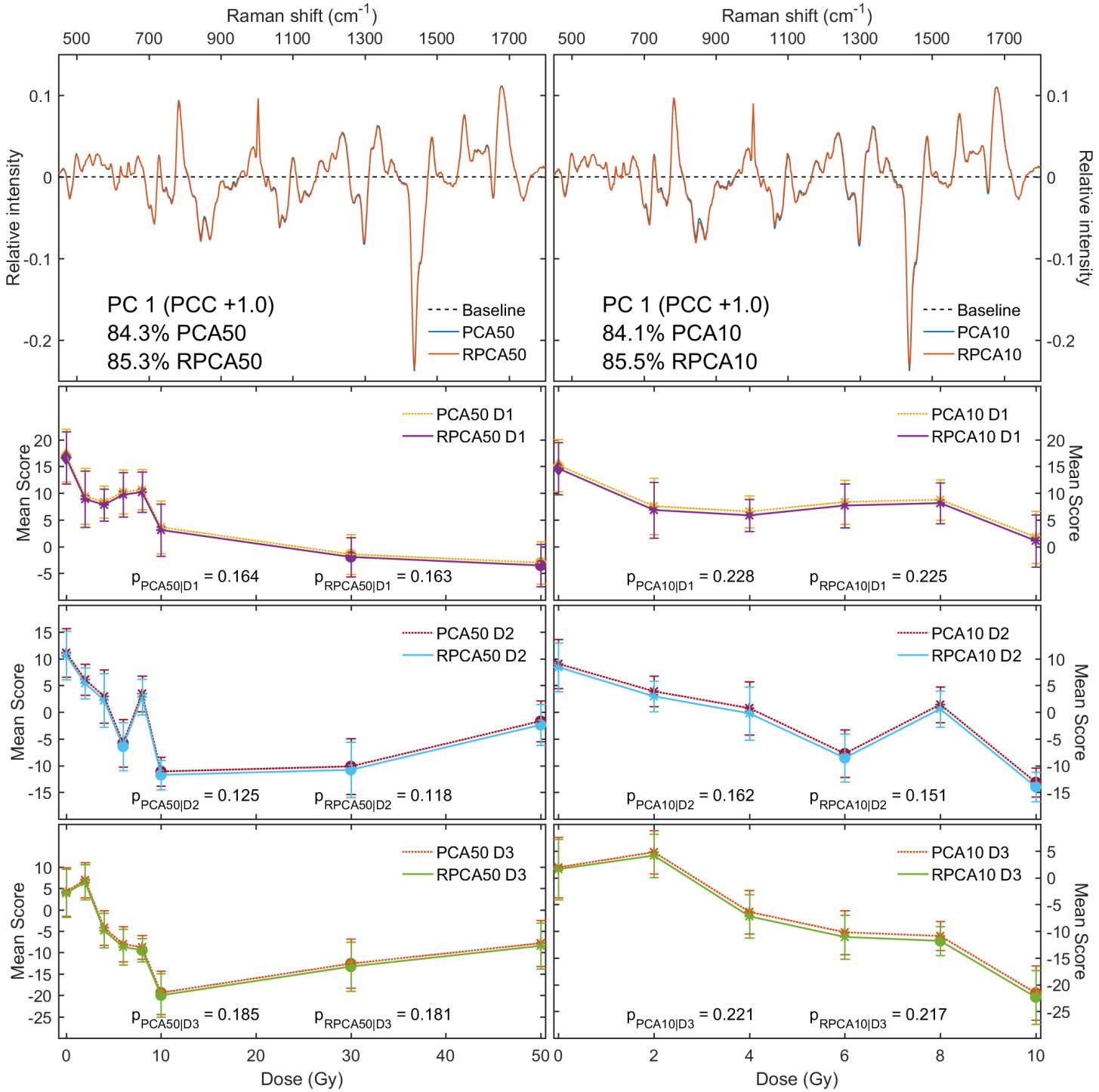


Figure 7.7: RPCA/PCA component 1 and respective scores for LNB 50Gy/10Gy datasets. Notation same as in figure 7.4, except that PCC is between PCA and RPCA.

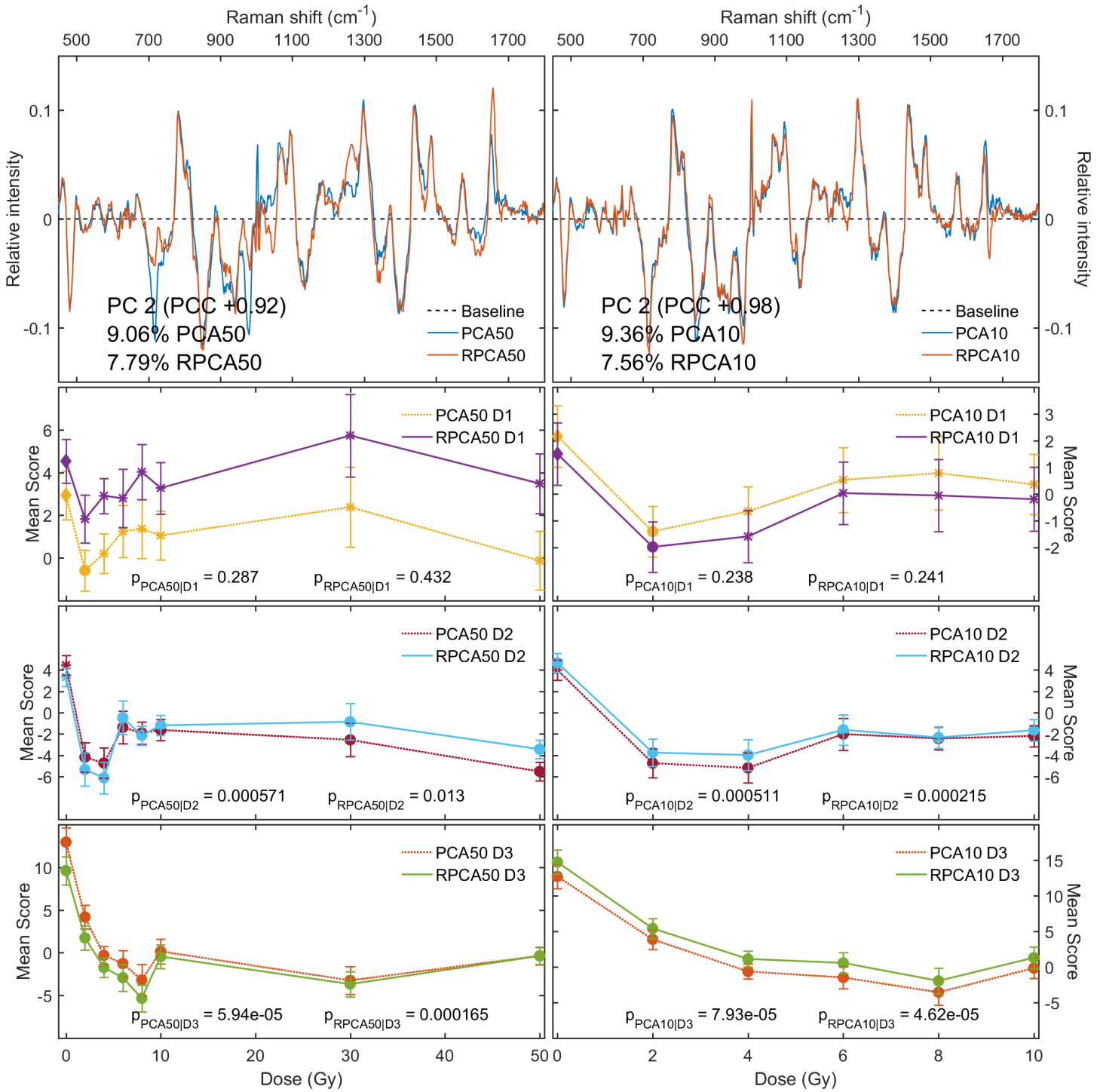


Figure 7.8: RPCA/PCA component 2 and respective scores for LNB 50Gy/10Gy datasets. Notation same as in figure 7.7.

Surprisingly, low dose set results differed markedly with ρ_{HL} of only 0.82 and none of the spectral changes observed previously. This was confirmed by IA correlations of $\rho_{IA50} = 0.92$ and $\rho_{IA10} = 0.98$, indicating better low dose agreement with PCA. Moreover, significance was regained for D1-2Gy point and p-values slightly lower than PCA ones were observed. One possible explanation for this behaviour is that high dose set RPCA has eliminated some part of radiation response as an outlier, since that dataset corresponds to steadier levels of choline signal resulting in it accounting for lower amounts of total variability.

7.2.4 Probabilistic PCA

Same outlier removal parameters were used as in H460 section for total rejection of 14.6%/14.7% of data in high and low dose sets respectively, slightly below H460B values as expected. Explained variances remained nearly the same and are shown in figure 7.9.

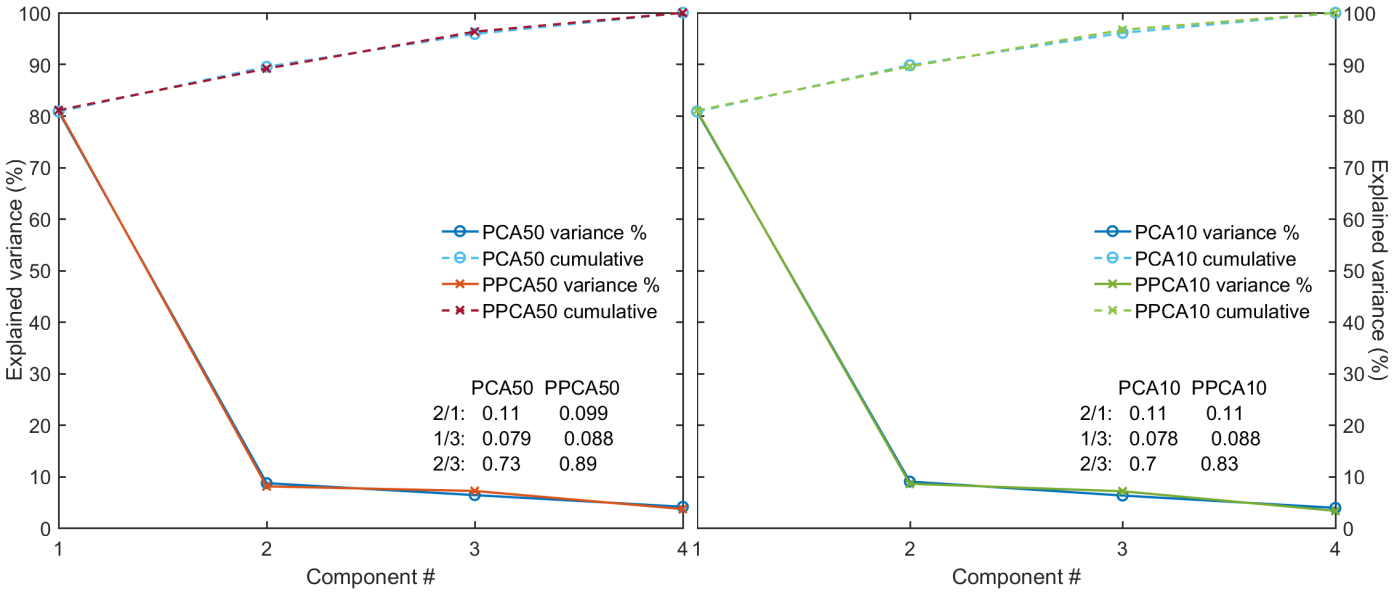


Figure 7.9: PPCA/PCA variances and respective cumulatives calculated with $k = 4$.

Component 1

As with RPCA above, results (figure 7.10) nearly matched PCA with slight but uniform PC attenuation, as is confirmed by ρ_{IA50} and ρ_{IA10} of 1.0. Two new significant results at D1-10Gy and D3-8Gy were observed, but these were clearly marginal. In

general, PPCA p-values were slightly lower but the difference was small. Score distances (figure figure B.24) had correspondingly low magnitudes, with slight downward trend in σ-distances but positive MM-distances. Low dose set showed no notable differences with $\rho_{HL} = 1.0$, same significant points, and slightly higher p-values. Overall, PPCA performance on component 1 was approximately equivalent.

Component 2

Larger changes were observed in component 2 (figure 7.11), with attenuation in 720 and 970-1000 cm^{-1} regions. No new peaks were noted, although this may be due to σ-scaling of the components confounding qualitative relative magnitude comparison (i.e. PPCA component is in general slightly smaller, but some peaks were not reduced as much as others). Two points became significant, D1-4Gy and D1-50Gy. Correspondingly lower p-values were observed, with increasing relative differences with time. Low dose set produced slightly noisier PC2, with ρ_{HL} of 0.99. Nonetheless, it exhibited the same slight attenuations with no new signal peaks. All significance results remained the same, and 2 of 3 p-values decreased. This again suggested that relatively more choline signal variation in the low dose set helped better elucidate radiation response, although since IA correlations were similar at $\rho_{IA10} = 0.97$ and $\rho_{IA50} = 0.98$, situation was not as clear as in RPCA case. In terms of score distances (figure B.25), σ-distances showed increasingly negative trends and MM-distances remained largely negative as well, consistent with better sensitivity. Overall, PPCA performed better on component 2.

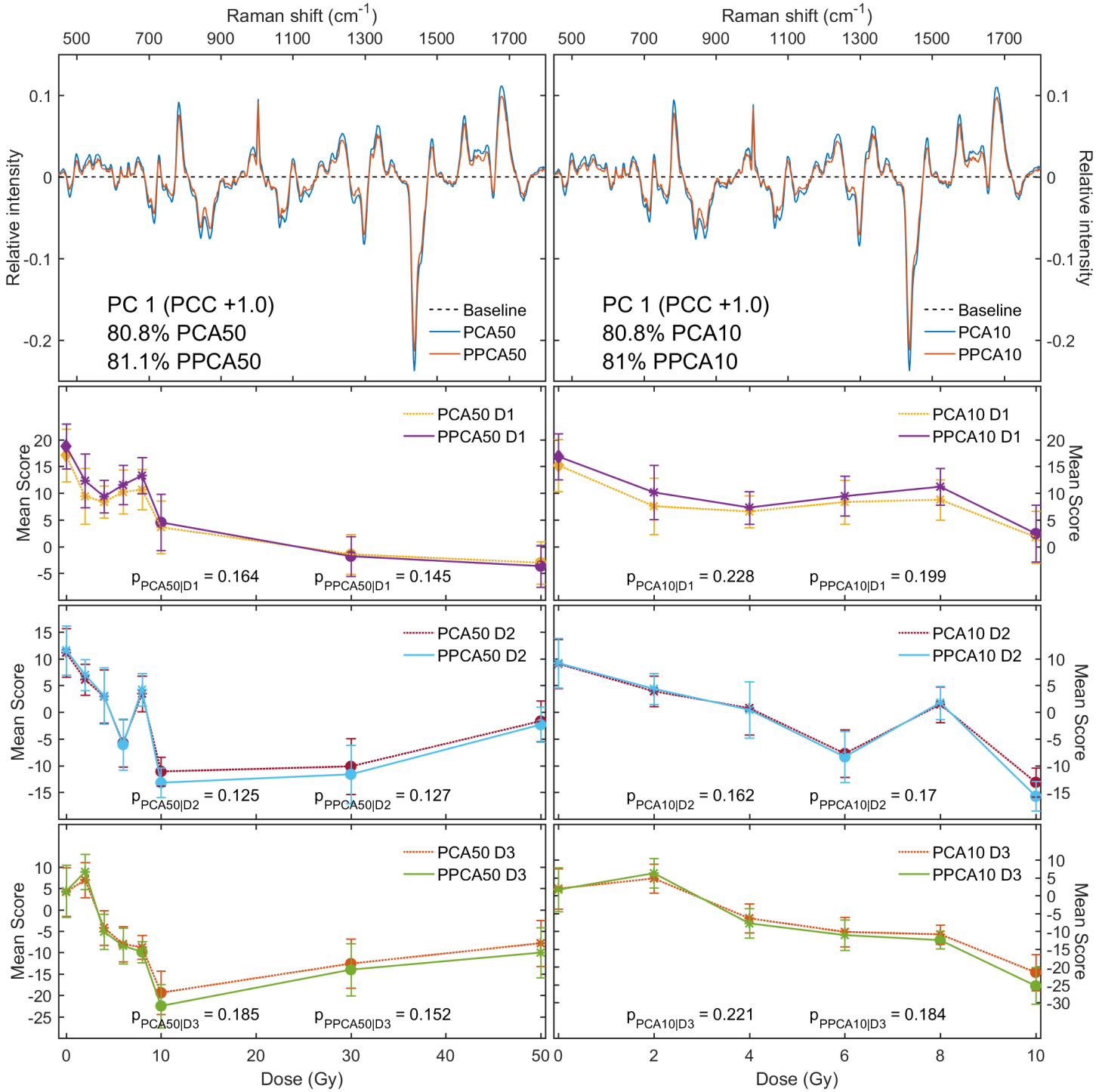


Figure 7.10: PPCA/PCA component 1 and respective scores for LNB 50Gy/10Gy datasets. Notation same as in figure 7.7. Note that 14.6%/14.7% of data was removed from 50Gy/10Gy sets respectively.

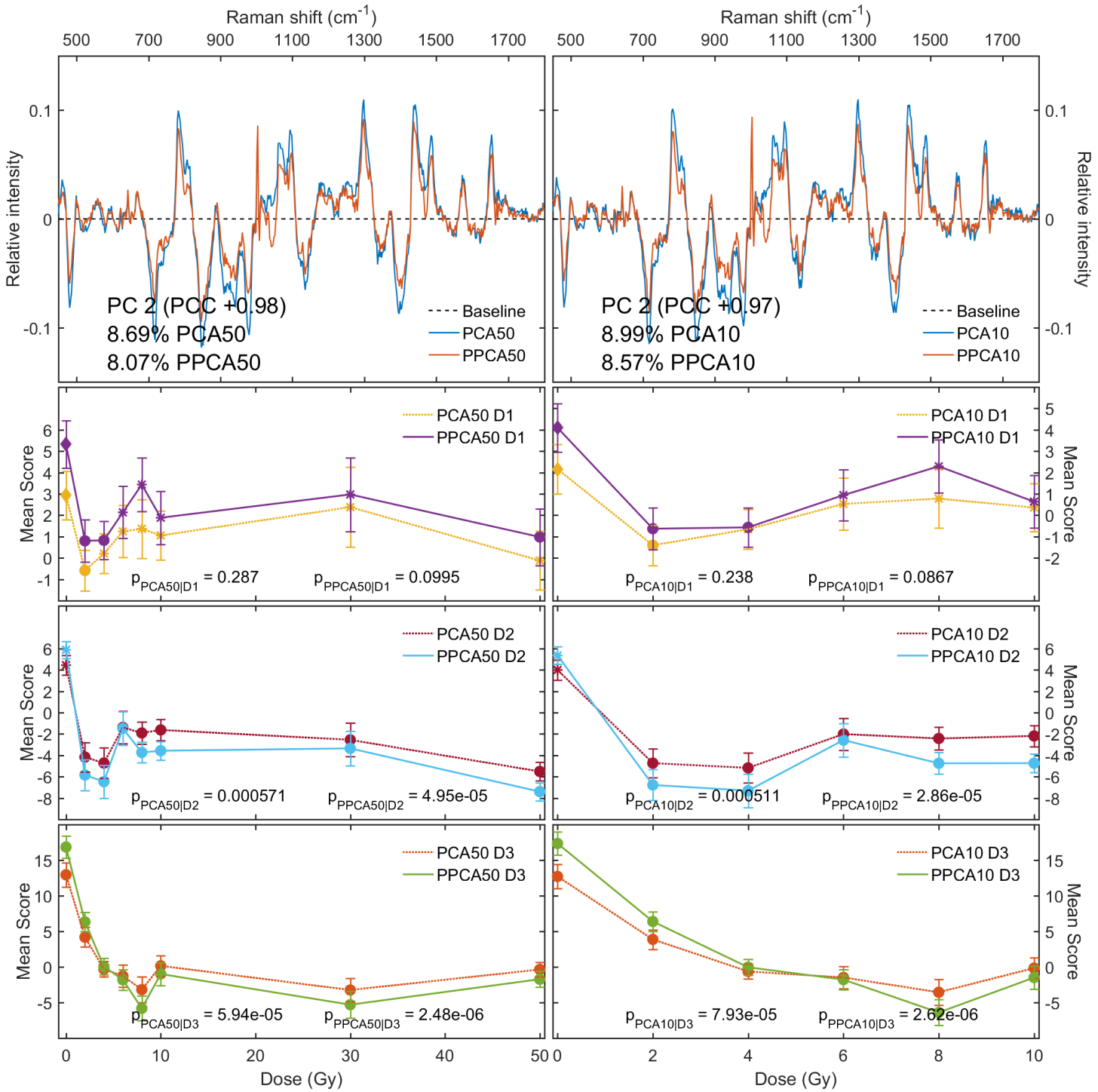


Figure 7.11: PPCA/PCA component 2 and respective scores for LNB 50Gy/10Gy datasets. Notation same as in figure 7.10.

