Spatial Resolution in Penalized-Likelihood Image Reconstruction

by

Joseph Webster Stayman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Electrical Engineering and Computer Science
(Electrical Engineering: Systems)
in The University of Michigan
2002

Doctoral Committee:

Associate Professor Jeffrey A. Fessler, Chair
Senior Associate Research Scientist Neal H. Clinthorne
Professor Alfred O. Hero
Emeritus Professor W. Leslie Rogers

This dissertation is dedicated to my mother and father.

ACKNOWLEDGEMENTS

I wish to thank my advisor, Jeff Fessler, for the guidance, friendship, and opportunities he has provided during my graduate career at the University of Michigan. I feel greatly indebted to Jeff. He is an outstanding person and has been the perfect advisor for me. I would also like to thank the rest of my committee for contributing their expertise for this work.

I also wish to thank my family for supporting me over the years. Thanks for all the care packages, letters, and late night counselling. And a special thank you to Aaron, which I must communicate telepathically, since it involves words I shouldn't put it in print.

Thanks and love to my best friend and companion, Yulia, who has cared for me and supported me (especially in times of kidney stones). She remains close to my heart and I can't imagine living without her.

Thanks to all my friends who have supported me in one way or another over the years. Scott, Jason, and Eric: I have missed you guys and the simple days before we moved apart. Thanks to all of my friends at Michigan including Tara, without whom I couldn't possibly have survived my first year, and the 714 Lawrence crew (Ketan, Pedram, Robert, Matt, Jenn, Gunther, Babak), without whom I would have played a lot less cards and had much fewer embarrassing Halloween photos. And thanks to those who continue the tradition with games of discovery and trade (Mark, Sasan, Rajiv).

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

The quality or performance of imaging systems is often quantified in terms of the spatial resolution of the images that those systems produce. A measure of resolution relates how well small features in the image can be resolved. If a true image has been convolved with a space-invariant filter, the blur filter is called the impulse response and completely specifies the resolution properties of the resultant image. However, many imaging systems do not produce images that can be represented in this form. For such systems, the resolution properties are space-variant and must be quantified locally.

Even systems that are intrinsically space-invariant can yield reconstructed images with space-variant resolution properties. This is because the resolution properties of an image are a function of both the imaging system and the image estimator. Regularized statistical estimators are often subject to space-variant effects, since the implicit data-weighting of the statistical model will induce different resolutions based on the local statistics. Thus, the resolution properties are image-dependent as well.

For images, space-variant resolution means that different image locations exhibit different blurs. These blurs can vary in both magnitude and shape. Anisotropic blur will preferentially smooth features of an image in certain directions. This can

complicate some tasks. For example, any task that involves extraction of the shape of image features can potentially suffer from anisotropic blur. Discs in a true image will appear elliptical after such a blur is applied. Similarly, object-dependent space-variant blur can complicate comparisons between images. If images are not resolution matched, detection tasks, registration, and boundary identification can yield misalignments or mismatches from the image-dependent shape distortions.

Thus, it is useful to have ways to measure the resolution properties of an image. The local impulse response is a generalization of the impulse response and is one tool for measuring space-variant resolution. The local impulse response has been widely used for space-variant resolution investigation. We provide a new derivation of the local impulse response in Chapter III, appropriate for a class of regularized statistical estimators, where a finite number of measurements are obtained from a continuous object and a discrete representation of the object is estimated.

In some cases, the local impulse response can be formulated as a function of the data measurements and the resolution properties can be predicted without performing any reconstruction. Such resolution predictors are also a function of the estimator parameters. Thus, resolution predictors may be used to select parameters like the level of regularization for a regularized statistical estimator. Since traditional regularization parameters are only obliquely related to resolution, predictors allow for more concrete resolution control. Specifically, resolution predictors can be used to design estimators with user-specified resolution properties. The idea of using resolution predictions to design estimators with specific resolution properties is investigated in detail in Chapter IV. Chapter V outlines many practical aspects of the implementation of the parameter design, with a concentration on the application to emission tomography estimators. We apply this estimator design technique with the goal of

uniform resolution properties (*i.e.*, space-invariant and isotropic) to positron emission tomography (PET) and single photon emission computed tomography (SPECT) systems in Chapter VI.

Such resolution control is important for comparing different estimators. However, resolution is only one of many potentially important factors in image quality. In fact, many estimators can be designed that specify arbitrarily "good" resolution. Generally, the trade-off is that for finer resolutions, estimated images exhibit increased noise. Thus, to make a fair comparison of different estimation methods, one should consider resolutions at a fixed image noise level, or, equivalently, noise levels at a fixed image resolution. We include such performance investigations in Chapter VII.

Resolution is only one of many measures of bias that is used in such bias versus variance comparisons. Other measures like bias-gradient length[54] have been used. Resolution is a convenient measure because it is easily interpreted and task-independent. Ultimately, the best performance comparisons are made by evaluating an estimator with a specific task in mind. For example, detection tasks can be evaluated by forming a receiver operating characteristic (ROC) curve, that shows the probability of detection for a given false alarm rate. There are also many extensions to the ROC curve that include localization in the detection task[121, 43]. These performance curves necessitate an observer to perform the detection task. Many computer observers have been developed which model human detectors[139]. These observer models generally require knowledge of the covariance of the reconstructed images. Typically, these covariance predictors require a large amount of computation. However, the same new techniques presented in Chapter V for rapid resolution predictions can also be used to yield rapid covariance predictions.

Thus, in the future, one logical extension of this work is to use fast resolution and

covariance predictors to design an estimator that is optimal for a specific task (*e.g.*, regularization that is tailored to a certain task). While it is unclear whether uniform resolution will play a part in the optimal estimator for certain tasks, the ability to make fast resolution and covariance predictions will remain important for situations where many resolution or covariance estimates are made.

The main contributions discussed in this work are:

- A new derivation of the local impulse response for a general class of imaging systems and noise models where a finite number of measurements are made from a continuous object and are reconstructed using a discrete object model and a penalized-likelihood estimator. [118]

- Development and validation of fast techniques for evaluating the local impulse response (and approximating covariance) in tomographic systems with particular attention to both shift-invariant and shift-variant emission tomography systems.[119, 120]

- Development of practical penalty design methods for penalized-likelihood estimators that allow the specification of user-defined resolution properties prior to image reconstruction. [118, 117, 116, 115, 114]

- A study of estimators that can provide uniform resolution. Specifically, which estimators can provide the most uniform resolution, and among resolution matched estimators, which estimators have the best noise performance.[118, 116]

While these topics are discussed in detail in this work, they have also been discussed in a number of conference and journal papers as indicated above.

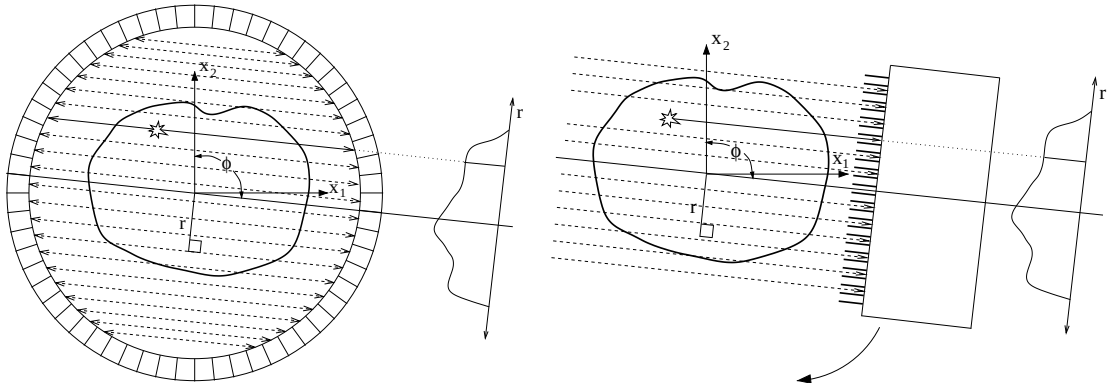
CHAPTER II

Background

While many contributions of this thesis can be applied generally to generic imaging problems, the initial impetus for this work arose from discussions with clinicians on the resolution properties of reconstructed images formed from emission tomography systems. Therefore, most of the studies and discussions presented in this work focus on emission tomography and incorporate physical effects appropriate for such systems into the system models. Thus, we begin the background introduction with a summary of some important aspects of emission tomography.

2.1 Emission Computed Tomography

Emission computed tomography (ECT) is a medical imaging modality that can provide unique functional information about physiological processes in the body. Typically, a small amount of a radioactive compound (or radiotracer) is introduced into a subject via injection or inhalation. Sometimes the radiotracer itself is of physiological interest as in ^{15}O imaging in the brain. In other situations, the radioisotope is attached to a molecule that is selectively taken up in different anatomical regions, such as ^{18}F labeled fluorodeoxyglucose (FDG) in tumors. After allowing the radiotracer to distribute throughout the body, an image of the radiotracer distribution can be made. This image indicates the concentration of the radiotracer in different



(a) Positron Emission Tomography (b) Single Photon Emission Computed Tomography

Figure 2.1: Idealized emission computed tomography systems.
(a) The PET system obtains projection data over a range of angles (ϕ) and radial positions (r) by detecting pairs of coincident gamma photons. (b) The SPECT system obtains projection data via collimation of gamma photons emitted from the object and rotation of the detector system around the object. Lines of response (LORs) are shown with dotted lines in both figures.

anatomical regions. This is important for diagnosis, for example, in cancer studies, since tumors tend to use more glucose than other regions in the body. Therefore, FDG images often show "hot spots" or regions of higher concentration where tumors lie. In other studies physicians are interested in "cold spots" due to improper blood circulation.

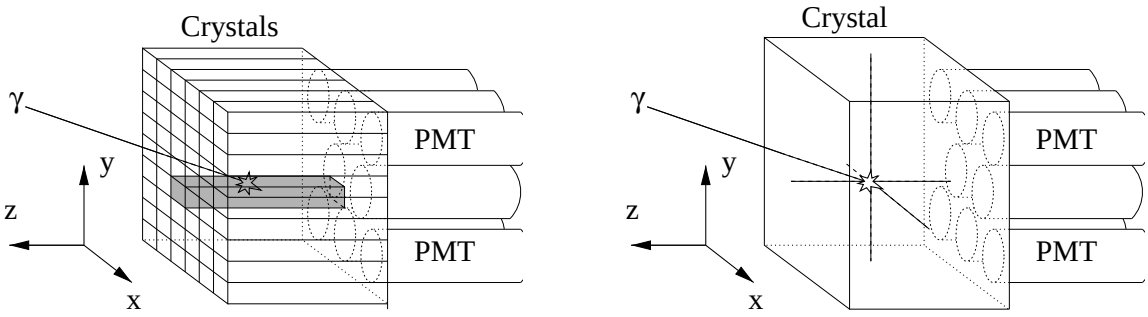
The two most common types of ECT are positron emission tomography (PET) and single photon emission computed tomography (SPECT). In PET imaging, as the radioisotope decays it emits positrons. These positrons travel a short distance before annihilating with an electron. (This effect, known as positron range[74] is usually small and therefore neglected.) Each annihilation event creates two 511 keV gamma photons that travel in opposite directions. If both of these photons travel coplanar with the coincidence detector ring that surrounds the patient and deposit energy in a pair of detectors, the number of events for that line of response (LOR) is incremented. In other words, we know the annihilation event took place somewhere along the line connecting the pair of coincidence detectors. See Figure 2.1a for a

simplified representation of a PET system. To determine which events are coincident, detectors look for a pair of events in a short duration coincidence window[122]. In this way, projection data are obtained over a range of angles and radial positions. (A single projection at angle ϕ is shown to the right of the simplified PET diagram in Figure 2.1a.) All projection data are collectively called a sinogram.

In SPECT imaging, the radioisotope that is used decays and emits a single gamma photon. These photons are detected by a rotating array of detectors. To obtain LOR information, a collimator is placed in front of the detectors so that the gamma photons can only enter at known angles. (Often a parallel hole collimator is used so that the photons enter perpendicularly.) As the detector array rotates around the object, a full range of angles may be obtained. See Figure 2.1b for a simplified representation of a SPECT system.

2.1.1 Detectors

The detectors and their geometry (and the associated detector components like the collimator in SPECT) are probably the most important items when it comes to system resolution properties. Therefore, for reconstruction from high resolution imaging systems an accurate detector model is important. A simple gamma ray detector is composed of a scintillating crystal, which is optically coupled to a photomultiplier tube (PMT). When a gamma photon interacts inside a scintillating crystal, such as sodium iodide (NaI) or bismuth germanate (BGO), a burst of many light photons is produced. This burst of light is converted into a short electrical pulse by the photomultiplier tube. These pulses may then be counted, recording the number of detected gamma photons. Simple detection is not sufficient to construct an image of the object. Localization of this detected gamma photon is also required, so that the LOR along which the gamma photon originated is known.



(a) Block Detector

(b) Anger Camera

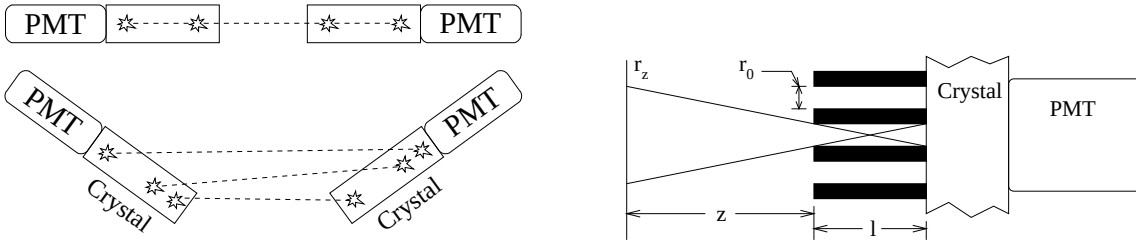
Figure 2.2: Localization in different detection systems.

In PET systems, a block detector is typically used (see Figure 2.2a). In a block detector a rectangular bundle of crystals is optically coupled to several PMTs. The crystal block is fabricated so that the light collected by the PMTs varies uniquely for each of the rectangular crystals. Therefore, the crystal in which a scintillation takes place can be identified by looking at the output of the PMTs.

In SPECT systems, the detectors are usually in the form of an Anger camera (shown in Figure 2.2b). In an Anger camera, a large crystal is optically coupled to many PMTs. PMTs closer to a scintillation event will gather more light; those further away will gather less light. Since light gathered by the PMTs is position-dependent, the (x, y) position in the crystal can be estimated from the PMT outputs.

There are a number of physical and geometric effects that can complicate event localization. These factors create an ambiguity in measuring the exact LOR for a particular event. For example, because scintillations often consist of one or more Compton scattering interactions before photoelectric absorption, the light created in a crystal has a spatial distribution.

In PET, the block detector has crystals of finite size (often on the order of a few millimeters) and can only localize an event to within a crystal and the probabilities



(a) PET Detectors

(b) SPECT Detection System

Figure 2.3: Detectors in PET and SPECT systems.
(a) In PET, the depth of interaction in the scintillating crystal causes an ambiguity in the LOR. The probability distribution of LORs for a given pair varies depending on the detector configuration. Generally, the greater the relative angle between a detector pair, the greater the range of LORs. (b) In SPECT, due to finite collimator depth and hole size, photons are detected within some acceptance angle even for an ideal collimator. This has the effect of increasing the area of accepted photons with increasing distance from the collimator.

of detection are spatially nonuniform within that crystal. These probabilities are highly dependent on the exact detector geometry and emission positions. In SPECT, Anger camera position estimates are related to the number of photons given off in a scintillation event and are subject to noise and thus have a spatial distribution (which is often approximated by a 2D Gaussian function). Such effects are often lumped together and called the detector response. The detector response is fully specified by a detector sensitivity pattern. In general this sensitivity pattern is a function of x, y, and z, and expresses the probability that a given detector[1] will detect gamma photons originating at a given position (x, y, z).

There are a number of effects that contribute to space-variant detector responses. That is, the sensitivity patterns for all detectors are not simple rotations and translations of a single detector's sensitivity pattern. In PET, the depth of interaction (DOI) in the crystal leads to space-variant detector responses. (This effect is illustrated in Figure 2.3a.) Because of the ring geometry in PET, pairs of detectors are

[1]Strictly speaking, in PET, any localization involves a pair of detectors. However, it is convenient to consider the pair of detectors a single LOR detector.

not necessarily parallel to each other. Since the detectors localize only in the x-y plane,[2] ambiguity in the z direction leads to an ambiguity of the LOR. This effect becomes worse for increasingly oblique detectors and leads to space-variant detector responses.

Additionally, there are other physical effects that can alter the magnitude or relative sensitivity of the detector response. This effect is known as detector efficiency and generally varies from detector to detector[108]. For example, since PMTs are relatively sensitive analog devices, the PMT sensitivity tends to change or drift over time[99]. Detector efficiency can be measured by performing a normalization scan[56, 53, 6]. Such scans are performed periodically with long acquisition times to obtain good estimates of detector efficiency.

The shape of the detector response is also important. As mentioned previously, in SPECT, a collimator is required in order to obtain LOR information. An enlarged view of a collimator and detector for a SPECT system is shown in Figure 2.3b. The collimator is typically made of lead and is meant to prevent oblique gamma photons from passing through to the scintillating crystal. It is obvious from the geometry that even if no photons pass through the lead, photons enter the detector over a range of angles. For a collimator with holes of diameter x_0 and depth l, the diameter of the circle over which photons are accepted at depth z is $r_0(1+2z/l)$. The result is a detector response whose size increases linearly with increasing distance. Therefore, there is increasing spatial ambiguity with increasing distance from the detector. In real collimators, gamma photons pass through the lead with a probability given by Beer's law. This is called penetration. These penetration effects tend to smooth out the detector response from the ideal detector response function.

[2]There has been much work in designing detectors that also measure depth of interaction[86, 42]. However, while these new detectors can minimize space-variant effects, in general the detector responses will still be space-variant due to ambiguity of the DOI estimate.

2.1.2 Physical Effects

There are other physical effects that are not directly related to the detectors themselves that also complicate the measured data. Gamma photons are prone to two dominant interactions: absorption and Compton scatter. Both absorption and scatter can have a similar effect. Many scattered photons are never detected (they are scattered out of the detector planes), and absorption in the object prevents the photons from reaching the detector. These effects are jointly termed attenuation.

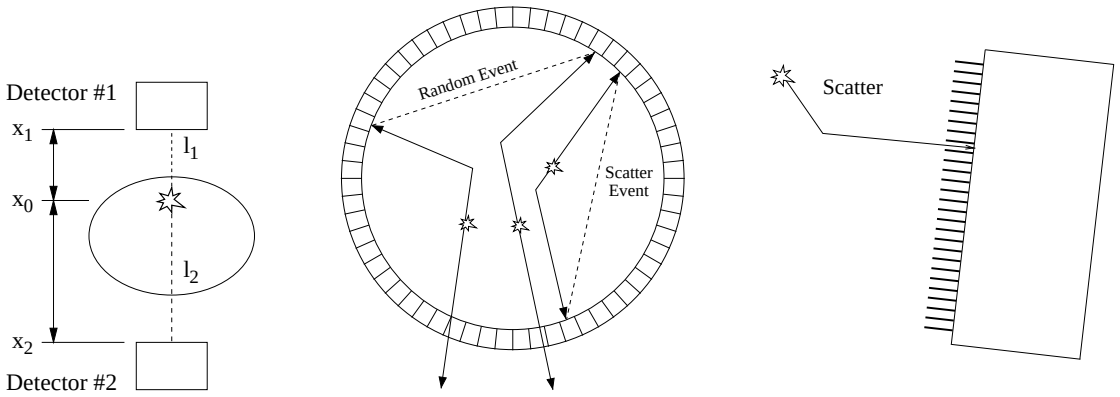
The survival probability[80] of a photon along a line l is given by

$$P_l = \exp\left\{-\int_l \mu(x)dx\right\}, \tag{2.1}$$

where $\mu(x)$ is the linear attenuation coefficient at position x. Therefore, for a PET system as in Figure 2.4a, where a pair of photons are emitted, the survival probability for both photons is

$$P = P_{l_1}P_{l_2} = \exp\left\{-\int_{x_1}^{x_0} \mu(x)dx\right\}\exp\left\{-\int_{x_0}^{x_2} \mu(x)dx\right\} = \exp\left\{-\int_{x_1}^{x_2} \mu(x)dx\right\}, \tag{2.2}$$

which is just the line integral along the LOR connecting the detector pair. Since the survival probability involves the complete line integral connecting the two detectors, attenuation can be incorporated as a ray-dependent scaling factor. This is not the case in SPECT. Consider the single detector system given by Figure 2.4a, when detector #2 is ignored. In this case the survival probability is given as P_{l_1}. This survival probability is depth-dependent and cannot be included as a simple ray-dependent factor. However, one can still include this effect in a system model by including attenuation in the detector model. In both PET and SPECT, in order to include attenuation effects, a transmission scan must be performed to estimate the attenuation map given by $\mu(\underline{x})$. To obtain the attenuation map, a radioactive



(a) Attenuation (b) PET: Randoms and Scatter (c) SPECT: Scatter

Figure 2.4: Physical effects in ECT systems.

rod source is typically rotated around the object prior to the emission scan in a transmission scan[72]. In addition, a blank scan is performed without any object in the scanner. The blank scan and transmission scan (or their ratio) are used to reconstruct the attenuation map[31, 44]. Once an attenuation map has been reconstructed, the emission system model can be modified to include these effects.

Sometimes, scattered photons are still detected. In PET, this means the photon paths are no longer colinear; and, in SPECT, this means the single photon has not originated along the detected LOR. This is illustrated in Figure 2.4b and c. Because scattered photons typically have less energy than unscattered photons, many scattered events can be eliminated by energy discrimination. For example, in ^{18}F PET imaging, the annihilation photons are emitted with 511 keV; and, in ^{99m}Tc SPECT imaging, the photons are at 140 keV. If photons are detected at lower energies (in either case), they can be assumed to be scatter and are rejected. Unfortunately, the detectors have finite energy resolution, and this scatter rejection is no longer completely straightforward. A number of energy windowing techniques[59, 79] and model-based methods[92, 131, 130] have been suggested to solve this problem. However, generally the measurement data will still contain a scatter component.

There is an additional concern in PET. This effect is called randoms or accidental coincidences and is illustrated in Figure 2.4b. Consider the case where one of the photons of a photon pair is scattered or attenuated so that it never reaches the detector. If this happens with two pairs of photons (such that only one photon of each pair is detected) within the duration of the coincidence window, the resulting event is termed an accidental coincidence[55]. These random events can be estimated by considering detected events in a delayed window. Pairs of photons detected with one photon in the coincidence (or prompt) window and one photon in the delayed window can only be due to accidental coincidences (since these photons travel at the speed of light and the windows are timed to include only events with the field of view of the scanner).

In addition, both PET and SPECT are susceptible to background radiation, that is, radiation from external sources such as naturally occurring radioactive decay. Much background radiation can also be eliminated through energy windowing methods; however, there will still be a background component present in the measurements.

2.2 Imaging System Models

Generally, any model of a real system must make certain approximations or neglect certain effects. Different models can have a number of different advantages. For example, more accurate models may contain parameters that are too numerous or difficult to measure, more complicated models are often difficult to analyze, and certain models will not lead to computationally feasible estimators, etc. For this reason we will present various models one might adopt for image reconstruction or analysis.

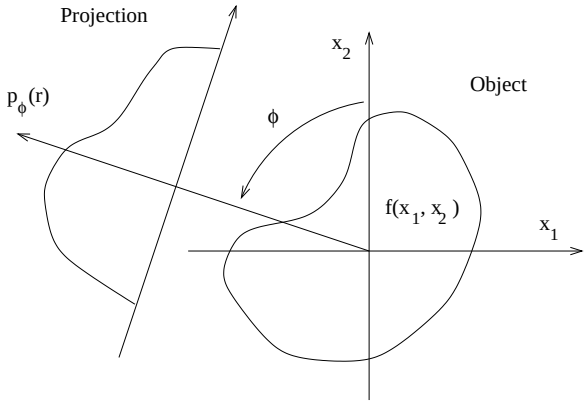


Figure 2.5: Generic continuous tomographic model and coordinate system.

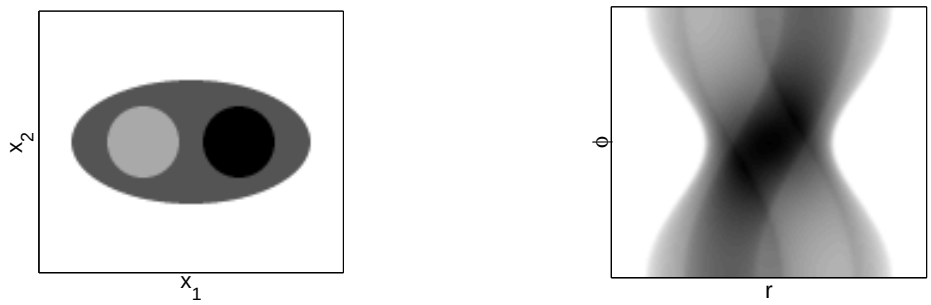
2.2.1 Idealized Continuous Model

In tomographic systems the measurements are projections of the object distribution onto a detector array. In Figure 2.1, all LORs for a given projection angle are shown for an idealized PET and SPECT system. A more generic simplified model is shown in Figure 2.5, which introduces the continuous model and its notation.

Let $f(x_1, x_2)$ denote an object intensity function defined over $\mathbb{R}^2$. Assuming a continuum of detectors oriented at an angle ϕ, a projection $p_\phi(r)$ is obtained which is a function of the radial distance, r. An ideal projection is related to the object distribution as a line integral along the line $L(r, \phi)$. However, we also include a (possibly) depth-dependent radial detector blur, $b(r, z)$, where z is defined as the perpendicular distance from an image location to the face of the detector. Thus, we write

$$\begin{aligned} p_\phi(r) &= \int_{-\infty}^{\infty} \int_{L(s,\phi)} b\left(r - s, z(l,s)\right) f(x_1, x_2)\ dl\, ds \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} b\left(r - s, z(l,s)\right) f(l\cos\phi + s\sin\phi,\ l\sin\phi - s\cos\phi)\ dl ds \end{aligned} \quad (2.3)$$

If $b(r, z) = \delta(r)$, there is no detector blur and the complete collection of projections $\{p_\phi(r) : \phi \in [0, \pi], r \in (-\infty, \infty)\}$ is known as the Radon transform[97] of the object,



(a) Object Distribution

(b) Sinogram

Figure 2.6: Example of a piecewise constant object and its Radon transform.

$f(x_1, x_2)$. We define the Radon transform operator, $\mathcal{P}$, where

$$p = \mathcal{P}f, \qquad \mathcal{P} : L_2(\mathbb{R}^2) \longrightarrow L_2([0, \pi] \times \mathbb{R}). \tag{2.4}$$

An example of an object, f, and samples of its Radon transform, g, are shown in Figure 2.6. The samples of the Radon transform are often called a sinogram due to shape of the projections of a single point. If $b(r, z)$ is not a delta function, then we will call the complete collection of projections the blurred Radon transform and will refer to the blurred Radon transform operator $\mathcal{P}_{\text{blur}}$.

We also identify the adjoint operation, backprojection, which transforms projections back into the image-domain:

$$f_b(x, y) = \int_0^{\infty} \int_{\infty}^{\infty} b(s) p_{\phi}(x_1 \cos\phi + x_2 \sin\phi - s) ds \, d\phi, \tag{2.5}$$

where the backprojection operators are denoted by $\mathcal{P}'$ and $\mathcal{P}'_{\text{blur}}$ for the ideal and blurred transforms, respectively.

The above continuous tomographic model is useful for analysis and insight. However, real systems yield discrete measurements; thus, there is an inherent model mismatch.