7.2.5 Nonlinear PCA

Component 1

Recall that for H460B, a tail was observed in high dose set projection which consisted of 30Gy/50Gy strong radiation response spectra. This was not the case for LNB, as is shown in figure 7.12. Points remained in a compact cluster for both sets, with high dose ones having roughly the same distribution as others. This was consistent with the majority of data, especially the high dose points, having no choline signal as well as roughly steady cell cycle component. The PC curves were near linear, with only slight curvature appearing at the edges (best seen in X-Y projection). These results suggested that linear dimensionality reduction methods should perform near-optimally on LNB dataset.

In terms of significances there were no new results and a marginal loss at D3-30Gy. Correspondingly, p-values were almost identical and remained so in low dose set. Score distances (figure B.28) gave no useful data due to very low magnitudes. Overall, NLPCA performance on component 1 was approximately equivalent to PCA.

Component 2

Component 2 results showed no changes in significance and roughly same p-values. For low dose set, p-values have increased marginally at D2 and D3, but remained close to PCA ones. Score distances (figure B.29) again gave little useful information. Overall, NLPCA performance on component 2 was approximately equivalent to PCA.

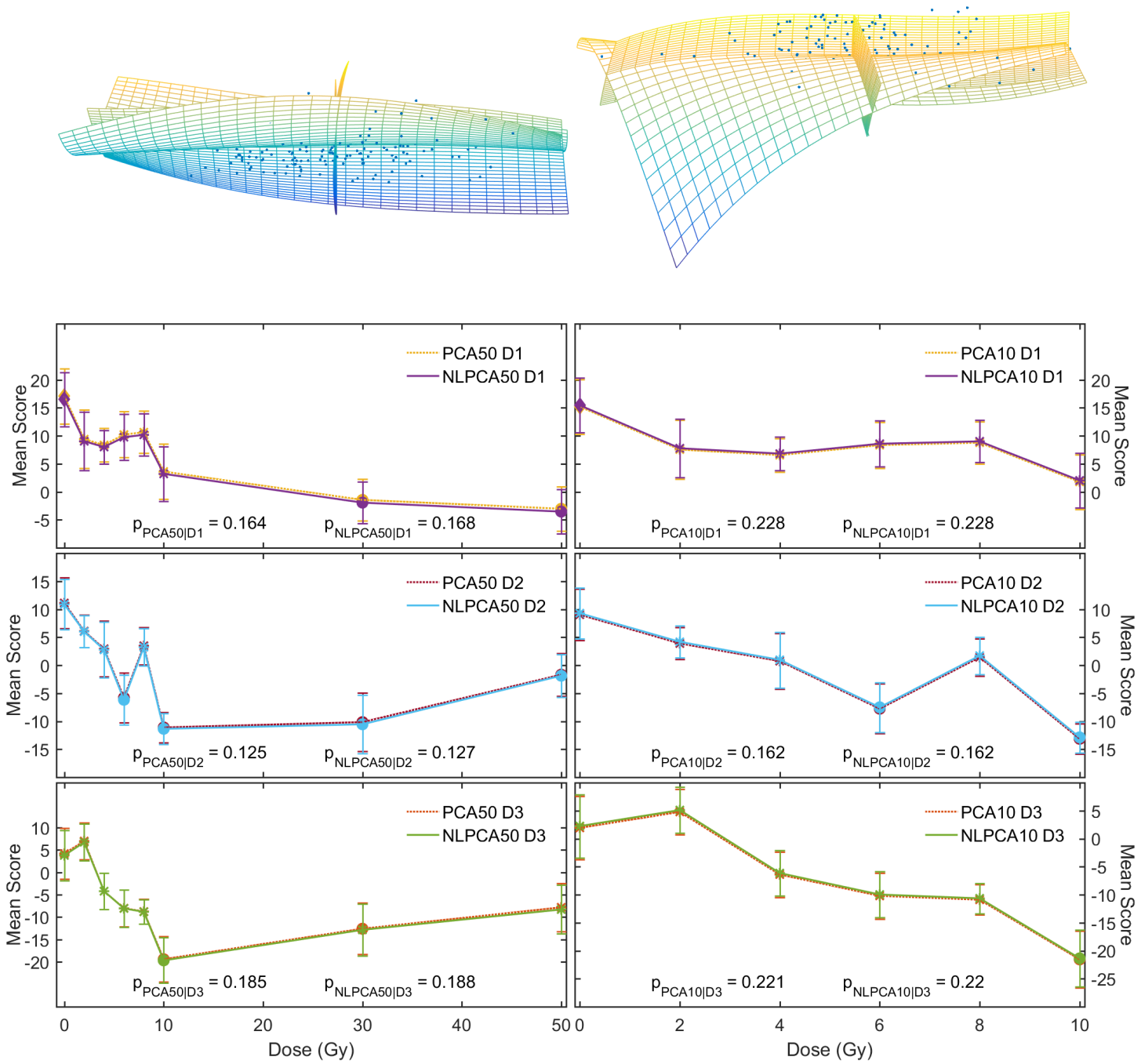


Figure 7.12: NLPCA/PCA projection and PC1 scores for LNB 50Gy/10Gy datasets. PC curves are projected into 3D principal subspace, and X-Z perspective used. Solid circles denote significant results, asterisks non-significant ones, and diamond the D1-0Gy reference point.

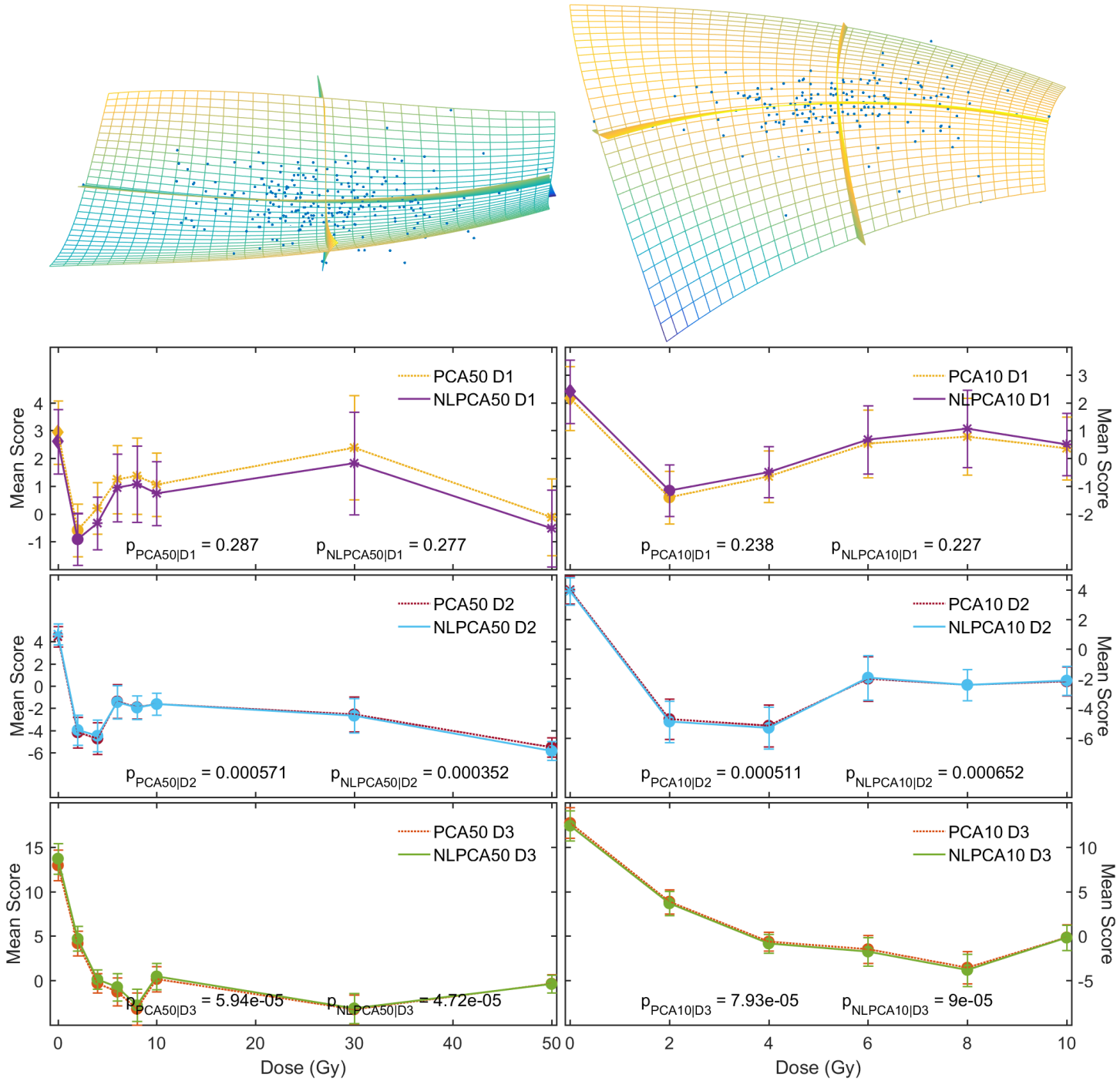


Figure 7.13: NLPCA/PCA projection and PC2 scores for LNB 50Gy/10Gy datasets. PC curves are projected into 3D principal subspace, and X-Y perspective used. Notation same as in figure 7.12.

7.3 Discussion of results

7.3.1 Component 1 performance

Since component 1 was almost identical to H460B PC2, its behaviour was expected to be very close to what was seen previously. This was in fact the case - generally either same and worse performance was observed but in all cases differences were minimal. Of the methods tested, WPCA had largest PC changes with correlations of under 0.9 for both dose sets but these did not result in notable score significance or p-value shifts. As such, it was concluded that PCA performance on component 1 was already near-optimal. This was likely due to PC1 dominating explained variability (and hence having strong signal) combined with low score dose dependence and few LNB outliers, all of which essentially reduced robust analysis methods to simple PCA.

7.3.2 Component 2 performance

Radiation response component results of LNB dataset were interesting due to combination of low explained variability and signal being associated with attenuation of chemical changes, both of which are the opposite of H460B PC1 case. Performance metrics are summarized in tables 7.1 and 7.2, and show that most methods did not achieve desired signal improvements. WPCA again had highest PC changes but they were smaller than in H460B case, consistent with more uniform data variance. Unfortunately, they also resulted in increased p-values and slightly reduced overall performance. RPCA behaviour was very interesting in that loss of signal in high dose set was recovered in low dose run, with p-values decreasing by almost an order of magnitude. This was associated with restoration of near-PCA PC2 as indicated by correlation change from 0.92 to 0.98. PPCA did not exhibit this behaviour, with significance and p-value improvements ($\sim 0.65\sigma$) in both sets as well as clear PC2 differences. It also removed less data than in H460B case, confirming lower number of outliers. Finally, NLPCA demonstrated near-linearity of first 3 principal components, and produced essentially same results as PCA.

Metric	PCA	WPCA	RPCA	PPCA	NLPCA
First detection D1	2Gy*	2Gy*	n/a	**2Gy**	2Gy*
First detection D2	2Gy	2Gy	2Gy	2Gy	2Gy
First detection D3	2Gy	2Gy	2Gy	2Gy	2Gy
Significant result Δ D1	n/a	0	-1	**+2**	0
Significant result Δ D2	n/a	0	0	0	0
Significant result Δ D3	n/a	0	0	0	0
Average p-value D1	0.29	0.44	0.43	**0.10**	**0.28**
Average p-value D2 ($\times 10^{-4}$)	5.7	24.2	130	**0.50**	**3.5**
Average p-value D3 ($\times 10^{-4}$)	0.59	1.36	1.65	**0.025**	**0.47**
Subjective examination	n/a	$-$	$--$	$++$	$\sim$same

Table 7.1: LNB 50Gy PC2 performance summary. Bolded values indicate results better than PCA, with green emphasis on best one, while red ones correspond to worse results. * indicates likely spurious detection due to insignificant higher doses. Subjective score (+/- meaning better/worse) refers to judgement based on listed metrics, score distance trends, and PC correlations.

Metric	PCA	WPCA	RPCA	PPCA	NLPCA
First detection D1	2Gy*	2Gy*	2Gy*	2Gy*	2Gy*
First detection D2	2Gy	2Gy	2Gy	2Gy	2Gy
First detection D3	2Gy	2Gy	2Gy	2Gy	2Gy
Significant result Δ D1	n/a	0	0	**+1**	0
Significant result Δ D2	n/a	0	0	0	0
Significant result Δ D3	n/a	0	0	0	0
Average p-value D1	0.24	0.44	0.24	**0.087**	**0.23**
Average p-value D2 ($\times 10^{-4}$)	5.1	15.6	**2.2**	**0.29**	6.5
Average p-value D3 ($\times 10^{-4}$)	0.79	1.6	**0.46**	**0.026**	0.90
Subjective examination	n/a	$-$	$\sim$same	$++$	$\sim$same

Table 7.2: LNB 10Gy PC2 performance summary. See 7.1 for notation.

High dose set

As with H460, performance criteria clearly showed PPCA to be the optimal method, being the only one to get any new significant results along with lower p-values.

Low dose set

Although RPCA and NLPCA were able to produce better results, PPCA was still the superior approach, albeit with slightly worse performance than in high dose case.

7.3.3 Comments

Radiation response vs actual signal

Radiation dependent attenuation of choline reduction for LNCaP cell line demonstrated weaknesses of certain robust approaches, especially RPCA. This was likely due to PCA and related algorithms seeking maximum explained variance. In H460 case, glycogen signal continued to change with increasing dose (and thus kept accounting for more and more variability), which allowed better signal detection in high dose set. However, for LNCaP the attenuation resulted in steady high dose signal which did not produce further variability. This forced full dataset analysis to place less weight on low dose radiation induced changes. In other words, for a dataset with more cells of same choline content, changes in it at low doses were not as important and in some cases mistakenly rejected. In contrast, PPCA performed well because it looked for outliers only within individual 20-point populations but did not reject any general low variability signals. Clearly, some robust methods have to be applied only to regions which contain sufficient radiation response variability so as to ensure that outliers are detected and rejected properly. As such, for LNCaP cell line further data at 0-8Gy would be beneficial but any >10Gy measurements detrimental.

7.3.4 Summary

For low signal dataset, it was found that all methods performed equivalently on dominant cell cycle component (PC1), indicative of already near optimal PCA performance. For radiation response component (PC2), only PPCA managed to produce improved results with all sets, while other methods generally performed worse or equally. This behaviour was attributed to low response strength, as well as the steady high dose signal attenuation confusing some outlier rejection methods.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

The primary goal of this work was to develop multivariate analysis techniques that would improve detection of radiation response signal. This was achieved by first exploring outlier presence and trends with formal rejection methods in chapter 5. The two studied cell lines were found to have markedly different characteristics, with H460 outlier counts showing strong dose and time dependence as well as problematic wavenumber windows, while LNCaP ones were approximately uniform, demonstrating the need for an adaptive robust approach.

Based on these observations, several promising modification of PCA were implemented and tested on two datasets with high and low strength of radiation response corresponding to different biochemical changes. In chapter 6 it was found that for high signal H460 cell line, all of proposed methods have performed better on 0-10Gy and 0-50Gy sets based on p-value, significance, and score distance metrics as defined in chapter 4. The best method, PPCA, has on average achieved 0.63σ improvement in p-values along with a lower detection threshold of 2Gy (vs 4Gy for PCA) while maintaining almost equivalent cell cycle signal metrics.

Application of above methods to low signal LNCaP cell line was explored in chapter 7, and has demonstrated that only PPCA was consistently better while other methods suffered due to the attenuation of radiation related chemical changes at higher doses as well as low radiation signal explained variability. Surprisingly, some methods such as RPCA obtained worse results in the 0-50Gy set as compared to 0-10Gy dose region. This was likely due to the low dose set having more strong radi-

ation response changes, which prevented erroneous outlier rejection. However, in the end only PPCA was able to achieve notable improvements, with p-value increase of 0.65σ but only 1-2 new significant points.

Based on above results, it was concluded that this work has succeeded in developing a viable robust multivariate analysis algorithm - PPCA. It was shown to perform better in a variety of conditions in terms of both signal strength and its dose dependence, while also maintaining good computational efficiency without numerical precision issues. Moreover, in further analyses with larger datasets, it is expected to perform even better to due its flexibility in outlier detection method selection and increasing statistical power at larger sample sizes.

8.2 Future work

8.2.1 Data visualization algorithms

Visualization algorithms represent a special class of dimensionality reduction techniques, in that unlike others they do not aim for quantitative separation but instead try to cluster information in most visually distinct way. They include isomaps, LLE, sammon mapping and others [166, 167]. One in particular stands out - t-distributed stochastic neighbour embedding (t-SNE), proposed in 2008 by Maaten and Hinton [168, 169]. It is a nonlinear dimensionality reduction algorithm that excels at mapping onto 2 or 3 dimensional spaces, trying to place similar inputs close to each other and dissimilar ones far away. Its results on multiple reference datasets (such as MNIST handwritten digit dataset) were consistently good, with distinct clusters and no crowding towards central regions. It has also been successfully used in bioinformatics and cancer research (ultrasound, MRI, and mammography) [170, 171] studies. As a simple demonstration, 2D t-SNE with default setting was applied on H460B 50Gy dataset and most visually appealing result of 10 runs chosen (figure 8.1).

Clear dose dependent clustering was observed and remained consistent between runs, as were outliers (for instance, at D3 one can see a single 2Gy point far within 50Gy group, verified to also be an outlier in PCA scores). While t-SNE does not provide useful quantitative results, it is fast and can be done for outlier removal purposes even during collection. Moreover, it provides an easy way to check for any trends in the data and can be used as a preliminary examination tool as well as a guide for more advanced analysis.

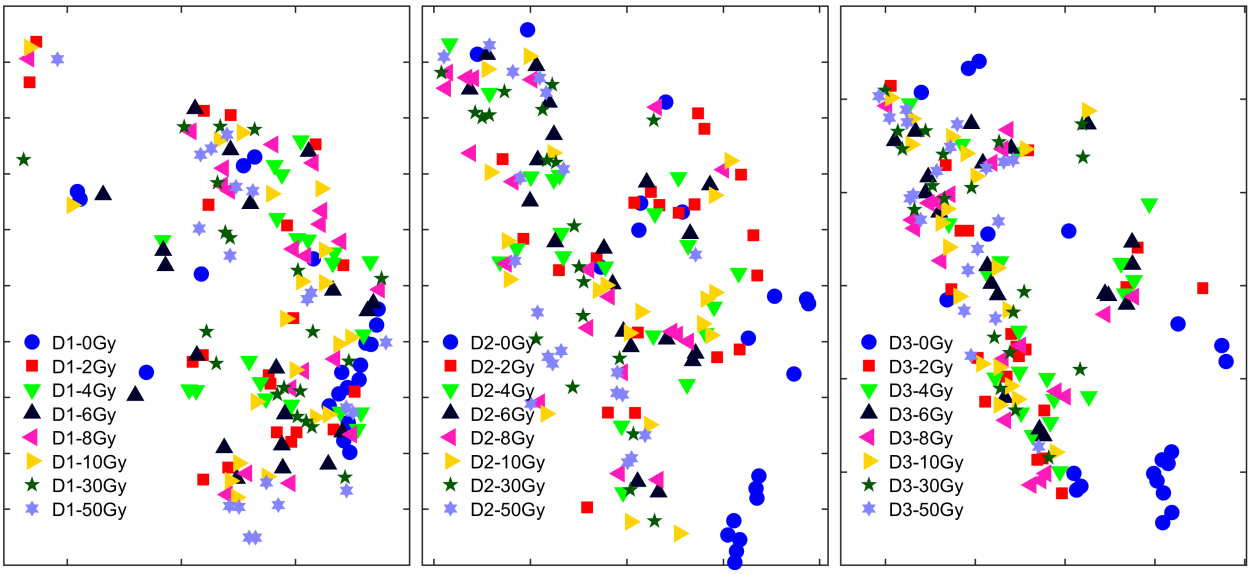


Figure 8.1: t-SNE of H460B 50Gy dataset separated by day.