2.2.2 A Generic Continuous-Discrete Measurement Model

We can put the emission tomography problem within the context of a wider set of imaging problems using a more generic system model, where a finite set of measurements are obtained from a system that makes measurements on a continuous object function. Let $\underline{Y} \in \mathbb{R}^N$ denote the measurement vector recorded by the imaging system. We will recognize these measurements are noisy. Therefore, $\underline{Y}$ denotes a random vector whose unknown means depend on a continuous-domain object function, $f(\underline{x})$, where $\underline{x} \in \mathbb{R}^2$ or $\mathbb{R}^3$. We assume that elements of mean vector, $\bar{\underline{Y}}$, have the following form:

$$\begin{aligned} \bar{Y}_i^\dagger(f) \triangleq E\left[Y_i\right] &= \tau_i^\dagger\left(\int h_i(\underline{x}) f(\underline{x}) d\underline{x}\right) \\ &= \tau_i^\dagger\left(\left[\mathcal{H}f\right]_i\right), \end{aligned} \tag{2.6}$$

where $h_i(\underline{x})$ is the system "sensitivity" function for the ith measurement and includes all detector response and attenuation effects. The $\tau_i^\dagger$ function denotes a transformation relating the weighted integral to the mean measurements. We write the collection of weighted integrals for all measurements concisely using the continuous-to-discrete operator, $\mathcal{H}$, which maps a continuous image into N (untransformed) measurements.

The model in (2.6) can be used for many imaging systems. For example, emission tomography systems have measurements that are linearly related to the object, thus

$$\tau_i^\dagger(l) = l + r_i, \tag{2.7}$$

where r_i represents the mean contribution of background, scatter, and/or randoms. A transmission tomography system includes an exponential transformation, such that

$$\tau_i^\dagger(l) = b_i \exp(-l) + r_i, \tag{2.8}$$

where b_i represents detector normalization factors.

The true $h_i(\underline{x})$ and $\tau_i^\dagger(\cdot)$ functions are rarely known exactly, thus the reconstruction model is inherently mismatched to the measurement model. Additionally, it is common to adopt a discrete reconstruction model, where the image is approximated using a linear combination of basis functions.

2.2.3 Discrete Reconstruction Model

A discrete reconstruction model is often adopted to simplify reconstruction, display, and storage of reconstructed images. In this case, both the measurement data and the object fall into discrete bins or are pixelized in some fashion. While the object discretization typically takes the form of pixels or voxels, which are easily displayed on computer systems, other basis functions[7, 81, 75] may also be used.

Let the object function be represented as a linear combination of P basis functions with coefficient vector $\underline{\theta} \in \mathbb{R}^P$. Elements of the vector of mean measurements, $\underline{\bar{Y}}$, are assumed to be related to the discretized object as follows:

$$\bar{Y}_i(\underline{\theta}) = \tau_i \left(\sum_{j=1}^{P} h_{ij}\theta_j \right) = \tau_i \left([\boldsymbol{H}\underline{\theta}]_i \right), \tag{2.9}$$

where the elements $\{h_{ij}\}$ collectively make up the system matrix $\boldsymbol{H}$, and $\tau_i(\cdot)$ is a function that relates weighted sums of image parameters to the mean measurements. The $N \times P$ system matrix, $\boldsymbol{H}$, is meant to approximate the action of the continuous-to-discrete operator, $\mathcal{H}$, and $\tau_i(\cdot)$ is meant to approximate the transformation $\tau_i^\dagger(\cdot)$.

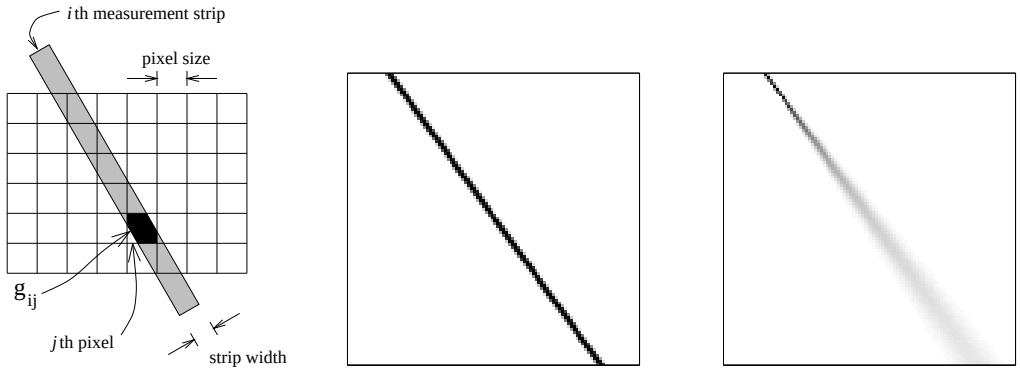
It is common to decompose the system matrix into several components that represent different physical aspects of the imaging system. For example, a reasonable formulation for PET or SPECT is

$$\bar{Y}_i(\underline{\theta}) = \tau_i \left(\sum_{j=1} c_i a_{ij} g_{ij} s_j \theta_j \right) = \tau_i \left(\left[\mathrm{diag}\{c_i\} \left(\boldsymbol{A} \odot \boldsymbol{G} \right) \mathrm{diag}\{s_j\} \, \underline{\theta} \right]_i \right), \tag{2.10}$$

where the c_i terms represent ray-dependent factors like detector efficiency or PET attenuation coefficients. The s_j terms represent pixel-dependent factors like spatial variations in sensitivity. The $\{a_{ij}\}$ terms or, equivalently, the matrix $\boldsymbol{A}$ represents factors that are dependent both on the detector and image position (*e.g.*, SPECT attenuation factors). And lastly, the collection of $\{g_{ij}\}$ terms or, equivalently, the matrix $\boldsymbol{G}$ denotes the geometric system model. This factorization is relatively generic and can be used to model many imaging systems.

For a typical PET model, the c_i terms are found by multiplying the attenuation factors specified by the survival probabilities in (2.2) with the detector efficiencies discussed in Section 2.1.1, which are derived from a normalization scan. Thus, the SPECT-type attenuation terms are typically eliminated (*i.e.*, $\boldsymbol{A} = \mathbf{1}$). Similarly, PET is usually modeled without pixel-dependent factors (*i.e.*, $s_j = 1$). The system geometry and detector response is modeled through the specification of the $\{g_{ij}\}$ terms. One simple modeling method is the strip-integral model. This method is illustrated in Figure 2.7a. The detector response is modeled as a rectangular strip and g_{ij} is proportional to the area of the intersection of ith measurement strip and the jth pixel. Thus, this discrete model approximates the continuous model in (2.3), with $b(r, z) = 1/w \operatorname{rect}(r/w)$, a rectangle function with width w.

For a typical SPECT model, the c_i terms represent detector efficiencies, which can be found with a uniformity scan, and $s_j = 1$. The attenuation terms, a_{ij}, are found by calculating the survival probability in (2.1), where the line segment connects the ith detector to the position j. The strip integral model may be appropriate for PET systems, but, as discussed previously, SPECT detectors generally have a depth-dependent response. In this case, the g_{ij} terms are often chosen to specify a depth-dependent response with a Gaussian profile.



(a) Strip Integral Model (b) PET Sensitivity Pattern (c) SPECT Sensitivity Pattern

Figure 2.7: Modeling the detector response in a discrete tomographic system model.

Sample sensitivity patterns for a single detector are shown for the strip integral model and a depth-dependent Gaussian model in Figures 2.7b and c, respectively.

2.2.4 Noise Models

Under ideal circumstances, emission tomography systems count individual photons, and the measurements should be Poisson distributed. However, due to other noise effects or data preprocessing, this is not always the case and other noise models are used. For example, for randoms-corrected PET data the measurements are more accurately represented by a shifted-Poisson model[140, 141]. In other cases one might adopt a Gaussian noise model[30] or other noise model[142].

Similarly, we would like the methods discussed in this paper to apply generally to other imaging systems. Thus, we will derive our main results for general noise models and will defer choosing a particular noise model until specific systems are investigated.

2.3 Space-Invariant versus Space-Variant Systems

In studying the resolution properties of an imaging system, one very important aspect is the intrinsic geometric response of the system. Of particular interest is

how this intrinsic response varies spatially. The intrinsic response is defined as the projection of an impulse followed by backprojection. Adopting the continuous model from Section 2.2.1, the backprojected projection of an image can be written as

$$f_b(x_1, x_2) = \mathcal{P}'\mathcal{P}f(x_1, x_2) = \frac{1}{r} ** f(x_1, x_2), \tag{2.11}$$

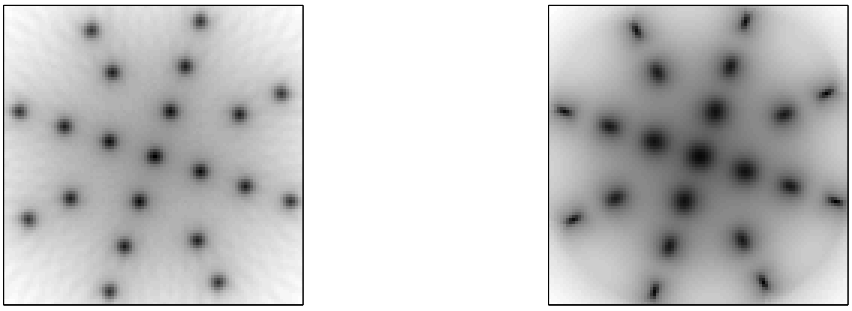
where $r = \sqrt{x_1^2 + x_2^2}$. This is the well-known $1/r$ space-invariant response of the Radon transform[80]. When the radial blur function in (2.3) and (2.5) is not depth-dependent, such that $b(r, z) = b(r)$; one can show that (see Appendix A)

$$f_b(x_1, x_2) = \mathcal{P}'_{\text{blur}}\mathcal{P}_{\text{blur}}f(x_1, x_2) = \frac{1}{r} ** \mathfrak{H}^{-1}\left\{B^2(\rho)\right\} ** f(x_1, x_2), \tag{2.12}$$

where $B(\rho)$ is the 1D Fourier transform of the blur function and $\mathfrak{H}^{-1}\{\cdot\}$ denotes the inverse Hankel transform[17]. Thus, ideal PET systems that can be modeled with a detector blur that is not depth-dependent have an intrinsic response that is space-invariant. When the detector response is depth-dependent, the response is space-variant, and one cannot represent the blur properties in convolution form.

For the discrete model, the geometric response is defined by the action of $\boldsymbol{G}'\boldsymbol{G}$. Therefore, the response is dependent on the exact choice of $\{g_{ij}\}$. Consider the strip integral model where the radial measurements are uniformly spaced. As the strip width and the pixel size are made arbitrarily small, the discrete model (see (2.10)) approaches the continuous model with $b(r) = 1/w \operatorname{rect}(r/w)$, where w is the strip width. Therefore, under the strip integral model, the response is approximately space-invariant except for discretization effects.

In contrast, consider a model that includes a depth-dependent detector response, such as the one shown in Figure 2.7c. Intuitively, one can see that this choice results in space-variant responses. Object points close to the detector will be measured with relatively high resolution and those far away will be measured with lower resolution.



(a) Strip Integral Model (b) Depth-Dependent Gaussian Model

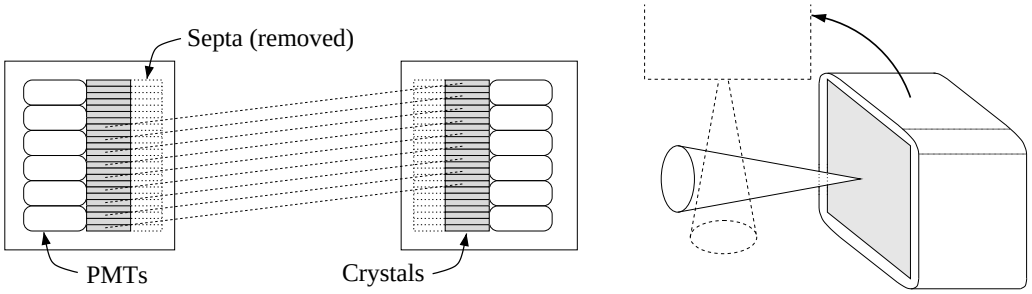
Figure 2.8: Comparison of space-invariant and space-variant geometric responses. (a) A set of space-invariant geometric responses and (b) a set of space-variant geometric responses.

For measurements covering 360° camera rotation, points in the center of the field of view will on average have worst resolution than those at the edges. Additionally, since the resolution changes with detector angle, anisotropic responses are expected.

We can illustrate the geometric response for these two discrete models by considering the operation of $\boldsymbol{G}'\boldsymbol{G}$ on discrete impulse functions. That is, we calculate $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$, where $\underline{e}^j$ is the jth unit vector, for several pixels positions, j. We calculated $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$ for 21 pixel locations using two geometric system models with detector responses that match the sensitivity patterns presented in Figures 2.7b and c. Figure 2.8 shows the superposition of these 21 responses for each model. Not only the does the average resolution differ, the responses for the depth-dependent Gaussian model in Figure 2.8b are anisotropic.

While the geometric response of an idealized PET system is space-invariant, a system response model that includes attenuation effects is rarely space-invariant. Adopting the discrete PET model in Section 2.2.3, the system response is $\boldsymbol{G}\text{diag}\{c_i^2\}\,\boldsymbol{G}$. Since the c_i terms are generally nonuniform, this response is also generally space-invariant. Similarly, incorporating the effects of attenuation and detector efficiencies generally increases the space-variance of SPECT system models.

The above examples (strip integrals and depth-dependent Gaussians) are only two



(a) Cross Section of a 3D PET System (b) 3D Detector Response in SPECT

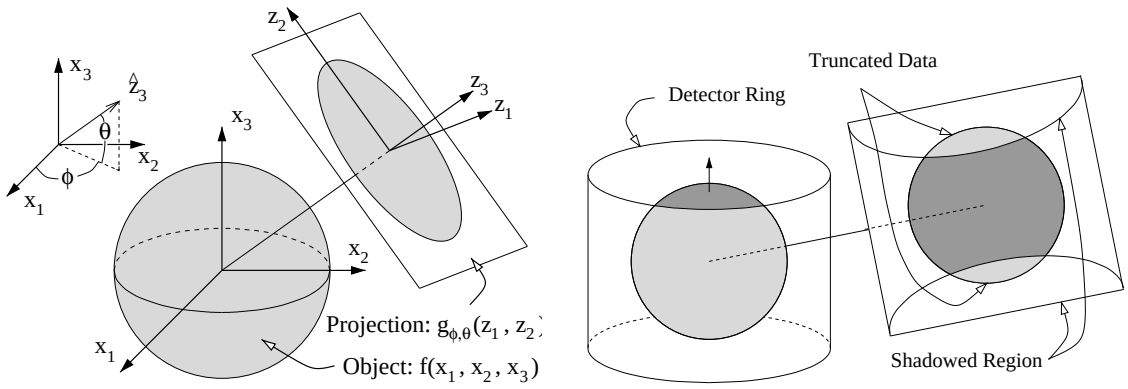
Figure 2.9: Three-dimensional PET and SPECT systems.

of many possible choices for a system model. Other detector effects, like depth of interaction and uneven sample spacing (note the radial sampling at the edges versus the center in Figure 2.1a), may also be incorporated into the system model[61, 102, 128, 62]. In general, such effects also contribute to the space-variance of a real PET system.

2.3.1 2D versus 3D

The systems described so far have all been two-dimensional. However, real PET systems are capable of three-dimensional acquisition and SPECT systems are inherently 3D because of the detector response.

Let us first consider the case of 3D SPECT. If the detector cross section shown in Figure 2.3b represents a collimator with round holes, the ideal detector response is given by a cone shown in Figure 2.9b. The 2D sensitivity map shown in Figure 2.7c is easily generalized to the 3D case, where each element of the system matrix, g_{ij}, represents the ith detector sensitivity at the jth voxel. The problem with this extension is that the system matrix becomes very large. Therefore, instead of storing the matrix, the projection and backprojection operators are often implemented as computer subroutines[26]. For the Gaussian model, it is possible to implement the depth-dependent response using Gaussian diffusion techniques[84, 82] or approximate



(a) 3D System Model with 2D Projections

(b) Truncated Projections in PET

Figure 2.10: Projections in fully three-dimensional systems.

fast methods like those in [46], which use the frequency-distance principle.

The geometric responses for 3D SPECT systems are shift-variant just as the 2D case because of the depth-dependent response. In the 3D case, one can see at the center of the field of view there will be decreased resolution not only in-plane but along the axis. This results in increased blur between slices at the center of the field of view. There is additional shift-variance at the axial extremes of the detector, since the first and last axial slice (and their neighbors) are "sampled" by fewer projections than at the center of the field of view.

PET systems can typically be operated in a 2D or 3D mode. In 2D mode, tungsten septa (annular rings placed in the scanner near the detectors) act as a collimator and isolate the axial planes from each other. In 3D mode, these septa are removed and interslice projections may be obtained. See Figure 2.9a for an illustration.

Figure 2.10a shows an ideal model for a fully 3D system. In this case, the projections are two-dimensional and are dependent on angular indices. For a continuous model, projections are given by

$$g_{\phi,\theta}(z_1, z_2) = \int b(z_1, z_2; z_3) * * f(x_1, x_2, x_3) dz_3, \tag{2.13}$$

with

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -\sin\phi & -\cos\phi\sin\theta & \cos\phi\cos\theta \\ \cos\phi & -\sin\phi\sin\theta & \sin\phi\cos\theta \\ 0 & \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}, \tag{2.14}$$

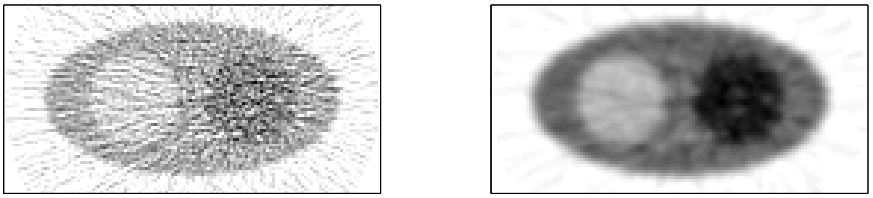
where the blur function is 2D, but possibly depth-dependent and a function of z_3. The complete set of projections (with no blur) over all ϕ and θ are termed the X-ray transform. Again, for discrete models the system matrix may be generalized to the 3D case so that each element of the geometric system matrix, g_{ij}, represents the ith detector sensitivity at the jth voxel.

With an ideal fully sampled projection model (*i.e.*: covering all ϕ and θ), one can show that the geometric response is shift-invariant, as in the ideal 2D case. However, since real PET systems generally have a (finite length) cylindrical geometry, one does not obtain a full range of θ angles. This results in truncated data. See Figure 2.10b. Certain regions within the scanner are effectively "shadowed" and projections are not obtained for some angle pairs. Such truncated data can complicate some reconstruction methods and generally lead to increased space-variance of the geometric system response.

2.4 Nonstatistical Reconstruction Methods

2.4.1 Filtered Backprojection

The classic reconstruction method for tomographic imaging is called filtered backprojection (FBP) and is based on a continuous model like the one discussed in Section 2.2.1. The main idea of FBP is to remove the $1/r$ blur in (2.11) caused by projection and backprojection. Since convolution with $1/r$ is equivalent to multiplication by $1/\rho$ in the frequency-domain, this can be accomplished by applying a 2D cone filter (ρ) to the 2D Fourier transform of the backprojected measurement



(a) Ramp Filtered Backprojection (b) Windowed Filtered Backprojection

Figure 2.11: Filtered backprojection with and without windowing.

data (followed by inverse 2D Fourier transformation). Equivalently, one may apply (in the radial direction) a 1D ramp filter $|\rho|$ to the 1D Fourier transform of the projections followed by inverse 1D Fourier transform, and then backprojecting the filtered projections. In practice, this process uses discretized samples and the latter method is used for computational speed. Additionally, for models that include non-depth-dependent detector blur, one may deconvolve the blur in the projection data, or equivalently modify the cone filter using (2.12).

In real systems the use of ramp or cone filters tends to yield unacceptable results. Any real system has a practical frequency limit imposed by the system geometry or noise. A pure ramp or cone filter will amplify high frequencies more than low frequencies without any cutoff frequency. Therefore, practical implementations require some kind of windowing. Many different windowing techniques may be applied which yield different overall responses, but each has the effect of imposing a cut-off frequency. Example reconstructions of a piecewise constant phantom with and without windowing are shown in Figure 2.11.

Different windowing methods induce different resolution properties in the reconstructed image. For a given windowing function, $W(\rho)$, we write the reconstructed image as

$$\tilde{f}(x_1, x_2) = \mathcal{P}'\mathfrak{F}_1^{-1}\left\{\mathfrak{F}_1\left\{g_\phi(r)\right\}|u|W(u)\right\}, \tag{2.15}$$

where $\mathfrak{F}_1\{\cdot\}$ and $\mathfrak{F}_1^{-1}\{\cdot\}$ denote the 1D Fourier transform and its inverse, and $g_\phi(r)$ denotes the ideal unblurred projections. One can derive the response (see Appendix B) due to this window function to be

$$\tilde{f}(x_1, x_2) = f(x_1, x_2) * * w(r), \tag{2.16}$$

where $w(r)$ is the inverse Hankel transform of the window function. For discrete implementations of FBP, one can use (2.16) to approximate the nearly space-invariant response.

Thus, for FBP we can easily specify the resolution properties of the reconstructed image. As we have seen in Figure 2.11, there is a noise-resolution trade-off that must be made in the reconstruction process. Figure 2.11a has reconstructed with single pixel resolution (*i.e.*: a pure ramp filter means $W(p) = 1$ and therefore $w(r) = \delta(r)$), and clearly there is more noise than in the windowed reconstruction in Figure 2.11b. Depending on the task involved, different reconstruction resolutions may be desired. However, it is relatively easy to find a direct relation between the resolution of the reconstructed image and a parameterized window function. For example, for a Gaussian window $W(\rho) = e^{-\pi(\rho/\rho_0)^2}$, the resulting impulse response is $w(r) = \rho_0^2 e^{-\pi(\rho_0 r)^2}$. The full-width half-maximum (FWHM) resolution of this response is $\frac{2}{\rho_0}\sqrt{\frac{\log 2}{\pi}}$. Therefore, the resolution may be specified prior to reconstruction.

One can apply Fourier methods in the three-dimensional case as well. For the complete data case, where the data covers the entire range of angles, one can apply 3D FBP[20]. Since 3D data contain redundancies (*i.e.*: acquisition of data with $\theta = 0$ only, is sufficient for reconstruction), there is no unique reconstruction filter unlike the 2D case. However, valid reconstruction filters must satisfy certain conditions[25]. This technique is attractive, since it provide shift-invariant resolution properties for a completely sampled (ideal) system.

2.4.2 Extensions for Real Systems

Unfortunately, real PET and SPECT systems rarely provide the ideal unblurred and appropriately sampled projections. Recall that PET systems have detectors in a cylindrical pattern, and FBP techniques generally assume parallel projections with a linear sampling. Thus, an arc-correction procedure to resample the data evenly is generally required. (There have also been some extensions to FBP that accommodate PET-type sampling[12].) Also recall that real 3D PET data are rarely completely sampled, which is a requirement for 3D FBP. Techniques have been developed that fill in missing data using a preliminary reconstruction and reprojection, and then reconstruct using 3D FBP. This method is known as 3D reprojection (3DRP)[100, 63].

Another method involves rebinning the 3D data into 2D projections followed by 2D FBP. There are a number of methods based on this idea: namely, single slice rebinning (SSRB)[21], multi-slice rebinning (MSRB)[77], and Fourier rebinning (FORE)[24]. These methods are generally faster than the above techniques since only 2D reconstructions are performed. However, methods such as SSRB suffer from significant off-axis geometric distortion from the rebinning procedure.

Even if appropriately sampled projections can be obtained or estimated, one must still correct for other physical effects. Additive factors like background radiation, randoms, and scatter can be subtracted from the projections. In PET, attenuation correction (and detector efficiencies) can be corrected by simply multiplying the projections by the c_i terms in (2.10). However, the same correction is not possible for SPECT attenuation.

In SPECT, attenuation is both a ray and pixel-dependent factor. Additionally, one cannot simply deconvolve the depth-dependent detector blur because of the inherent depth-dependence. A number of methods have been developed to deal with these

issues. Approximate attenuation correction factors may be applied to reconstructions of uncompensated projections[19]; however, these factors cannot completely compensate for nonuniform attenuation.

There has been much work on so-called "exact" methods, which replace FBP with a reconstruction method that is inherently suited to a more realistic projection transformation. For example, reconstruction methods have been derived for the *attenuated Radon transform*, where either uniform attenuation[9, 83] or nonuniform attenuation[64, 45] is incorporated directly into the projection transformation. Unfortunately these methods do not incorporate the depth-dependent blur of SPECT detectors.

However, approximate methods based on the frequency-distance principle can be used to partially compensate for the SPECT detector blur with an appropriate projection-domain filtering operation[76, 135, 138, 93, 47]. Similarly, van Elmbt and Walrand have developed an approximate technique which compensates for both Gaussian detector response and uniform attenuation[126]. Appledorn[5] derived an "exact" method for cases when the depth-dependent blur is a Cauchy function. Others have expanded this derivation to accommodate uniform attenuation[107].

Unfortunately, to date an "exact" method has not been derived that can incorporate both nonuniform attenuation and a realistic SPECT detector response. Additionally, these "exact" methods generally do not incorporate any noise model in the reconstruction.

2.4.3 Nonstatistical Iterative Methods

Another alternative is to adopt the discrete measurement model like the one discussed in Section 2.2.3. If the transformation $\tau(\cdot)$ is linear, then (2.9) is a linear system of equations that relates the object parameters to the measurements. Given

measurements, one can solve these equations to reconstruct the object. Such an estimator is identified as the algebraic reconstruction technique (ART) and generally requires an iterative method to find the solution because of the number of object parameters and measurements[49, 52, 48].

2.5 Statistical Image Reconstruction

One problem with all the methods discussed so far is that the noise model is not taken into account. It is also difficult for many of the methods to completely model all of the physical aspects of the system. For these reasons, one often uses a statistical reconstruction technique. Using the discrete reconstruction model from Section 2.2.3 and choosing an appropriate noise model, one can construct a maximum-likelihood estimator (MLE) for image reconstruction. Mathematically,

$$\underline{\hat{\theta}}_{ML}(\underline{Y}) = \arg\max_{\underline{\theta}\in\underline{\Theta}} l(\underline{\theta}, \underline{Y}), = \arg\max_{\underline{\theta}\in\underline{\Theta}} L(\underline{\theta}, \underline{Y}), , \tag{2.17}$$

where $\underline{\Theta}$ denotes the set of feasible images, $\underline{Y}$ is a single realization of the random vector $\underline{Y}$, and $l(\underline{\theta}, \underline{Y})$ represents the likelihood function. Equivalently, we may maximize the log-likelihood function $L(\underline{\theta}, \underline{Y})$. Under the assumption of independent measurements, we may write the log-likelihood as a sum of marginal log-likelihoods,

$$L(\underline{\theta}, \underline{Y}) = \sum_{i}^{N} L_i(Y_i, \bar{Y}_i(\underline{\theta})), \tag{2.18}$$

where the marginal log-likelihood, $L_i(\cdot, \cdot)$, is a two-dimensional function of the ith measurement, Y_i, and its mean, $\bar{Y}_i$. The system model enters (2.18) through the model for the mean measurements in (2.9).

Many noise models can fit into the framework provided by (2.18), including those discussed in Section 2.2.4. When the measurements are well modeled with Poisson

noise, the marginal log-likelihoods are

$$L_i^{\text{Poisson}} = Y_i \log \bar{Y}_i(\underline{\theta}) - \bar{Y}_i(\underline{\theta}) - \log Y_i!, \tag{2.19}$$

which may be plugged into (2.17) and (2.18). Unfortunately, there is no closed form for the maximizer of the objective in (2.17) and one typically uses an iterative algorithm to find the solution. There has been a good deal of work on iterative algorithms. Section 2.5.2 discusses many of the algorithms utilized for our work.