8.2.2 Predictive model design

In this work all supervised algorithms were excluded but there is a wealth of information such as viability counts and cell cycle ratios that is available in addition to Raman spectra. For clinical applications a development of predictive model will eventually be required to estimate patient radiation response, based only on a spectral dataset (at several doses) and classification labels. Finding and validating candidates for such a model should be a big focus of future research.

Bibliography

[1] R. Lozano et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380(9859):2095–2128, 2012.

[2] R. P. Symonds, C. Deehan, C. Meredith, and J. A. Mills. *Walter and Miller's Textbook of Radiotherapy: Radiation Physics, Therapy and Oncology*. Churchill Livingstone, 7th edition, 2012.

[3] N. Howlader, A. M. Noone, M. Krapcho, D. Miller, K. Bishop, S. F. Altekruse, and C. L. Kosary. SEER Cancer Statistics Review, 1975-2013. Technical report. National Cancer Institute. Bethesda, MD, 2016.

[4] N. P. Ploquin and P. B. Dunscombe. The cost of radiation therapy. *Radiother. Oncol.*, 86(2):217–223, 2008.

[5] J. Hayman, J. Weeks, and P. Mauch. Economic analyses in health care: an introduction to the methodology with an emphasis on radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.*, 35(4):827–841, 1996.

[6] J. Ragaz, I. A. Olivotto, J. J. Spinelli, N. Phillips, S. M. Jackson, K. S. Wilson, M. A. Knowling, C. M. Coppin, L. Weir, and K. Gelmon. Locoregional radiation therapy in patients with high-risk breast cancer receiving adjuvant chemotherapy: 20-year results of the British Columbia randomized trial. *J. Natl. Cancer Inst.*, 97(2):116–126, 2005.

[7] J. D. Cox, J. Stetz, and T. F. Pajak. Toxicity criteria of the radiation therapy oncology group (RTOG) and the European organization for research and treatment of cancer (EORTC). *Int. J. Radiat. Oncol. Biol. Phys.*, 31(5):1341–1346, 1995.

[8] K. Fu, T. F. Pajak, A. Trotti, C. U. Jones, S. A. Spencer, T. L. Phillips, A. S. Garden, J. A. Ridge, J. S. Cooper, K. Ang, et al. A Radiation Therapy Oncology Group (RTOG) phase III randomized study to compare hyperfractionation and two variants of accelerated fractionation to standard fractionation radiotherapy for head and neck squamous cell carcinomas: first report of RTOG 9003. *Int. J. Radiat. Oncol. Biol. Phys.*, 48(1):7–16, 2000.

[9] D. I. Thwaites and J. B. Tuohy. Back to the future: the history and development of the clinical linear accelerator. *Phys. Med. Biol.*, 51(13):343–362, 2006.

[10] J. M. Galvin, G. Ezzell, A. Eisbrauch, C. Yu, B. Butler, Y. Xiao, I. Rosen, J. Rosenman, M. Sharpe, L. Xing, P. Xia, T. Lomax, D. A. Low, and J. Palta. Implementing IMRT in clinical practice: a joint document of the American Society for Therapeutic Radiology and Oncology and the American Association of Physicists in Medicine. *Int. J. Radiat. Oncol. Biol. Phys.*, 58(5):1616–1634, 2004.

[11] A. Bertelsen, C. R. Hansen, J. Johansen, and C. Brink. Single Arc Volumetric Modulated Arc Therapy of head and neck cancer. *Radiother. Oncol.*, 95(2):142–148, 2010.

[12] J. R. Cunningham and H. E. Johns. *Physics of Radiology, Fourth Edition.* Charles C Thomas Pub Ltd, 1983.

[13] F. M. Khan. *Khan's Lectures: Handbook of the Physics of Radiation Therapy.* Wolters Kluwer, 2011.

[14] K. Luby-Phelps. Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. *Int Rev Cytol.*, 192:189–221, 2000.

[15] G. van Meer, D. R. Voelker, and G. W. Feigenson. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.*, 9(2):112–124, 2008.

[16] R. Milo. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, 35(12):1050–1055, 2013.

[17] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[18] B. Alberts, A. Johnson, and J. Lewis. *Molecular Biology of the Cell, 4th edition.* Garland Science, 2002.

[19] C. M. Yibing Wang and D. Mittar. *Cell Cycle and DNA Content Analysis Using the BD Cycletest Assay on the BD FACSVerse System*. BD Biosciences, 2011.

[20] M. Frankenberg-Schwager and D. Frankenberg. DNA Double-strand Breaks: Their Repair and Relationship to Cell Killing in Yeast. *Int. J. Radiat. Biol.*, 58(4):569–575, 1990.

[21] J. Ward. DNA Damage Produced by Ionizing Radiation in Mammalian Cells: Identities, Mechanisms of Formation, and Reparability. In, *Progress in Nucleic Acid Research and Molecular Biology*. Volume 35, pages 95–125. Academic Press, 1988.

[22] M. Dizdaroglu, P. Jaruga, M. Birincioglu, and H. Rodriguez. Free radical-induced damage to DNA: mechanisms and measurement. *Free Radical Biology and Medicine*, 32(11):1102–1115, 2002.

[23] L. F. Povirk. DNA damage and mutagenesis by radiomimetic DNA-cleaving agents: bleomycin, neocarzinostatin and other enediynes. *Mutat. Res.*, 355(1):71–89, 1996.

[24] L. J. Marnett. Oxyradicals and DNA damage. *Carcinogenesis*, 21(3):361–370, 2000.

[25] J. Wondergem, editor. *Radiation Biology: A Handbook for Teachers and Students*. International Atomic Energy Agency, 2010.

[26] U. Hagen. Mechanisms of induction and repair of DNA double-strand breaks by ionizing radiation: some contradictions. *Radiat. Environ. Biophys.*, 33(1):45–61, 1994.

[27] J. F. Ward. The yield of DNA double-strand breaks produced intracellularly by ionizing radiation: a review. *Int. J. Radiat. Biol.*, 57(6):1141–1150, 1990.

[28] D. J. Brenner. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Semin. Radiat. Oncol.*, 18(4):234–239, 2008.

[29] J. P. Kirkpatrick, J. J. Meyer, and L. B. Marks. The linear-quadratic model is inappropriate to model high dose per fraction effects in radiosurgery. *Semin. Radiat. Oncol.*, 18(4):240–243, 2008.

[30] M. Astrahan. Some implications of linear-quadratic-linear radiation dose-response with regard to hypofractionation. *Med. Phys.*, 35(9):4161–4172, 2008.

[31] M. Zaider and L. Hanin. Tumor control probability in radiation treatment. *Med. Phys.*, 38(2):574–583, 2011.

[32] B. Emami, J. Lyman, A. Brown, L. Coia, M. Goitein, J. E. Munzenrider, B. Shank, L. J. Solin, and M. Wesson. Tolerance of normal tissue to therapeutic irradiation. *Int. J. Radiat. Oncol. Biol. Phys.*, 21(1):109–122, 1991.

[33] L. B. Marks, E. D. Yorke, A. Jackson, R. K. Ten Haken, L. S. Constine, A. Eisbruch, S. M. Bentzen, J. Nam, and J. O. Deasy. Use of normal tissue complication probability models in the clinic. *Int. J. Radiat. Oncol. Biol. Phys.*, 76(3 Suppl):10–19, 2010.

[34] A. C. Begg. Predicting response to radiotherapy: evolutions and revolutions. *Int. J. Radiat. Biol.*, 85(10):825–836, 2009.

[35] G. J. Kutcher and C. Burman. Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method. *Int. J. Radiat. Oncol. Biol. Phys.*, 16(6):1623–1630, 1989.

[36] Bentzen and Overgaard. Patient-to-patient variability in the expression of radiation-induced normal tissue injury. *Semin. Radiat. Oncol.*, 4(2):68–80, 1994.

[37] S. L. Tucker and H. Thames Jr. The effect of patient-to-patient variability on the accuracy of predictive assays of tumor response to radiotherapy: a theoretical evaluation. *Int. J. Radiat. Oncol. Biol. Phys.*, 17(1):145–157, 1989.

[38] B. Fertil and E. P. Malaise. Inherent cellular radiosensitivity as a basic concept for human tumor radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.*, 7(5):621–629, 1981.

[39] R. M. Hoffman. In vitro sensitivity assays in cancer: a review, analysis, and prognosis. *J. Clin. Lab. Anal.*, 5(2):133–143, 1991.

[40] G. Browman et al. The clonogenic assay as a reproducible in vitro system to study predictive parameters of treatment outcome in acute nonlymphoblastic leukemia. *Am. J. Hematol.*, 15(3):227–235, 1983.

[41] M. Nordsmark, M. Overgaard, and J. Overgaard. Pretreatment oxygenation predicts radiation response in advanced squamous cell carcinoma of the head and neck. *Radiother. Oncol.*, 41(1):31–39, 1996.

[42] W. J. Koh, K. S. Bergman, J. S. Rasey, L. M. Peterson, M. L. Evans, M. M. Graham, J. R. Grierson, K. L. Lindsley, T. K. Lewellen, and K. A. Krohn. Evaluation of oxygenation status during fractionated radiotherapy in human nonsmall cell lung cancers using [F-18]fluoromisonidazole positron emission tomography. *Int. J. Radiat. Oncol. Biol. Phys.*, 33(2):391–398, 1995.

[43] U. Sunar, H. Quon, T. Durduran, J. Zhang, J. Du, C. Zhou, G. Yu, R. Choe, A. Kilger, R. Lustig, L. Loevner, S. Nioka, B. Chance, and A. G. Yodh. Non-invasive diffuse optical measurement of blood flow and blood oxygenation for monitoring radiation therapy in patients with head and neck tumors: a pilot study. *J. Biomed. Opt.*, 11(6), 2006.

[44] T. Nakano and K. Oka. Differential values of ki-67 index and mitotic index of proliferating cell population. An assessment of cell cycle and prognosis in radiation therapy for cervical cancer. *Cancer*, 72(8):2401–2408, 1993.

[45] S. W. Lowe, E. M. Schmitt, S. W. Smith, B. A. Osborne, and T. Jacks. p53 is required for radiation-induced apoptosis in mouse thymocytes. *Nature*, 362(6423):847–849, 1993.

[46] S. W. Lowe, S. Bodis, A. McClatchey, L. Remington, H. E. Ruley, D. E. Fisher, D. E. Housman, T. Jacks, et al. p53 status and the efficacy of cancer therapy in vivo. *Science*, 266(5186):807–807, 1994.

[47] S. L. Kerns, H. Ostrer, and B. S. Rosenstein. Radiogenomics: using genetics to identify cancer patients at risk for development of adverse effects following radiotherapy. *Cancer Discov*, 4(2):155–165, 2014.

[48] A. K. Das, M. H. Bell, C. S. Nirodi, M. D. Story, and J. D. Minna. Radiogenomics predicting tumor responses to radiotherapy in lung cancer. *Semin. Radiat. Oncol.*, 20(3):149–155, 2010.

[49] J. Alsner, C. N. Andreassen, and J. Overgaard. Genetic markers for prediction of normal tissue toxicity after radiotherapy. *Semin. Radiat. Oncol.*, 18(2):126–135, 2008.

[50] S. M. Bentzen, M. Parliament, J. O. Deasy, A. Dicker, W. J. Curran, J. P. Williams, and B. S. Rosenstein. Biomarkers and surrogate endpoints for normal-tissue effects of radiation therapy: the importance of dose-volume effects. *Int. J. Radiat. Oncol. Biol. Phys.*, 76(3 Suppl):145–150, 2010.

[51] O. Popanda, J. U. Marquardt, J. Chang-Claude, and P. Schmezer. Genetic variation in normal tissue toxicity induced by ionizing radiation. *Mutat. Res.*, 667(1-2):58–69, 2009.

[52] C. M. West and G. C. Barnett. Genetics and genomics of radiotherapy toxicity: towards prediction. *Genome Medicine*, 3(8):1–15, 2011.

[53] C. M. L. West, M. J. McKay, T. Holscher, M. Baumann, I. J. Stratford, R. G. Bristow, M. Iwakawa, T. Imai, S. M. Zingde, M. S. Anscher, J. Bourhis, A. C. Begg, K. Haustermans, S. M. Bentzen, and J. H. Hendry. Molecular markers predicting radiotherapy response: report and recommendations from an International Atomic Energy Agency technical meeting. *Int. J. Radiat. Oncol. Biol. Phys.*, 62(5):1264–1273, 2005.

[54] T. Rattay and C. J. Talbot. Finding the genetic determinants of adverse reactions to radiotherapy. *Clin Oncol (R Coll Radiol)*, 26(5):301–308, 2014.

[55] Y. Mardor, R. Pfeffer, R. Spiegelmann, Y. Roth, S. E. Maier, O. Nissim, R. Berger, A. Glicksman, J. Baram, A. Orenstein, J. S. Cohen, and T. Tichler. Early detection of response to radiation therapy in patients with brain malignancies using conventional and high b-value diffusion-weighted magnetic resonance imaging. *J. Clin. Oncol.*, 21(6):1094–1100, 2003.

[56] M. Diehn, C. Nardini, D. S. Wang, S. McGovern, M. Jayaraman, Y. Liang, K. Aldape, S. Cha, and M. D. Kuo. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. U. S. A.*, 105(13):5213–5218, 2008.

[57] K. I. Shoghi, J. Xu, Y. Su, J. He, D. Rowland, Y. Yan, J. R. Garbow, Z. Tu, L. A. Jones, R. Higashikubo, K. T. Wheeler, R. A. Lubet, R. H. Mach, and M. You. Quantitative receptor-based imaging of tumor proliferation with the sigma-2 ligand [(18)F]ISO-1. *PLoS One*, 8(9), 2013.

[58] M. Rafat, R. Ali, and E. E. Graves. Imaging radiation response in tumor and normal tissue. *Am J Nucl Med Mol Imaging*, 5(4):317–332, 2015.

[59] G. J. Puppels, F. F. de Mul, C. Otto, J. Greve, M. Robert-Nicoud, D. J. Arndt-Jovin, and T. M. Jovin. Studying single living cells and chromosomes by confocal Raman microspectroscopy. *Nature*, 347(6290):301–303, 1990.

[60] L. J. Goeller and M. R. Riley. Discrimination of bacteria and bacteriophages by Raman spectroscopy and surface-enhanced Raman spectroscopy. *Appl. Spectrosc.*, 61(7):679–685, 2007.

[61] Z. Movasaghi, S. Rehman, and D. I. U. Rehman. Raman Spectroscopy of Biological Tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007.

[62] I. Notingher, S. Verrier, S. Haque, J. M. Polak, and L. L. Hench. Spectroscopic study of human lung epithelial cells (A549) in culture: Living cells versus dead cells. *Biopolymers*, 72(4):230–240, 2003.

[63] S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang. Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue. *Br. J. Cancer*, 98(2):457–465, 2008.

[64] D. I. Ellis and R. Goodacre. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst*, 131(8):875–885, 2006.

[65] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22(5):245–252, 2004.

[66] C. W. Freudiger, W. Min, B. G. Saar, S. Lu, G. R. Holtom, C. He, J. C. Tsai, J. X. Kang, and X. S. Xie. Label-Free Biomedical Imaging with High Sensitivity by Stimulated Raman Scattering Microscopy. *Science*, 322(5909):1857–1861, 2008.

[67] S. J. Bauman, A. A. Darweesh, and J. B. Herzog. Surface-enhanced Raman spectroscopy substrate fabricated via nanomasking technique for biological sensor applications. *Proc. SPIE*, 9759, 2016.

[68] Z. Huang, A. McWilliams, H. Lui, D. I. McLean, S. Lam, and H. Zeng. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. *Int. J. Cancer*, 107(6):1047–1052, 2003.

[69] N. Stone, C. Kendall, J. Smith, P. Crow, and H. Barr. Raman spectroscopy for identification of epithelial cancers. *Faraday Discuss.*, 126:141–157, 2004.

[70] J. Surmacki, J. Musial, R. Kordek, and H. Abramczyk. Raman imaging at biological interfaces: applications in breast cancer diagnosis. *Mol. Cancer*, 12:48, 2013.

[71] R. E. Kast, S. C. Tucker, K. Killian, M. Trexler, K. V. Honn, and G. W. Auner. Emerging technology: applications of Raman spectroscopy for prostate cancer. *Cancer Metastasis Rev.*, 33(2-3):673–693, 2014.

[72] K. Kong, C. Kendall, N. Stone, and I. Notingher. Raman spectroscopy for medical diagnostics – from in-vitro biofluid assays to in-vivo cancer detection. *Adv Drug Deliv Rev*, 89:121–134, 2015.

[73] J. E. McGeehan, P. Carpentier, A. Royant, D. Bourgeois, and R. B. Ravelli. X-ray radiation-induced damage in DNA monitored by online Raman. *Journal of Synchrotron Radiation*, 14(1):99–108, 2007.

[74] S. R. Panikkanvalappil, M. A. Mackey, and M. A. El-Sayed. Probing the unique dehydration-induced structural modifications in cancer cell DNA using surface enhanced Raman spectroscopy. *J. Am. Chem. Soc.*, 135(12):4815–4821, 2013.

[75] M. S. Vidyasagar, K. Maheedhar, B. M. Vadhiraja, D. J. Fernendes, V. B. Kartha, and C. M. Krishna. Prediction of radiotherapy response in cervix cancer by Raman spectroscopy: A pilot study. *Biopolymers*, 89(6):530–537, 2008.

[76] Q. Matthews, A. Jirasek, J. J. Lum, and A. G. Brolo. Biochemical signatures of in vitro radiation response in human lung, breast and prostate tumour cells observed with Raman spectroscopy. *Phys. Med. Biol.*, 56(21):6839, 2011.

[77] Q. Matthews, A. G. Brolo, J. J. Lum, X. Duan, and A. Jirasek. Raman spectroscopy of single human tumour cells exposed to ionizing radiation in vitro. *Phys. Med. Biol.*, 56(1):19–38, 2011.

[78] S. J. Harder, Q. Matthews, M. Isabelle, A. G. Brolo, J. J. Lum, and A. Jirasek. A Raman spectroscopic study of cell response to clinical doses of ionizing radiation. *Appl. Spectrosc.*, 69(2):193–204, 2015.

[79] Q. Matthews, M. Isabelle, S. J. Harder, J. Smazynski, W. Beckham, A. G. Brolo, A. Jirasek, and J. J. Lum. Radiation-induced glycogen accumulation detected by single cell Raman spectroscopy is associated with radioresistance that can be reversed by metformin. *PLoS ONE*, 10(8):1–15, 2015.

[80] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via Outlier Pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

[81] S. Verboven and M. Hubert. LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75(2):127–136, 2005.

[82] A. Hyviirinen, J. Karhunen, and E. Oja. Independent component analysis. *Wileyand Sons*, 2001.

[83] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, 43(1):59–69, 1982.

[84] A. D. Back and A. S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *Int. J. Neural Syst.*, 8(4):473–484, 1997.

[85] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[86] *Introduction to Raman Spectroscopy*. Thermo Electron Scientific Instruments, 2008.

[87] C. V. Raman and K. S. Krishnan. A New Type of Secondary Radiation. *Nature*, 121, 1928.

[88] A. M. Committee. Raman spectroscopy in cultural heritage: Background paper. Technical report. 2015, pages 4844–4847.

[89] M. Ji et al. Rapid, label-free detection of brain tumors with stimulated Raman scattering microscopy. *Science Translational Medicine*, 5(201):119, 2013.

[90] F. Siebert and P. Hildebrandt. *Theory of Infrared Absorption and Raman Spectroscopy*. In. *Vibrational Spectroscopy in Life Science*. Wiley-VCH Verlag, 2008. part 2, pages 11–61.

[91] E. Smith and G. Dent. *The Theory of Raman Spectroscopy*. In. *Modern Raman Spectroscopy, A Practical Approach*. John Wiley & Sons, Ltd, 2005, pages 71–92.

[92] R. Tabaksblat, R. J. Meier, and B. J. Kip. Confocal Raman Microspectroscopy: Theory and Application to Thin Polymer Samples. *Appl. Spectrosc.*, 46(1):60–68, 1992.

[93] H. E. Ian R. Lewis, editor. *Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line*. CRC Press, 2001.

[94] N. J. Everall. Modeling and Measuring the Effect of Refraction on the Depth Resolution of Confocal Raman Microscopy. *Appl. Spectrosc.*, 54(6):773–782, 2000.

[95] C. A. Lieber and A. Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological Raman spectra. *Appl. Spectrosc.*, 57(11):1363–1367, 2003.

[96] I. Notingher, S. Verrier, H. Romanska, A. E. Bishop, J. M. Polak, and L. L. Hench. In situ characterisation of living cells by Raman spectroscopy. *Spectroscopy*, 16(2), 2002.

[97] S. Dochow, C. Krafft, U. Neugebauer, T. Bocklitz, T. Henkel, G. Mayer, J. Albert, and J. Popp. Tumour cell identification by means of Raman spectroscopy in combination with optical traps and microfluidic environments. *Lab. Chip*, 11:1484–1490, 8, 2011.

[98] J. M. Smulko, N. C. Dingari, J. S. Soares, and I. Barman. Anatomy of noise in quantitative biological Raman spectroscopy. *Bioanalysis*, 6(3):411–421, 2014.

[99] P. Heraud, B. R. Wood, J. Beardall, and D. McNaughton. Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells. *J. Chemom.*, 20(5):193–197, 2006.

[100] P. Lasch. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems*, 117:100–114, 2012.

[101] H. J. Butler et al. Using Raman spectroscopy to characterize biological materials. *Nat. Protocols*, 11(4):664–687, 2016.

[102] O. Ryabchykov, T. Bocklitz, A. Ramoji, U. Neugebauer, M. Foerster, C. Kroegel, M. Bauer, M. Kiehntopf, and J. Poppi. Automatization of spike correction in Raman spectra of biological samples. *Chemometrics and Intelligent Laboratory Systems*, 155:1–6, 2016.

[103] Y. Hu, T. Jiang, A. Shen, W. Li, X. Wang, and J. Hu. A background elimination method based on wavelet transform for Raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 85(1):94–101, 2007.

[104] M. C. Grimbergen, C. F. van Swol, C. Kendall, R. M. Verdaasdonk, N. Stone, and J. L. Bosch. Signal-to-noise contribution of principal component loads in reconstructed near-infrared Raman tissue spectra. *Appl. Spectrosc.*, 64(1):8–14, 2010.

[105] F. Gan, G. Ruan, and J. Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):59–65, 2006.

[106] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, and H. Zhou. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *J. Raman Spectrosc.*, 41(6):659–669, 2010.

[107] C. Camerlingo, F. Zenone, G. M. Gaeta, R. Riccio, and M. Lepore. Wavelet data processing of micro-Raman spectra of biological samples. *Meas. Sci. Technol.*, 17(2):298, 2006.

[108] W. J. Bruno, G. Ullah, D.-O. D. Mak, and J. E. Pearson. Automated Maximum Likelihood Separation of Signal from Baseline in Noisy Quantal Data. *Biophys. J.*, 105(1):68–79, 2013.

[109] A. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[110] M. AlMahamdy and B. Riley. Performance study of different denoising methods for ECG signals. *Procedia Computer Science*, 37:325–332, 2014.

[111] S. Guo, T. Bocklitz, and J. Popp. Optimization of Raman-spectrum baseline correction in biological application. *Analyst*, 141:2396–2404, 8, 2016.

[112] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst*, 137(14):3202–3215, 2012.

[113] M. E. Keating, H. Nawaz, F. Bonnier, and H. J. Byrne. Multivariate statistical methodologies applied in biomedical Raman spectroscopy: assessing the validity of partial least squares regression using simulated model datasets. *Analyst*, 140:2482–2492, 7, 2015.

[114] B. Berry, J. Moretto, T. Matthews, J. Smelko, and K. Wiltberger. Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis. *Biotechnol. Prog.*, 31(2):566–577, 2015.

[115] S. M. Ascencio, C. Choe, M. C. Meinke, R. H. Miller, G. V. Maksimov, W. Wigger-Alberti, J. Lademann, and M. E. Darvin. Confocal Raman microscopy and multivariate statistical analysis for determination of different penetration abilities of caffeine and propylene glycol applied simultaneously in a mixture on porcine skin ex vivo. *European Journal of Pharmaceutics and Biopharmaceutics*, 104:51–58, 2016.

[116] K. M. Sorensen, C. Westley, R. Goodacre, and S. B. Engelsen. Simultaneous quantification of the boar-taint compounds skatole and androstenone by surface-enhanced Raman scattering (SERS) and multivariate data analysis. *Anal. Bioanal. Chem.*, 407(25):7787–7795, 2015.

[117] V. Vrabie, C. Gobinet, O. Piot, A. Tfayli, P. Bernard, R. Huez, and M. Manfait. Independent component analysis of Raman spectra: Application on paraffin-embedded skin biopsies. *Biomedical Signal Processing and Control*, 2(1):40–50, 2007.

[118] W. Wang and T. Adali. Constrained ICA and its application to Raman spectroscopy. In *IEEE antennas and propagation society international symposium.* Volume 4. IEEE; 1999, 2005, page 109.

[119] Y. Ozeki, W. Umemura, Y. Otsuka, S. Satoh, H. Hashimoto, K. Sumimura, N. Nishizawa, K. Fukui, and K. Itoh. High-speed molecular spectral imaging of tissue with stimulated Raman scattering. *Nat Photon*, 6(12):845–851, 2012.

[120] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

[121] K. Buckley, J. G. Kerns, A. W. Parker, A. E. Goodship, and P. Matousek. Decomposition of in vivo spatially offset Raman spectroscopy data using multivariate analysis techniques. *J. Raman Spectrosc.*, 45(2):188–192, 2014.

[122] M. Alessio and C. V. Cannistraci. *Nonlinear Dimensionality Reduction by Minimum Curvilinearity for Unsupervised Discovery of Patterns in Multidimensional Proteomic Data*. In. *2-D PAGE map analysis: methods and protocols*. E. Marengo and E. Robotti, editors. Springer New York, 2016, pages 289–298.

[123] A. N. Gorban and A. Zinovyev. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. Neural Syst.*, 20(03):219–232, 2010.

[124] H. Yang, I. R. Lewis, and P. R. Griffiths. Raman spectrometry and neural networks for the classification of wood types. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 55(14):2783–2791, 1999.

[125] J. Shlens. *A Tutorial on principal component analysis*, 2003.

[126] I. Jolliffe. *Principal Component Analysis, 2nd ed.* Springer-Verlag New York, 2002.

[127] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, and J. Dongarra. LAPACK Users' Guide (3rd ed.) Technical report. Society for Industrial and Applied Mathematics., 1991.

[128] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.*, 11:1957–2000, 2010.

[129] S. Roweis. EM Algorithms for PCA and SPCA. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*. In NIPS '97. MIT Press, Denver, Colorado, USA, 1998, pages 626–632.

[130] O. Tamuz, T. Mazeh, and S. Zucker. Correcting systematic effects in a large set of photometric light curves. *Monthly Notices of the Royal Astronomical Society*, 356(4):1466–1470, 2005.

[131] L. Delchambre. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society*, 446(2):3545–3555, 2014.

[132] K. R. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498, 1979.

[133] T. Bouwmans and E. H. Zahzah. Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014.

[134] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. *J. ACM*, 58(3):1–37, 2011.

[135] S. Hou and P. Wentzell. Fast and simple methods for the optimization of kurtosis used as a projection pursuit index. *Analytica Chimica Acta*, 704(12):1–15, 2011.

[136] W. F. de Carvalho Rocha, R. Nogueira, G. Estev, J. B. da Silva, S. M. Queiroz, and G. F. Sarmanho. A comparison of three procedures for robust PCA of experimental results of the homogeneity test of a new sodium diclofenac candidate certified reference material. *Microchemical Journal*, 109:112–116, 2013.

[137] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[138] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Comput.*, 11(2):443–482, 1999.

[139] M. Scholz. Validation of Nonlinear PCA. *Neural Processing Letters*, 36(1):21–30, 2012.

[140] M. Scholz, F. Kaplan, C. L. Guy, J. Kopka, and J. Selbig. Non-linear PCA: a missing data approach. *Bioinformatics*, 21(20):3887–3895, 2005.

[141] M. Scholz, M. Fraunholz, and J. Selbig. *Nonlinear Principal Component Analysis: Neural Network Models and Applications*. In. *Principal Manifolds for Data Visualization and Dimension Reduction*. A. N. Gorban, B. Kegl, D. C. Wunsch, and A. Y. Zinovyev, editors. Springer Berlin Heidelberg, 2008, pages 44–67.

[142] M. Scholz and R. Vigario. Nonlinear PCA: a new hierarchical approach. In *Proceedings of 10th European Symposium on Artificial Neural Networks*, 2002.

[143] S. Harder. Raman spectroscopy of tumour cells exposed to clinically relevant doses of ionizing radiation. Master's thesis. University of Victoria, 2013.

[144] Q. Matthews. Single-cell Raman spectroscopy of irradiated tumour cells. PhD thesis. University of Victoria, 2011.

[145] M. Brower, D. N. Carney, H. K. Oie, A. F. Gazdar, and J. D. Minna. Growth of cell lines and clinical specimens of human non-small cell lung cancer in a serum-free defined medium. *Cancer Res.*, 46(2):798–806, 1986.

[146] NCI-H460 Culture Method. ATCC. URL: `http://www.atcc.org/products/all/HTB-177.aspx`.

[147] IEEE. IEEE Standard for Floating-Point Arithmetic. *IEEE Std 754-2008*:1–70, 2008.

[148] A. M. Yair. *Accelerating MATLAB Performance: 1001 tips to speed up MATLAB programs*. CRC Press, 2014.

[149] ADVANPIX: Multiprecision Computing Toolbox. 2015. URL: `http://advanpix.com`.

[150] G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, and M. W. Blades. Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation. *Applied Spectroscopy*, 59(5):545–574, 2005.

[151] G. E. P. Box, J. S. Hunter, and W. G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery, 2nd edition.* Wiley-Interscience, 2005.

[152] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[153] A. Ghasemi and S. Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab*, 10(2):486–489, 2012.

[154] F. Marmolejo-Ramos and J. Gonzalez-Burgos. A power comparison of various tests of univariate normality on ex-Gaussian distributions. *Methodology*, 9(4):137–149, 2013.

[155] N. Razali and Y. B. Wah. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[156] J. Rochon, M. Gondan, and M. Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12(1):81, 2012.

[157] NIST/SEMATECHlessio. e-Handbook of Statistical Methods. URL: `http://www.itl.nist.gov/div898/handbook/`.

[158] R. B. Dean and W. J. Dixon. Simplified Statistics for Small Numbers of Observations. *Anal. Chem.*, 23(4):636–638, 1951.

[159] B. Rosner. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, 25(2):165–172, 1983.

[160] F. E. Grubbs. Sample criteria for testing outlying observations. *Ann. Math. Statist.*, 21(1):27–58, 1950.

[161] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:2004, 2004.

[162] K. I. Penny and I. T. Jolliffe. A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(3):295–307, 2001.

[163] S. Shekhar, C.-T. Lu, and P. Zhang. A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, 7(2):139–166, 2003.

[164] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[165] S. Engelen, M.Hubert, and K. V. Branden. A comparison of three procedures for robust PCA in high dimensions. *Austrian Journal of Statistics*, 34(2):117–126, 2005.

[166] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, 44(7):1540–1551, 2011.

[167] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[168] L. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[169] L. Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

[170] I. Wallach and R. Lilien. The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. *Bioinformatics*, 25(5):615–620, 2009.

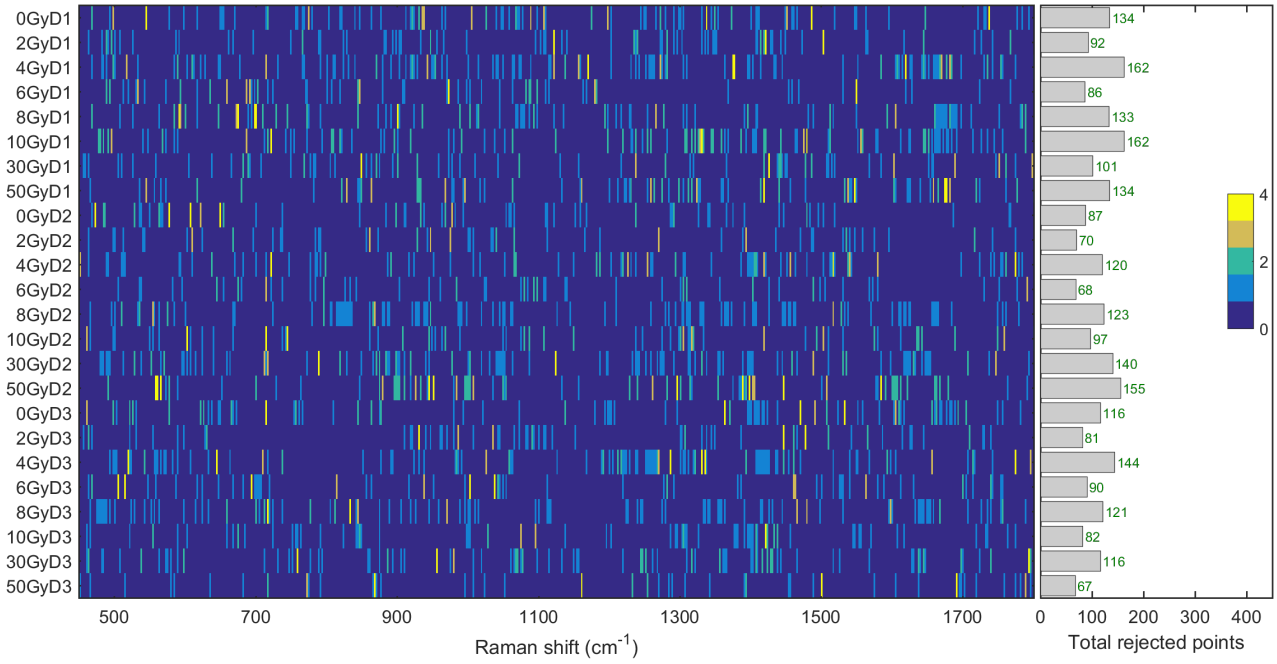
[171] A. R. Jamieson, M. L. Giger, K. Drukker, H. Li, Y. Yuan, and N. Bhooshan. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and t-SNE. *Med. Phys.*, 37(1):339–351, 2010.

Appendices

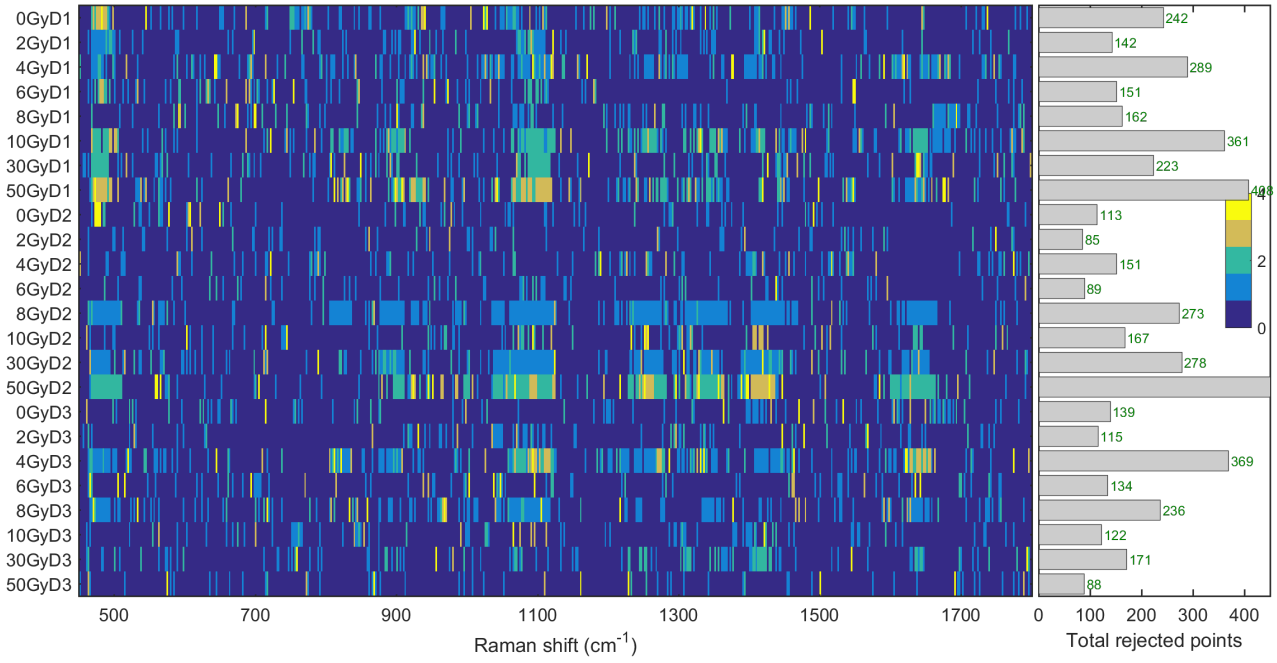
Appendix A

Outlier rejection maps of other datasets

H460 A



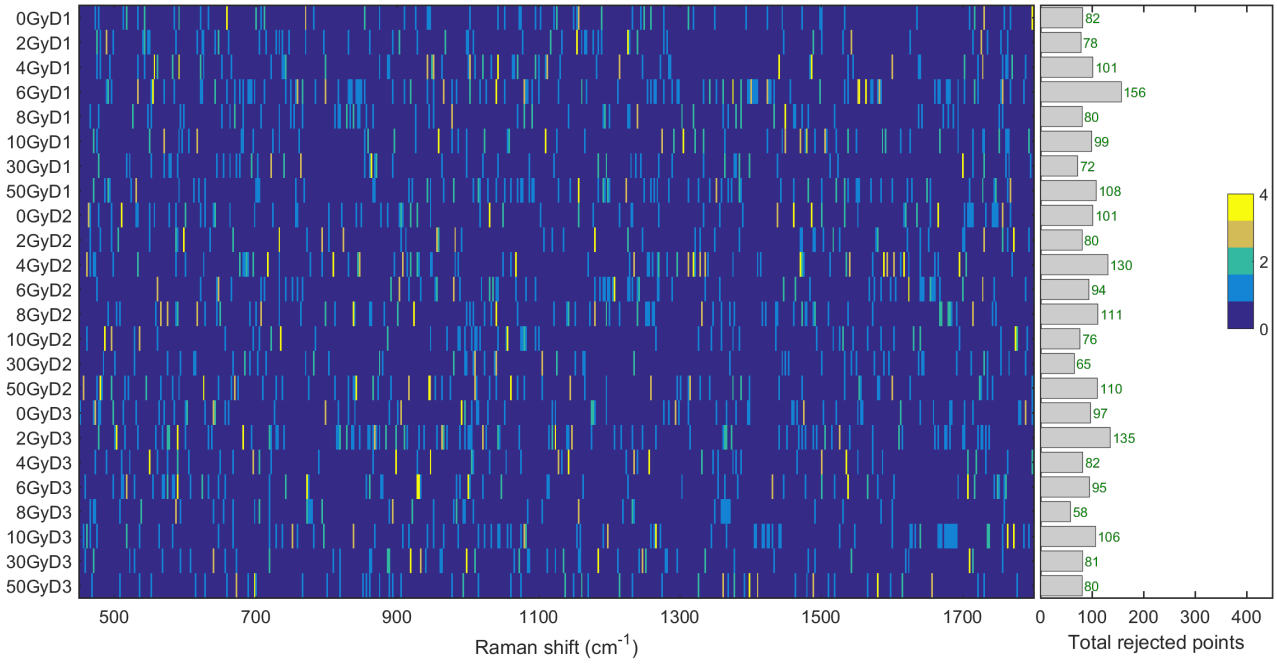
(a) Rejection of normally distributed data only.