When the Gaussian noise model is adopted and the mean measurements are a linear function of the object (*i.e.*, $\tau_i(l) = l + r_i$), the maximum-likelihood estimator does have a closed form. The log-likelihood for the Gaussian model is

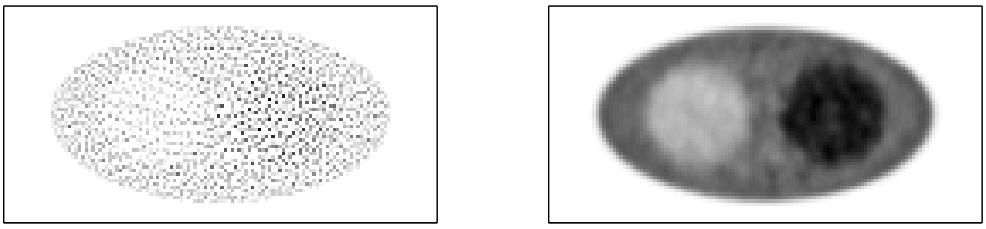
$$\begin{aligned} L^{\text{Gaussian}}(\underline{\theta}, \underline{Y}) &= \sum_{i=1}^{N} -\frac{1}{2}\frac{[Y_i - \bar{Y}_i(\underline{\theta})]^2}{\sigma_i} - \log \sigma_i \sqrt{2\pi} & (2.20) \\ &= -\frac{1}{2}[\underline{Y} - \underline{\bar{Y}}(\underline{\theta})]' \boldsymbol{\Sigma}^{-1} [\underline{Y} - \underline{\bar{Y}}(\underline{\theta})] - \sum_{i=1}^{N} \log \sigma_i \sqrt{2\pi}, & (2.21) \end{aligned}$$

where $\boldsymbol{\Sigma} = \text{diag}\{\sigma_i\}$ and σ_i represents the variance of measurement i. Therefore, dropping constant terms, we can write the estimator as

$$\begin{aligned} \hat{\underline{\theta}}_{ML}^{\text{Gaussian}}(\underline{Y}) &= \arg\min_{\underline{\theta}} [\underline{Y} - \boldsymbol{H}\underline{\theta} - \underline{r}]' \boldsymbol{\Sigma}^{-1} [\underline{Y} - \boldsymbol{H}\underline{\theta} - \underline{r}] & (2.22) \\ &= [\boldsymbol{H}' \boldsymbol{\Sigma}^{-1} \boldsymbol{H}]^{-1} \boldsymbol{H}' \boldsymbol{\Sigma}^{-1} (\underline{Y} - \underline{r}), & (2.23) \end{aligned}$$

where the vector $\underline{r}$ represents additive terms in the linear model. This is the well-known weighted least-squares estimator[113]. While the Gaussian noise model leads to a closed-form estimator, in practice (2.23) is still solved iteratively since the system matrix is typically quite large and the above estimator involves a matrix inverse.

Unfortunately, since the image reconstruction problem is ill-conditioned and there is noise in real systems, MLEs tend to produce overly noisy images, much in the same way exact FBP reconstruction with a pure ramp filter produces poor images. An



(a) Maximum Likelihood Reconstruction (b) Regularized Likelihood Reconstruction

Figure 2.12: Statistical reconstruction with and without regularization.
Reconstruction of measurements obtained for a piecewise constant image are plagued by noise in a pure maximum-likelihood reconstruction (left), but the noise can be greatly reduced by regularizing the reconstruction problem (right).

example image from a MLE is shown in Figure 2.12a. There are a number techniques that are used to improve image quality. Many of these methods fall under the general topic of regularization. One example of a regularized reconstruction is shown in Figure 2.12b. Note the obvious decrease in noise.

2.5.1 Noise Reduction Techniques and Regularization

As demonstrated above, the image reconstruction problem requires some form of noise reduction; otherwise, overly noisy images result. There are a wide range of possibilities when it comes to reducing noise. Each method comes with certain advantages and disadvantages. The following is a brief discussion of several popular choices.

Truncated Iterations

One of the simplest forms of noise reduction is to stop the iterative algorithm before convergence. If one initializes an iterative algorithm with a uniform image, the image starts very smooth and high-frequency components tend to increase with iteration. Visually, the images start smooth and become noisy with increasing iteration. Therefore, stopping the algorithm before convergence acts as a kind of regularization. Much work has been done on this type of noise reduction, including selection

of stopping criterion[127, 50]. This method is implemented relatively easily. However, in terms of resolution control and flexibility this method is very limited. The resolution properties depend on the iterative algorithm that is used and how far it is iterated, as well as the imaging system and object that is being imaged.

Post-Smoothed Maximum Likelihood

Another simple alternative is to simply post-smooth a maximum-likelihood image that has been iterated until convergence[105]. Since maximum-likelihood attempts to find a perfect reconstruction with a delta impulse response, the post-filtering operation can be customized to any desired resolution properties. Another advantage is that a variety of filters may be applied and only a single iterative solution needs to be found. However, there are also disadvantages. Unregularized MLEs are ill-conditioned and tend to take many iterations to converge to a solution. Therefore, reconstructions tend to take longer.

Sieves

Yet another technique that is used is the method of sieves[105, 104]. The main idea of this method is to constrain the estimate, $\underline{\hat{\theta}}$, to a subset of feasible images (called a sieve) that are smooth. Typically, this is accomplished by defining a kernel sieve through which the emission image is related to a "pre-emission image." The iterative maximization algorithm estimates this pre-image, then the kernel is applied to yield the emission image estimate. Resolution properties of the resulting image estimate are controlled easily through the selection of the kernel sieve. One disadvantage of this method is the large number of iterations required for convergence[85, 78]. By adding the kernel sieve to the estimator, the estimator must essentially perform additional deconvolution of the measurement data, making an ill-conditioned prob-

lem even more ill-conditioned, thus increasing the number of iterations. This additional deconvolution is removed by the final application of the kernel sieve. There are other disadvantages as well. One cannot find appropriate kernels for arbitrary combinations of desired resolution properties and system models. Additionally, a space-invariant sieve cannot provide uniform resolution properties for space-variant systems.

Blob-Type Image Discretization

A similar approach is to adopt smooth basis functions for the image discretization like the so-called blob bases[81]. If implemented poorly, this approach can be computationally expensive since the blobs overlap. Much like the sieve approach, this modification can increase the ill-conditioning of the reconstruction problem. Thus, in general, more iterations are required with blobs than voxels.

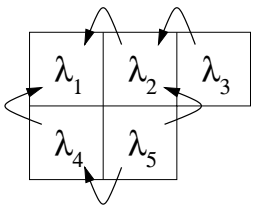
Penalized-Likelihood Estimation

The penalized-likelihood estimator (PLE) is another popular technique that yields images with decreased noise. In a Bayesian framework, these estimators are also known as maximum *a posteriori* (MAP) estimators[87, 69]. These techniques involve changing the maximum-likelihood objective function to a different objective where *a priori* information about the object is included. Thus, one forms a prior distribution for the object and then finds a MAP solution. In a penalty framework, one includes a roughness penalty in the objective function. Mathematically, the two frameworks often produce similar forms, and the differences are largely semantic.

We concentrate on PLEs which can be written in the following form:

$$\hat{\underline{\theta}}_{PL}(\underline{Y}) = \arg\max_{\underline{\theta}\in\underline{\Theta}} \Phi(\underline{\theta}, \underline{Y}) = \arg\max_{\underline{\theta}\in\underline{\Theta}} L(\underline{\theta}, \underline{Y}) - R(\underline{\theta}). \tag{2.24}$$

The penalty term $R(\underline{\theta})$ should yield large values for undesirable images (*i.e.*: noisy



$$\boldsymbol{C}\underline{\lambda} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} \lambda_2 - \lambda_1 \\ \lambda_3 - \lambda_2 \\ \lambda_5 - \lambda_4 \\ \lambda_4 - \lambda_1 \\ \lambda_5 - \lambda_2 \end{bmatrix}$$

$$R(\underline{\lambda}) = \psi(\lambda_2 - \lambda_1) + \psi(\lambda_3 - \lambda_2) + \psi(\lambda_5 - \lambda_4) + \psi(\lambda_4 - \lambda_1) + \psi(\lambda_5 - \lambda_2)$$

Figure 2.13: Illustration of the construction of a roughness penalty.
Illustration of the construction of $\boldsymbol{C}$ and the resulting roughness penalty $R(\underline{\lambda})$. The picture on the left represents a small image with arrows showing pairs of penalized pixels.

images). Therefore, the objective function, $\Phi(\underline{\theta}, \underline{Y})$, specifies the image estimate should fit the data model (the likelihood term), but should be balanced by avoiding noisy images.

The penalty term can take a wide variety of forms. One class of penalty functions is given by

$$R(\underline{\theta}) = \beta \sum_k \psi([\boldsymbol{C}\underline{\theta}]_k), \quad \text{where} \quad [\boldsymbol{C}\underline{\theta}]_k = \sum_{j=1}^{p} c_{kj}\theta_j. \tag{2.25}$$

This model is fairly general and includes most popular penalties. (Exceptions include line-site models [103, 60] and the median root prior [4].) The β term is called the regularization parameter (or sometimes the hyperparameter) and controls the noise-resolution trade-off. A large β puts a heavier weight on the roughness penalty and thus yields smoother reconstructed images. A smaller β yields higher resolution but noisier images. In the case where $\beta = 0$, we have a maximum-likelihood estimator.

Equation (2.25) allows for functions, $\psi(t)$, of linear combinations of pixels, $\boldsymbol{C}\underline{\theta}$. These combinations of pixels are defined by the elements of $\boldsymbol{C}$. For many practical penalties, a given pixel will be combined with only a few of its neighbors. This pixel set is called a neighborhood. Figure 2.13 shows an example of how the matrix $\boldsymbol{C}$ can specify the penalty. In this example, only the horizontal and vertical neighbors are included in the penalty. For the case where $\boldsymbol{C}$ is an identity matrix and a quadratic

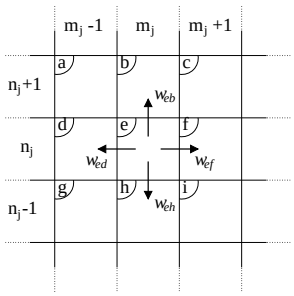


Figure 2.14: A pairwise roughness penalty with first-order neighborhood on a small image area.

penalty is adopted (*i.e.*: $\boldsymbol{C} = \boldsymbol{I}$ and $\psi(t) = \frac{1}{2}t^2$), one has the classic Tikhonov-Miller regularization [123].

If one chooses to penalize only the differences between pairs of pixels, we may write the penalty as

$$R(\underline{\theta}) = \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p} w_{jk}\psi(\theta_j - \theta_k), \tag{2.26}$$

where the w_{jk} terms are the interpixel weighting strengths between pixel j and pixel k, and $w_{jk} = w_{kj}$. The w_{jk} terms absorb the β term mentioned earlier. Weights between pixels not in each other's neighborhoods are set to zero.

Typically, penalties are chosen that are nonnegative definite functions. This assures that (2.24) has a unique solution when the likelihood term is convex. For the penalty in (2.26) it is common to adopt the sufficient condition where interpixel weightings are constrained to be nonnegative.

Consider the roughness penalty shown in Figure 2.14. This penalty uses only a pixel's vertical and horizontal neighbors, known as a first-order penalty (a second-order penalty includes the diagonal neighbors). Looking at pixel e, for terms of the form $w_{e\star}$, only w_{eb}, w_{ed}, w_{ef}, and w_{eh} have nonzero values. If the four interpixel weights are identical regardless of the pixel location, the $R(\underline{\theta})$ is called a space-invariant penalty. If the weights are space-invariant and are identical for all pixel

pairs in a neighborhood, the penalty is called a uniform penalty.[3] For example, a conventional uniform first-order penalty is to choose the w_{jk} terms equal to β for the horizontal and vertical neighbors and zero otherwise. A uniform second-order penalty adds $w_{jk} = \beta/\sqrt{2}$, for the diagonal neighbors (where the $\sqrt{2}$ is a distance scaling).

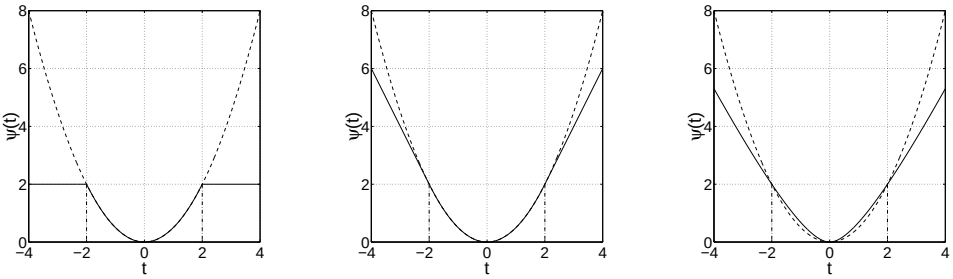
The selection of the $\psi(t)$ functions has not yet been discussed. These functions are very important in determining the resolution properties of the reconstructed image. It is natural to choose functions that are symmetric about $t = 0$. One of the simplest choices is to use a quadratic penalty. The pairwise quadratic penalty has the advantage of having the simple matrix form:

$$R(\underline{\theta}) = \frac{1}{2}\underline{\theta}'\boldsymbol{R}\underline{\theta} \qquad \text{where} \qquad \boldsymbol{R}_{jk} = \begin{cases} \sum_{l=1}^{p} \frac{1}{2}(w_{lj} + w_{jl}), & k = j \\ -w_{jk}, & k \neq j. \end{cases} \tag{2.27}$$

For the conventional uniform quadratic penalty (*i.e.*: the w_{jk} terms equal to β), we may write the penalty matrix, $\boldsymbol{R}$, as a simple scaling, $\boldsymbol{R} = \beta\boldsymbol{R}_0$. We refer to this $\boldsymbol{R}$ as a scaled penalty matrix. The matrix $\boldsymbol{R}_0$ specifies unit interpixel weightings and a particular neighborhood size.

One property of quadratic penalties is that increasing pixel differences are penalized with increasing weight. Because there are large pixel differences even in noiseless images (*e.g.*: edges), the quadratic choice discourages edges in the reconstructed image. This is often interpreted as oversmoothing. Because of this effect, many nonquadratic penalties have also been proposed which have edge-preserving effects. In the case of a truncated quadratic penalty [10], the penalty becomes constant for differences greater than some value. Unfortunately, this yields a nonconvex objective that requires more complicated maximization algorithms[90]. Additionally, noncon-

[3]Note that a uniform penalty is uniform only in the sense that it penalizes uniformly. A uniform penalty, as we shall see in Chapter III, does *not* imply uniform resolution properties.



(a) Truncated Quadratic (b) Huber Penalty (c) Lange Penalty

Figure 2.15: Examples of edge-preserving penalty functions.
The quadratic penalty is shown with a dashed line for comparison. The penalty functions are scaled, or parameters are chosen so that they all pass through (0,0) and (2,2).

vex functions can add a data-dependence, where small changes in measurements can yield large changes in the image[14]. Such data sensitivity may be undesirable, particularly in medical imaging applications. Therefore, many practical penalties become linear (or more linear) after some point, penalizing less steeply than the quadratic function, thus preserving edges, while maintaining a convex objective. Two such penalties are the penalties of Huber [57] and Lange [69]:

$$\psi_{\text{Huber}}(t) \triangleq \begin{cases} t^2/2, & |t| \leq \delta \\ \delta|t| - \delta^2/2, & |t| > \delta \end{cases} \qquad \psi_{\text{Lange}}(t) \triangleq |t| - \ln(1 + |t|). \tag{2.28}$$

Plots of these penalties are shown in Figure 2.15.

Penalized-likelihood methods have advantages over other regularization techniques. These methods improve the conditioning of the reconstruction problem and tend to increase the convergence rates of iterative algorithms. PLEs converge faster than MLEs and estimators using sieves. Penalty functions also allow for a wide range of resolution control, including edge-preserving penalties[69, 51], the inclusion of anatomical information through modification of the interpixel weightings[73, 36, 16], and space-variant regularizations for other goals, such as contrast optimization[95].

There are some possible disadvantages with penalized-likelihood methods, most

notably the nonintuitive relation between the penalty function and the resolution properties of the reconstructed image. However, such limitations may only be problems with conventional penalty functions. In this work we choose to use specially designed penalty functions to correct for space-variant resolution properties inherent in real imaging systems and statistical estimators. This topic is the central focus of the research presented in this thesis and will be discussed in-depth beginning in Chapter III.

2.5.2 Algorithms

There are a wide variety of algorithms for performing the objective function maximization in (2.24). In general, these algorithms depend on the exact form of the objective. Perhaps most important is what form of the likelihood portion of the objective is chosen. Under the Gaussian model, the estimates have a closed-form solution and iterative techniques are used because of size of system matrix and the infeasibility of matrix inversion. Under the Poisson model, there is no closed form and thus an iterative technique is required.

For reconstruction methods that wish to maximize (or minimize) an objective function,[4] the algorithm itself does not effect the solution. Only the speed at which the solution is found is controlled by the algorithm. This assumes, of course, that the algorithm eventually reaches the objective's minimum. Certain algorithms such as ordered-subsets [58] are not guaranteed to converge. (Although, new modified algorithms have been developed that do converge[3].) Ordered subsets algorithms are often used due to their increased "convergence" rates and can always be followed by an application of a convergent algorithm if convergence is an issue. In practice, we have used coordinate ascent[15], conjugate gradient methods[37], and paraboloidal

[4]This eliminates techniques that use stopping criterion to impose image smoothness.

surrogate routines[109, 28, 27] for the penalized weighted least-squares objective. For the penalized-likelihood (Poisson) objective we have used the SAGE algorithm[38], ordered-subsets versions of De Pierro's algorithms[23, 22], and paraboloidal surrogate routines[29, 27, 28]. We have used expectation-maximization[70] (EM) and ordered-subsets EM (OSEM) for unpenalized maximum-likelihood solutions.

2.6 Summary of Methods with Practical Resolution Control

Different estimators offer varying degrees of resolution control. With some estimators, one can easily specify the exact blur function that defines the global resolution properties of the reconstructed image; in others the resolution control is more oblique and only an average global resolution can be specified. Table 2.1 summarizes the degree of resolution control for various methods. Specifically, the table identifies qualitatively how well (space-invariant) resolution properties can be controlled on an ideal PET system with an inherently space-invariant (SI) response, and on a space-variant (SV) SPECT system. We assume the true attenuation maps, scatter, randoms, etc., are known and available to the estimators.

In addition to the methods discussed in Sections 2.4 and 2.5, we present a few more reconstruction methods. These are penalized unweighted least-squares (PULS), FBP with post-reconstruction filtering, and statistical sinogram deblurring. PULS is equivalent to a penalized-likelihood estimator under the assumption of white Gaussian measurement noise. Post-reconstruction Kalman filtering has been applied to SPECT FBP reconstruction in [13], in an attempt to correct for the space-variant blur. Lastly, statistical sinogram deblurring has been demonstrated in [65] that can provide nearly uniform resolution properties.

Moving from the top to the bottom of the list in Table 2.1, we discuss the rel-

Table 2.1: Relative controllability of resolution for different estimators.

				Resolution Control	
	Reconstruction Method	Relative Speed	Noise Model	SI PET	SV SPECT
Nonstatistical	Plain FBP	Fast	None	Excellent	Poor
	3DRP	Fast	None	Good	N/A
	3D-2D Rebinning	Fast	None	Fair	N/A
	FDP-Corrected FBP	Fast	None	N/A	Good
	Appledorn + ext. FBP	Fast	None	N/A	Good
	"Exact" Methods	Fast-Moderate	None	Very Good	Fair/Good
	Post-Filtered ART	Slow	None	Excellent	Excellent
Statistical	PULS	Slow	Incorrect	Good	Poor
	Post-Recon. Filtering	Fast-Moderate	Incorrect	N/A	Fair
	Sinogram Deblurring	Moderate	Yes	Good	Unknown
	Post-Filtered ML	Very Slow	Yes	Excellent	Excellent
	Inter-Update Filtering	Slow	Yes	Fair	Poor
	Sieves	Very Slow	Yes	Very Good	N/A
	ML Blob Bases	Very Slow	Yes	Excellent	Excellent
	Truncated OSEM	Moderate	Yes	Poor	Poor
	Conventional PL	Slow	Yes	Fair	Poor

ative controllability of the estimators. First, we discuss nonstatistical estimators. Ordinary FBP is well suited to the SI problem (2D or untruncated 3D) and the blur function can be specified exactly. However, for the SV problem, FBP is inherently mismatched with the system and will yield very nonuniform resolution. The 3DRP routine can probably be implemented so that the resolution properties are anisotropic; however, this will require careful design of both the initial 2D reconstruction filter and the final 3D reconstruction filter. 3D-to-2D rebinning techniques will generally introduce nonuniform axial distortion for the PET problem. For SPECT, the frequency-distance principle (FDP)-based corrections and Appledorn-type corrections can compensate for the space-variant response. However, this compensation is not complete due to the inherent model mismatches or approximations made by these methods. Similarly, an "exact" method that completely models SPECT detector response and nonuniform attenuation has not yet been developed. Thus, resolution control will generally not be complete. For PET, "exact" methods like Chebyshev-domain FBP[12] should be very well matched to the system and reso-

lution properties are well specified by the filtering operations that are applied to control noise. The resolution properties of ART techniques are highly dependent on the exact noise reduction method that is adopted. For unregularized ART methods applied to fully determined systems, fully converged ART solutions should have "perfect" single pixel resolutions. Thus, if noise control is applied via post-filtering, the resolution properties are fully specified by the post-filter.

The following two estimators are statistical, but are based on incorrect noise models. A PULS estimator is a linear estimator, but is generally not space-invariant due to attenuation effects and the system geometry. Similarly, resolution is often controlled with a regularization parameter that is only obliquely related to the resolution properties. However, if the tomographic data comes from an intrinsically shift-invariant system and is first precorrected for attenuation and other ray-dependent effects, PULS can be used to provide uniform resolution, with the FWHM resolution controlled through the regularization parameter. Post-reconstruction filtering has been used in [13] for correcting for the space-variant response in FBP reconstructed SPECT images. Unfortunately, as implemented in [13], the space-variant blurs are measured using point sources that ignore attenuation effects, and the noise model of the FBP reconstructed images is unknown and must be estimated using small neighborhoods of pixels. Thus, the post-filtering generally cannot completely compensate for the space-variant blurs.

Lastly, we identify the resolution control in statistical estimators. Statistical sinogram deblurring has been applied by [65] for PET-type geometries, followed by FBP reconstruction. The resulting resolution properties are highly uniform and good control is provided by the FBP filtering. Post-filtered ML provides excellent resolution control for fully determined systems, since the resolution properties are fully spec-

ified by the post-filter. Sometimes this filtering is applied between iterations and is referred to as inter-update filtering. Because resolution properties are iteration-dependent and a function of the data, these methods generally provide only coarse resolution control. Sieves can yield very good resolution control when the desired sieve kernel exists. This is possible for many desired resolutions and space-invariant PET geometries. However, space-variant kernels for SPECT systems have not yet been developed. Alternative discretizations like the "blobs" of [81], can provide very good resolution control, assuming the system model is fully determined and the solution is fully converged. Most versions of maximum-likelihood reconstruction that use truncated iterations, including truncated OSEM, will yield space-variant resolution properties that are object-dependent, iteration-dependent, and system-dependent. Thus, resolution control is generally poor. Lastly, penalized-likelihood estimators with conventional space-invariant penalties yield images with space-variant resolution properties[32, 41]. Rough control of resolution is specified through the regularization parameter.

Having reviewed many reconstruction techniques, while there are many estimators that provide good resolution control, there appears to be no single estimator that allows good resolution control, incorporates a noise model, and provides fast estimates. Among the statistical reconstruction techniques, we have chosen to investigate the penalized-likelihood approach in detail to see if good resolution control can be incorporated into the estimator.

In the following sections we investigate in detail why penalized-likelihood estimators have space-variant resolution properties, what kind of resolution properties they yield, and ultimately how to control the resolution properties of such estimators.

CHAPTER III

Quantifying Resolution

To investigate or control resolution properties of an estimator, one needs effective methods for quantifying resolution. In this chapter we discuss how to measure resolution, including a new derivation of the local impulse response for penalized-likelihood estimators using the continuous-discrete measurement model of Section 2.2.2 and the discrete reconstruction model of Section 2.2.3. We use the local impulse response and other techniques to investigate resolution properties of conventional penalized-likelihood estimators.

3.1 Prior Work in Quantifying Resolution

One of the simplest techniques for investigating the resolution properties of an imaging system is to propagate known signals through the system. For example, images like radial "spoke" test patterns are often used to identify what frequencies are passed through optical imaging systems. In nuclear imaging systems, test phantoms can be prepared with cylinders of varying sizes. As smaller cylinders approach the resolution limit, the contrast of those rods decreases.

Imaging of test patterns and phantoms is an important aspect of resolution investigation. Since the images themselves are often the point of interest, these investigations directly show which features can be resolved. Unfortunately, as we will show in

following sections of this chapter, the resolution properties of an image are not only system- and estimator-dependent, they are also object-dependent. Since the test patterns or phantoms are often very different from the typical objects that are being imaged, sample reconstructions often cannot offer a great deal of predictive value. Similarly, because resolution properties are typically shift-variant, it is difficult to fully investigate the shift-variant properties with a single phantom.

In systems with shift-invariant resolution properties, one can fully represent the resolution properties of a system with a shift-invariant convolutional filter. This filter is called the impulse response function, since it represents how an impulse function would be imaged by the shift-invariant system.

Many investigators have extended this idea of an impulse response by looking at reconstructions with impulses added to an object of interest[112, 133]. These responses depend on the location of the impulse and are referred to as local point responses, or local impulse responses. Strictly speaking, for nonlinear estimators, the response is also dependent on the magnitude of the added impulse; however, for locally linear estimators the responses are relatively insensitive to this scaling. Wilson focused on the resolution properties of EM as a function of iteration and found that the resolution improves as a function of iteration, but is generally shift-variant. Such impulse addition methods are powerful, since they take the estimator-, system-, and object-dependence into account. However, it would be advantageous to have analytic forms so that resolution properties may be predicted without performing reconstructions.

There has been much work analyzing the image properties as a function of iteration. Barrett *et al.* demonstrated methods showing how the variance or covariance may be estimated by identifying how noise is propagated through EM iterations[8].

These methods have been extended to identify the mean and variance properties of MAP[129] and OSEM[106] estimates as a function of iteration.

Rather than focusing on the properties of an estimator as a function of iteration, another approach is to concentrate on the properties of solution that maximizes the objective function. When the local impulse response is defined in terms of the mean reconstruction, it is possible to develop unbiased estimators to find the resolution properties[54, 125, 124]. However, these methods generally involve performing many reconstructions. Analytic forms based on a locally linear approximation for the mean and variance[33] and local resolution properties[32, 41] have been derived for penalized-likelihood estimators. In Section 3.3, we extend the methods of [32, 41] to the continuous-discrete measurement model of Section 2.2.2 and the generic penalized-likelihood estimator discussed in Section 2.5.1.

3.2 Resolution Investigation via Phantom Reconstruction

Before we discuss analytic forms for the local impulse response, we would like to provide further motivation for the investigation of resolution properties. In this section we show sample reconstructions in which the nonuniform resolution properties are readily apparent. Specifically, we demonstrate the emission tomography reconstructions from noiseless projections. Without noise, we can show the resolution properties of the reconstructed image by comparing the reconstructed image and the original true emission distribution. We demonstrate that nonuniform resolution properties can arise even for intrinsically shift-invariant systems and simple noise models.