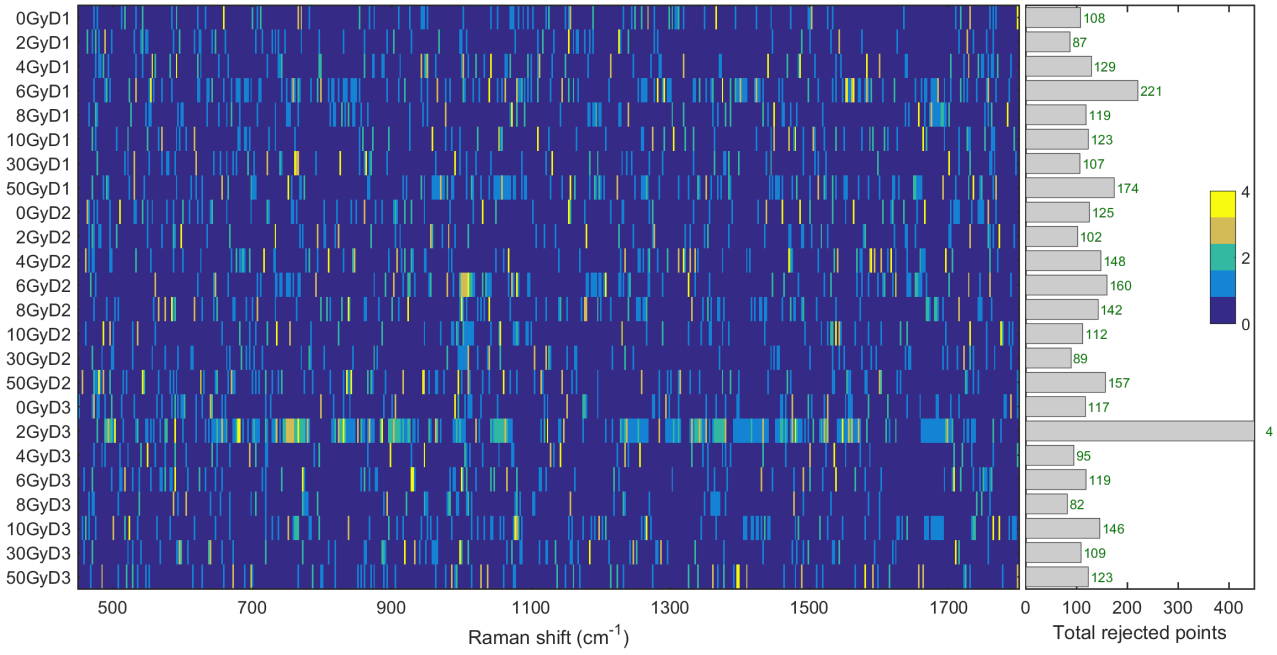
(b) Rejection of all data.

Figure A.1: TT + gESD outlier removal in H460A dataset at $\alpha = 5\%$ mapped into 2D colormap by batch (y), pixel number (x) and number of outliers removed (z, color).

LNCaP A



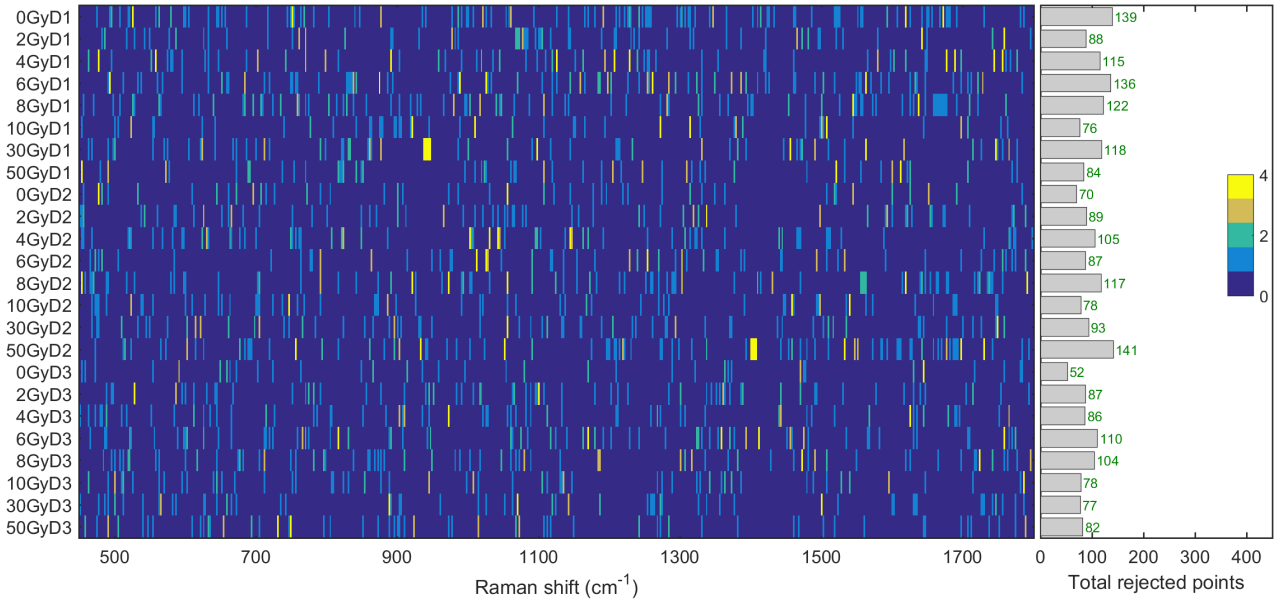
(a) Rejection of normally distributed data only.



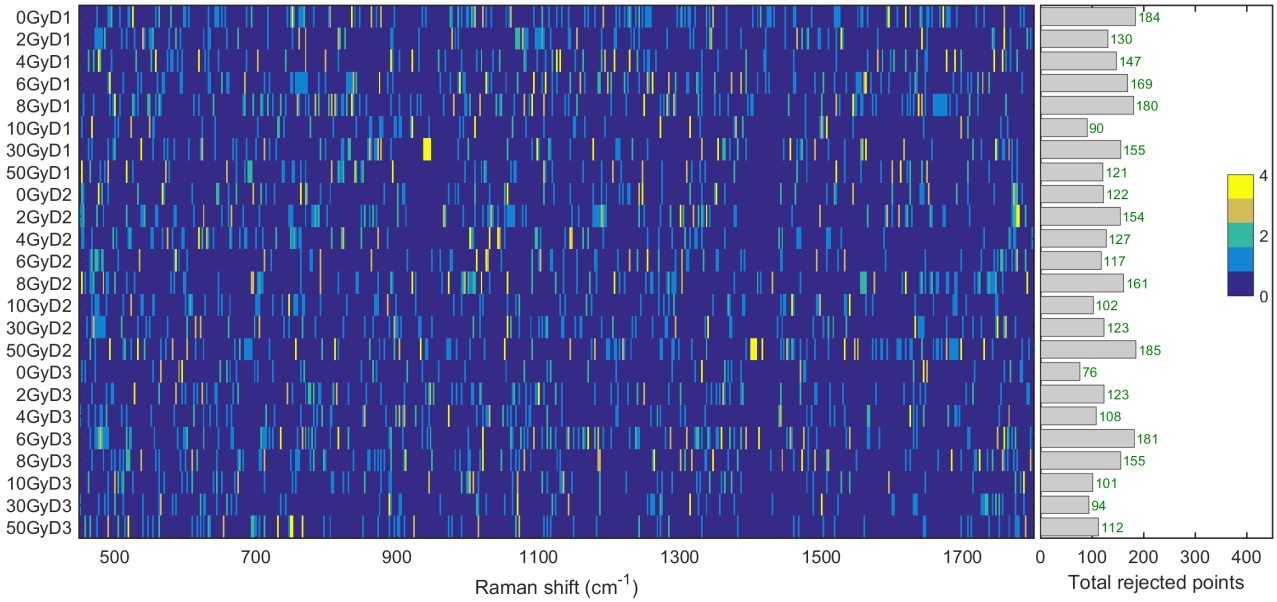
(b) Rejection of all data.

Figure A.2: TT + gESD outlier removal in LNA dataset at $\alpha = 5\%$ mapped into 2D colormap by batch (y), pixel number (x) and number of outliers removed (z, color).

LNCaP B



(a) Rejection of normally distributed data only.



(b) Rejection of all data.

Figure A.3: TT + gESD outlier removal in LNB dataset at $\alpha = 5\%$ mapped into 2D colormap by batch (y), pixel number (x) and number of outliers removed (z, color).

Appendix B

Additional components and score distributions

B.1 H460 B

PCA

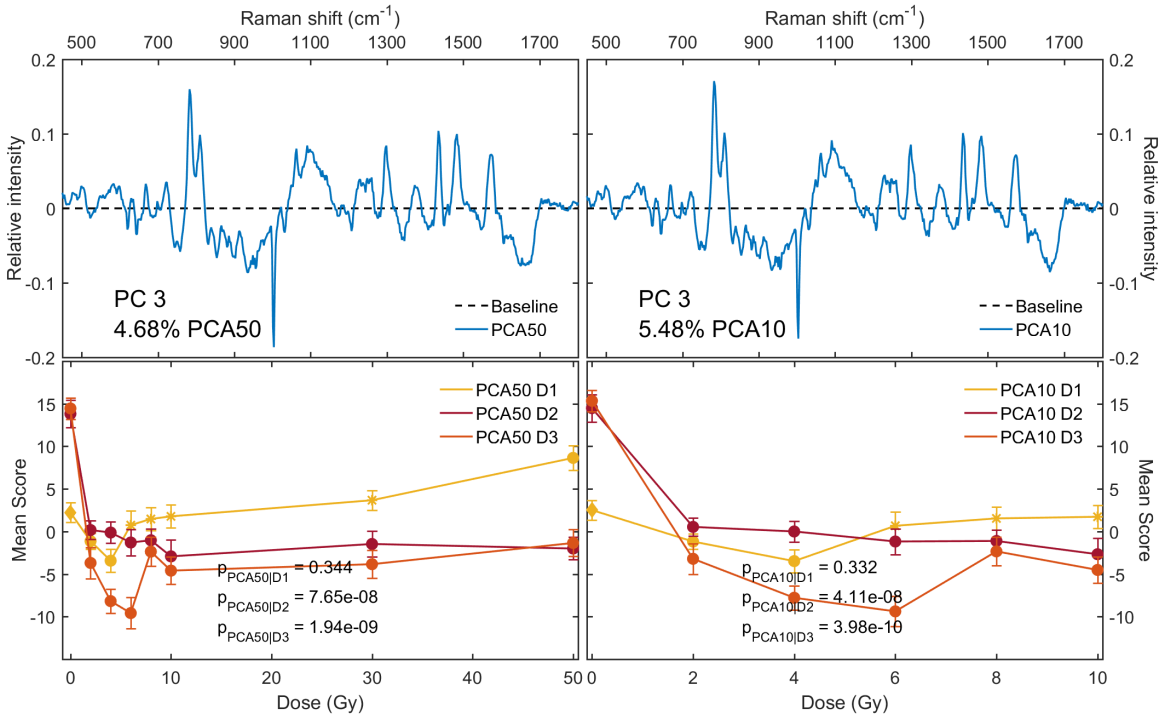


Figure B.1: PCA component 3 and respective scores of H460B 50Gy/10Gy datasets.

WPCA

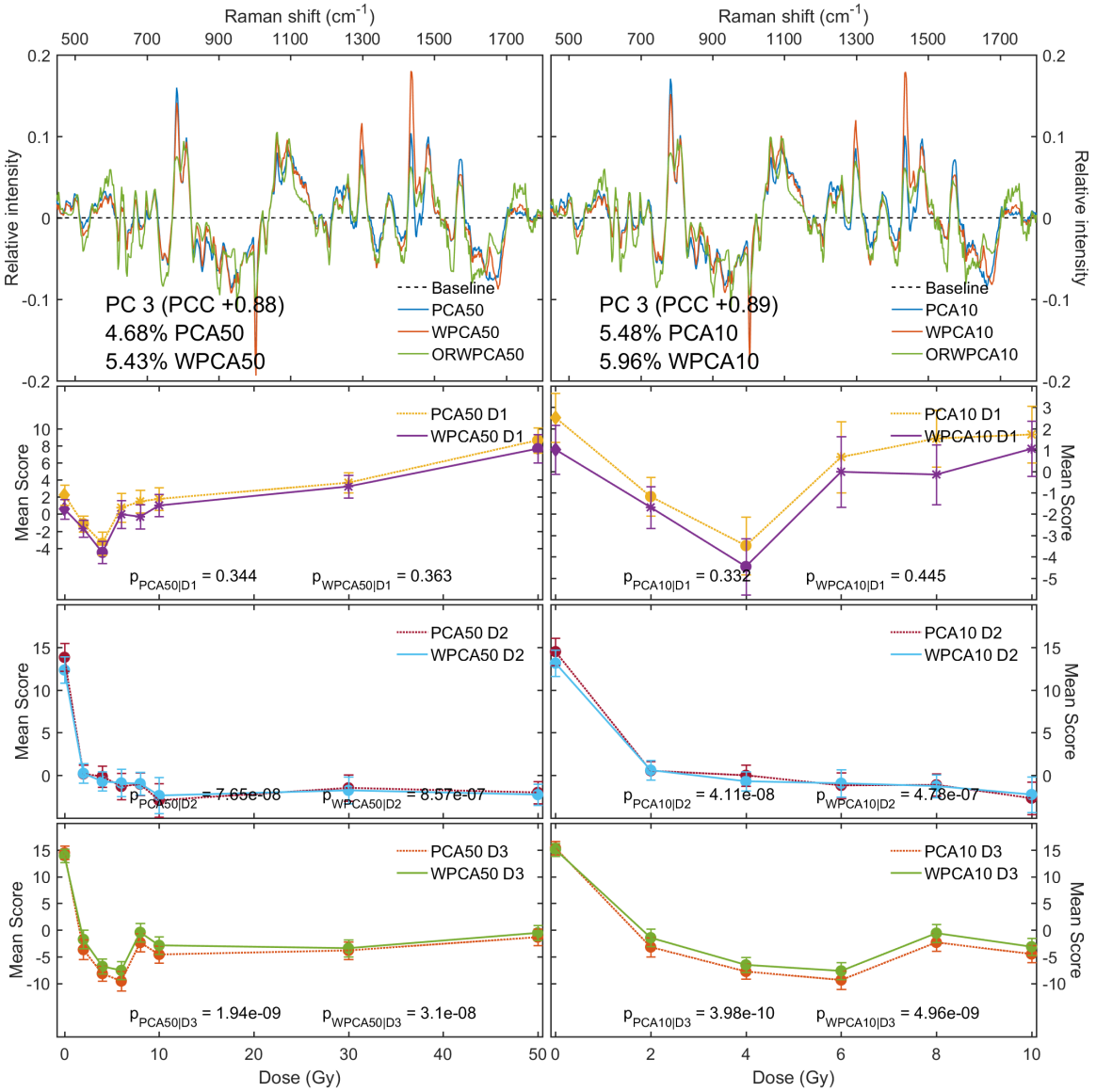


Figure B.2: WPCA/PCA component 3 and respective scores for H460B 50Gy/10Gy datasets.

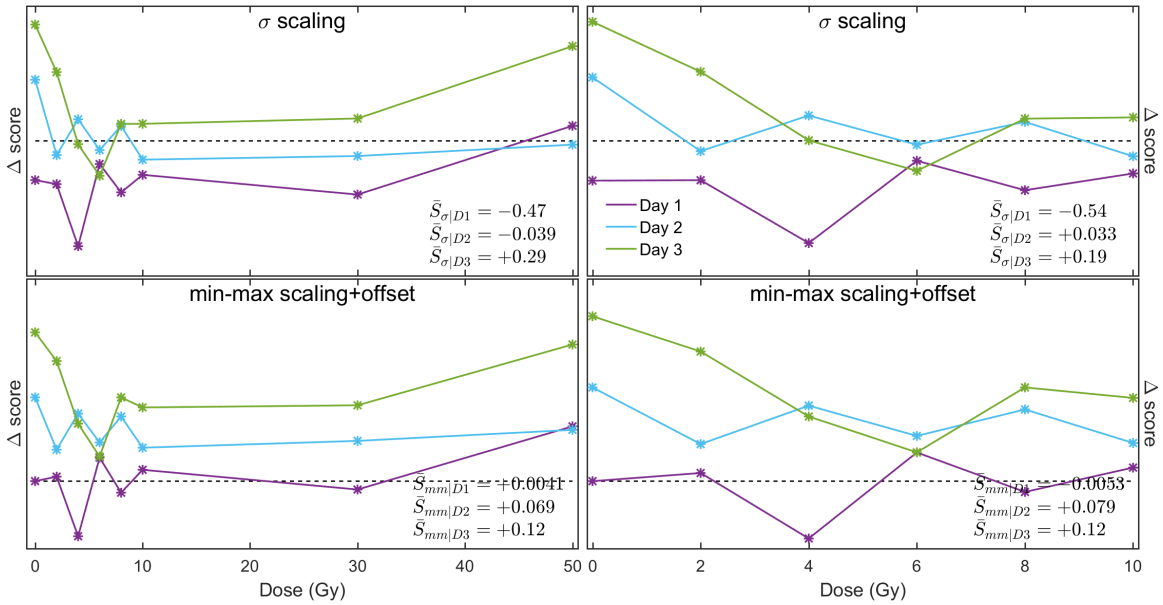


Figure B.3: WPCA/PCA H460B PC2 score distances.

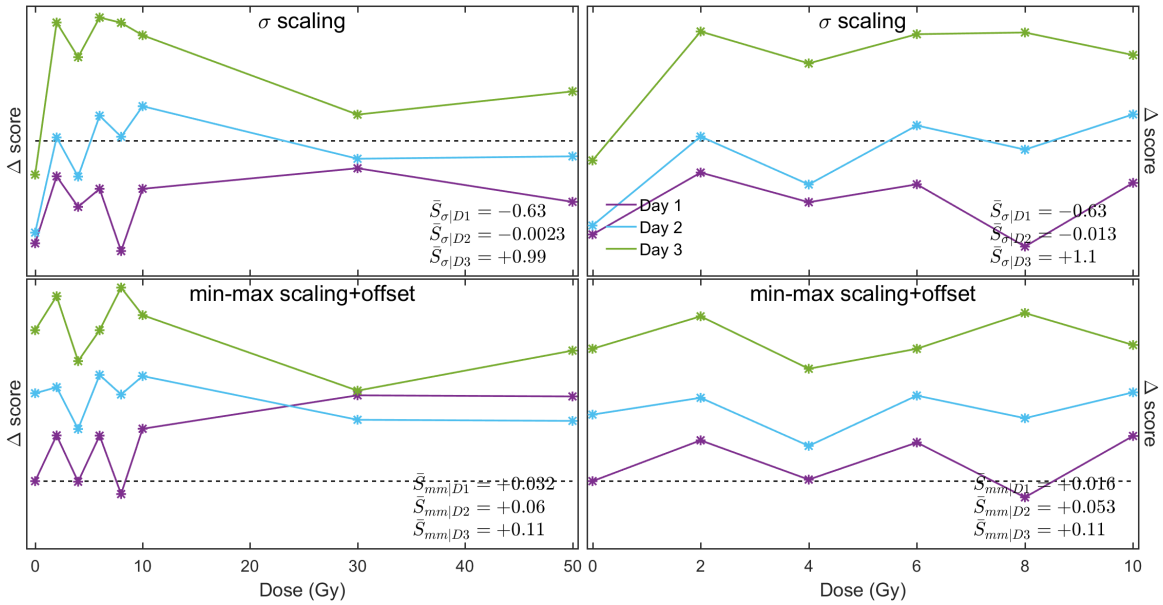


Figure B.4: WPCA/PCA H460B PC3 score distances.