3.2.1 Nonuniform Resolution in Ideal ECT

Consider the case of a penalized-likelihood estimator with a linear measurement model and the Gaussian log-likelihood given in (2.21). Adopting the quadratic penalty of (2.27) with a conventional shift-invariant first-order scaled penalty matrix, we may write this estimator as

$$\hat{\underline{\theta}}_{PWLS} = [\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H} + \beta\boldsymbol{R}_0]^{-1}\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}(\underline{Y} - \underline{r}), \tag{3.1}$$

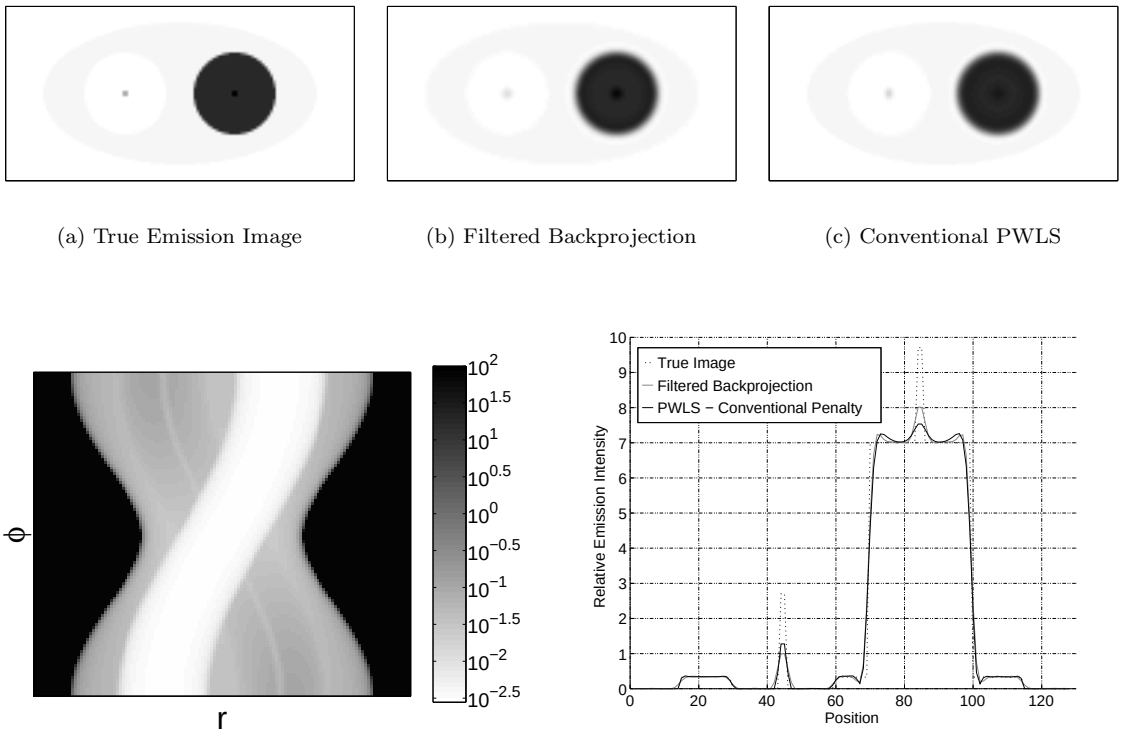
where $\boldsymbol{\Sigma} = \mathrm{Cov}(\underline{Y})$, the covariance of the measurements. As with (2.23), the estimator is linear and may be evaluated using standard iterative approaches. For emission tomography, the Gaussian model is sometimes adopted for precorrected data that no longer obey a Poisson model. Typically, the variance of the (independent) measurement noise is assumed to be the same as the mean. Thus, $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}\{1/\bar{Y}_i\}$. In practice, one often uses $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}\{1/\max\{\bar{Y}_i, t_c\}\}$ for some small positive value t_c to prevent inordinate weighting for measurements with means near zero.

Recall the system matrix factorization presented in (2.10). Using this factorization with $\boldsymbol{H} = \mathrm{diag}\{c_i\}\,(\boldsymbol{A} \odot \boldsymbol{G})\mathrm{diag}\{s_j\}$, consider the case of an idealized PET system where $\boldsymbol{A} = \boldsymbol{1}$, $s_j = 1$, and the geometric response, $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$, is shift-invariant (except for small discretization effects). For this PET system, we may write the estimator as

$$\hat{\underline{\theta}}_{PWLS} = [\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G} + \beta\boldsymbol{R}_0]^{-1}\boldsymbol{G}'\boldsymbol{W}(\underline{Y} - \underline{r}), \tag{3.2}$$

with $\boldsymbol{W} = \mathrm{diag}\{c_i/\bar{Y}_i\}$.

For the first sample reconstruction, we will further simplify the PET system by eliminating attenuation ($c_i = 1$) and any additive terms ($r_i = 0$). Consider the emission distribution shown in Figure 3.1a. The object has a hot circular region on the right, a cold circular region on the left, a cool background ellipse. Additionally,



(a) True Emission Image (b) Filtered Backprojection (c) Conventional PWLS

(d) Weights: $1/\bar{Y}_i$ (e) Image Profiles

Figure 3.1: Nonuniformities in ideal ECT phantom reconstruction.
Nonuniform resolution properties arise even in the case of an idealized shift-invariant ECT system model without attenuation, detector efficiencies, and randoms.

a small hot spot is placed in each circular region. From this emission distribution, we will obtain noiseless projections, $\underline{\bar{Y}}$, simulated with 1 million total mean counts using a discrete measurement model that matches the reconstruction model. Reconstructions are performed on those noiseless projections.

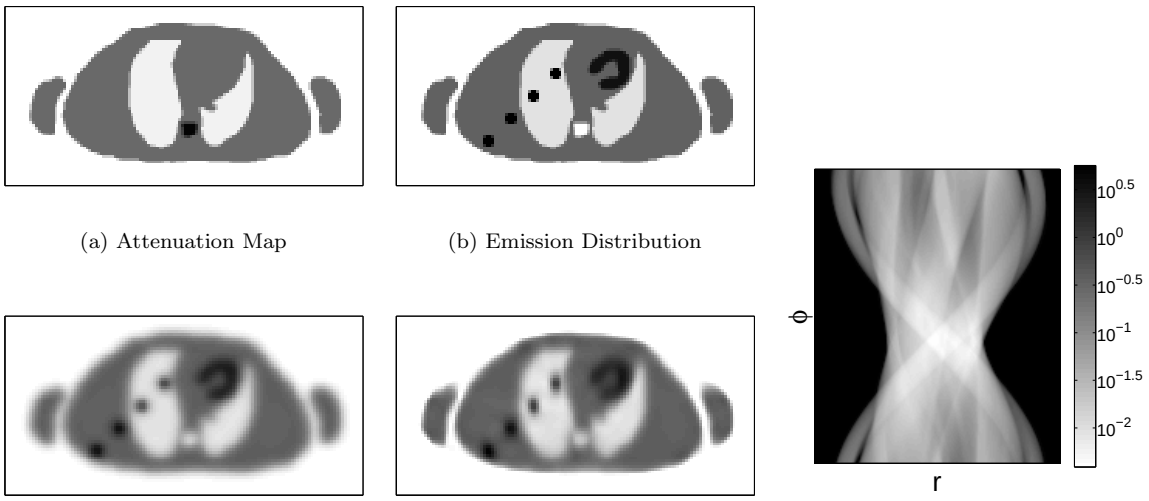
A filtered-backprojection reconstruction is shown in Figure 3.1b. Since the system is shift-invariant, FBP yields shift-invariant resolution properties given by (2.16). All regions of the image are smoothed identically. A PWLS reconstruction is shown in Figure 3.1c. The elements of the diagonal weighting matrix, $\boldsymbol{W}$, for this PWLS estimator are represented in image form in Figure 3.1d. Comparing the FBP and PWLS reconstructions, we see a decreased contrast in the small hot spot (in the

hot region) in the PWLS reconstruction. This is an indication of the nonuniform resolution properties even in this very simplified imaging system. The effect is more apparent if we look at the profiles of the reconstructions in Figure 3.1e. The profiles of the two hot spots in the FBP image are nearly identical (both have a relative intensity of +1 compared to the local background and nearly identical shape). In comparison the hot spots in the PWLS reconstruction have different heights, even though we have selected β to match resolution with FBP at the center of the image. In the cold region the height is roughly +1.3, as opposed to +0.5 for the hot region. Consequently, the FWHM resolution at these two points is markedly different. The resolution is much lower in the hot region than in the cold region due to the nonuniform diagonal weighting, $\boldsymbol{W}$.

This particular system and object do not demonstrate the possible anisotropy of the local smoothing properties. While the resolution is nonuniform, the hot spots appear fairly symmetric, indicating relatively isotropic smoothing. In more realistic systems, attenuation effects contribute to less uniform weightings and more anisotropic resolution properties.

3.2.2 Nonuniform Resolution in PET

Consider a PWLS estimator of the form in (3.2) that includes the PET attenuation factors c_i. We use this estimator on an object whose attenuation map and emission distribution are shown in Figure 3.2a and Figure 3.2b, respectively. This digital phantom is anthropomorphic, representing a human thorax. The linear attenuation coefficients are chosen to be appropriate for 512 keV photons and are 0.001 mm^{-1} for the lungs, 0.016 mm^{-1} for the spine, and 0.0096 mm^{-1} for the remaining soft tissue. The relative emission intensities of the lungs is 0.4, the spine is 0.0, the heart is 3.0, and the remaining soft tissue is 2.0. Four radially symmetric hot spots have been



(a) Attenuation Map (b) Emission Distribution

(c) Filtered Backprojection (d) Conventional PWLS (e) Weights: $c_i^2/\bar{Y}_i$

Figure 3.2: Nonuniformities in an anthropomorphic torso phantom PET reconstruction.

added in the left half of the emission image, each with a relative emission intensity of 4. The PET system model represents the CTI 921 ECAT EXACT system which includes 160 radial bins and 192 angles evenly spaced over 180°, with 3.375 mm strip integrals (3.375 mm center-to-center spacing) and 4.21875 mm square pixels.

Again, we perform FBP and PWLS reconstruction on noiseless projections to obtain the images given in Figure 3.2c and Figure 3.2d. The weights for the PWLS estimator are for data with a mean of 1 million counts and are shown in Figure 3.2e, and β and the FBP cutoff frequency are chosen to match resolutions at the object center.

FBP will still have uniform resolution properties. (The measurements must be precorrected for attenuation effects, however.) The FBP image appears uniformly smoothed for all points in the image. The hot spots appear radially symmetric indicating isotropic smoothing in those regions. Compare this with the PWLS reconstruction in Figure 3.2d. The most striking nonuniformities in this image are the

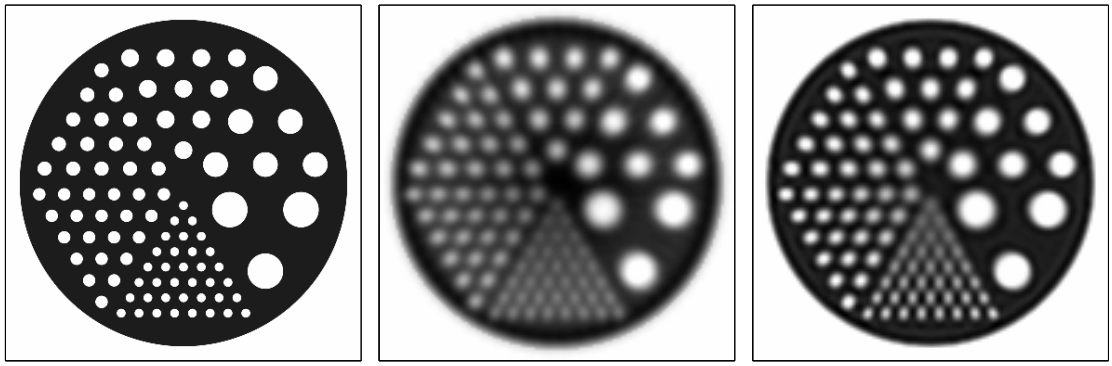
two hot spots in the left lung of the image. These hot spots are smoothed nonuniformly with greater blur in the vertical direction than in the horizontal direction, making the circular hot spots appear elliptical. Other resolution nonuniformities are also apparent. The outer edges of the object are smoothed much less than the internal edges of the phantom. (Compare the air-soft tissue boundary with the lung-soft tissue boundary and the difference between the appearance of the arms in the FBP and PWLS images.)

3.2.3 Nonuniform Resolution in SPECT

While some standard reconstruction techniques yield resolution nonuniformities in PET due to the noise and attenuation coefficients, many more methods have difficulty obtaining uniform resolution in SPECT due to the depth-dependent detector response. We perform a brief investigation showing sample reconstruction for such a shift-variant SPECT system.

We adopt a SPECT (reconstruction) system model whose circular orbit contains a field of view of 128×128 2 mm pixels. The detector head rotates at a radius of 12.8 cm, and collects data for 110 projection angles over 360° with 128 evenly spaced 2 mm radial bins. The system response is modeled after a high resolution collimator with a linearly varying depth-dependent Gaussian response that has a 1.75 mm FWHM at face of the collimator and a slope of 0.044, which corresponds to about 7.4 mm FWHM at the center of the field of view. We model the true projections (*i.e.*, the $\mathcal{H}$ operator) using a discrete system model that is upsampled by a factor of three. That is, the image-domain support contains 384×384 pixels for the true projector. The projections and reconstruction models are matched in all other respects.

We simulated a 23 cm diameter cold rod phantom with uniform attenuation coef-



(a) Emission Distribution (b) Filtered Backprojection (c) Conventional PL

Figure 3.3: Nonuniformities in a cold rod phantom SPECT reconstruction.

ficient of 0.015 mm^{-1} (the approximate attenuation coefficient of water at 140 keV) and rod diameters of 6.4, 9, 10.25, 12.8, 17.9, and 25.6 mm. The emission image for this object is shown in Figure 3.3a. To represent scatter, the model includes 5% uniformly distributed background events and 10 million counts total.

Figure 3.3b and 3.3c show sample reconstructions of noiseless data using filtered backprojection and a conventional penalized likelihood estimator, respectively. As is typical of many SPECT reconstructions, there is coarser resolution at the center of the field of view than at the edges. Moreover, there are additional reconstruction artifacts in the filtered-backprojection reconstruction (*e.g.*, the slight bulging at the center of the field of view), despite using a Chang-type attenuation correction[19]. This is because the depth-dependent response and attenuation factors are not modeled in the FBP backprojection. Close inspection of Figure 3.3c shows the nonuniform resolution properties are also anisotropic with increased radial blur. (Note that the rods appear slightly elliptical.)

While these sample reconstructions are helpful in seeing the actual results of a particular estimator, it is difficult to identify the local resolution properties quantita-

tively. Therefore, we can also look at the local impulse response at various locations to quantify the smoothing properties.

3.3 The Local Impulse Response

In [41] an approximate local impulse response was derived for discrete object models. Here, we extend those derivations for the continuous projection model described in Section 2.2.1 and for the discrete reconstruction model in Section 2.2.3. That is, we present the local impulse response for the discrete reconstruction of a finite number of measurements, which arise from a continuous object.

The local impulse response is defined in terms of the mean reconstruction,

$$\mu(f) = E_f[\underline{\hat{\theta}}(\underline{Y})] = \int \hat{\theta}(\underline{Y}) p(\underline{Y}; f) d\underline{Y}, \tag{3.3}$$

where $p(\underline{Y}; f)$ is the probability density function of the measurements. The local impulse response is the limiting difference between mean reconstructions of an image and reconstructions of a perturbed image, which we define mathematically at spatial location $\underline{x}_0$ as

$$l(\underline{x}_0) \triangleq \lim_{\varepsilon \to 0} \frac{\mu\left(f + \varepsilon \delta_{\underline{x}_0}\right) - \mu\left(f\right))}{\varepsilon}, \tag{3.4}$$

where $\delta_{\underline{x}_0} \triangleq \delta(\underline{x} - \underline{x}_0)$ is a Dirac impulse at position $\underline{x}_0$. This formulation is similar to the influence function introduced by Hampel, which is used a heuristic tool in robust statistics[57].

Many investigators have noted that the ensemble mean of likelihood-based estimators is approximately equal to the likelihood-based reconstruction of noiseless data[8, 134, 18]. Mathematically, one may write $\mu(f) \approx \underline{\hat{\theta}}(\underline{\bar{Y}}^{\dagger}(f))$. One can obtain the same approximation by finding the first-order Taylor approximation about the

noiseless measurements[41]. Substituting this approximation into (3.4), we obtain

$$\begin{aligned} l(\underline{x}_0) &= \lim_{\varepsilon\to 0} \frac{\hat{\underline{\theta}}\left(\bar{\underline{Y}}^{\dagger}(f+\varepsilon\delta_{\underline{x}_0})\right) - \hat{\underline{\theta}}\left(\bar{\underline{Y}}^{\dagger}(f)\right)}{\varepsilon} \\ &= \nabla_{\underline{Y}}\hat{\underline{\theta}}\left(\bar{\underline{Y}}^{\dagger}(f)\right)\cdot\left(\lim_{\varepsilon\to 0}\frac{\bar{\underline{Y}}^{\dagger}(f+\varepsilon\delta_{\underline{x}_0}) - \bar{\underline{Y}}^{\dagger}(f)}{\varepsilon}\right) \\ &= \nabla_{\underline{Y}}\hat{\underline{\theta}}\left(\bar{\underline{Y}}^{\dagger}(f)\right)\cdot\triangle\bar{\underline{Y}}^{\dagger}(f;\underline{x}_0), \end{aligned} \tag{3.5}$$

where $\nabla_{\underline{Y}} = [\frac{\partial}{\partial Y_1}\cdots\frac{\partial}{\partial Y_N}]$ and $\triangle\underline{\bar{Y}}$ denotes the variational derivative of $\underline{Y}(f)$ with respect to f. We evaluate $\triangle\bar{\underline{Y}}^{\dagger}(f;\underline{x}_0)$ using the mean measurements in (2.6) and applying the chain rule. Assuming the transformation $\tau(\cdot)$ is differentiable, the ith element is

$$\begin{aligned} \left[\triangle\bar{\underline{Y}}^{\dagger}(f;\underline{x}_0)\right]_i &= \dot{\tau}_i^{\dagger}\left(\langle h_i, f\rangle\right)\lim_{\varepsilon\to 0}\frac{\langle a_i, f+\varepsilon\delta_{\underline{x}_0}\rangle - \langle h_i, f\rangle}{\varepsilon} \\ &= \dot{\tau}_i^{\dagger}\left(\langle h_i, f\rangle\right)h_i(\underline{x}_0) \\ &= \dot{\tau}_i^{\dagger}\left([\mathcal{H}f]_i\right)\left[\mathcal{H}\delta_{\underline{x}_0}\right]_i, \end{aligned} \tag{3.6}$$

where $\dot{\tau}_i^{\dagger}(l) = \frac{\partial}{\partial l}\tau_i^{\dagger}(l)$.

Following [41], we may find an equation for $\nabla_{\underline{Y}}\hat{\underline{\theta}}\left(\underline{Y}\right)$. Adopting the implicitly defined estimator in (2.24) and disregarding the nonnegativity constraints, the solution to that estimator must satisfy the following equation,

$$\frac{\partial}{\partial\theta_j}\Phi(\underline{\theta},\underline{Y})\bigg|_{\underline{\theta}=\hat{\underline{\theta}}(\underline{Y})} = 0, \quad j = 1,\ldots,p \tag{3.7}$$

for any $\underline{Y}$. This equation can be written concisely in vector form as

$$\nabla^{10}\Phi\left(\hat{\underline{\theta}}(\underline{Y}),\underline{Y}\right) = \underline{0}, \quad \forall\underline{Y}, \tag{3.8}$$

where $\nabla^{10} = \left[\frac{\partial}{\partial\theta_1}\ldots\frac{\partial}{\partial\theta_P}\right]$ is the row gradient operator, which returns a vector of partial derivatives with respect to the first argument of Φ. Differentiating with respect to $\underline{Y}$ and using the chain rule, we write

$$\nabla^{20}\Phi\left(\hat{\underline{\theta}}(\underline{Y}),\underline{Y}\right)\nabla_{\underline{Y}}\hat{\theta}(\underline{Y}) + \nabla^{11}\Phi\left(\hat{\theta}(\underline{Y}),\underline{Y}\right) = \mathbf{0}, \tag{3.9}$$

where the ∇^{20} operator yields a matrix whose (j,k)th element is $\frac{\partial^2}{\partial\theta_j\partial\theta_k}$, and the ∇^{11} operator yields a matrix with the (j,i)th element equal to $\frac{\partial^2}{\partial\theta_j\partial Y_i}$. Assuming $-\nabla^{20}\Phi\left(\hat{\underline{\theta}}(\underline{Y}),\underline{Y}\right)$ is positive definite, we may rearrange (3.9) to obtain:

$$\nabla_{\underline{Y}}\hat{\underline{\theta}}(\underline{Y}) = \left[-\nabla^{20}\Phi\left(\hat{\underline{\theta}}(\underline{Y}),\underline{Y}\right)\right]^{-1}\nabla^{11}\Phi\left(\hat{\underline{\theta}}(\underline{Y}),\underline{Y}\right). \tag{3.10}$$

Recalling the form of the objective function, $\Phi(\underline{\theta},\underline{Y})$, in (2.24), and the form of the log-likelihood in (2.18), we may write (3.10) as

$$\begin{aligned}\nabla_{\underline{Y}}\hat{\underline{\theta}}(\underline{Y}) &= \left[-\sum_{i=1}^{N}\nabla^{20}L_i\left(\underline{Y},\bar{\underline{Y}}(\hat{\underline{\theta}}(\underline{Y}))\right)\nabla^{20}R(\hat{\underline{\theta}}(\underline{Y}))\right]^{-1}\cdot \\ &\qquad \left[\sum_{i=1}^{N}\nabla^{11}L_i\left(\underline{Y},\bar{\underline{Y}}(\hat{\underline{\theta}}(\underline{Y}))\right)\right],\end{aligned} \tag{3.11}$$

where we have used the fact that the penalty function is not a function of $\underline{Y}$, which means $\nabla^{11}R(\underline{\theta}) = \mathbf{0}$. Recalling the reconstruction measurement model in (2.9), the derivatives of the marginal log-likelihoods may be written as

$$\begin{aligned}\left[\nabla^{20}L_i\left(\underline{Y},\bar{\underline{Y}}(\underline{\theta})\right)\right]_{jk} &= L_i^{02}\left(Y_i,\tau_i\left(\sum_l h_{il}\theta_l\right)\right)\left[\dot{\tau}\left(\sum_l h_{il}\theta_l\right)\right]^2 h_{ij}h_{ik} \\ &+ L_i^{01}\left(Y_i,\tau_i\left(\sum_l h_{il}\theta_l\right)\right)\left[\ddot{\tau}\left(\sum_l h_{il}\theta_l\right)\right]h_{ij}h_{ik} \qquad (3.12)\\ \left[\nabla^{11}L_i\left(\underline{Y},\bar{\underline{Y}}(\underline{\theta})\right)\right]_{ji} &= L_i^{11}\left(Y_i,\tau_i\left(\sum_l h_{il}\theta_l\right)\right)\left[\dot{\tau}\left(\sum_l h_{il}\theta_l\right)\right]h_{ij}, \qquad (3.13)\end{aligned}$$

where the derivatives of $L_i(u,v)$ and $\tau_i(l)$ are defined as

$$\begin{aligned} L_i^{01}(u,v) &= \tfrac{\partial}{\partial v}L_i(u,v) & \qquad \dot{\tau}_i(l) &= \tfrac{\partial}{\partial l}\tau_i(l)\\ L_i^{02}(u,v) &= \tfrac{\partial^2}{\partial v^2}L_i(u,v) & \qquad \ddot{\tau}_i(l) &= \tfrac{\partial^2}{\partial l^2}\tau_i(l).\\ L_i^{11}(u,v) &= \tfrac{\partial^2}{\partial u\partial v}L_i(u,v) & & \end{aligned} \tag{3.14}$$

Thus, we may write the local impulse response in (3.5) using (3.6) and (3.11) with (3.12) and (3.13) as

$$\underline{l}(\underline{x}_0) = [\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}(\breve{\underline{\theta}})]^{-1}\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_0}, \tag{3.15}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are the following $N \times N$ diagonal matrices:

$$\begin{aligned}
[\boldsymbol{D}_1]_{ii} &= -L_i^{02}\left(\bar{Y}_i^{\dagger}(f), \bar{Y}_i(\breve{\underline{\theta}})\right)\left[\dot{\tau}_i([\boldsymbol{H}\breve{\underline{\theta}}]_i)\right]^2 \\
&\quad -L_i^{01}\left(\bar{Y}_i^{\dagger}(f), \bar{Y}_i(\breve{\underline{\theta}})\right)\left[\ddot{\tau}_i([\boldsymbol{H}\breve{\underline{\theta}}]_i)\right] \qquad (3.16) \\
[\boldsymbol{D}_2]_{ii} &= L_i^{11}\left(\bar{Y}_i^{\dagger}(f), \bar{Y}_i(\breve{\underline{\theta}})\right)\left[\dot{\tau}_i([\boldsymbol{H}\breve{\underline{\theta}}]_i)\right]\cdot\left[\dot{\tau}_i^{\dagger}([\mathcal{H}f]_i)\right], \qquad (3.17)
\end{aligned}$$

and where $\breve{\underline{\theta}} \triangleq \hat{\underline{\theta}}(\bar{\underline{Y}}^{\dagger}(f))$ denotes the estimate of $\underline{\theta}$ using the mean data, and $\boldsymbol{R}(\theta)$ denotes the Hessian of the penalty function. In the typical cases where $g(\cdot)$ and $\tau^{\dagger}(\cdot)$ are invertible functions, we can write the diagonal matrices (3.16) and (3.17) as functions of the mean measurements. Later, we shall see that this observation that the local impulse response is dependent on the object only through its measurements (except possibly for the penalty term) is very important for penalty design. Specifically, we write

$$\begin{aligned}
[\boldsymbol{D}_1]_{ii} &= -L_i^{02}\left(\bar{Y}_i^{\dagger}(f), \bar{Y}_i(\breve{\underline{\theta}})\right)\left[\dot{\tau}_i\left(\tau^{-1}\left(\bar{Y}_i(\breve{\underline{\theta}})\right)\right)\right]^2 \\
&\quad -L_i^{01}\left(\bar{Y}_i^{\dagger}(f), \bar{Y}_i(\breve{\underline{\theta}})\right)\left[\ddot{\tau}_i\left(\tau^{-1}\left(\bar{Y}_i(\breve{\underline{\theta}})\right)\right)\right] \qquad (3.18) \\
[\boldsymbol{D}_2]_{ii} &= L_i^{11}\left(\bar{Y}_i^{\dagger}(f), \bar{Y}_i(\breve{\underline{\theta}})\right)\left[\dot{\tau}_i\left(\tau^{-1}\left(\bar{Y}_i(\breve{\underline{\theta}})\right)\right)\right] \\
&\quad \cdot\left[\dot{\tau}_i^{\dagger}\left(\tau_i^{\dagger -1}\left(\bar{\underline{Y}}_i^{\dagger}(f)\right)\right)\right]. \qquad (3.19)
\end{aligned}$$

When the mean measurements and Hessian of the penalty are known, the local impulse response in (3.15) may be evaluated with iterative techniques. (Note that (3.15) has the same form as the solution to a linear system of equations.)

Strictly speaking, to calculate the impulse response, one must substitute (3.18) and (3.19) into (3.15). However, when the system model, $\boldsymbol{H}$ and $\tau_i(x)$, closely approximates the actual system, $\mathcal{H}$ and $\tau_i^{\dagger}(x)$, the means, $\bar{Y}_i^{\dagger}(f)$ and $\bar{Y}_i(\breve{\underline{\theta}})$ are often very similar to each other. Such is the case in tomography where the smoothing inherent to the tomographic model often dominates over the blur due to the estimator. Thus, we can use the same estimate of $\bar{Y}_i$ for both arguments of the derivatives

of L_i in (3.18) and (3.19). Typically, in cases where the mean measurements are unknown, a simple plug-in technique where we replace $\bar{Y}_i$ by Y_i yields very good approximations[41].

Because we will generally be evaluating derivatives of $L_i(u,v)$ with $u = v$, it is interesting to note a few properties of L_i under this condition. First, $L_i^{01}(v,v)$ often equals zero. Such is the case when $L_i(v,u) \leq L_i(v,v), \forall u$. Recall that the second term of L_i represents the mean measurements, and L_i is the log-likelihood for the ith measurement. Thus, this case is satisfied when the log-likelihood attains a peak at its mean. For such noise models, the second term of (3.18) disappears. Similarly, for many practical noise models like those in Table 3.1, $L_i^{11}(v,v) = -L_i^{02}(v,v)$. Thus, when $\tau_i(l) = \tau_i^{\dagger}(l), \forall i$, the diagonal matrices, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, are equal, and the local impulse response simplifies to

$$\begin{aligned} \underline{l}(\underline{x}_0) &= \left[\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} + \boldsymbol{R}(\breve{\underline{\theta}})\right]^{-1} \boldsymbol{H}'\boldsymbol{D}\mathcal{H}\delta_{\underline{x}_0} \\ \boldsymbol{D}_{ii} &= L_i^{11}(\bar{Y}_i, \bar{Y}_i)\left[\dot{\tau}_i\left(g^{-1}(\bar{Y}_i)\right)\right]^2, \end{aligned} \tag{3.20}$$

where $\boldsymbol{D} \triangleq \boldsymbol{D}_1 = \boldsymbol{D}_2$.

Using (3.20), one can estimate local impulse responses for many imaging systems. For example, for the emission tomography problem which fits the linear measurement model given in (2.7) and Poisson noise, it is straightforward to calculate the diagonal matrix in (3.20) as

$$\boldsymbol{D}_{\text{emis}} = \text{diag}\left\{\frac{1}{Y_i}\right\}, \tag{3.21}$$

where we have used the simple plug-in technique for unknown means. In contrast, adopting a Gaussian noise model (see Table 3.1) and the transmission tomography model in (2.8), one may write

$$\boldsymbol{D}_{\text{trans}} = \text{diag}\left\{\frac{(Y_i - r_i)^2}{\sigma_i^2}\right\}. \tag{3.22}$$

Table 3.1: Derivatives of marginal log-likelihoods under various noise models.
(Additive constants not important for the maximization of the penalized-likelihood objective have been dropped.)

Distribution	$L_i(u,v)$	$L_i^{01}(u,v)$	$-L_i^{02}(u,v)$	$L_i^{11}(u,v)$
Gaussian	$-\frac{1}{2\sigma_i^2}(u-v)^2$	$\frac{1}{\sigma_i^2}(u-v)$	$\frac{1}{\sigma_i^2}$	$\frac{1}{\sigma_i^2}$
Generalized Gaussian*	$-\left(\frac{\lvert u-v\rvert}{\beta_i}\right)^{\alpha_i}$	$-\mathrm{sgn}(u-v)\frac{\alpha_i}{\beta_i}\left(\frac{\lvert u-v\rvert}{\beta_i}\right)^{\alpha_i-1}$	$\frac{(\alpha_i-\alpha_i^2)}{\beta_i^2}\left(\frac{\lvert u-v\rvert}{\beta_i}\right)^{\alpha_i-2}$	$\frac{(\alpha_i-\alpha_i^2)}{\beta_i^2}\left(\frac{\lvert u-v\rvert}{\beta_i}\right)^{\alpha_i-2}$
Poisson	$u\log(v)-v$	$\frac{u}{v}-1$	$\frac{u}{v^2}$	$\frac{1}{v}$
Shifted Poisson	$(u+a_i)\log(v+a_i)-v$	$\frac{u+a_i}{v+a_i}-1$	$\frac{u+a_i}{(v+a_i)^2}$	$\frac{1}{v+a_i}$

*For the generalized Gaussian distribution, $\alpha_i \in (1,\infty)$ and $\beta_i \in (0,\infty)$.

Thus, we may evaluate (3.20) for various locations, imaging systems, and estimator parameters. In some simulation studies or approximations, we may also be interested in the "artificial" case of a discrete projection model and a discrete reconstruction model. It is straightforward to show that the local impulse response in that case has the similar form:

$$\underline{l}^j = \left[\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} + \boldsymbol{R}(\breve{\underline{\theta}})\right]^{-1} \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j, \tag{3.23}$$

where $\boldsymbol{H}\underline{e}^j$ is the discrete projection of the jth unit vector, and $\underline{l}^j$ is discrete response centered at position j.

3.4 Resolution Properties of Tomographic Reconstructions

3.4.1 Sample Local Impulse Responses for PET

Let us return to the sample reconstruction presented in Section 3.2.2. We may write the local impulse response for the PWLS estimator in (3.1) as

$$\underline{l}^j = [\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} + \beta\boldsymbol{R}]^{-1}\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j, \tag{3.24}$$

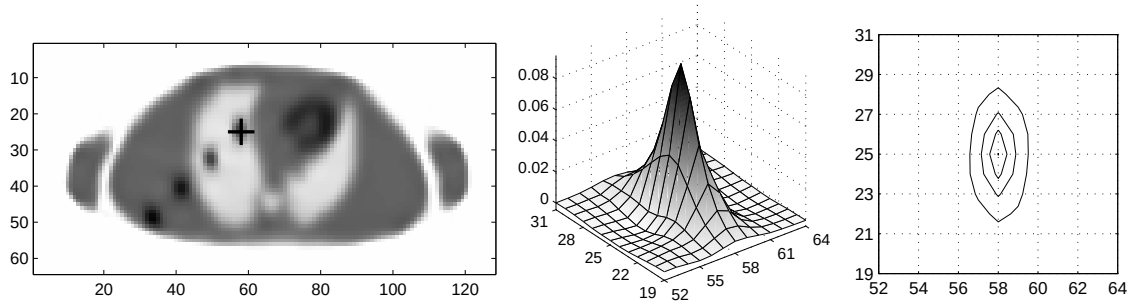
with $\boldsymbol{D} = \mathrm{diag}\{1/Y_i\}$. For the factorized PET model of Section 3.2.2, the PWLS local impulse response can be written:

$$\underline{l}^j = [\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G} + \beta\boldsymbol{R}_0]^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}\underline{e}^j, \tag{3.25}$$

where $\boldsymbol{W} = \mathrm{diag}\{c_i/Y_i\}$. While this expression is closed-form, because of the size of the system matrix, the matrix inverse is typically impractical to compute. One alternative is to use iterative methods to approximate a solution.[1]

Considering the torso phantom introduced in the Section 3.2.2, we can use (3.25) to investigate the resolution properties at different points in the image. Selecting the

[1] Note that (3.25) is in the same form as the solution to a linear system of equations, $[\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}+\beta\boldsymbol{R}_0]\underline{l}^j = \boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}\underline{e}^j$. There are a number of iterative techniques that may be used to approximate a solution. We often use a coordinate ascent or conjugate gradient algorithm.



(a) Conventional PWLS Reconstruction (b) Local Impulse Response (c) Response Contours

Figure 3.4: A 2D local impulse response for the torso phantom and a PET model with attenuation.

The PWLS reconstruction with the conventional penalty is shown in (a) and the location of the local impulse response is marked with a + mark. The local impulse response and its contours are shown in (b) and (c), respectively.

pixel position j, so that it lies within one of the hot spots (shown in Figure 3.4a.), we may evaluate (3.25) to obtain the local impulse response at that position. That 2D function is shown in Figure 3.4b. Since it is difficult to see some of the anisotropic effects in this figure, we present contours of the local impulse response in Figure 3.4c.

As we have seen in the sample reconstruction in Figure 3.2d, there is more vertical smoothing than horizontal smoothing at this image location. This is confirmed by looking at the local impulse response. Note that the contours of the local impulse response are much wider in the vertical direction than the horizontal in Figure 3.4c. The local impulse response allows us to quantify resolution properties at different image locations. For example, from the local impulse response we can specify vertical and horizontal FWHM resolution. Since the local impulse responses are not generally aligned with the Cartesian axes, we can use other measures of resolution, including the maximum and minimum FWHM resolution. For example, after one obtains the half-maximum contour, one finds the minimum or maximum contour diameter. Two measures we find useful in quantifying the uniformity of a local im-

pulse response are the mean and standard deviation of the contour radius. (For an isotropic response, the standard deviation should be zero.) Similarly, we can find the mean absolute deviation from some desired resolution. For example, if the desired response is isotropic with 4.0-pixel FWHM resolution, the mean absolute deviation is given by the average over all FWHM contour radii of $|4.0 - \text{radius}|$. For an exactly matched response, this deviation is zero.

For a complete investigation of resolution uniformity, one would have to consider the local impulse response at all pixel locations. This tends to be impractical, since the local impulse response calculation involves iterative approximation and is the same dimension as the imaging problem itself. However, since the local impulse responses generally vary slowly with position, a subsampling of image positions is often sufficient for investigating the global resolution properties.

Consider the digital phantom in Figure 3.5. This is the same phantom presented in [41]. This 128×64 phantom is composed of warm background ellipse, a cold left disc, and a hot right disc with relative emission intensities of 2, 1, and 3, and attenuation coefficients 0.0096, 0.003, and 0.013 mm^{-1}, respectively. Sample positions are represented by the cross-hairs that are arranged in a grid pattern. The system model specifies projection data with 128 radial bins and 110 angles uniformly spaced over 180° with 3 mm pixels, 6 mm wide strip integrals (3 mm center-to-center spacing), and detector efficiencies with a pseudo-random log normal variance with $\sigma = 0.3$ to model detector efficiency effects. We simulate data with 1 million mean total counts.

Choosing a PLE with a Poisson noise model, matched discrete measurement and reconstruction system models, and a conventional quadratic shift-invariant penalty, leads to local impulse responses that are written as

$$\underline{l}^j = [\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} + \beta \boldsymbol{R}_0]^{-1} \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j, \tag{3.26}$$

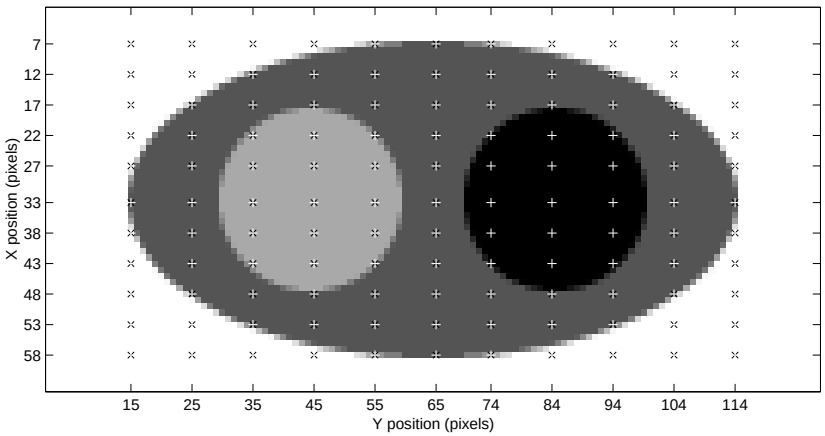


Figure 3.5: Emission distribution and sample positions for the local impulse response investigation.

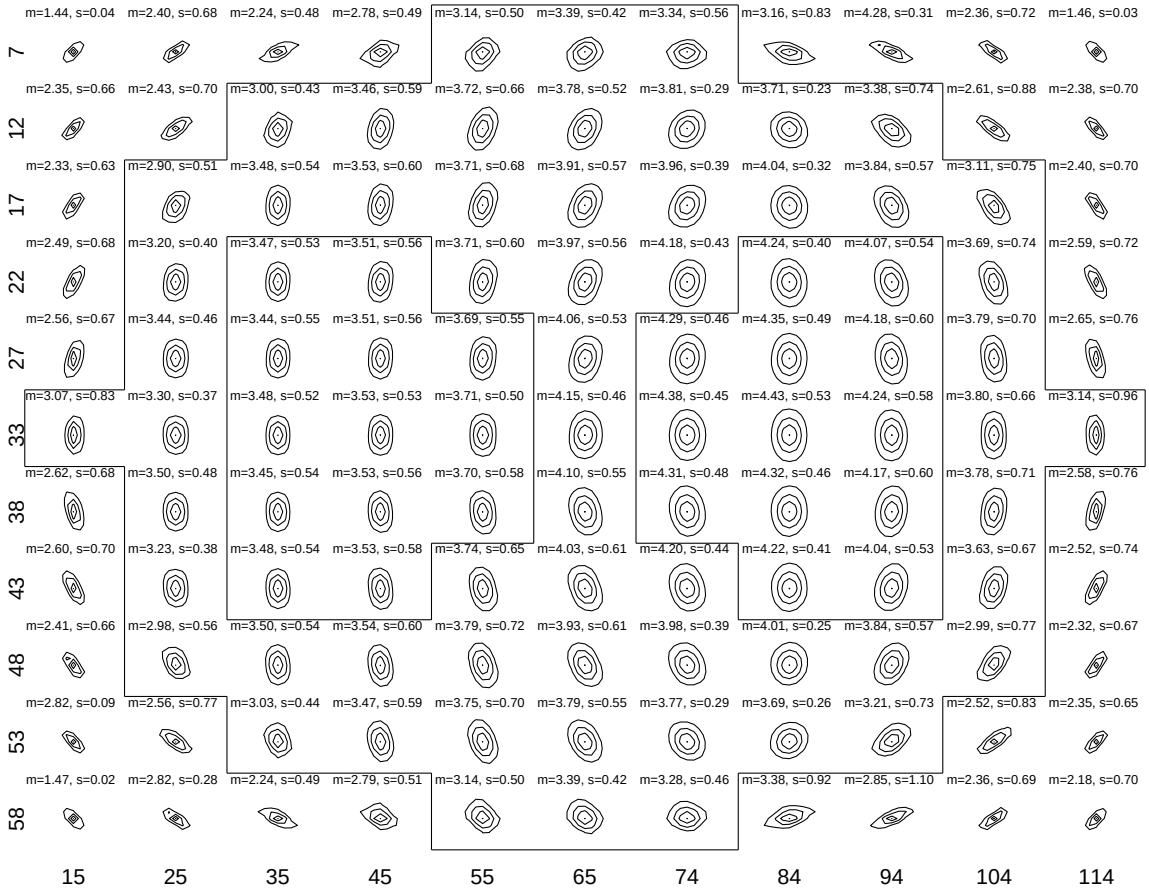


Figure 3.6: Local impulse response map for a PLE with conventional penalty.

with $\boldsymbol{D} = \text{diag}\{1/Y_i\}$. Note that this form is identical to the PWLS local impulse response in (3.24), thus (3.25) is the local impulse response under the Poisson model for PET systems. For each location the local impulse response for the PLE with conventional first-order penalty is evaluated, and contours are formed at 25%, 50%, 75%, and 99% of the peak value. (The regularization parameter β for this particular estimator was chosen to yield 4.0 pixels FWHM resolution at the center of the image.) These contours are shown in Figure 3.6. The contours are arranged in the same order as the sampling grid so that the spatial positioning of the local impulse response roughly identifies the position of the local impulse response in the phantom. Additionally, we have indicated the boundaries of the uniform emission regions on this map. Above each response, the mean (m) and standard deviation (s) of the radius of the FWHM contour is presented.

This investigation demonstrates the shift-variant resolution for a conventional PLE and a shift-invariant system. We see the greater smoothing in high count regions, as demonstrated by the broader responses in the hot disc region. These responses are anisotropic and shift-variant. Even though the estimator was designed with a 4.0 pixel FWHM target resolution, one can see the variation of mean resolution with location in looking at the different m values.

3.4.2 Sample Local Impulse Responses for SPECT

We may perform the same kind of local impulse response investigation for the SPECT system of Section 3.2.3. Recall from Section 3.2.3 that we have used a mismatched true projection model and reconstruction model. Thus, we use the form of the local impulse in (3.20).

Returning to the cold rod phantom presented in Figure 3.3, we present a subsampling of local impulse responses for the FBP and PL estimators in Figures 3.7A

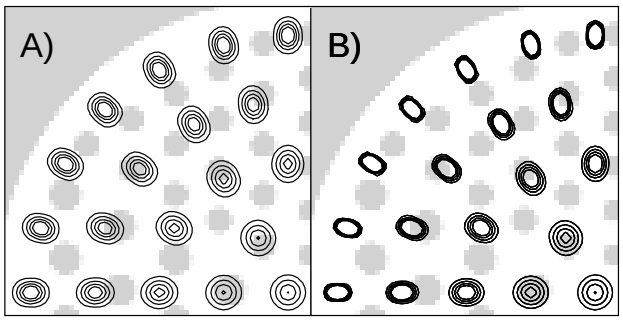


Figure 3.7: Local impulse responses for shift-variant SPECT reconstructions.
This figure shows responses the upper left quadrant of the phantom in Figure 3.3 for A) FBP with Chang attenuation correction and B) Conventional PL with space-invariant penalty.

and 3.7B, respectively. These are the same methods used in the FBP and PL reconstruction shown in Figure 3.3B and 3.3C, respectively. Because FBP (with Chang correction) is a linear reconstruction technique, we may form the local impulse responses for this method simply by propagating true projections of impulse functions through the estimator.

As one might expect, due to the depth-dependent detector response, the local impulse responses for both the standard FBP method and the penalized-likelihood approach are broader in the center of the field of view and narrower at the edges. Moreover, the anisotropic blur at different locations in the image is immediately obvious.

3.4.3 Analysis of the Local Impulse Response

While evaluating the local impulse response serves as an important tool for resolution investigation, one can learn much by considering the form of the local impulse response. Consider the case of an ECT system in which the measurements have homoscedastic measurement noise. If the measurement noise has variance, σ^2, then

the local impulse response for the conventional PLE in (3.26) may be written[41] as

$$\begin{aligned}
\underline{l}^j &= [\boldsymbol{H}'\mathrm{diag}\{1/\sigma^2\}\,\boldsymbol{H} + \beta\boldsymbol{R}_0]^{-1}\boldsymbol{H}'\mathrm{diag}\{1/\sigma^2\}\,\boldsymbol{H}\underline{e}^j \\
&= \left[\frac{1}{\sigma^2}\boldsymbol{H}'\boldsymbol{H} + \beta\boldsymbol{R}_0\right]^{-1}\frac{1}{\sigma^2}\boldsymbol{H}'\boldsymbol{H}\underline{e}^j \\
&= [\boldsymbol{H}'\boldsymbol{H} + \sigma^2\beta\boldsymbol{R}]^{-1}\boldsymbol{H}'\boldsymbol{H}\underline{e}^j. \qquad (3.27)
\end{aligned}$$

Thus, for different noise levels, the regularization parameter is effectively scaled by the variance term. Since the variance equals the mean for Poisson data, higher count data will have higher variance and thus be smoothed more than low count data. From a Bayesian standpoint, higher noise levels require more reliance on the prior, which leads to increased smoothness. Thus, this intuitively explains Figure 3.1, where high count regions appear to be smoother than low count regions. (Recall that this tomographic system has a shift-invariant response $\boldsymbol{H}'\boldsymbol{H}\underline{e}^j$.)

For the PET models, where we may factor $\boldsymbol{H}$ into geometric and ray-dependent factors, we may write an analogous equation to (3.27) with constant ray-dependent factors as

$$\underline{l}^j = \left[\boldsymbol{G}'\boldsymbol{G} + \frac{\sigma^2}{c^2}\beta\boldsymbol{R}_0\right]\boldsymbol{G}'\boldsymbol{G}\underline{e}^j. \qquad (3.28)$$

Thus, we expect that increased PET attenuation factors lead to decreased smoothing. Therefore, rays in directions that have more attenuation should exhibit finer resolutions. This is exactly what we see in Figure 3.6. There is less attenuation for vertical rays since the object's major axis is along the x-axis, and we see a general tendency for increased local impulse response width in the vertical direction. Similarly, at the edges of the field of view, we see that the local impulse responses tend to show increased tangential smoothing, since rays that are oblique to the object are subject to less attenuation than those rays that intersect a large portion of the object.

In summary, the resolution properties of images formed from penalized-likelihood reconstruction are generally nonuniform (*i.e.*, shift-variant and anisotropic). This is true even for shift-invariant systems when the measurements have different noisse levels. The local impulse response can be used to predict the resolution properties of images without performing reconstructions, and is an important tool for analyzing and quantifying nonuniform resolution.

CHAPTER IV

Quadratic Penalty Design

We have demonstrated that resolution nonuniformities are an inherent aspect of penalized-likelihood methods with space-invariant penalties. Because these nonuniformities are noticeable in reconstructions and can potentially distort the shape and quantitation of image features, it would be advantageous to have a penalized-likelihood technique that can produce images without such nonuniformities. In this chapter, we discuss how to design penalties for penalized-likelihood estimators for user-specified resolution properties like uniform resolution.

4.1 Prior Work in Penalty Design

It has been widely recognized that the specification of the penalty function in a penalized-likelihood estimator can be used to produce images with desirable properties. As discussed in Section 2.5.1, penalty functions can be chosen that preserve edges or other features. Thus, performing penalized-likelihood image reconstruction with a modified penalty for user-specified resolution properties is a natural approach for controlling the resolution. However, since we have found that the local impulse response is typically shift-variant and a function of the data and the image being estimated, we expect that we will need to locally modify the penalty. That is, we believe simply specifying a certain global $\psi(\cdot)$ function in (2.25) is *not* sufficient to

adequately control resolution properties. (We discuss this further in Section 4.2.1.) Thus, generally we will be performing a shift-variant penalty design.

Shift-variant regularization has previously been used to provide user-specified space-variant resolution properties[98, 91] or for edge-preservation by locally modifying penalty weights[73, 36, 16]. However, these techniques are not concerned with the exact form of the local impulse response functions, and yield only coarse resolution control.

Qi has used local impulse response functions and variance estimates to locally adapt penalty weights for contrast optimization[95]. While this is a form of resolution control, these methods are not trying to specify the exact shapes of the local impulse responses. Therefore, reconstructions using the regularization techniques of [95] typically will exhibit shift-variant and anisotropic resolution properties.

However, some work has been done that uses the local impulse response to provide user-specified resolution properties like uniform resolution. Mustafovich has used local filter design techniques to attempt to provide uniform resolution for likelihood-based estimators with inter-update filtering[88]. In [41], Fessler and Rogers developed a certainty-based approach that attempts to locally modify the penalty weights to provide uniform resolution. The certainty-based approach has been an important starting point for us and we have used this method as a baseline in comparison with the methods developed in this dissertation. Thus, we review the certainty-based approach in the following section.

4.1.1 Certainty-Based Penalty

The local impulse responses given in (3.23) and (3.26) are dependent on the Fisher information matrix, which for the PET-style factorization (*i.e.*, (2.10) with $\boldsymbol{A} = \mathbf{1}$)

may be written as

$$\begin{aligned} \boldsymbol{F}(\underline{\theta}) \triangleq \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} &= \operatorname{diag}\{s_j\}\,\boldsymbol{G}'\operatorname{diag}\left\{c_i^2 d_i(\underline{\theta})\right\}\boldsymbol{G}\operatorname{diag}\{s_j\} \\ &= \operatorname{diag}\{s_j\}\,\boldsymbol{G}'\operatorname{diag}\{q_i(\underline{\theta})\}\,\boldsymbol{G}\operatorname{diag}\{s_j\}\,, \end{aligned} \tag{4.1}$$

where $d_i(\underline{\theta}) = [\boldsymbol{D}]_{ii}$, the ith element of the diagonal matrix, $\boldsymbol{D}$, and $q_i(\underline{\theta}) = c_i^2 d_i(\underline{\theta})$. Recall in (3.18) and (3.19) that elements of $\boldsymbol{D}$ are object-dependent, through a function of the measurements $\underline{Y}(\underline{\theta})$. The certainty-based approach is applicable to systems where $\boldsymbol{G}'\boldsymbol{G}$ is approximately a shift-invariant operator and, therefore, we adopt the above factorization of the Fisher information matrix. The nonuniformity of the $q_i(\underline{\theta})$ terms makes $\boldsymbol{F}(\underline{\theta})$ a shift-variant operator even for shift-invariant systems. The diagonal elements of $\boldsymbol{F}(\underline{\theta})$ are given by

$$\boldsymbol{F}_{jj}(\underline{\theta}) = s_j^2 \sum_i g_{ij}^2 q_i(\underline{\theta}) = \kappa_j^2(\underline{\theta}) \sum_i g_{ij}^2, \qquad j = 1, \ldots, p \tag{4.2}$$

where

$$\kappa_j(\underline{\theta}) \triangleq s_j \sqrt{\sum_i g_{ij}^2 q_i(\underline{\theta}) \Big/ \sum_i g_{ij}^2}\,. \tag{4.3}$$

These $\kappa_j(\underline{\theta})$ terms are a kind of normalized backprojection of q_i, which is related to the inverse of the variance of the ith corrected measurement. Thus, $\kappa_j(\underline{\theta})$ quantifies the aggregate *certainty* of the measurements intersecting the jth pixel. Since the inherent response in tomographic systems is $1/r$, $\boldsymbol{F}_{jj}(\underline{\theta})$ concentrates along the diagonal, and one can make the following approximation:

$$\operatorname{diag}\{s_j\}\,\boldsymbol{G}'\operatorname{diag}\{q_i(\underline{\theta})\}\,\boldsymbol{G}\operatorname{diag}\{s_j\} \approx \operatorname{diag}\{\kappa_j(\underline{\theta})\}\,\boldsymbol{G}'\boldsymbol{G}\operatorname{diag}\{\kappa_j(\underline{\theta})\} \tag{4.4}$$

$$\boldsymbol{F}(\underline{\theta}) \approx \boldsymbol{D}_\kappa \boldsymbol{G}'\boldsymbol{G}\boldsymbol{D}_\kappa, \tag{4.5}$$

where $\boldsymbol{D}_\kappa = \operatorname{diag}\{\kappa_j(\underline{\theta})\}$. This approximation is exact for the diagonal elements of $\boldsymbol{F}(\underline{\theta})$. The off-diagonal elements would be exact if the q_i terms were uniform. This

approximation tends to be accurate even for nonuniform q_i because of the implicit smoothing in (4.3) and the fact that the response at pixel j is mainly affected by the q_i terms for measurements intersecting the pixel j.

Substituting (4.5) into the expression for the local impulse response in (3.23) yields

$$\begin{aligned} \underline{l}^j &\approx [\boldsymbol{D}_\kappa \boldsymbol{G}'\boldsymbol{G}\boldsymbol{D}_\kappa + \boldsymbol{R}(\breve{\underline{\theta}})]^{-1}\boldsymbol{D}_\kappa \boldsymbol{G}'\boldsymbol{G}\boldsymbol{D}_\kappa \underline{e}^j \\ &= \boldsymbol{D}_\kappa^{-1}[\boldsymbol{G}'\boldsymbol{G} + \boldsymbol{D}_{\underline{\theta}}^{-1}\boldsymbol{R}(\breve{\underline{\theta}})\boldsymbol{D}_\kappa^{-1}]^{-1}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{D}_\kappa \underline{e}^j \\ &= \kappa_j(\underline{\theta})\boldsymbol{D}_\kappa^{-1}[\boldsymbol{G}'\boldsymbol{G} + \boldsymbol{D}_\kappa^{-1}\boldsymbol{R}(\breve{\underline{\theta}})\boldsymbol{D}_\kappa^{-1}]^{-1}\boldsymbol{G}'\boldsymbol{G}\underline{e}^j, \end{aligned} \tag{4.6}$$

since $\boldsymbol{D}_\kappa \underline{e}^j = \kappa_j(\underline{\theta})\underline{e}^j$. Since the local impulse response is highly localized about position j, (4.6) may be further approximated with the following form:

$$\underline{l}^j \approx [\boldsymbol{G}'\boldsymbol{G} + 1/\kappa_j^2(\underline{\theta})\boldsymbol{R}(\breve{\underline{\theta}})]^{-1}\boldsymbol{G}'\boldsymbol{G}\underline{e}^j. \tag{4.7}$$

The approximations in (4.6) and (4.7) suggest the following modified form for pairwise roughness penalties:

$$R^\star(\underline{\theta}) = \frac{1}{2}\sum_{j=1}^{p}\sum_{j=1}^{p} w_{jk}\hat{\kappa}_j\hat{\kappa}_k\psi(\theta_j - \theta_k), \tag{4.8}$$

where $\hat{\kappa}_j$ denotes an estimate of the certainty $\kappa_j(\underline{\theta})$, typically formed through an application of (4.3) to estimates $\hat{q}_i$ of q_i. Since the Hessian of this modified penalty is given by

$$\boldsymbol{R}^\star_{jk}(\underline{\theta}) = \begin{cases} \displaystyle\sum_{l\neq j} w_{jl}\hat{\kappa}_j\hat{\kappa}_l\ddot{\psi}(\theta_j - \theta_l), & j = k \\ -w_{jk}\hat{\kappa}_j\hat{\kappa}_k\ddot{\psi}(\theta_j - \theta_k), & j \neq k, \end{cases}$$

and $\mathrm{diag}\{\hat{\kappa}_j\} \approx K_{\underline{\theta}}$, we may write

$$\boldsymbol{R}^\star(\underline{\theta}) \approx K_{\underline{\theta}}\boldsymbol{R}(\underline{\theta})K_{\underline{\theta}}. \tag{4.9}$$

Substituting the Hessian for the modified penalty into the approximate local impulse given in (4.6), we get

$$\underline{l}^j(\underline{\theta}) \approx \kappa_j(\underline{\theta}) K_{\underline{\theta}}^{-1} [\boldsymbol{G}'\boldsymbol{G} + \boldsymbol{R}(\breve{\underline{\theta}})]^{-1} \boldsymbol{G}'\boldsymbol{G}\underline{e}^j. \tag{4.10}$$

Since $\kappa_j(\underline{\theta})/\kappa_k(\underline{\theta}) \approx 1$ when j and k represent nearby pixels, for narrow impulse responses which cover a relatively small set of pixels, we can ignore the first term in (4.10). For systems where $\boldsymbol{G}'\boldsymbol{G}$ represents a shift-invariant system response and $\boldsymbol{R}(\underline{\theta})$ is chosen to be a uniform quadratic penalty, the above local impulse response is approximately shift-invariant. Nonquadratic penalties generally introduce object-dependent nonuniformities (like edge-preservation); however, this penalty attempts to remove those nonuniformities that are due to the interaction of the nonuniform measurement statistics and the penalty function.