RPCA

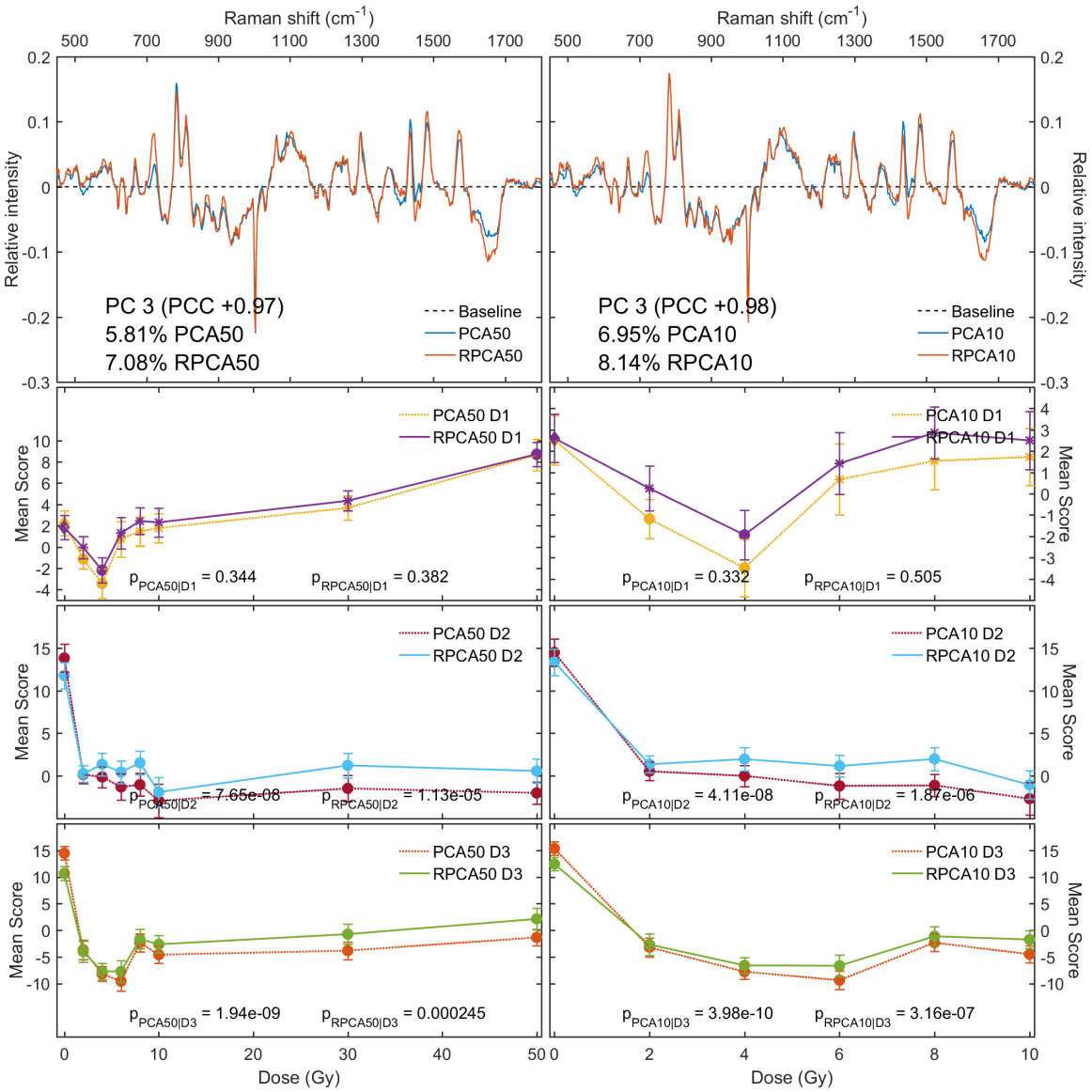


Figure B.5: RPCA/PCA component 3 and respective scores for H460B 50Gy/10Gy datasets.

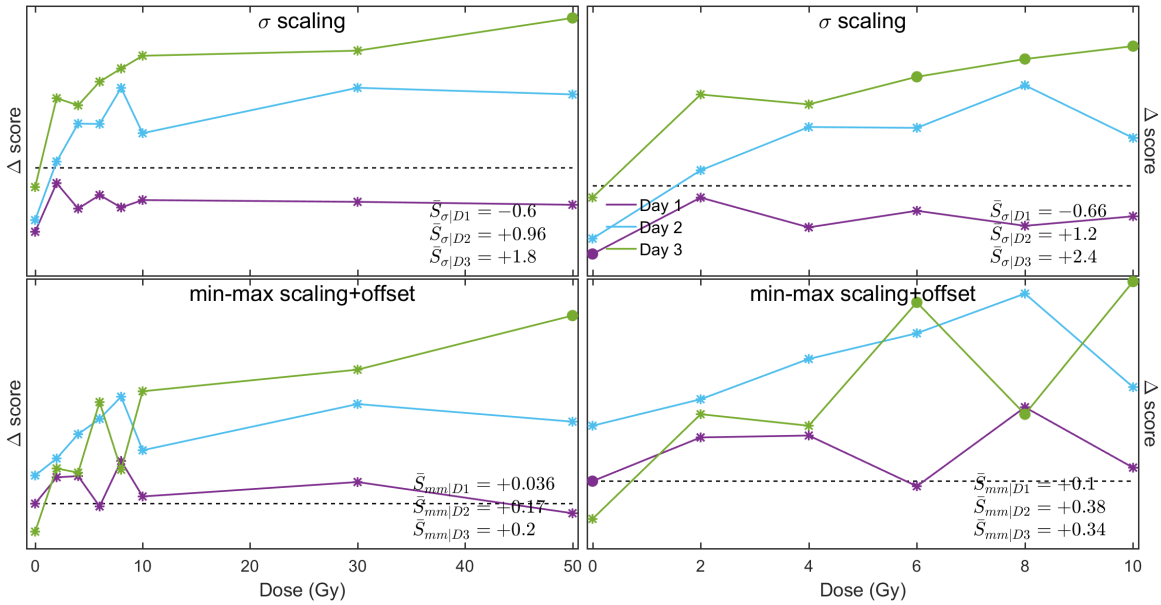


Figure B.6: WPCA/PCA H460B PC2 score distances.

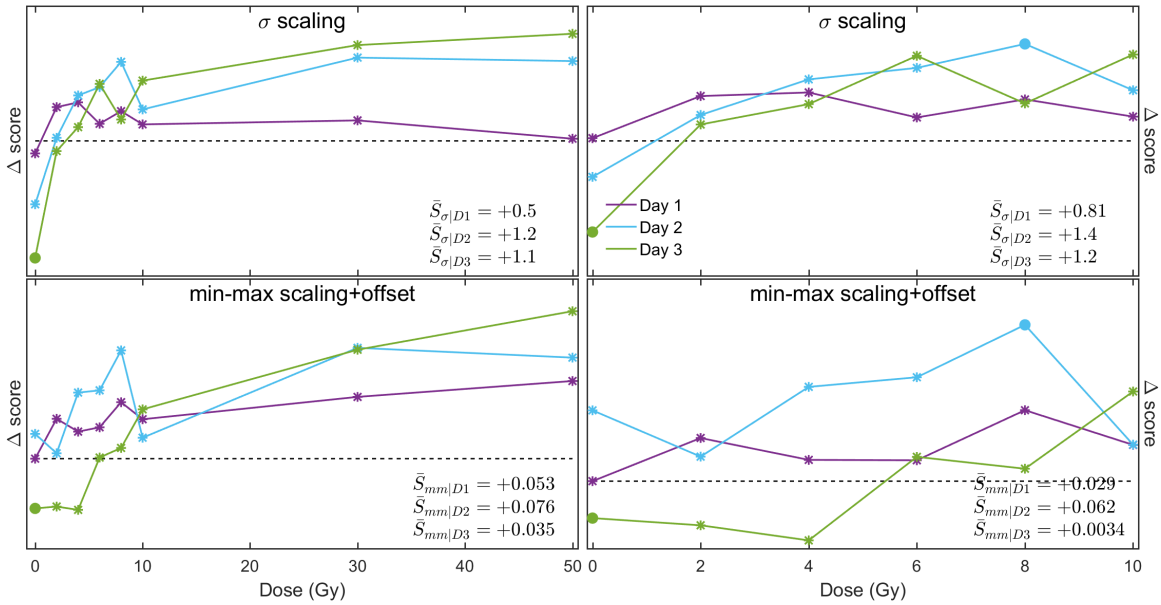


Figure B.7: WPCA/PCA H460B PC3 score distances.

PPCA

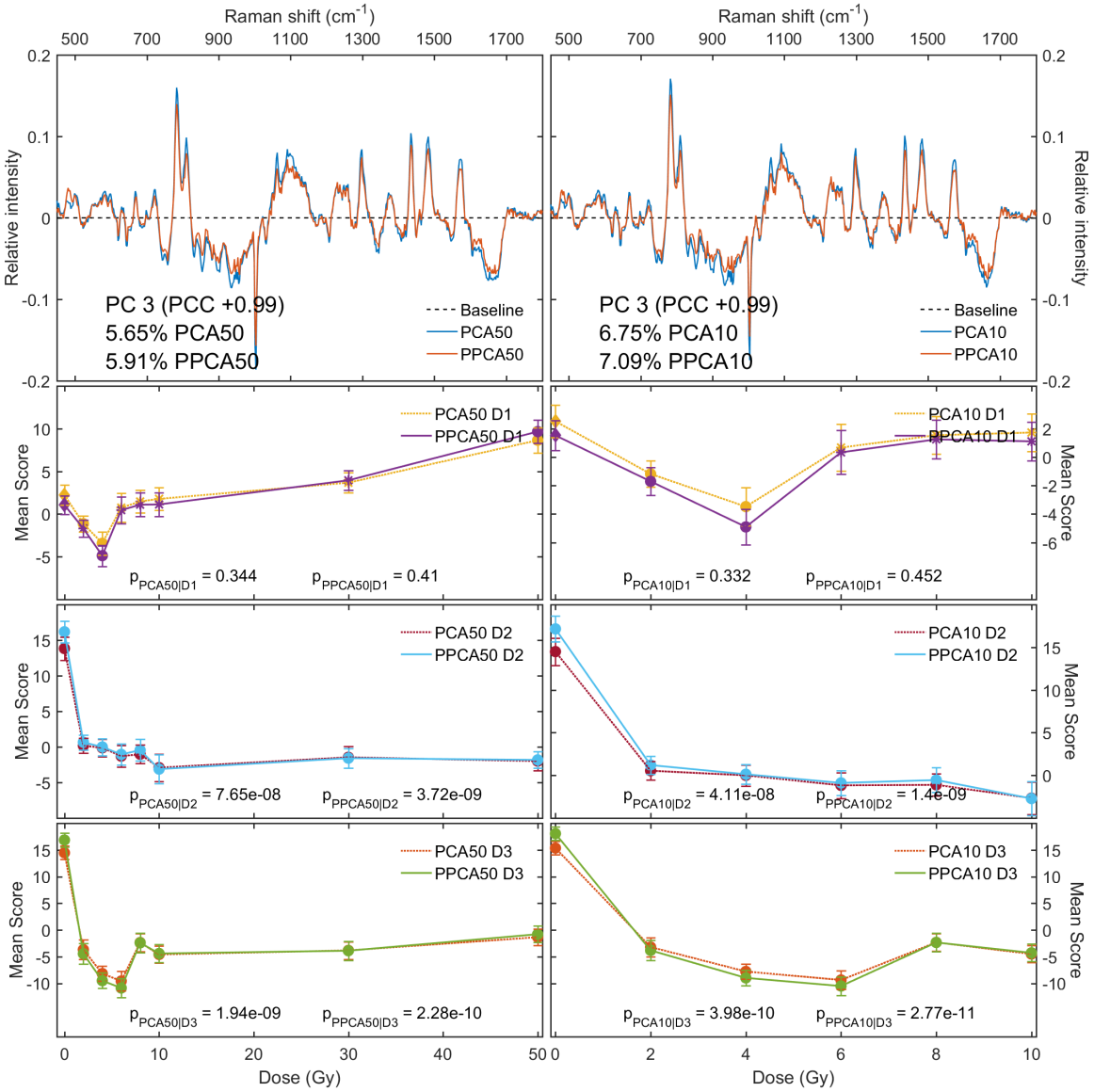


Figure B.8: PPCA/PCA component 3 and scores of H460B 50Gy/10Gy datasets.

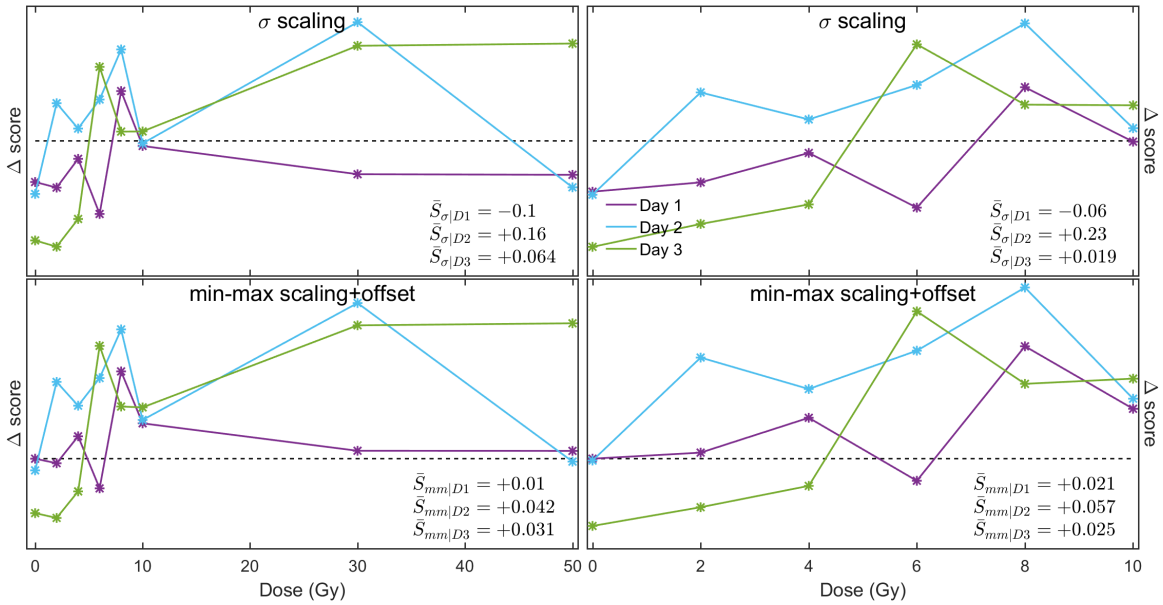


Figure B.9: PPCA/PCA H460B PC2 score distances.

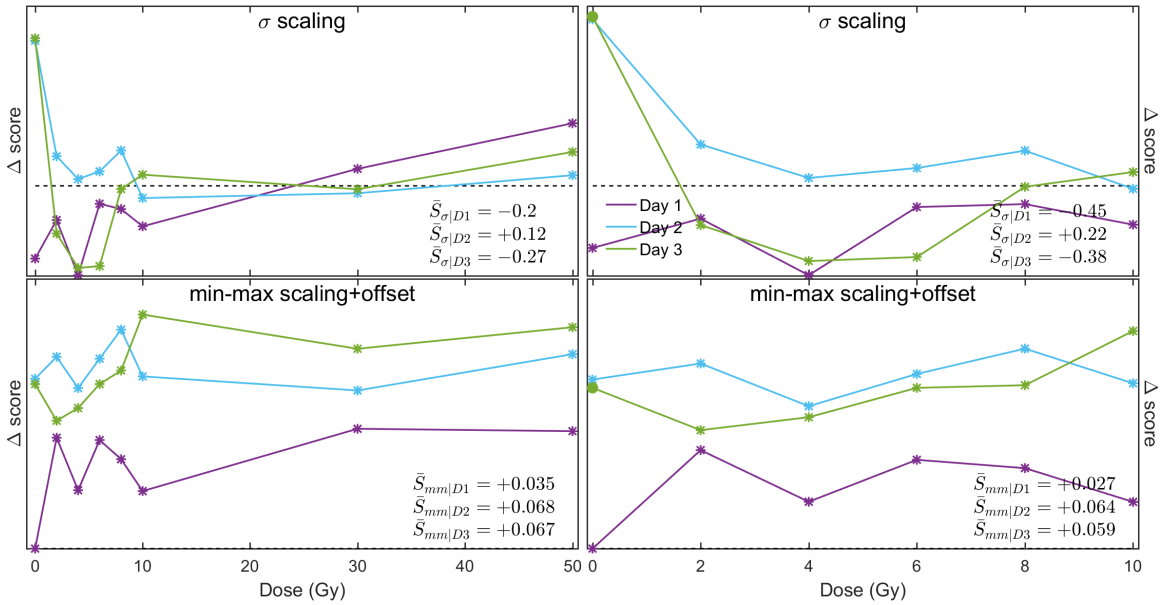


Figure B.10: PPCA/PCA H460B PC3 score distances.

NLPCA

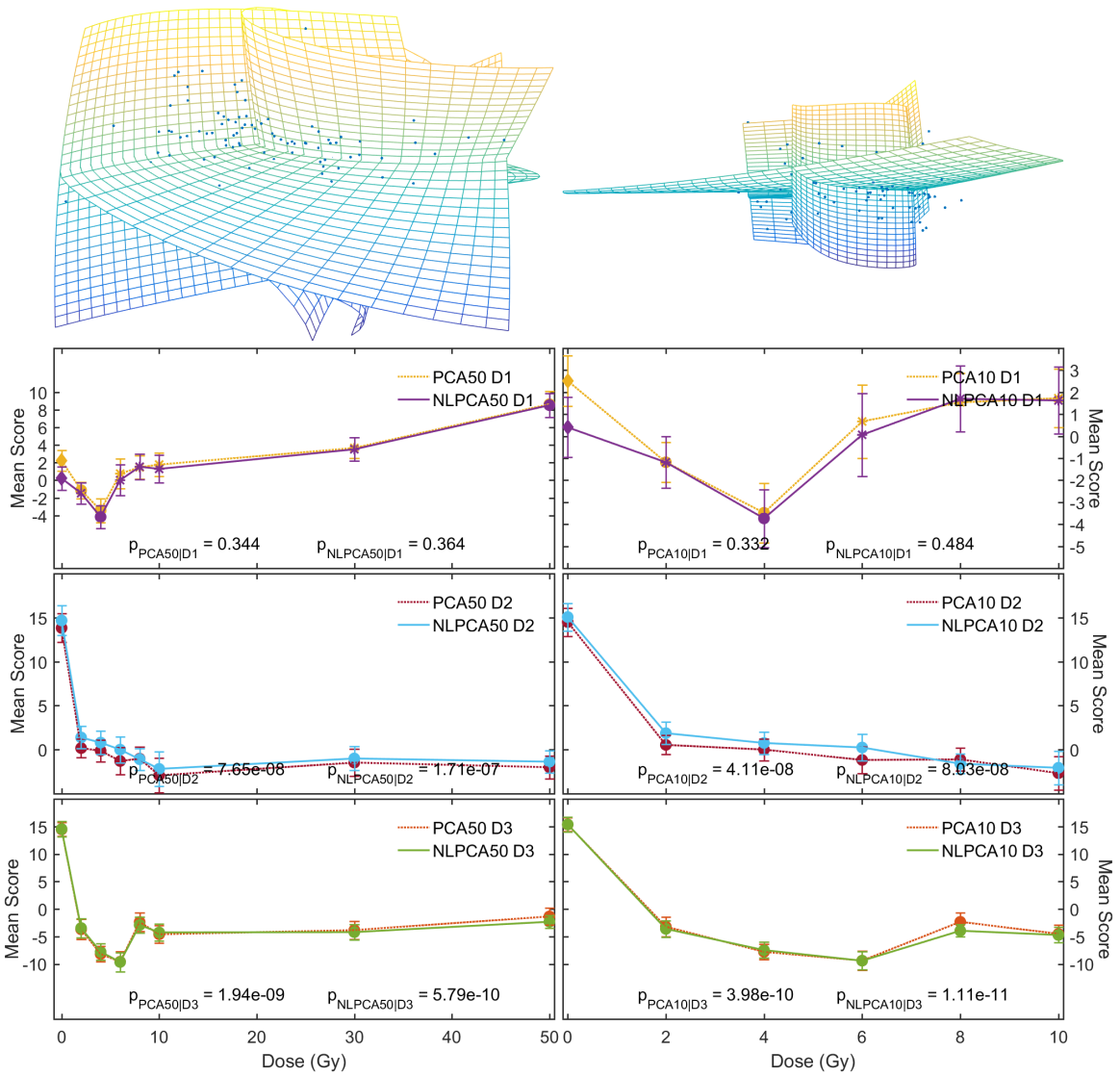


Figure B.11: NLPCA projection and PC3 scores for H460B 50Gy/10Gy datasets. PC curves are projected into 3D principal subspace, and Y-Z perspective used.