Problems with the Certainty-Based Penalty

The certainty-based penalty generally improves the spatial uniformity; however, there are still problems with this penalty. Because $\kappa_j(\underline{\theta})K_{\underline{\theta}}^{-1}$ cannot be ignored in general, this technique tends to produce asymmetric responses due to the additional scaling of the local impulse response. Similarly, the Fisher information approximation in (4.5) can be inaccurate for nonuniform $q_i(\underline{\theta})$.

The limitations of the certainty-based method can be shown simply by considering the form of the penalty in (4.8). Since the κ_j terms vary relatively slowly with position, $\kappa_j \approx \kappa_k$ for neighboring j and k. This means locally, the certainty-based penalty acts as multiplicative factor on the local penalty (much like the regularization parameter β does on the entire image). In fact, we see in the local impulse response approximation with conventional penalty in (4.7) there is an "effective" smoothing parameter of β/κ_j^2. The certainty-based penalty, in effect, adjusts this

"effective" parameter locally in an attempt to yield uniform resolution properties. Simply changing the local smoothing parameter tends to increase or decrease the average resolution at that position. Generally, the shape of the response varies relatively little with such parameter scaling. Therefore, the certainty-based penalty should be able to provide resolution properties whose radially averaged FWHM resolution is nearly equal, but will not be able to eliminate the anisotropy in the local impulse responses.

Returning to the test phantom and PET model illustrated in Figure 3.5, we computed the local impulse responses over the same subsampling of image locations. For the certainty-based penalty, we have adopted a penalty of the form in (4.9), where $\boldsymbol{R}(\underline{\theta}) = \beta \boldsymbol{R}_0$, a scaled quadratic first-order penalty matrix. The parameter β is chosen to yield reconstructions of 4.0 pixels FWHM resolution at the image center. Figure 4.1 presents the results of this local impulse response investigation.

In the local impulse responses in this figure much of the variability in the size of the local impulse response has decreased compared to the local impulse response map shown in Figure 3.6. Further evidence of this improvement in the mean resolution can be seen in the mean FWHM resolutions stated above each response. These values are generally much closer to the 4.0 pixel FWHM target resolution. However, the anisotropy is still clear throughout the image. This is obvious from looking at the stretched responses and by noting the nonzero radial standard deviations (s) stated above each response.

The certainty-based penalty was originally developed for shift-invariant PET systems and may be calculated very quickly using (4.3), which is roughly equivalent to a single backprojection operation. While one might be able to extend these techniques to shift-variant systems by calculating position-dependent κ_j terms, this has

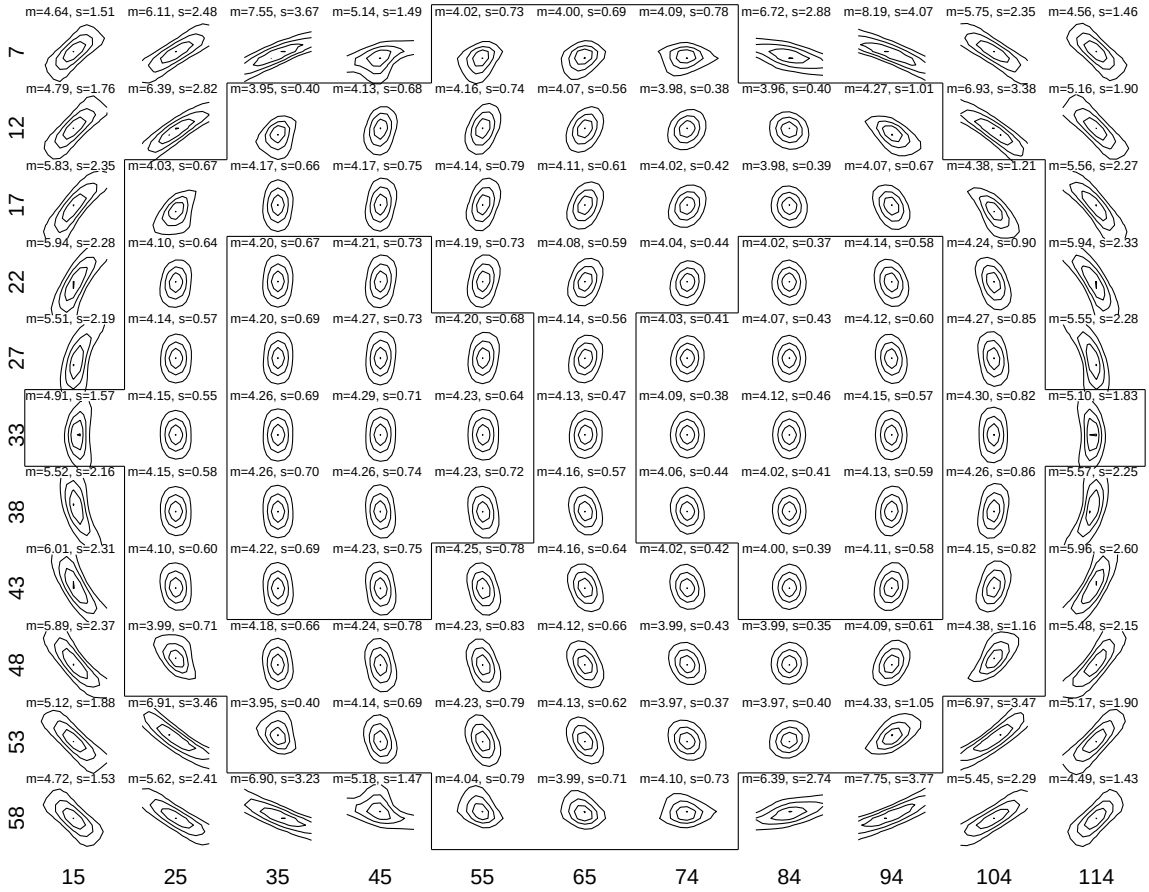


Figure 4.1: Local impulse response map for a PLE with certainty-based penalty.

not yet been thoroughly investigated, and would presumably take significantly more computation.

As mentioned before, the certainty-based penalty simply scales the penalty function locally. As seen in the local impulse responses in Figure 3.6, this scaling is insufficient for providing uniform resolution properties. In the following sections we develop a penalty where directional weightings are modified to control resolution.

4.2 Choosing and Parameterizing the Penalty Function

4.2.1 Choosing a Family of Penalty Functions

We wish to find a penalty function that induces user-specified resolution properties like uniform resolution. We believe it is unlikely that simply choosing a specific global $\psi(\cdot)$ function in (2.25) will suffice, since the resolution properties are intimately tied to physical system aspects like attenuation and detector response. However, certain penalty choices may be advantageous because they can provide uniform resolution under some kind of ideal circumstances, or because they are easy to approximate or evaluate.

Since uniform resolution is one class of user-specified resolutions that we would like to accommodate, we shall not adopt any of the edge-preserving penalties described in Figure 2.15. (The concept of edge preservation is inherently nonuniform, since edges will have higher resolution than other portions of the image.)

Square-Root Penalty

It has been suggested that the penalizing the square-root of pixel values in an image will lead to more uniform resolution. For example, in the statistical sinogram deblurring work of La Riviere[65], a square-root penalty was used successfully to provide nearly uniform resolution properties.

Consider the square-root penalty which is written as

$$R(\underline{\theta}) = \beta \sum_j \sum_k w_{jk} \left(\sqrt{\theta_j} - \sqrt{\theta_k} \right)^2 . \tag{4.11}$$

It is straightforward to find the Hessian of this penalty. The (j,k)th element of the

Hessian is

$$[\boldsymbol{R}(\underline{\theta})]_{jk} = \begin{cases} \beta \sum_l (w_{jl} + w_{lj}) \frac{\sqrt{\theta_l}}{2\theta_j^{\frac{3}{2}}}, & j = k \\ \frac{-\beta(w_{jk} + w_{kj})}{2\sqrt{\theta_j \theta_k}}, & j \neq k. \end{cases} \tag{4.12}$$

If the image $\underline{\theta}$ is locally flat, the Hessian of the penalty is

$$\boldsymbol{R}(\underline{\theta}) \approx \frac{\beta}{\theta_j} \boldsymbol{R}, \quad \text{where} \quad \boldsymbol{R} = \begin{cases} \sum_l \frac{1}{2}(w_{jl} + w_{lj}), & j = k \\ -\frac{1}{2}(w_{jk} + w_{kj}), & j \neq k. \end{cases} \tag{4.13}$$

Recall the homoscedastic measurement noise example in Section 3.4.3 where the noise model leads to an effective regularization parameter, $\sigma^2\beta$. Extending this analysis to the square-root penalty, we find that in regions that are nearly constant the square-root penalty has an effective regularization parameter of $\sigma^2\beta/\theta_j$.

The square-root penalty generally will not completely compensate for the nonuniform smoothing induced by the estimator for the general imaging problem. However, for the image restoration problem, where measurements are a slightly blurred version of the image, the square-root penalty can yield relatively uniform resolution. Under the Poisson noise model the variance equals the mean measurement. Thus, $\sigma^2 \approx \theta_j$, and the two terms approximately cancel out, making the effective regularization parameter uniform across the image. Similarly, one can apply such a penalty to statistical sinogram deblurring methods as La Riviere did in [65]. La Riviere also noted the greater resolution uniformity of reconstructions from deblurred sinograms when the square root penalty is applied.

Because the local impulse response is a function of all the physical effects in $\boldsymbol{H}$, including attenuation and detector response, it seems that any penalty that is applied in an effort to control resolution must be a function of these components as well.

Desirable Properties of a Penalty Function

While there are many choices of penalty functions, some have better properties than others for penalty design. For example, it would be advantageous to have a penalty that results in a simple form for the local impulse response. Since the local impulse response is a function of the Hessian of the penalty (see (3.15)), we would like penalties whose Hessians have simple forms. For example, if we choose the quadratic penalty in (2.27), the Hessian is simply $\boldsymbol{R}$. Thus, the local impulse response in (3.15) depends on the object only through its projections, and an estimate of the image is not required to find the local impulse response. We can use the local impulse response in turn to find an appropriate quadratic penalty.

Strictly speaking, this means that the penalty is dependent on the measurements. Thus, the second term in (3.10) requires that the gradient of the objective (and, therefore, the penalty term as well) with respect to $\underline{Y}$ be calculated. However, even though we will eventually design a penalty that is dependent on the projection data, we have found that ignoring the dependence of R on $\underline{Y}$ nevertheless leads to good estimates of the local impulse response. In other words, the derivatives of the penalty with respect to $\underline{Y}$ are sufficiently small as to be disregarded when evaluating (3.10) and (3.11).

4.2.2 The Quadratic Penalty

We chose to adopt the quadratic penalty since it leads to an object-independent form for the local impulse response and since the linearized response derived in Section 3.3 is a good predictor of the local resolution properties. (For some penalties, the linearized response may not be a good predictor.) Recalling (2.25), we may write

the quadratic penalty as

$$R(\underline{\theta}) = \beta \sum_k \frac{1}{2}([\boldsymbol{C}\underline{\theta}]_k)^2 = \beta \frac{1}{2}\underline{\theta}'\boldsymbol{C}'\boldsymbol{C}\underline{\theta} = \frac{1}{2}\underline{\theta}'\boldsymbol{R}\underline{\theta}. \tag{4.14}$$

We note that this form is slightly more general than the pairwise quadratic penalty, since one can include a nonzero magnitude penalty on θ_j^2. That is, if one adds a magnitude penalty to the pairwise penalty of (2.26) with $\psi(t) = \frac{1}{2}t^2$, then

$$\begin{aligned}
& \frac{1}{2}\sum_j \sum_k w_{jk}\frac{1}{2}(\theta_j - \theta_k)^2 + \frac{1}{2}\sum_j w_j\theta_j^2 \\
= \; & \frac{1}{2}\sum_j \sum_k -w_{jk}\theta_j\theta_k + \frac{1}{2}\sum_j \left(\sum_k \frac{w_{jk}}{2}\right)\theta_j^2 + \frac{1}{2}\sum_k \left(\sum_j \frac{w_{jk}}{2}\right)\theta_k^2 + \frac{1}{2}\sum_j w_j\theta_j^2 \\
= \; & \frac{1}{2}\sum_j \sum_{k\neq j} \underbrace{-w_{jk}}_{r_{jk},j\neq k}\theta_j\theta_k + \frac{1}{2}\sum_j \underbrace{\left(w_j + \sum_{k\neq j}\frac{1}{2}(w_{jk}+w_{kj})\right)}_{r_{jj}}\theta_j^2 \\
= \; & \frac{1}{2}\sum_j \sum_k r_{jk}\theta_j\theta_k = \frac{1}{2}\underline{\theta}'\boldsymbol{R}\underline{\theta}.
\end{aligned} \tag{4.15}$$

Given the above equalities and using the bracketed equations above as a definition for the elements of $\boldsymbol{R}$, we find that the pairwise and magnitude penalties completely span all quadratic penalties. That is, a set of pairwise weights, $\{w_{jk}\}$, and magnitude penalty weights, $\{w_j\}$, can be represented by a corresponding set of $\{r_{jk}\}$ (the elements of $\boldsymbol{R}$), and vice versa.

The quadratic penalty is completely specified by the matrix $\boldsymbol{R}$. One may use asymmetric penalty matrices; however, only the symmetric portion of the matrix is important in the penalized-likelihood objective. This is because the penalty evaluates to a scalar value and $\underline{\theta}'\boldsymbol{R}\underline{\theta} = (\underline{\theta}'\boldsymbol{R}\underline{\theta})' = \underline{\theta}'\boldsymbol{R}'\underline{\theta}$. Similarly, if one evaluates the Hessian of a quadratic penalty specified by an asymmetric penalty matrix:

$$\nabla^{20} R(\underline{\theta}) = \frac{1}{2}(\boldsymbol{R} + \boldsymbol{R}') \triangleq \boldsymbol{R}_{\text{sym}}. \tag{4.16}$$

While the penalty matrix form provides an elegant form for writing the local impulse response, $\boldsymbol{R}$ is typically very sparse since only small pixel neighborhoods are used. Thus, it is helpful to be able to specify the penalty with a reduced number of coefficients.

4.2.3 Parameterizing the Quadratic Penalty

The sparsity of $\boldsymbol{R}$ scales with the size of the neighborhood used in the penalty function. A given column of $\boldsymbol{R}$ contains roughly the same number of values as the size of the neighborhood and should be able to be represented with the same number of coefficients.

Recalling (4.15), we may write a column of a symmetric penalty matrix as

$$\begin{aligned}
\boldsymbol{R}\underline{e}^j &= \begin{bmatrix} -w_{1,j} \\ \vdots \\ -w_{j-1,j} \\ w_j + \sum_{l \neq j} w_{l,j} \\ -w_{j+1,j} \\ \vdots \\ -w_{p,j} \end{bmatrix} = w_j \underline{e}^j + \sum_{l \neq j} w_{lj}(\underline{e}^j - \underline{e}^l) && (4.17) \\
&= w_j \text{vec}\left\{\delta(m - m_j, n - n_j)\right\} + \sum_{q=1}^{B-1} w_{l_{jq},j} \text{vec}\left\{b_q(m - m_j, n - n_j)\right\} && (4.18) \\
&= \boldsymbol{B}^j \underline{w}^j. && (4.19)
\end{aligned}$$

In (4.17), we demonstrate that a column of the penalty matrix may be written as a weighted sum of differences of unit vectors plus a weighted unit vector for the magnitude penalty. We may rewrite this weighted sum as a weighted sum of basis functions in (4.18). For a 2D imaging problem these basis functions have the form:

$$b_q(m,n) = \delta(m,n) - \delta(m + m_q, n + n_q), \quad (4.20)$$

where (m,n) are image coordinates and (m_q, n_q) are coordinate offsets for the qth

neighbor (basis). There is one additional basis function for the magnitude weighting that is simply a discrete delta function. Thus, in (4.18) we write the parameterization as a weighted sum over the B bases, which are lexicographically reordered into vector form (as denoted by vec$\{\cdot\}$). The bases are shifted by (m_j, n_j) which represent the coordinates of the jth pixel. In (4.18), l_{jq} represents the vector position which corresponds to the pixel identified by position (m_j, n_j) and the offset (m_q, n_q). This sum may also be written succinctly, as in (4.19), in a matrix form using a $P \times B$ basis matrix, $\boldsymbol{B}^j$, and a vector $\underline{w}^j$ that is composed of the weights $\{w_{l_{jq},j}\}_{q=1}^{B-1}$ and w_j.

The basis representation is an important form for representing $\boldsymbol{R}$, since one can think of the local weightings, $\underline{w}^j$, as local filter coefficients. Penalty design can then be thought of as local filter design, rather than the design of a large $\boldsymbol{R}$ matrix. For example, the Hessian of the conventional 2D first-order shift-invariant penalty can be represented using the following filter:

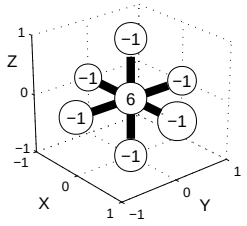
$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{4.21}$$

This filter can be represented using the following first-order basis set:

$$b_{(-1,0)}=\begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \; b_{(1,0)}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \; b_{(0,-1)}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \end{bmatrix} \; b_{(0,1)}=\begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{4.22}$$

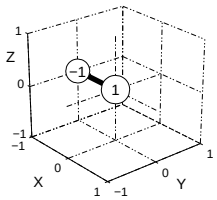
and equal weightings. If $\boldsymbol{B}^j$ is formed from the bases in (4.22), then the conventional shift-invariant penalty can be represented by $\{\boldsymbol{B}^j \underline{1}\}_{j=1}^{p}$. This parameterization is easily extended to 3D penalties. For example, the conventional 3D first-order shift-

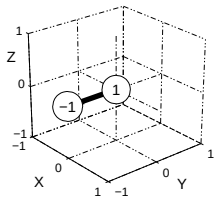
invariant penalty can be represented with the following filter:

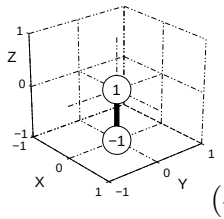


. (4.23)

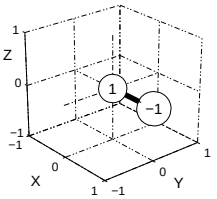
Thus, the following bases can be used to specify 3D penalties over a first-order neighborhood:

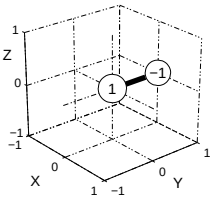
$b_{(-1,0,0)} =$ 

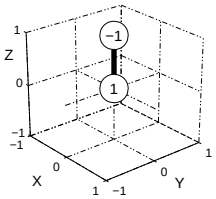
$b_{(0,-1,0)} =$ 

$b_{(0,0,-1)} =$ 

(4.24)

$b_{(1,0,0)} =$ 

$b_{(0,1,0)} =$ 

$b_{(0,0,1)} =$ 

These basis sets can specify a possibly asymmetric penalty. However, recall that only the symmetric portion of a quadratic penalty is important in the penalized-likelihood objective. If a symmetric penalty matrix is required (*e.g.*, for a specific algorithm), one can always obtain the symmetric portion simply by using (4.16).

4.3 Penalty Design for Resolution Control

Having selected the quadratic penalty and a simple way of specifying the penalty coefficients, we next develop a technique for finding those coefficients. Because our local impulse response approximation (3.15) is a function of the measurements, $\underline{Y}$, but not of the object $\underline{\theta}$, we can find approximate local impulse responses *prior* to image reconstruction. Thus, we can use the impulse response estimates to generate a penalty matrix, $\boldsymbol{R}$, that yields user-specified resolution properties.

4.3.1 An Explicit Form for the Penalty Matrix

Recalling the form of the local impulse response in (3.15), we wish to select a penalty matrix such that the local impulse responses equal a desired response, $\underline{l}_0^j$. Although one can evaluate local impulse responses in (3.15) or (3.20) for any spatial coordinates denoted by the continuous variable $\underline{x}_0$, for penalty design we would like to implement a design over a finite set of positions. For example, for a pixel basis representation of the object, we can consider a single local impulse response for each pixel. Thus, for each location ideally we would like to choose $\boldsymbol{R}$ such that

$$\underline{l}(\underline{x}_j) \triangleq [\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}_{\text{sym}}]^{-1}\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j} = \underline{l}_0^j, \tag{4.25}$$

where $\underline{x}_j$ denotes the coordinates of the j voxel. Letting $\boldsymbol{L}_0$ denote the matrix of desired responses for all voxel locations such that $\boldsymbol{L}_0\underline{e}^j = \underline{l}_0^j$, and assuming that $\boldsymbol{H}\underline{e}^j \approx \mathcal{H}\delta_{\underline{x}_j}$, we may write

$$[\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}_{\text{sym}}]^{-1}\boldsymbol{H}'\boldsymbol{D}_2\boldsymbol{H} = \boldsymbol{L}_0. \tag{4.26}$$

Following [40], we may solve for the penalty matrix that yields the collection of desired responses. Assuming the appropriate matrix inverses exist, we may write

$$\begin{aligned}
\boldsymbol{H}'\boldsymbol{D}_2\boldsymbol{H} &= [\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}_{\text{sym}}]\boldsymbol{L}_0 \\
\boldsymbol{H}'\boldsymbol{D}_2\boldsymbol{H} - \boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\boldsymbol{L}_0 &= \boldsymbol{R}_{\text{sym}}\boldsymbol{L}_0 \\
\boldsymbol{R}_{\text{sym}}^{\text{soln}} &\triangleq \boldsymbol{H}'\boldsymbol{D}_2\boldsymbol{H}\boldsymbol{L}_0^{-1} - \boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}.
\end{aligned} \tag{4.27}$$

While (4.27) is a theoretically attractive closed form for specifying the penalty matrix for a given set of desired responses, there are a number of problems. First, the solution in (4.27) will only exist if the right-hand side is symmetric and the appropriate invertibility conditions hold. Thus, there are some desired responses that are

impossible to attain. Second, due to the size of the matrices involved, (4.27) does not generally represent a practically implementable design. (Although, one might be able to derive an iterative technique to solve for $\boldsymbol{R}_{\text{sym}}$.) Similarly, practical penalties have small order neighborhoods so that the penalty portion of the penalized-likelihood objective function is not too expensive to calculate. No sparsity conditions are imposed on $\boldsymbol{R}_{\text{sym}}$ in (4.27). Therefore, while the above explicit formula is attractive for theoretical investigations, it does not represent practical penalty design.

4.3.2 Fitting a Desired Response

Another approach is to set up an objective function, where one attempts to fit a desired response for a given parameterization of the penalty matrix. Thus, rather than attempting to find an impractically large penalty matrix, we may choose weights in a parameterized penalty that best fit the user-specified desired responses.

Mathematically, if we consider the local impulse response to be a function of the penalty matrix, $\boldsymbol{R}$, we would like to find

$$\hat{\boldsymbol{R}} = \arg\min_{\boldsymbol{R}\geq\boldsymbol{0}} \sum_{j=1}^{p} d(\underline{l}^j(\boldsymbol{R}), \underline{l}_0^j), \tag{4.28}$$

where

$$\underline{l}^j(\boldsymbol{R}) \triangleq \underline{l}(\underline{x}_j; \boldsymbol{R}) = [\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}_{\text{sym}}]^{-1}\boldsymbol{H}\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j} \tag{4.29}$$

and $d(\underline{l}^j, \underline{l}_0^j)$ is some measure of disparity between the local impulse response, $\underline{l}^j$, and a desired space-invariant response,[1] $\underline{l}_0^j$. Moreover, we have constrained this minimization such that the penalty matrix is nonnegative definite ($\boldsymbol{R} \geq \boldsymbol{0}$). As discussed in Section 2.5.1, a nonnegative definite penalty will guarantee unique solutions when used with convex likelihood functions. For quadratic penalties this translates into a

[1] One might choose a space-variant $\underline{l}_0^j$ for user-specified nonuniform resolution properties. For a desired space-invariant response $\underline{l}_0^j$ is a function of the pixel position j only in that the desired response must be centered at pixel j. That is, since the local impulse response at pixel j is centered at pixel j, we must shift the desired response to that location for comparison using $d(\cdot,\cdot)$.

nonnegative definite constraint on the penalty matrix, $\boldsymbol{R}$, which may also be stated as a (complicated) set of constraints on the elements of $\boldsymbol{R}$.

As mentioned in Section 4.3.1, some desired responses are unachievable. Similarly, if one constrains $\boldsymbol{R}$ to represent only a small order neighborhood, the set of achievable responses is also reduced. Similarly, in trying to fit a desired response, some desired responses are easier to fit than other responses. While it may be fairly easy to match penalized-likelihood responses throughout an image with the response at the center of the field of view, it may be difficult to match an arbitrary response without using very large neighborhoods. In Chapter V we will discuss a class of "natural" responses that are relatively easy to fit.

Adopting the penalty parameterization of Section 4.2.3, in which the penalty matrix is fully specified by local coefficients, $\{\underline{w}^j\}$, means we may write the penalty design as

$$\{\hat{\underline{w}}^k\}_{k=1}^p = \arg \min_{\{\underline{w}^k\}_{k=1}^p \geq \underline{0}} \sum_{j=1}^p d(\underline{l}^j(\{\underline{w}^k\}_{k=1}^p), \underline{l}_0^j). \tag{4.30}$$

This minimization has the advantage of reducing the dimension of the design problem. However, the design objective is slightly different than the one posed in (4.28). In (4.30), a nonnegative definite penalty matrix, $\boldsymbol{R}$, is enforced by requiring the local penalty coefficients to be nonnegative. This is a sufficient constraint but not a necessary constraint. (One can see how this forces the penalty to be nonnegative by looking at the first line of (4.15). Only negative weights could possibly lead to a negative overall penalty.) Simply restricting the weights to be nonnegative is a constraint that is easy to implement; however, we discuss other constraints in Section 4.4 that might lead to better fits in the penalty design.

Unfortunately, solving either (4.28) or (4.30) appears to be computationally intractable, since all coefficients (or equivalently the entire penalty matrix) must be

solved for simultaneously. It may be feasible to develop iterative routines to solve these objectives (particularly in the case of (4.30) where the constraints are local). However, it would be advantageous to find an approximate method that yields results more quickly.

Similarly, although the local impulse response in (4.29) may be evaluated using iterative techniques[41], we would like to evaluate local responses over many locations and would prefer a faster approximate technique for the purpose of penalty design.

4.3.3 Penalty Design using a Circulant Approximation

Because $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}$ is generally *locally* space-invariant, we use the following circulant approximation (as in [116] and [96]) to the local impulse response at the jth pixel:

$$\underline{l}^j \approx \underline{l}^j_{\text{circ}} \triangleq \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\{\underline{e}^j\} \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\}}{\mathfrak{F}\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\} + \mathfrak{F}\{\boldsymbol{R}_{\text{sym}}\underline{e}^j\}}\right\}, \tag{4.31}$$

where $\odot$ denotes element-by-element multiplication and the division is an element-by-element division. The function $\mathfrak{F}\{\cdot\}$ takes the 2D or 3D Fourier transform of its argument. Such circulant approximations have been used for preconditioning iterative algorithms, and rely on the fact that the 2D or 3D Fourier transform of a column of a matrix diagonalizes a matrix that is doubly or triply block circulant. Thus, one can find the eigenvalues of a circulant matrix and easily perform operations like matrix inversion.