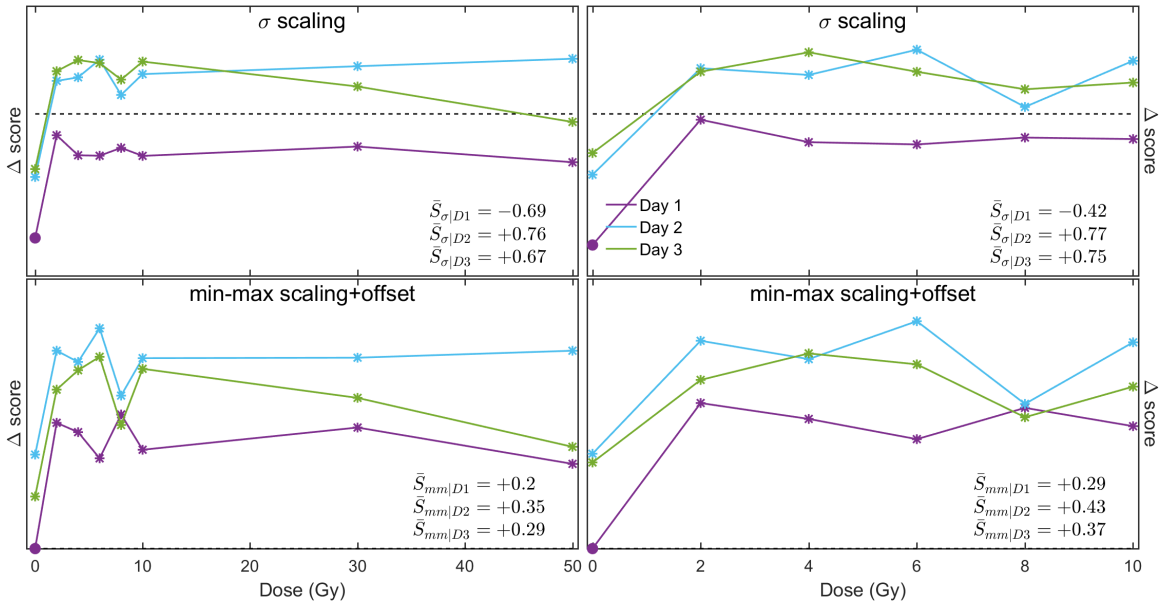


Figure B.12: NLPCA/PCA H460B PC2 score distances.

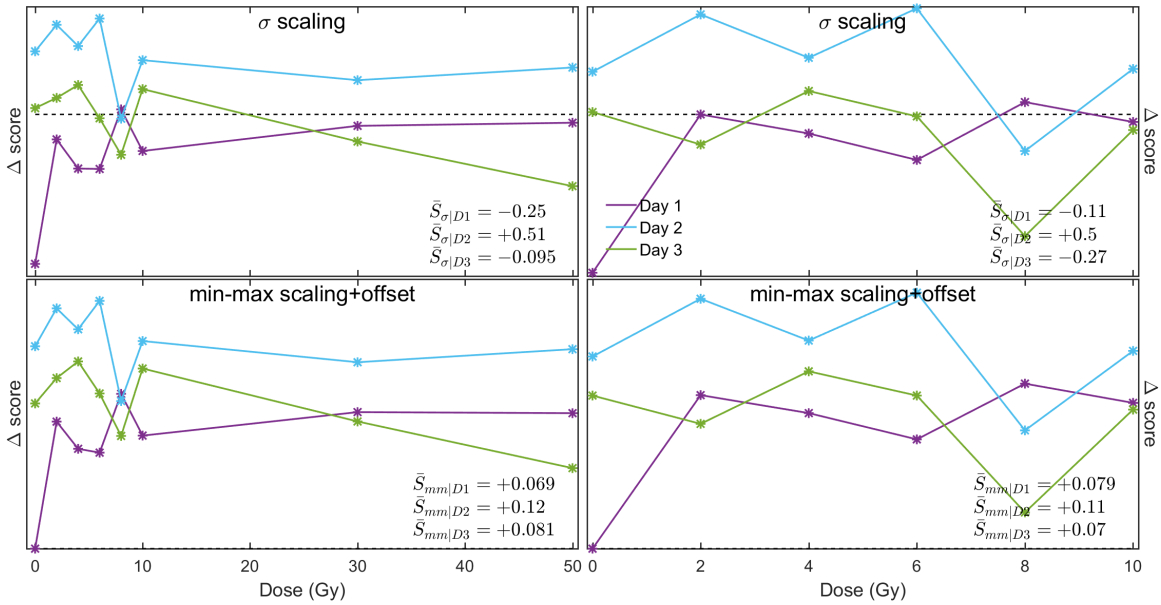


Figure B.13: NLPCA/PCA H460B PC3 score distances.

B.2 LNCaP B

PCA

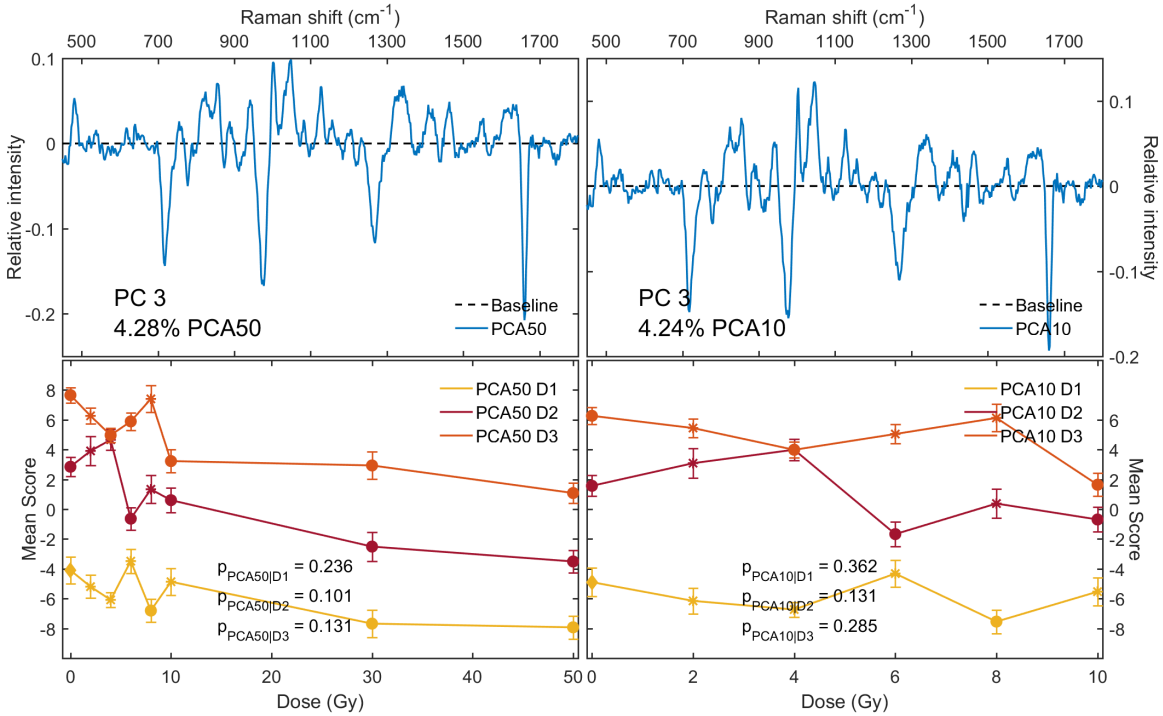


Figure B.14: PCA component 3 and respective scores of LNB 50Gy/10Gy datasets.

WPCA

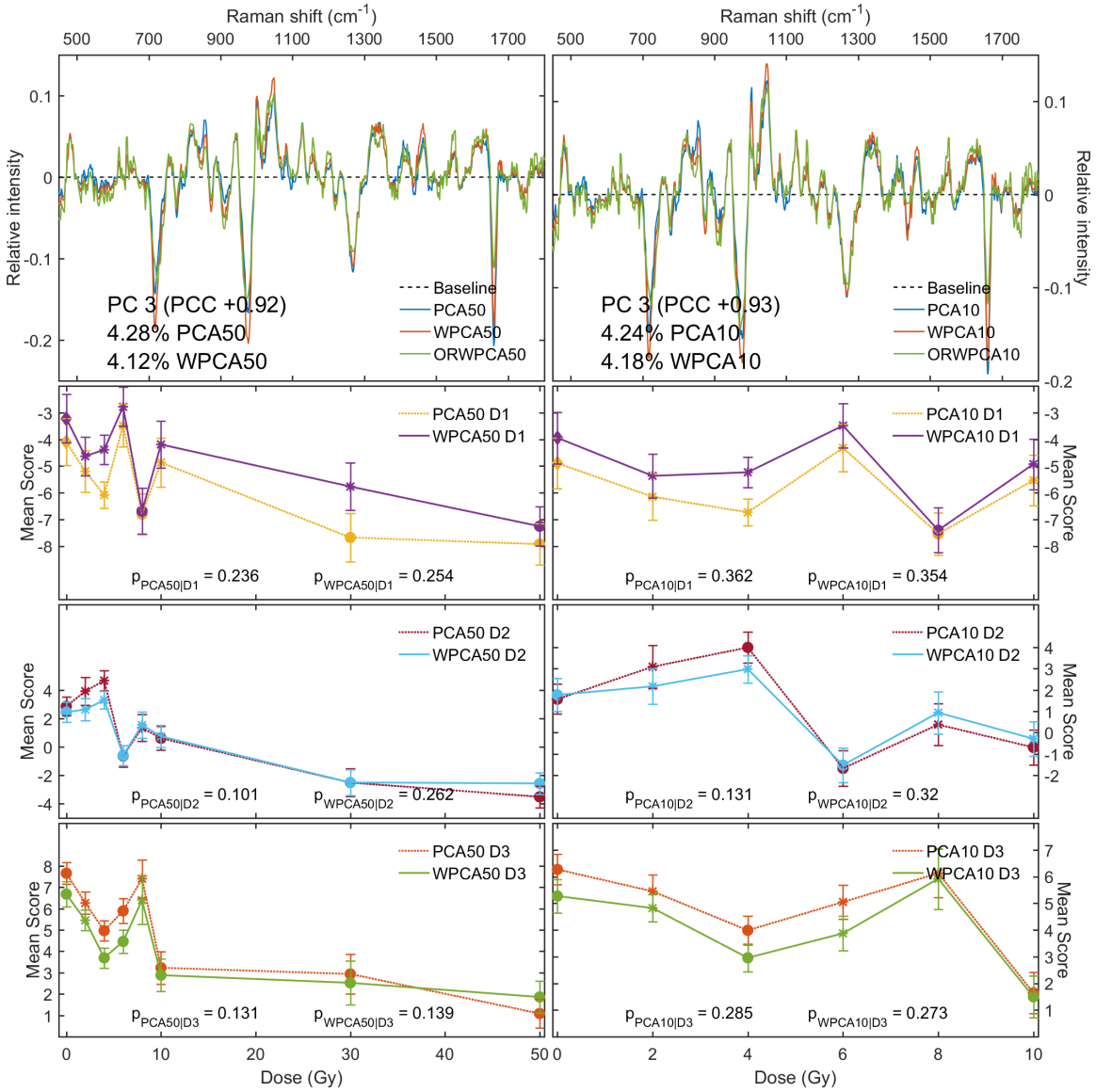


Figure B.15: WPCA/PCA component 3 and respective scores for LNB 50Gy/10Gy datasets.

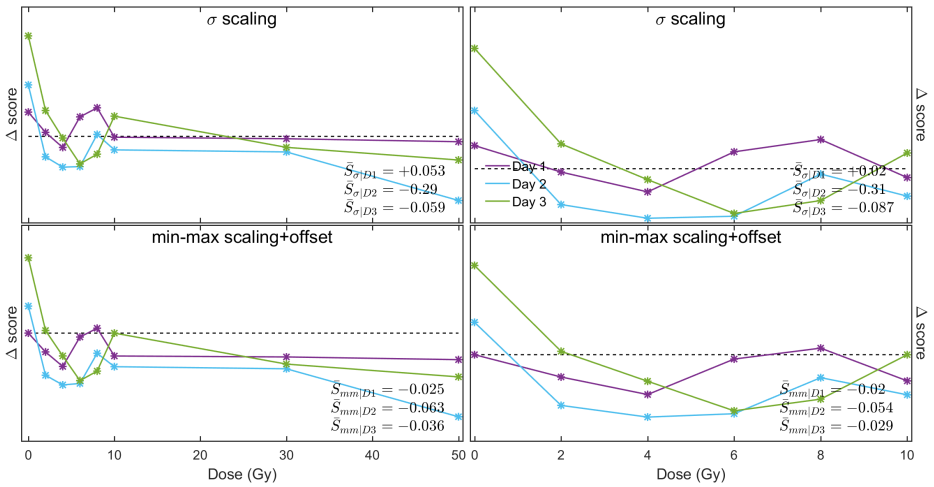


Figure B.16: WPCA/PCA LNB PC1 score distances.

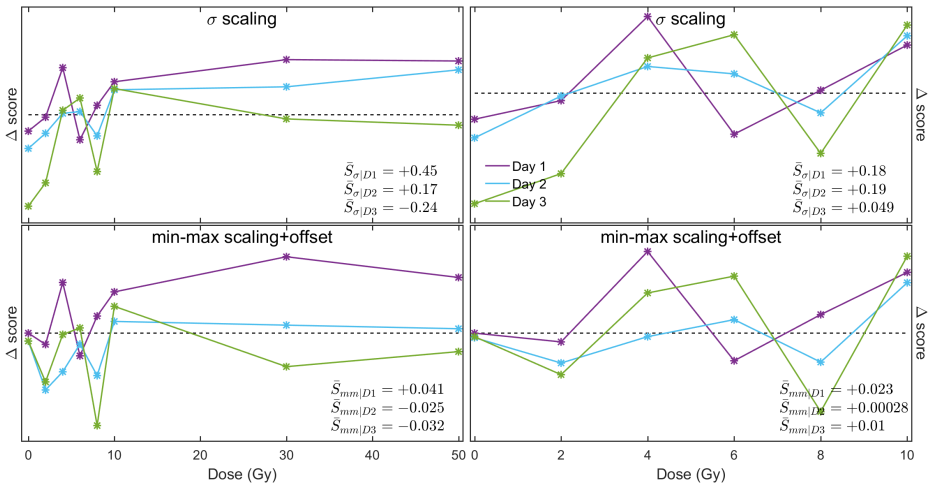


Figure B.17: WPCA/PCA LNB PC2 score distances.

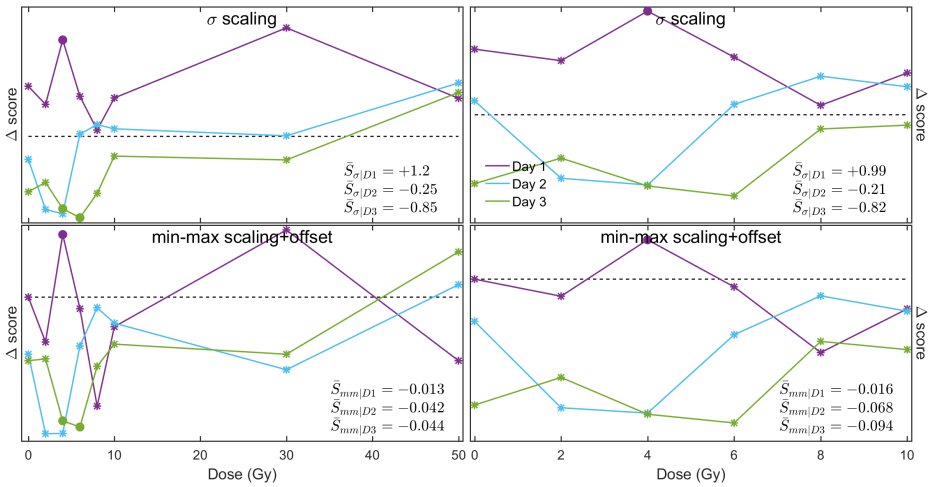


Figure B.18: WPCA/PCA LNB PC3 score distances.

RPCA

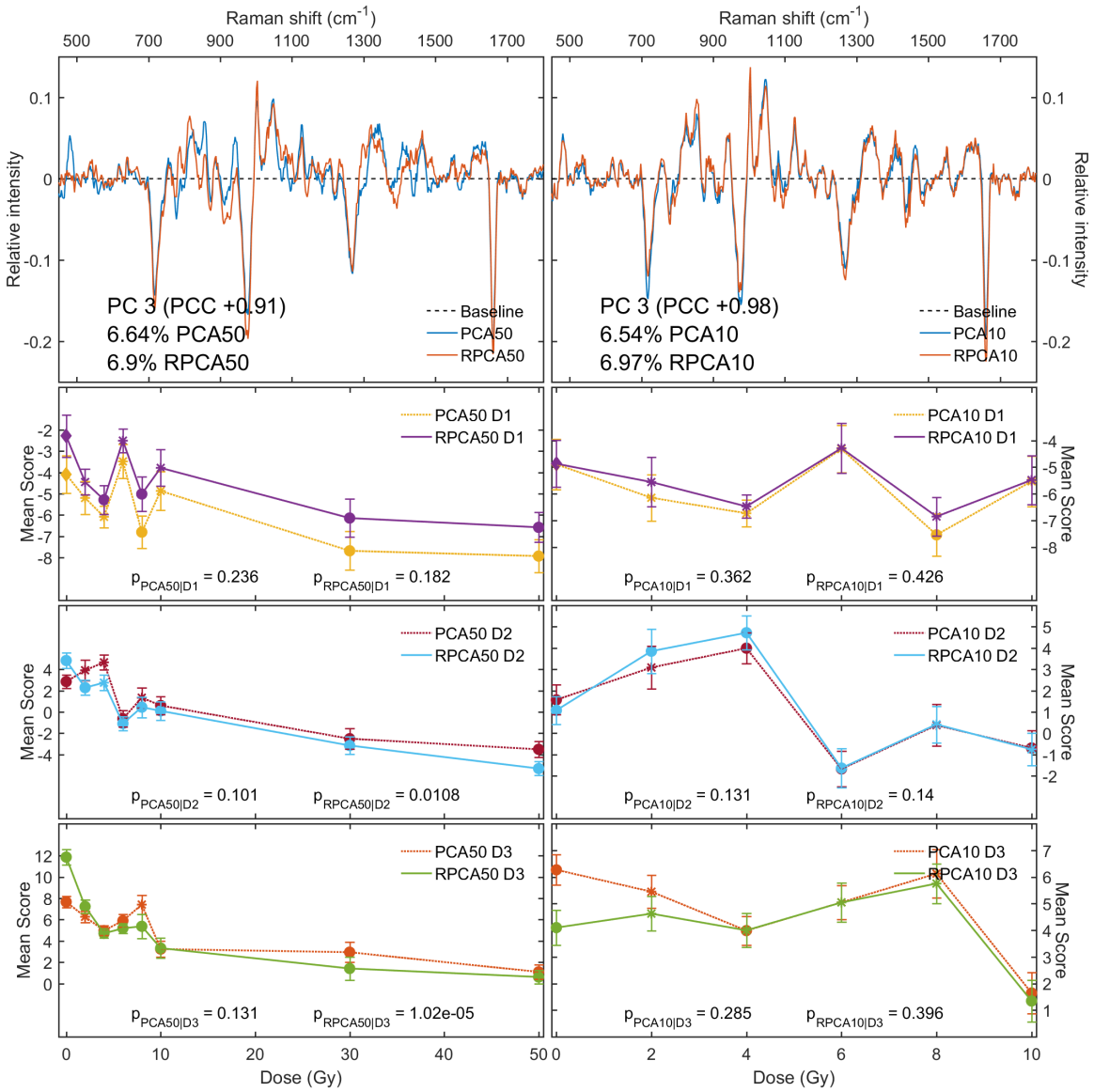


Figure B.19: RPCA/PCA component 3 and respective scores for LNB 50Gy/10Gy datasets.

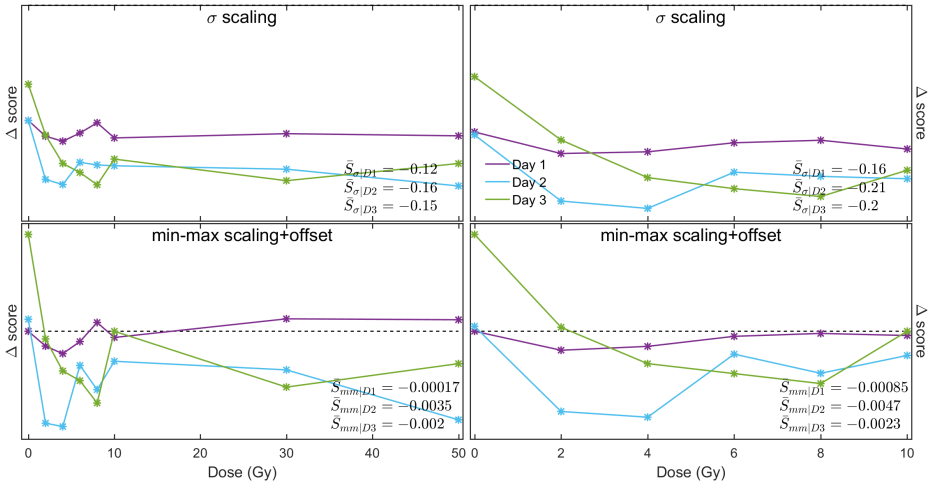


Figure B.20: RPCA/PCA LNB PC1 score distances.

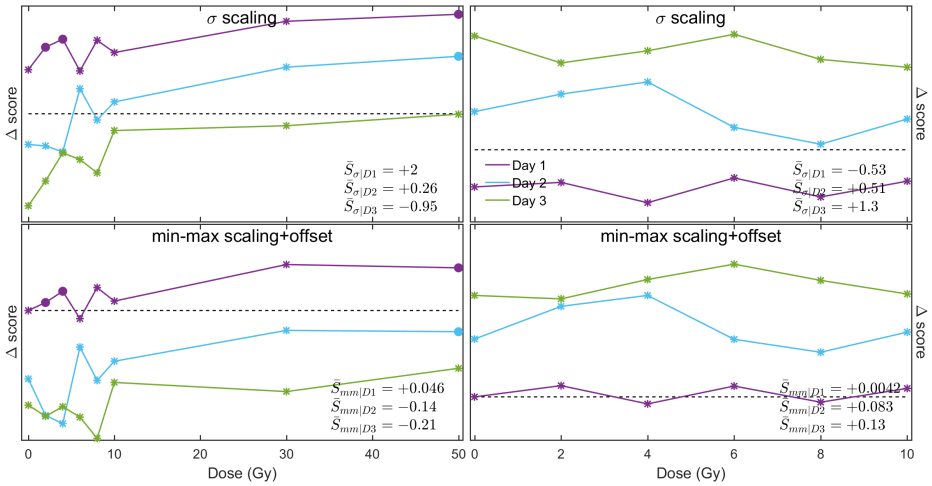


Figure B.21: RPCA/PCA LNB PC2 score distances.

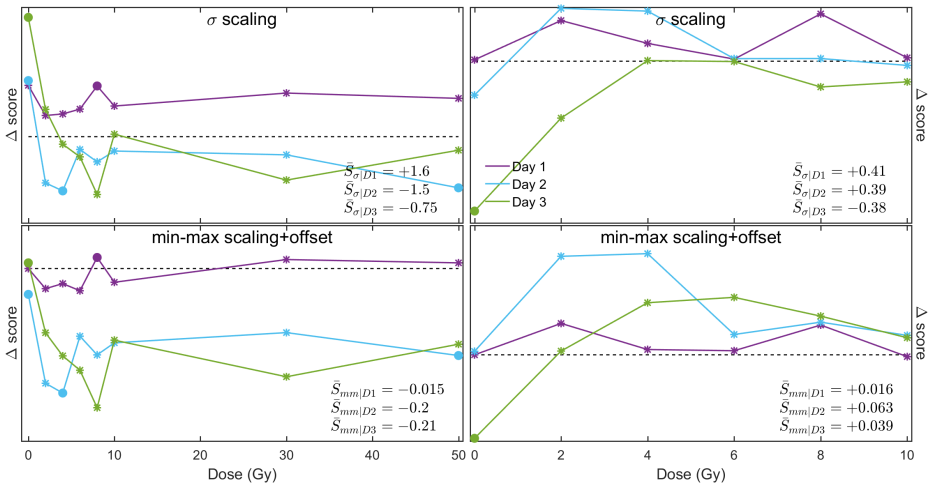


Figure B.22: WPCA/PCA LNB PC3 score distances.

PPCA

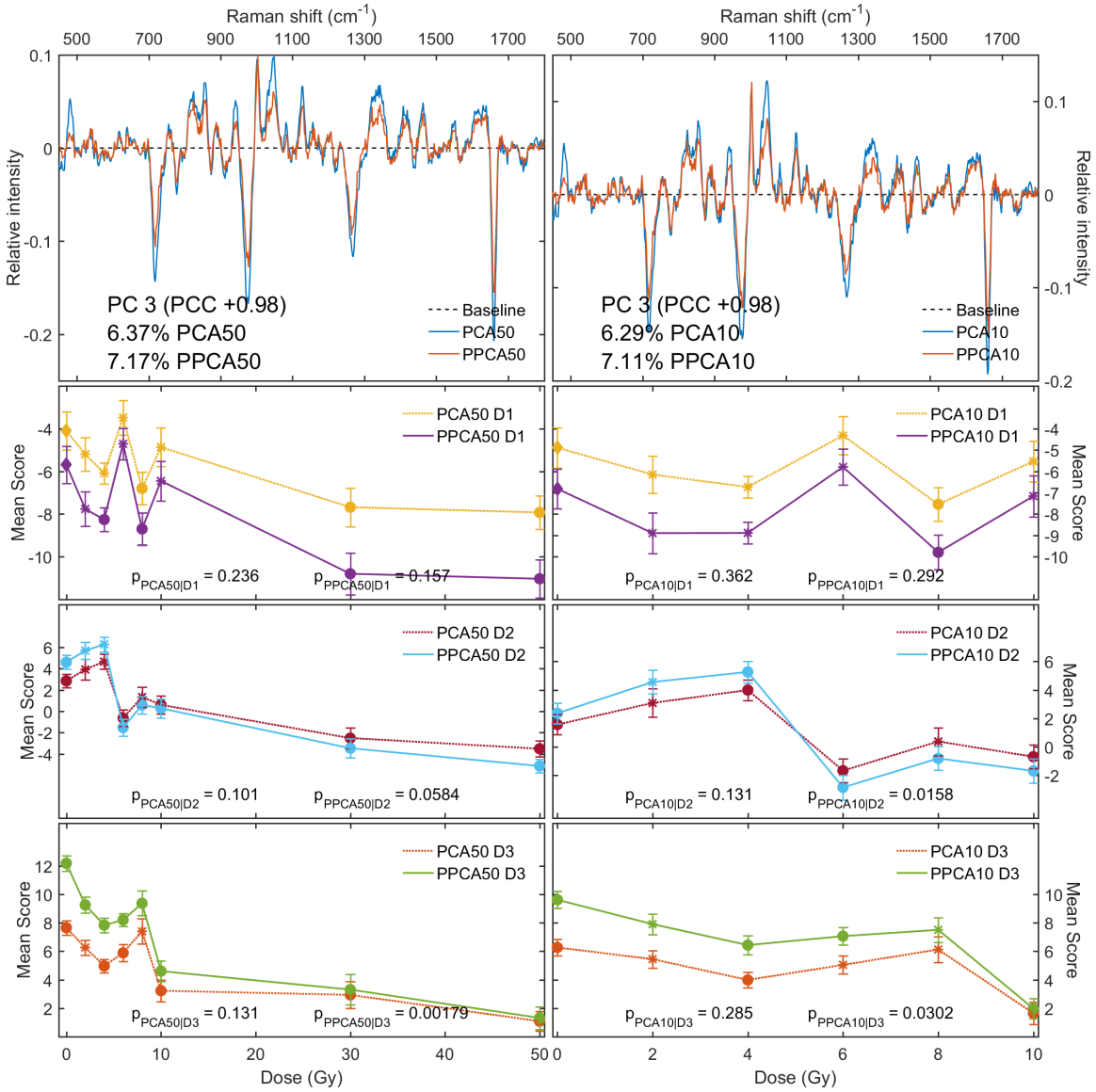


Figure B.23: PPCA/PCA component 3 and scores of LNB 50Gy/10Gy datasets.

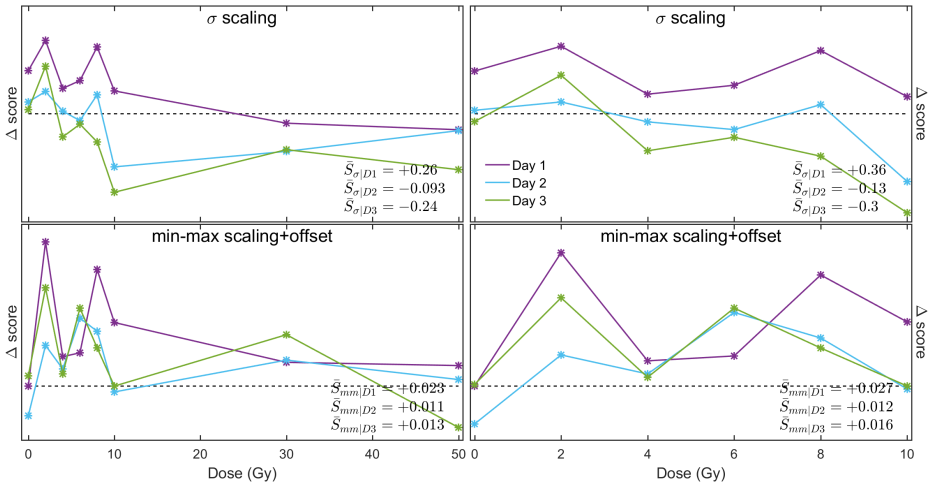


Figure B.24: PPCA/PCA LNB PC1 score distances.

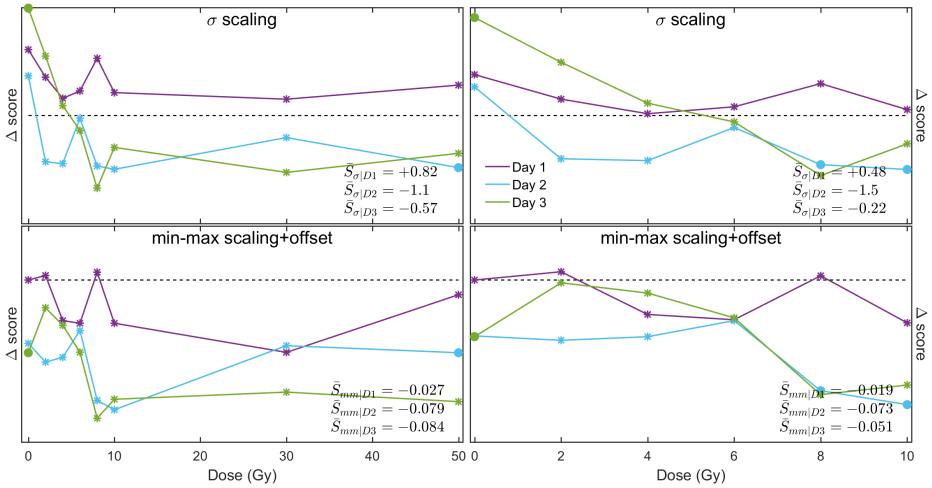


Figure B.25: PPCA/PCA LNB PC2 score distances.

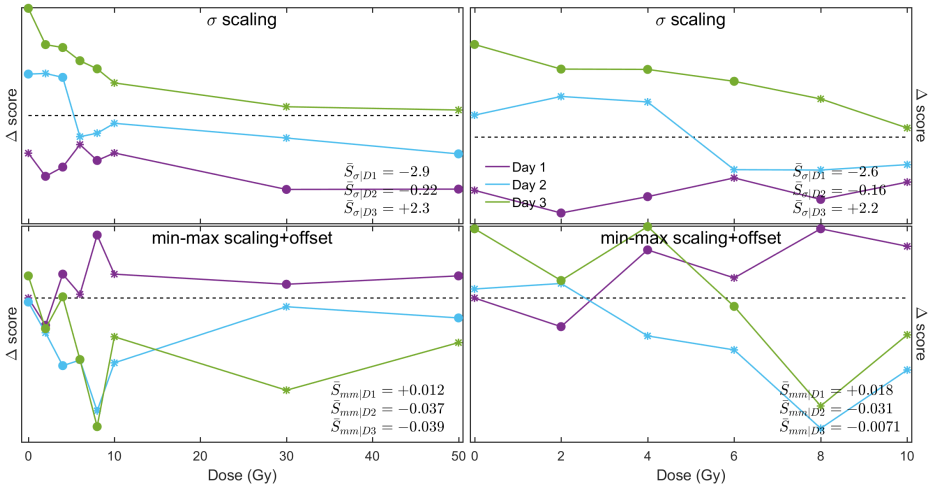


Figure B.26: PPCA/PCA LNB PC3 score distances.

NLPCA

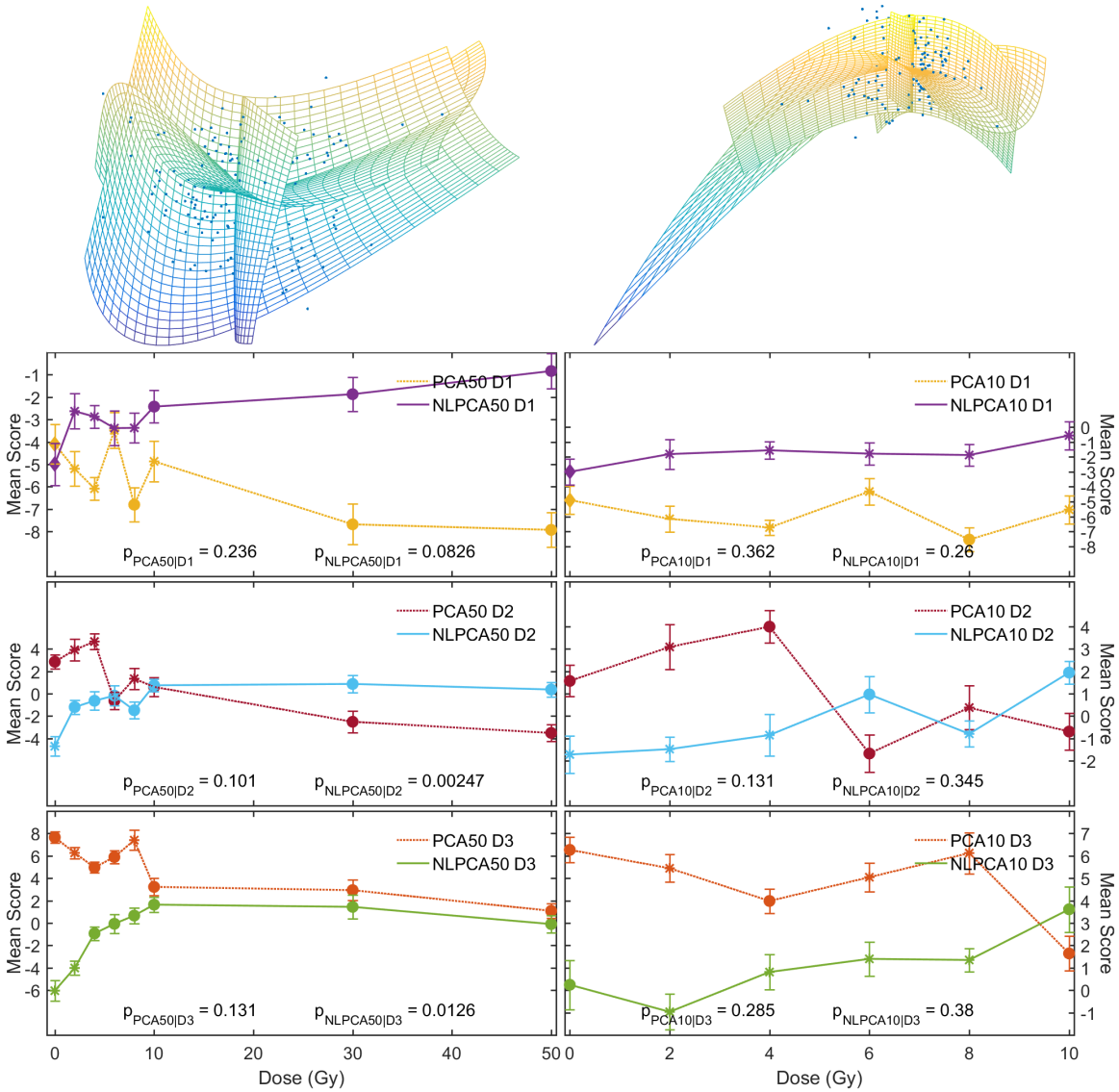


Figure B.27: NLPCA PC3 projection and scores for LNB 50Gy/10Gy datasets. PC curves are projected into 3D principal subspace, and Y-Z perspective used.

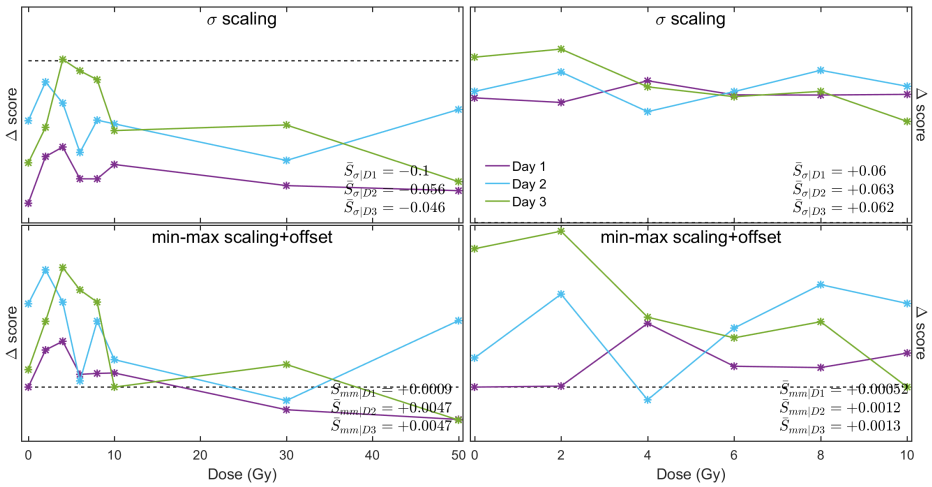


Figure B.28: NLPCA/PCA LNB PC1 score distances.

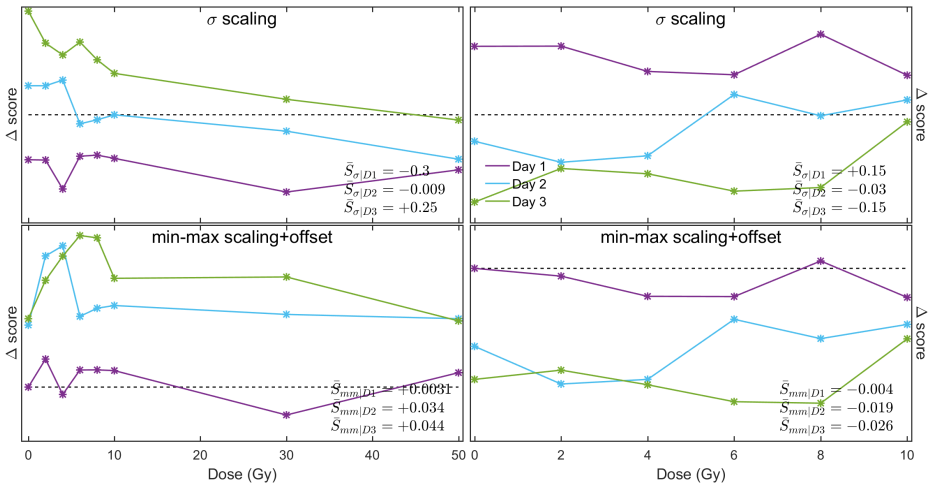


Figure B.29: NLPCA/PCA LNB PC2 score distances.

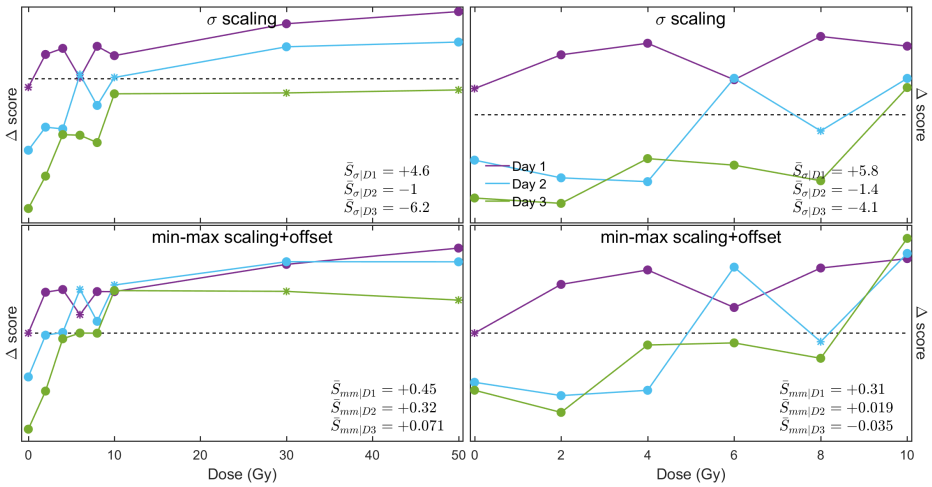


Figure B.30: NLPCA/PCA LNB PC3 score distances.