Returning briefly to the PET phantom investigated in Section 3.4.1, we compare local impulse responses calculated iteratively and from the circulant approximation in (4.31). The results of this investigation are summarized in Figure 4.2. The contour lines of the responses calculated by the two methods are nearly identical for many of the response locations. The relatively minor mismatches for some responses are concentrated at the edge and outside the object.

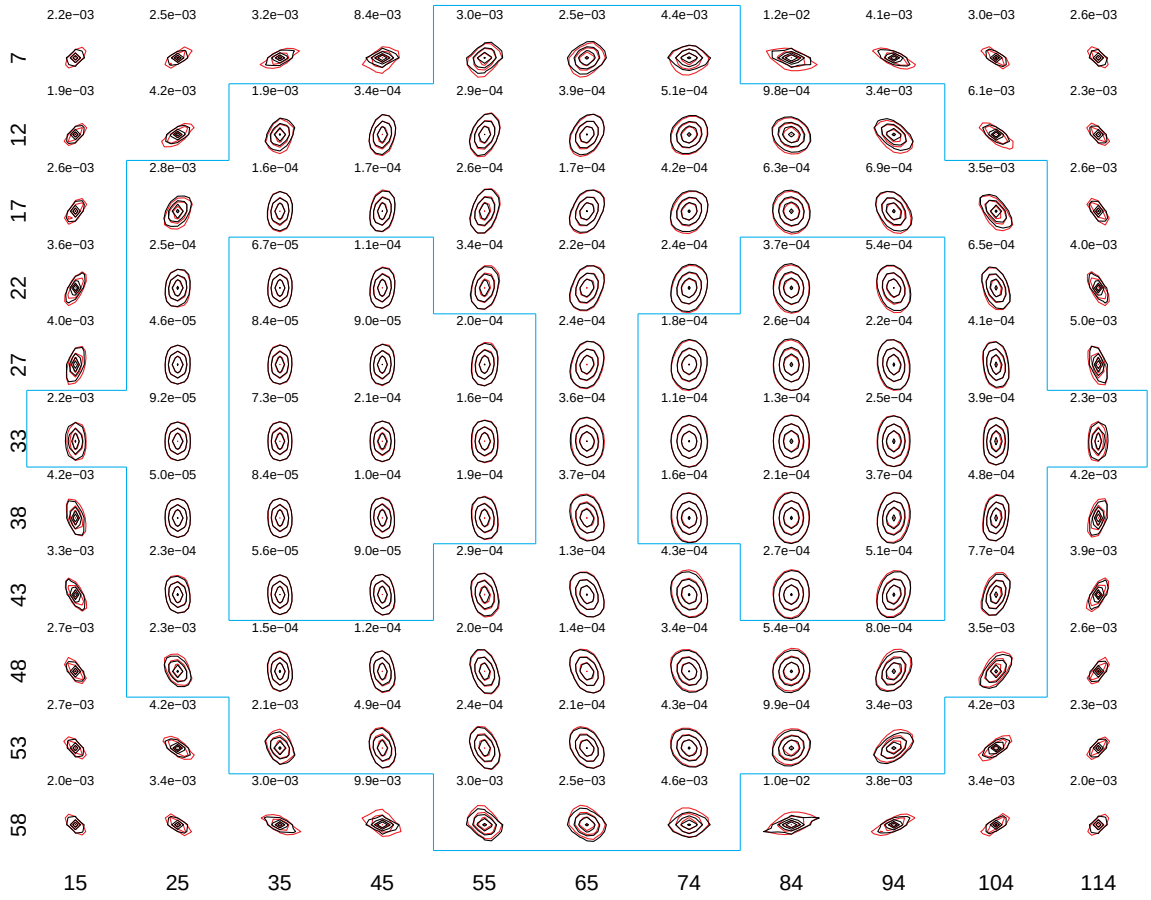


Figure 4.2: A comparison between iteratively evaluated local impulse responses and responses calculated using the circulant approximation.

This figure reproduces the iteratively evaluated local impulse responses shown in Figure 3.6 with the contours shown in red. In addition, the same responses calculated using the circulant approximation in (4.31) are shown in black. Not only are the responses very similar with overlapping contour lines, the mean squared differences between the two estimates (shown above each response) are remarkably small.

One can calculate (4.31) relatively quickly for any j using fast Fourier transforms (FFTs). This circulant approximation includes the term, $\mathfrak{F}\{\underline{e}^j\}$, which includes the appropriate complex exponentials such that the response is centered at the jth position, and $\underline{l}^j_{\text{circ}} \approx \underline{l}^j$.

Recalling the basis representation of the penalty matrix in (4.19), we use the substitution[2] $\boldsymbol{R}_{\text{sym}}\underline{e}^j \approx \boldsymbol{B}^j\underline{w}^j$ in (4.31). We may then write the local impulse response

[2]Strictly speaking, this substitution does not yield a symmetric $\boldsymbol{R}$. However, one may calculate a symmetric $\boldsymbol{R}$ after the design, or simply apply an asymmetric $\boldsymbol{R}$, since only the symmetric portion is important for penalties of

as a function of only the local weightings, $\underline{w}^j$, and use this in a separable design objective. Specifically, we may design penalty coefficients locally using

$$\hat{\underline{w}}^j = \arg\min_{\underline{w}^j \geq \underline{0}} d(\underline{l}^j_{\text{circ}}(\underline{w}^j), \underline{l}^j_0), \tag{4.32}$$

with

$$\underline{l}^j_{\text{circ}}(\underline{w}^j) = \mathfrak{F}^{-1}\left\{ \frac{\mathfrak{F}\{\underline{e}^j\} \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\}}{\mathfrak{F}\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\} + \mathfrak{F}\{\boldsymbol{B}^j\underline{w}^j\}} \right\}. \tag{4.33}$$

Because (4.32) may be evaluated successively for each position j, the penalty matrix can be formed from these individual designs. Thus, (4.32) represents a computationally feasible design technique. For example, choosing the distance metric to be the sum of the squared differences between the responses, one can solve a constrained nonlinear least-squares problem for every pixel or voxel to design the entire quadratic penalty. We denote this penalty design as the constrained nonlinear least-squares (CNLLS) design.

While performing these nonlinear designs is feasible, in practice it typically takes more time than we would like. Since the computation time scales directly with the number of pixels, we would like the local optimization problems to be evaluated as quickly as possible. While we do investigate the performance of the CNLLS design in Section 6.2.1, there are a number of simplifications that can help to reduce the computational burden significantly.

4.3.4 Linearized Penalty Design

Because the designs specified by (4.32) are nonlinear, iterative methods generally must be used to solve the minimization. We would prefer to use a linear least-squares objective that is much easier to solve.

---

the form $\frac{1}{2}\underline{\theta}'\boldsymbol{R}\underline{\theta}$.

Using the circulant approximation to the local impulse response in (4.33), we would like to perform a design (choose $\underline{w}^j$) such that

$$\underline{l}^j_{\text{circ}}(\underline{w}^j) = \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\left\{\underline{e}^j\right\} \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\}}{\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\} + \mathfrak{F}\left\{\boldsymbol{B}^j\underline{w}^j\right\}}\right\} \approx \underline{l}^j_0. \tag{4.34}$$

As introduced in [116], we move the $\underline{w}^j$ terms out of the denominator by Fourier transforming both sides of (4.34) and cross-multiplying to obtain:

$$\underline{L}^j_0 \odot \mathfrak{F}\left\{\boldsymbol{B}^j\underline{w}^j\right\} \;\approx\; \mathfrak{F}\left\{\underline{e}^j\right\} \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\} - \underline{L}^j_0 \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\}, \tag{4.35}$$

where $\underline{L}^j_0 = \mathfrak{F}\left\{\underline{l}^j_0\right\}$ represents the Fourier transform of the desired response, $\underline{l}^j_0$. The form of (4.35) suggested that we could design the local penalty weights, $\underline{w}^j$, using the following constrained, weighted least-squares approach[116]:

$$\underline{\hat{w}}^j = \arg\min_{\underline{w}^j \geq \underline{0}} \left|\left|\boldsymbol{\Phi}^j\underline{w}^j - \underline{\alpha}^j\right|\right|^2, \tag{4.36}$$

with

$$\boldsymbol{\Phi}^j \;\triangleq\; \boldsymbol{W}^j\text{diag}\left\{\underline{L}^j_0\right\}\mathfrak{F}\left\{\boldsymbol{B}^j\right\} \tag{4.37}$$

$$\underline{\alpha}^j \;\triangleq\; \boldsymbol{W}^j\text{diag}\left\{\mathfrak{F}\left\{\underline{e}^j\right\}\right\}\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\} - \boldsymbol{W}^j\text{diag}\left\{\underline{L}^j_0\right\}\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\}, \tag{4.38}$$

where $\boldsymbol{W}^j$ represents a user-selected, nonnegative-definite, least-squares weighting that could possibly be space-variant. The penalty design in (4.36) is a linearly constrained linear least-squares problem, which may be solved using the nonnegative least-squares (NNLS) algorithm[71]. This is an iterative algorithm; however, it is guaranteed to converge in a finite number of iterations. (Specifically, in $\leq 2^B$ iterations, where B is the number of constrained coefficients.) One could also use suboptimal approaches, like the one in Table 4.1, that find approximate solutions more quickly.

Table 4.1: Suboptimal greedy routine used to constrain penalty coefficients.

Set $\mathbf{\Psi}^j = [\mathbf{\Phi}^j]'\mathbf{\Phi}^j$ and $\gamma^j = [\mathbf{\Phi}^j]'\underline{\alpha}^j$.
Let $\underline{\hat{w}}^j = [\mathbf{\Psi}^j]^{-1}\gamma^j$.
while $\underline{\hat{w}}^j$ contains negative values,
    Let i equal the index of the minimum value of $\underline{\hat{w}}^j$.
    Remove the ith row and ith column of $\mathbf{\Psi}^j$.
    Remove the ith row from γ^j.
    Set the ith element of $\underline{\hat{w}}^j$ to zero.
    Find the remaining elements of $\underline{\hat{w}}^j$ by $[\mathbf{\Psi}^j]^{-1}\gamma^j$.
end

The penalty design described in (4.36) represents the final form in developing a least-squares design. However, for typical applications, straightforward evaluation of (4.36) for every pixel generally requires significantly more computation time than it takes to solve the actual image reconstruction problem. Therefore, for practical use, it is desirable to find an efficient procedure for computing the penalty. The practical implementation of the penalty design is discussed at length in Chapter V. However, before we discuss the practical implementation, we discuss some potential improvements to the least-squares design in (4.36).

4.4 Relaxed Design Constraints

While the previous sections discuss a practical design, they all rely on a design constraint that forbids negative coefficients. Using these methods that individually constrain the interpixel weightings to be nonnegative, we found that in many cases (particularly for large neighborhood penalties) the nonnegativity constraint can be quite active. In looking at the fits of actual local impulse responses to the target responses, we find that, in some cases, even when a particular interpixel weighting is zero, the estimator can still induce too much smoothness between pixels. Thus, we have investigated so-called relaxed design constraints that increase the size of the feasible parameter space in an attempt to perform more flexible designs[117].

Recall from Section 4.3, the penalty design is stated in terms of the minimization of some objective function. We would like to ensure the convexity of the penalty function $R(\underline{\theta})$, so that the image reconstruction problem has unique solutions. For quadratic penalties this can be expressed as a nonnegative definiteness constraint on $\boldsymbol{R}$, or equivalently as a nonnegativity constraint on the eigenvalues of $\boldsymbol{R}$. Thus, for a particular penalty design objective function, $\Upsilon(\boldsymbol{R})$, we may write

$$\hat{\boldsymbol{R}} = \arg \min_{\text{eig}(\boldsymbol{R}) \geq \underline{0}} \Upsilon(\boldsymbol{R}). \tag{4.39}$$

This eigenvalue constraint does not preclude negative pairwise weights between pixels and will yield increased design freedom over the individual nonnegativity constraints adopted in the previous section.

The main problem with the formulation presented in (4.39) is that the minimization will generally be impractical due to the size of $\boldsymbol{R}$. Evaluation of the constraint, $\text{eig}(\boldsymbol{R})$, and possibly the cost, $\Upsilon(\boldsymbol{R})$, will generally be too computationally intensive for most applications.

In the case of a shift-invariant penalty, $\boldsymbol{R}$ is block circulant and its eigenvalues may be computed using fast Fourier transforms. It is straightforward to formulate a shift-invariant "toy" design problem where the *Fourier constraints* may be applied. Specifically, we have formed a shift-invariant problem using the circulant approximation in (4.31). That is, for any given location j, there is an analogous shift-invariant problem to which (4.31) corresponds.

Figure 4.3 shows a summary of this "toy" problem investigation. We have purposely chosen a location in the shift-variant problem where we know the nonnegatively constrained penalty design of (4.36) has difficulty achieving the desired uniform response. Thus, the shift-invariant "toy" problem should have similar problems. The right half of this figure shows contours of the resulting impulse response functions

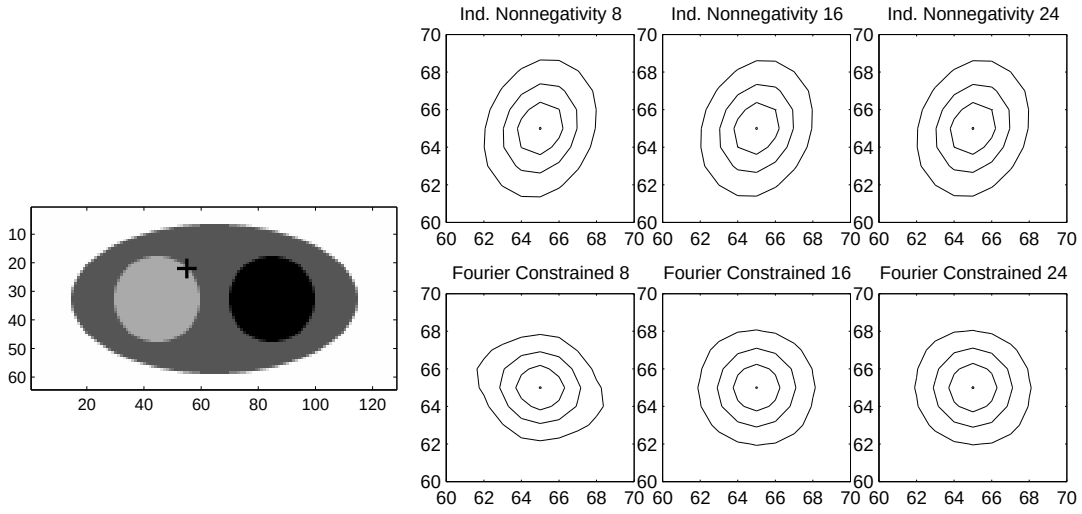
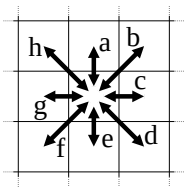


Figure 4.3: Design of a circulant penalty for a "toy" PET problem.
At left the figure shows a difficult design location in the PET phantom study of Figure 3.5. The local circulant approximation at this location is used to construct a difficult shift-invariant "toy" design problem. At right we show local impulse responses arising from penalty designs using the simple nonnegativity constraints and the Fourier constraints, with penalties using 8, 16, and 24 neighboring pixels. The target response for the design is isotropic.

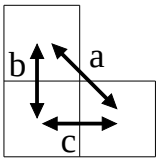
designed using the individual nonnegativity constraints and the Fourier constraints. Two things are immediately evident in this figure: (1) the Fourier constraints lead to more uniform responses, and (2) increasing the size of the penalty neighborhood does not improve uniformity for the individually applied nonnegativity constraints. Clearly the incorporation of negative weights has allowed for greater design freedom.

The Fourier constraints are just one way of constraining a shift-invariant $\boldsymbol{R}$. An alternative is to use simpler constraints such as those derived by Lakshmanan for 2D Gaussian Markov random fields[66, 67]. Unfortunately, it is unclear how to extend either these constraints or the Fourier constraints to the shift-variant case. More general methods that bound the distance to the nearest singular methods for perturbations of single components of $\boldsymbol{R}$ have been developed[101]. However, these bounds generally depend on all of the elements of $\boldsymbol{R}$ and will typically lead to impractical constraints.



$$\begin{array}{lll} a \geq 0 & b \geq 0 & c \geq 0 \\ d \geq 0 & e \geq 0 & f \geq 0 \\ g \geq 0 & h \geq 0 & \end{array}$$

Figure 4.4: Pointwise constraints for a single pixel with eight neighbors/interpixel weights.



$$\boldsymbol{R} = \begin{bmatrix} a+b & -a & -b \\ -a & a+c & -c \\ -b & -c & b+c \end{bmatrix}$$

Figure 4.5: A three pixel image and its penalty matrix.

As in Section 4.3.3, one typically wants to develop a shift-variant penalty by performing a local design. That is, one would like to determine the weights in a pixel-by-pixel fashion, rather than all weights simultaneously. For example, at a given pixel, one would like to determine all the weights between that pixel and its neighbors (see Figure 4.4). Unfortunately, the only pointwise constraint is the individual nonnegativity constraint. Thus, one needs to consider groups of pixels to incorporate negative weights.

Consider the small three pixel image shown in Figure 4.5. There are three weights associated with the three pixel pairs, labeled a, b, and c. (We have restricted ourselves to interpixel weightings without any magnitude penalty.) The penalty matrix for this image is also shown. Finding the characteristic polynomial for $\boldsymbol{R}$ and applying the Routh-Hurwitz criterion[143], it is straightforward to derive the following constraints on the weights themselves:

$$\begin{aligned} a + b + c &\geq 0 \\ ab + bc + ac &\geq 0. \end{aligned} \tag{4.40}$$

These constraints allow for at most one of the weights to be arbitrarily negative, as

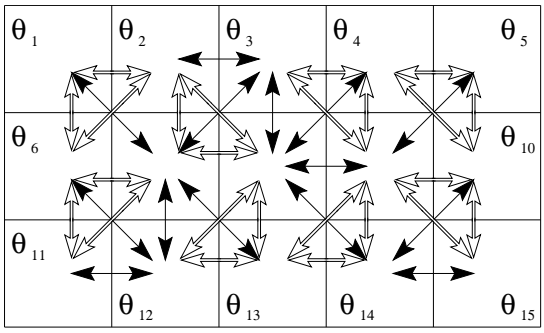


Figure 4.6: Illustration of how constraints over loops of three weights may be used to ensure nonnegative definiteness of the penalty matrix.

long as the other two weights are sufficiently large and positive.

While one would rarely deal with an image this small, the constraints found here can still be quite useful. Since the sum of nonnegative definite functions is nonnegative definite, one can break the summation in (4.15) into more manageable portions and satisfy nonnegative definiteness constraints locally. Specifically, using the constraints in (4.40), one can satisfy a nonnegative definiteness constraint on any sum of three weights in a large image, provided they form a loop.

A sample application of the constraints in (4.40) applied to a larger image is shown in Figure 4.6. All of the weights represented by white arrows form loops of three weights and must satisfy the constraints in (4.40). The remaining weights (black arrows) are not part of a loop constraint and, thus, must satisfy the usual individual nonnegativity constraint. Thus, the nonnegative definiteness of $\boldsymbol{R}$ can be guaranteed, the weights are locally constrained (allowing some form of local design), and negative weights are allowed. (Again, we have ignored any magnitude penalty. However, those too can easily be incorporated using simple nonnegativity constraints.)

These constraint loops may be chosen somewhat arbitrarily, as long as each weight is constrained exactly once (using either (4.40) or the simple individual nonnegativity constraint). Clearly, the number of ways to choose these loops increases tremendously

with the number of pixels. The (impractical) optimal solution is to optimize over all possible loop configurations and select the set that yields the best $\boldsymbol{R}$ according to the penalty design objective, $\Upsilon(\boldsymbol{R})$.

The best way to choose these loops will be dependent on the specific penalty design goal. We would like to be able to perform the penalty design using the least-squares objective in (4.36). Specifically,

$$\underline{\hat{w}}^j = \arg\min_{\underline{w}^j \in \mathfrak{C}} ||\boldsymbol{\Phi}^j \underline{w}^j - \underline{\alpha}^j||^2, \tag{4.41}$$

where $\mathfrak{C}$ denotes the feasible region using whatever particular combination of loop constraints and nonnegativity constraints have been adopted.

Thus, to choose loops we have adopted the following heuristics:

- Calculate the unconstrained local solution, $\underline{w}^j_{\text{uc}}$, to (4.41) for each pixel j.
- Choose only from loops that include the most negative element of $\underline{w}^j_{\text{uc}}$.
- Select from remaining loops by finding the loop that allows for the most negative weight. (Plug in the unconstrained solutions for the two positive values in (4.40) and find the bounds on the remaining weight.)

While these heuristics do not necessarily yield an optimal choice for the weight constraints, such choices should generally increase design flexibility and allow for the most important negative (*i.e.*, the most negative weight in the unconstrained problem) to go negative in the constrained problem. Because neighboring pixel locations share interpixel weighting, this selection task is somewhat difficult. In practice, we have used a greedy approach where the above heuristics are applied over a grid of locations with no shared weightings, then a "second pass" is performed where the remaining constraints are chosen using the same heuristics, but also as to not interfere with the previous constraint choices.

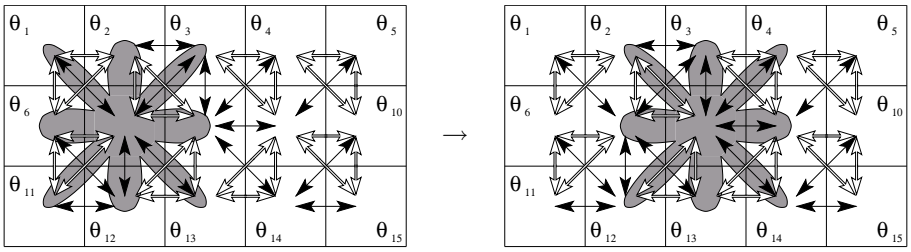


Figure 4.7: An illustration of the update approach for penalty design with relaxed constraints. In the first image a pointwise design is used to update only the interpixel weights lying in the gray region. Nearby weights (not in the pointwise design) that are used to constrain the design are held constant. In the following image, the pointwise design is applied to the next pixel in the sequence, cycling through all pixel positions.

Once a set of constraints has been chosen (*i.e.*, a "map" such as Figure 4.6 is available), one must still find $\boldsymbol{R}$. Unfortunately, the minimization shown in (4.41) couples all $\{\underline{w}^j\}$ through the constraints. Thus, this minimization should technically be solved simultaneously for all pixel positions. We have opted to use an iterative approach where all non-local weights are held constant, the constrained (4.41) is minimized using a sequential quadratic programming algorithm[111, 110], and the local solution is used as an update to the current estimate of $\boldsymbol{R}$. This approach is illustrated in Figure 4.7. We cycle through all pixel positions until the weightings appear to have sufficiently "converged."

This method has not been proven to converge. However, consider Figure 4.8. If we initialize the above procedure using the nonnegatively constrained solution (point B), and step slowly toward the solution to the relaxed constraint design (point C) at a given position as we cycle through pixel positions, we should increase the chances that monotonic updates to the weights are applied. In practice this method appears to "converge" to solutions with lower costs than the design in (4.36) with the individual nonnegativity constraints. Thus, these relaxed constraints may be used for more flexible penalty design. We demonstrate their use in Section 6.2.1. Unfortunately, these methods significantly complicate the penalty design as compared with

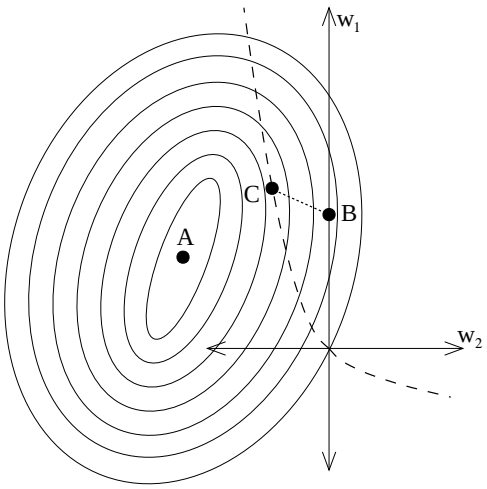


Figure 4.8: Illustration of the solution to the relaxed constraint design.
Because the penalty design objective function in (4.36) is a least-squares cost function, the unconstrained solution (A) lies at the bottom of a parabola. While the nonnegatively constrained solution (B) can be found relatively easily, the relaxed constraint solution (C) generally requires iterative methods.

(4.36), because of the difficulty in specifying the constraints and the nonlinearities introduced into the optimization.

We note that even more relaxed constraints may be obtained by selecting a subimage of weightings that is larger than those presented in Figure 4.5, and finding the corresponding characteristic polynomial and applying the Routh-Hurwitz criterion. However, for larger numbers of weights, the constraints analogous to those in (4.40) quickly become very complicated, which makes the selection of those constraints and the optimization even more difficult.

4.5 Summary

In this chapter, we have discussed and developed a number of penalty design approaches to control the resolution properties of the reconstructed images from penalized-likelihood estimators. These techniques are based on a circulant approximation of the local impulse response that allows one to write the local resolution properties as a function of local penalty parameters. A global penalty is designed

by systematically (over all image locations) choosing the local penalty penalty parameters so that the local impulse responses are approximately equal to some set of desired responses.

We have identified three specific design techniques: 1) the CNLLS design which is based on a nonnegatively constrained nonlinear least-squares design objective, 2) the nonnegatively constrained linearized design expressed in (4.36), and 3) the linearized design with relaxed design constraints from Section 4.4. We apply these design approaches in Chapter VI and discuss the relative performance of these and other methods, when the goal is uniform resolution.

While the mathematical forms of these penalty design methods have been developed in this chapter, these forms generally do not relate directly to computationally practical techniques. In the next chapter, we discuss methods that can be used (particularly for tomographic systems) that allow for fast penalty design.

CHAPTER V

Rapid Calculation of Resolution and Covariance

While the local impulse response derived in Section 3.3 serves as an important formula for predicting resolution, typically it must be evaluated iteratively. Thus, when many resolution prediction need to be made, or when resolution predictors need to be made repeatedly for a given system (as with the penalty design discussed in Chapter IV), it is important to have fast routines for evaluating these predictions.

Because covariance predictors are very similar in form to the resolution predictors, many fast techniques that are applicable to one predictor are applicable to the other. Therefore, we briefly review covariance prediction and apply our fast prediction methods to the covariance predictor as well.

In this chapter, we show that rapid calculation of resolution and covariance is intimately tied to fast calculation of weighted projection-backprojections of an impulse. We develop efficient routines for calculating these weighted responses and show how they may be applied to rapid resolution prediction, fast covariance prediction, and efficient penalty design for resolution control. These routines involve precomputing and storing components of the geometric system response and are applicable to 2D and 3D systems that include nonuniform attenuation effects and complicated geometries and detector responses.

5.1 Covariance in Reconstructed Images

Predicting the covariance between pixels in reconstructed images has a history that has closely paralleled resolution predictions. As with resolution, there have been many investigations on the noise properties of images as a function of various estimator and system parameters. For iterative methods, noise has been investigated as function of iteration. For example, Wilson *et al.* studied the image covariance properties for MLEM[132] by reconstructing a large set of emission images, and calculating the empirical sample covariance. Similar noise studies as a function of iteration have been performed for OSEM[68]. Several groups have developed analytic methods that formulate how noise is propagated from iteration to iteration. These studies include techniques for analyzing noise propagation in EM[8], MAP-EM[129], and OSEM[106].

However, as with resolution, instead of focusing on how subsequent iterations effect the image, one can analyze the covariance properties of the solution of an iterative technique. In [33], Fessler derived an expression for the covariance of an implicitly defined estimator like the one defined in (2.24). Adopting our previously defined notation, we restate that predictor here:

$$\mathrm{Cov}\{\hat{\underline{\theta}}\} \;=\; \left[\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}(\breve{\underline{\theta}})\right]^{-1} \boldsymbol{H}'\boldsymbol{D}_3\boldsymbol{H} \left[\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H} + \boldsymbol{R}(\breve{\underline{\theta}})\right]^{-1}, \qquad (5.1)$$

where $\boldsymbol{D}_3 = \boldsymbol{D}_2\mathrm{Cov}\{\underline{Y}\}\boldsymbol{D}_2$ is a diagonal matrix if the measurements are independent.

Because the covariance matrix in (5.1) is generally too large to compute in its entirety, people often focus on calculating a single row or column of the covariance matrix. This row or column is denoted as $\mathrm{Cov}^j\{\hat{\underline{\theta}}\}$, and can be interpreted as the covariance function at position j. Iterative methods to calculate the variance or covariance function at position j were discussed in [33].

5.2 Prior Work in Rapid Resolution and Covariance Prediction

While the explicit formula for the local impulse responses derived in Section 3.3 or a row of the covariance in (5.1) can be evaluated quickly by developing fast iterative methods, we concentrate on noniterative methods that can be used to evaluate the resolution or covariance predictions quickly.

5.2.1 Circulant Approximation

We have already discussed one such approximation. Specifically, the circulant approximation in (4.31). This approximation has been used to provide fast resolution predictions[116], fast covariance predictions[94, 11], and fast predictions of contrast[95]. We assume that we can model the continuous-to-discrete projection well with the discrete model,[1] *i.e.*, $\mathcal{H}\delta_{\underline{x}_j} = \boldsymbol{H}\underline{e}^j$. Then as in (4.31), we may write the circulant approximation of the resolution predictor as

$$\underline{l}^j \approx \underline{l}^j_{\text{circ}} \triangleq \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\left\{\underline{e}^j\right\} \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\boldsymbol{H}\underline{e}^j\right\}}{\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\} + \mathfrak{F}\left\{\boldsymbol{R}(\breve{\underline{\theta}})\underline{e}^j\right\}}\right\}. \tag{5.2}$$

Similarly, using the circulant approximation and applying it to (5.1) yields the following predictor for the covariance function at position j:

$$\text{Cov}^j\{\hat{\underline{\theta}}\} \approx \text{Cov}^j_{\text{circ}}\{\hat{\underline{\theta}}\} \triangleq \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\left\{\underline{e}^j\right\} \odot \mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_3\boldsymbol{H}\underline{e}^j\right\}}{\left|\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\} + \mathfrak{F}\left\{\boldsymbol{R}(\breve{\underline{\theta}})\underline{e}^j\right\}\right|^2}\right\}, \tag{5.3}$$

where the complex exponentials represented by $\left[\mathfrak{F}\left\{\underline{e}^j\right\}\right]^2$ term incorporate the appropriate shifts so that the covariance function is "centered" at location j.

Although (5.2) and (5.3) may be calculated relatively quickly using fast Fourier transform operations, the repeated calculation of the weighted projection-backprojections of a unit vector, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$, can be quite computationally expensive. For example, in

[1] The methods discussed in Section 5.3 can apply equally well to predictors that rely on $\boldsymbol{H}'\boldsymbol{D}\mathcal{H}\delta_{\underline{x}_j}$. However, we develop the methods for $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ only. It is straightforward to obtain the analogous methods for $\boldsymbol{H}'\boldsymbol{D}\mathcal{H}\delta_{\underline{x}_j}$.

fully 3D SPECT, where the matrix $\boldsymbol{H}$ is implemented as an "on-the-fly" routine, the projections and backprojections greatly outweigh the Fourier transforms in terms of the time required to evaluate a resolution or covariance prediction. Similarly, the frequency-domain multiplications and divisions correspond to very little of the total evaluation time. (The multiplication by the complex exponentials like $[\mathfrak{F}\{\underline{e}^j\}]$ can be eliminated by appropriately shifting (permuting) the image-domain vectors.) Thus, for fast predictions it is crucial to be able to provide rapid evaluation of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$.

5.2.2 Approximations Based on "Outer" Diagonalization

We have already discussed one approximation involving $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}$, when $\boldsymbol{H}$ has a PET-style factorization. This approximation arises from the certainty-based penalty developed in [32, 41] and reviewed in Section 4.1.1. When $\boldsymbol{H} = \mathrm{diag}\{c_i\}\,\boldsymbol{G}\mathrm{diag}\{s_j\}$, with $\boldsymbol{G}'\boldsymbol{G}$ approximately block circulant, the following approximation can be made:

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} \approx \boldsymbol{D}_\kappa \boldsymbol{G}'\boldsymbol{G}\boldsymbol{D}_\kappa, \tag{5.4}$$

where $\boldsymbol{D}_\kappa = \mathrm{diag}\{\kappa_j\}$ and

$$\kappa_j \triangleq s_j \sqrt{\frac{\sum_i g_{ij}^2 c_i^2 d_i}{\sum_i g_{ij}^2}}, \tag{5.5}$$

with d_i denoting the ith diagonal element of $\boldsymbol{D}$. Thus, the weighted response is approximated as

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{D}_\kappa \boldsymbol{G}'\boldsymbol{G}\boldsymbol{D}_\kappa \underline{e}^j = \kappa_j \boldsymbol{D}_\kappa \boldsymbol{G}'\boldsymbol{G}\underline{e}^j. \tag{5.6}$$

We categorize this method as an "outer" diagonalization approximation, since all of the object-dependences are moved into pre- and post-multiplications by the diagonal matrix, $\boldsymbol{D}_\kappa$.

Since $\boldsymbol{G}'\boldsymbol{G}$ is approximately shift-invariant for PET systems, one may calculate $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$ for a single position j_0 (perhaps at the center of the image) and the remaining

geometric responses may be formed from shifted versions of $\boldsymbol{G}'\boldsymbol{G}\underline{e}^{j_0}$. Thus, calculating (5.5) once for a given diagonal weighting $\boldsymbol{D}$ and set of PET attenuation factors, $\{c_i\}$, allows one to use (5.6) to form an approximate weighted response, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$, at any location from a shifted $\boldsymbol{G}'\boldsymbol{G}\underline{e}^{j_0}$. This technique is very fast since (5.5) is equivalent to a single backprojection and $\boldsymbol{G}'\boldsymbol{G}\underline{e}^{j_0}$ may be precomputed and stored for a given system geometry. This technique was used successfully to provide approximate variance predictors in [34].

Unfortunately this technique is only very fast for systems with a shift-invariant geometric response, and does not incorporate SPECT-style attenuation factors. Xing *et al.* has endeavored to extend these methods to resolution and variance prediction in SPECT[137, 136], where the SPECT factorization, $\boldsymbol{H} = \mathrm{diag}\{c_i\}\,(\boldsymbol{A} \odot \boldsymbol{G})\mathrm{diag}\{s_j\}$, applies. In this case,

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{D}_{\tilde{\kappa}}\boldsymbol{S}'\boldsymbol{S}\boldsymbol{D}_{\tilde{\kappa}}\underline{e}^j = \tilde{\kappa}_j\boldsymbol{D}_{\tilde{\kappa}}\boldsymbol{S}'\boldsymbol{S}\underline{e}^j, \tag{5.7}$$

where

$$\tilde{\kappa}_j \triangleq \sqrt{\frac{\sum_i h_{ij}^2 d_i}{\sum_i s_{ij}^2}} = s_j\sqrt{\frac{\sum_i g_{ij}^2 a_{ij}^2 c_i^2 d_i}{\sum_i s_{ij}^2}}, \tag{5.8}$$

and $\boldsymbol{S}'\boldsymbol{S}\underline{e}^j$ is an object-independent approximation of the response. For example, one could choose $\boldsymbol{S} = \boldsymbol{G}$, so that one can multiply the geometric response, $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$, by the appropriate $\tilde{\kappa}_j$ to incorporate the effects of the diagonal weighting and SPECT attenuation terms. This is an interesting extension to (5.6), whose computational requirements depend on the exact choice of $\boldsymbol{S}$. (Xing has suggested using an $\boldsymbol{S}$ operator that produces projections and backprojections using the frequency-distance principle[46] to approximate the SPECT detector response.) We will discuss the performance of this approximation in more detail in Section 5.3.2.

5.3 Rapid Calculation of Weighted Projection-Backprojections

In this section we present an alternate way to approximate the weighted response, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$. As with the approximations discussed in Section 5.2.2, the key to rapid evaluation of the weighted response is to separate object-dependent portions of the response from geometric components that may be precalculated or evaluated quickly, and to be able to recombine these separate components rapidly to approximate the response.

In this section we describe a series of approximations that allow many operations to be precomputed for the evaluation of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$, when the system matrix fits a PET or SPECT model. (Portions of this work were originally presented in [118] and [119].) In the special case of a space-invariant system and a PET-style attenuation model, the results simplify to the methods we presented in [116]. These methods should apply to other imaging systems that have a system matrix factorization similar to the object-dependent and object-independent factors discussed in (2.10).

5.3.1 Linear Operators

One important property of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ used in [116] is that it is linear in terms of the diagonal elements of $\boldsymbol{D}$. That is, we may write

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j = \sum_{i=1}^{N} m_i^j [\boldsymbol{D}]_{ii} = \boldsymbol{M}^j \underline{d}, \tag{5.9}$$

where $\underline{m}_i^j$ are position-dependent vectors that are related to $\boldsymbol{H}$. Similarly, we may write this linear combination in terms of a $P \times N$ matrix, $\boldsymbol{M}^j = [\underline{m}_1^j \ldots \underline{m}_N^j]$, and a vector of the diagonal elements of $\boldsymbol{D}$, which are denoted as $\underline{d}$ with $[\underline{d}]_i = [\boldsymbol{D}]_{ii}$.

One could construct $\boldsymbol{M}^j$ using the superposition principle. Specifically, $\underline{m}_i^j$ may be found by applying diagonalized unit vectors for each measurement such that

$$\underline{m}_i^j = \boldsymbol{H}' \mathrm{diag}\left\{\underline{e}^i\right\} \boldsymbol{H}\underline{e}^j. \tag{5.10}$$

In principle, if $\{\boldsymbol{M}^j\}_{j=1}^P$ could be precalculated, then $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ can be evaluated quickly for different diagonal matrices. Unfortunately, there are several problems with this kind of precomputation. Perhaps the most significant problem is that, even if one were to calculate all $\{\boldsymbol{M}^j\}_{j=1}^P$ operators, these linear operators are object-dependent because the SPECT system matrix depends on the attenuation properties of the object. Thus, any such "precalculation" would need to be performed for every object. While one might be able to use a generic attenuation model in cases like brain imaging where there is less variability, we would like to develop an efficient technique that applies to a wide range of attenuating objects.

Another problem is the sheer size of $\{\boldsymbol{M}^j\}_{j=1}^P$. One must be able to store these precomputed linear operators to exploit any computational speed-up. Recall that each operator $\boldsymbol{M}^j$ is $P \times N$ in size. Generally it would be infeasible to store all P operators since they have a similar degree of sparsity as the system matrix, $\boldsymbol{H}$.

We address these issues in the following sections.

5.3.2 Attenuation Approximations

To use the linear operator technique effectively we must eliminate the object-dependence from the precomputed portions of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$. Consider the following factorization of the system matrix:

$$\boldsymbol{H} = \boldsymbol{D}_c \left(\boldsymbol{A} \odot \boldsymbol{G} \right), \tag{5.11}$$

where $\boldsymbol{D}_c$ is a diagonal matrix of the ray-dependent factors, c_i, and $\boldsymbol{A}$ and $\boldsymbol{G}$, are collections of the (object-dependent) attenuation terms, a_{ij}, and (object-independent) geometric terms, g_{ij}, as described in Section 2.2. This factorization is slightly different than the one in (2.10) in that the pixel-dependent factors, s_j, factors have

been eliminated.[2] (Moreover, we do *not* require that geometric operator $\boldsymbol{G}'\boldsymbol{G}$ be a shift-invariant operator.) This factorization isolates all of the object-dependence in the $\boldsymbol{A}$ and $\boldsymbol{D}_c$ terms. Recall that PET attenuation may be modeled in $\boldsymbol{D}_c$ and SPECT attenuation in $\boldsymbol{A}$.

Let $\boldsymbol{F} = \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}$ denote the entire weighted projection-backprojection operator. Using the factorization in (5.11), we make the following sequence of observations regarding the (k, j)th element of $\boldsymbol{F}$:

$$\begin{aligned}
[\boldsymbol{F}]_{kj} &= (\underline{e}^k)'\boldsymbol{F}\underline{e}^j \\
&= (\underline{e}^k)'\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \\
&= (\underline{e}^k)'\left[\boldsymbol{D}_c(\boldsymbol{A}\odot\boldsymbol{G})\right]'\boldsymbol{D}\left[\boldsymbol{D}_c(\boldsymbol{A}\odot\boldsymbol{G})\right]\underline{e}^j \\
&= \left[(\boldsymbol{A}\underline{e}^k)'\odot(\boldsymbol{G}\underline{e}^k)'\right]\boldsymbol{D}_c\boldsymbol{D}\boldsymbol{D}_c\left[(\boldsymbol{A}\underline{e}^j)\odot(\boldsymbol{G}\underline{e}^j)\right] \\
&= (\boldsymbol{G}\underline{e}^k)'\mathrm{diag}\left\{\boldsymbol{A}\underline{e}^k\right\}\boldsymbol{D}_c\boldsymbol{D}\boldsymbol{D}_c\mathrm{diag}\left\{\boldsymbol{A}\underline{e}^j\right\}\boldsymbol{G}\underline{e}^j \\
&= (\underline{e}^k)'\boldsymbol{G}'\boldsymbol{D}^{jk}\boldsymbol{G}\underline{e}^j, \qquad (5.12)
\end{aligned}$$

where the diagonal matrix, $\boldsymbol{D}^{jk}$, has the following elements:

$$\begin{aligned}
\left[\boldsymbol{D}^{jk}\right]_{ii} &= [\boldsymbol{A}\underline{e}^k]_i[\boldsymbol{D}]_{ii}[\boldsymbol{D}_c]^2_{ii}[\boldsymbol{A}\underline{e}^j]_i \\
&= c_i^2 a_{ij}a_{ik}[\boldsymbol{D}]_{ii}. \qquad (5.13)
\end{aligned}$$

Because $\boldsymbol{A}\underline{e}^k$ generally varies relatively smoothly with changing k, and $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ is fairly concentrated about the pixel position j, we use (5.12) and (5.13) to make the following approximation:

$$\boldsymbol{F}\underline{e}^j = \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{G}'\boldsymbol{D}^j\boldsymbol{G}\underline{e}^j, \qquad (5.14)$$

with elements of the diagonal matrix, $\boldsymbol{D}^j$, defined as

$$\left[\boldsymbol{D}^j\right]_{ii} = \left[\boldsymbol{D}^{jj}\right]_{ii} = c_i^2 a_{ij}^2[\boldsymbol{D}]_{ii}. \qquad (5.15)$$

[2] If the s_j terms are object-independent, they may be easily absorbed into $\boldsymbol{G}$. Otherwise, these terms may be placed in the object-dependent $\boldsymbol{A}$.

Thus, we approximate $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ using only the geometric model $\boldsymbol{G}$ and a position-dependent diagonal weighting $\boldsymbol{D}^j$. This approximation is exact at location j and yields very good results for the neighborhood around j. Because the SPECT attenuation terms, $\{a_{ij}\}$, are formed from the integral in (2.1), even discontinuous attenuation maps will yield smoothly varying attenuation terms. Thus, the above approximation can perform well for nonuniformly attenuating objects. For most PET systems, $\boldsymbol{A} = \mathbf{1}$, and (5.14) is an equality, not an approximation.

Brief Aside: "Inner" versus "Outer" Diagonalizations

In contrast to the "outer" diagonalizations discussed in Section 5.2.2, the approximation for $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ shown in (5.14) moves all of object-dependence into the "inner" diagonal term. We perform a brief investigation of these two different kinds of approximation. Figure 5.1 shows the results of this investigation. Specifically we have compared the "inner" diagonal approximation[3] in (5.14) and the "outer" diagonal approximation of (5.7) to the unapproximated response, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$.

In this study we used the SPECT model and emission map used in Section 3.2.3 and Figure 3.3. We used a nonuniform attenuation map with linear attenuation coefficients appropriate for water for the background disc, and appropriate for air for the cold "rods." Weighted responses for three different positions in the phantom are shown in Figure 5.1. The left column shows the unapproximated responses, $\boldsymbol{H}'\boldsymbol{W}\boldsymbol{H}\underline{e}^j$, the center column shows the responses as calculated using the "inner" diagonalization in (5.14), and the right column shows the responses approximated by the "outer" diagonalization method in (5.7).

While both approximation methods exactly match the magnitude of the response

[3]Strictly speaking, we have also used the projection-constant approximation discussed in Section 5.3.4, which should typically degrade the accuracy of the approximation. However, despite using this additional approximation, the comparison between the "inner" diagonal approach and the actual response remains quite close.

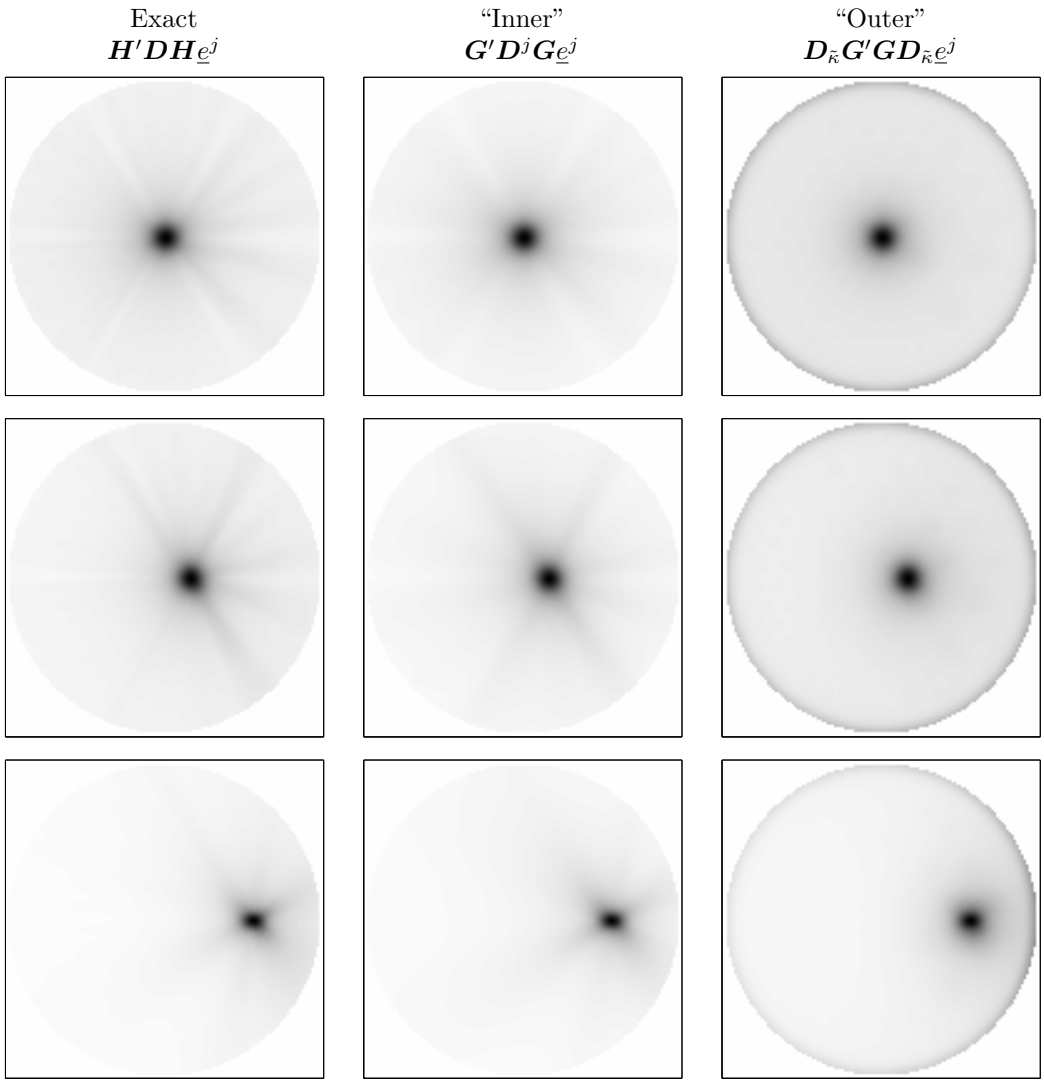


Figure 5.1: Comparison of "inner" and "outer" diagonalization approximations for calculating weighted responses for SPECT.
The left column shows the unapproximated weighted responses for three locations for a simulated phantom with the same emission distribution as the one in Figure 3.3, but with nonuniform attenuation. The center column of responses is computed using the "inner" diagonalization of (5.14), and the right column is computed using the "outer" diagonalization in (5.7). Images in each row use an identical colormap window.

peak, the "inner" diagonalization yields a better approximation of the overall response. For example, much of the anisotropy in the weighted response is not captured by the "outer" diagonalization method. This is particularly noticeable for the two off-center responses which have a higher degree of anisotropy due to attenuation and the diagonal weighting, $\boldsymbol{D}$.

Return to Linear Operators

Returning to (5.14), since the right-hand side of the approximation is a linear function of the diagonal elements of $\boldsymbol{D}^j$, we may now calculate approximate precomputed operators, $\boldsymbol{M}^j$, whose columns are given by

$$\underline{m}_i^j = \boldsymbol{G}' \mathrm{diag}\left\{\underline{e}^i\right\} \boldsymbol{G}\underline{e}^j. \tag{5.16}$$

Since these $\underline{m}_i^j$ vectors depend only on the system geometry $\boldsymbol{G}$, but not on the object itself, we can precompute the object-independent portion of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$. These operators may be applied to form the approximation given in (5.14) as

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{M}^j \underline{d}^j, \tag{5.17}$$

where $\underline{d}^j$ is a vector constructed from the elements of $\boldsymbol{D}^j$ in (5.15).

It is reasonable to include a_{ij} terms in the calculation of $\underline{d}^j$, since these factors generally must be computed for the reconstruction method that is chosen to estimate the SPECT image. In fact, while $\boldsymbol{G}$ is often too large to precompute and store for 3D-SPECT, if $\boldsymbol{A}$ is modeled with the simple line integral model of (2.1), $\boldsymbol{A}$ is very sparse with only a single value per row. Thus, it may be possible to compute and store $\boldsymbol{A}$ for a given object for both estimation of the SPECT image, and for evaluation of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$. As mentioned previously, $\boldsymbol{A} = \boldsymbol{1}$ in PET and would not be stored at all.

Equation (5.17) represents an approximation that allows for precomputation of a portion of $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ using the linear operator technique of [116]. If $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$ represents a shift-invariant response, only a single operator, $\boldsymbol{M}^{j_0}$, is required. (This is discussed in more detail in Section 5.4.2.) The remaining operators may be found by applying the appropriate shifts. Such shift-invariance holds for PET systems near the center of the field of view. However, for large field of view PET systems like

small animal PET scanners, and for most SPECT systems, the geometric response is shift-variant. Thus, it appears that to use (5.17) one would need to calculate very many linear operators. Specifically, without further simplifications, one would need to compute, store, and use one $P \times N$ matrix for each voxel. In the following sections we demonstrate ways to reduce both storage requirements and computation time.

5.3.3 Image-Domain Simplifications

There are a number of observations and approximations that allow us to reduce the computation and storage requirements to practical levels. We break these simplifications into two groups: 1) Image-domain simplifications, that reduce either the number of operators that are stored, or the number of rows in each of the matrices. 2) Projection-domain simplifications, that reduce the number of columns required for each $\boldsymbol{M}^j$, and consequently the number of diagonal weighting elements (*i.e.*, a smaller $\boldsymbol{D}$). We discuss the image-domain simplifications in this section, and discuss projection-domain simplifications in Section 5.3.4.

For each approximation, we first describe the basic principle in words, and then give an explicit mathematical representation. Since matrices in the following sections represent operations on 3D projections or images, care should be taken in interpreting the mathematical forms.

Single Slice Sampling

Because $\boldsymbol{H}$ is object-dependent due to attenuation, there are generally few symmetries that would allow one to reduce computation and storage requirements. However, because we are utilizing (5.18), which requires only the geometric model, $\boldsymbol{G}$, we can take advantage of symmetries in the PET or SPECT geometry.

For many tomographic systems there are a number of symmetries in the imaging

system that can simplify our goals. For example, in SPECT, most parallel hole and fan collimators have a detector response that is essentially shift-invariant for axial shifts of the detector, excluding magnitude scaling factors like detector efficiency (*i.e.*, the c_i terms). Similarly, PET systems operating in 2D mode (with septa in place) are made of rings or blocks of detectors where the detector response changes little with axial shifts. Thus, if one varies j only in the transaxial direction, $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$ changes only by a transaxial shift. Similarly, for the same j, the columns of our precalculated $\boldsymbol{M}^j$ in (5.16) would differ only by transaxial shifts.

Therefore, it is not necessary to compute (5.16) for all j. A single slice is sufficient. Thus, we let

$$\underline{m}_i^{(x_j,y_j)} = \boldsymbol{G}'\mathrm{diag}\left\{\underline{e}^i\right\}\boldsymbol{G}\underline{e}^{(x_j,y_j,z_0)}, \tag{5.18}$$

where x_j and y_j denote the x and y-coordinates of the jth voxel, and z_0 reflects the transaxial coordinate of the center slice. Consequently,

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{S}^{z_j}\boldsymbol{M}^{(x_j,y_j)}[\boldsymbol{S}_P^{z_j}]^{-1}\underline{d}^j, \tag{5.19}$$

where $\boldsymbol{S}^{z_j}$ shifts an image from the center slice to the z-coordinate of the j voxel, $\boldsymbol{M}^{(x_j,y_j)}$ is formed from columns of (5.18), and $\boldsymbol{S}_P^{z_j}$ is the projection-domain analogue of $\boldsymbol{S}^{z_j}$, which shifts projection values along the transaxial direction. In terms of storage, we may now store $P_x \cdot P_y$ operators instead of $P = P_x \cdot P_y \cdot P_z$.

Even in systems that do not have axially shift-invariant detector responses like 3D PET, there are often partial symmetries. Recall that 3D PET generally has shift-variant detector responses due to the truncated projections discussed in Section 2.3.1. However, since PET systems are generally cylindrical, there is typically a symmetry through the center of the cylinder axis and only half the slices need to be stored.

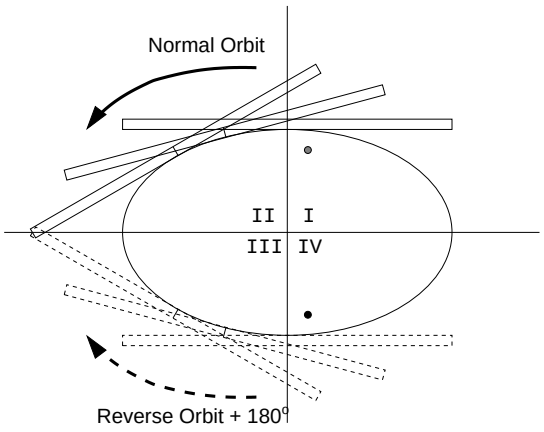


Figure 5.2: Symmetries in elliptical orbit SPECT.

In fact, as we shall see in the following sections, one can use other symmetries to further reduce storage.

Partial Orbital Sampling

We can also take advantage of symmetries in the SPECT detector orbit or, equivalently, the PET ring geometry. Consider the 360° elliptical orbit SPECT system shown in Figure 5.2. Suppose that we may only compute weighted projection-backprojections for points in quadrant IV. Because this elliptical orbit is symmetric about both axes, we may compute weighted projection-backprojections for the remaining quadrants. For example, consider the black point in quadrant IV in Figure 5.2. The gray point in quadrant I is "seen" by the rotating detector head with the same detector response as the black point, if the orbit direction is reversed and is started 180° from the normal orbit's starting point. In other words, if one has obtained the projections for the black point in quadrant IV, one can obtain the projections for the gray point in quadrant I simply by reordering the projection images. Similarly, if one may only backproject projections obtained from points in quadrant IV, one can obtain projection-backprojections for points in the other quadrants

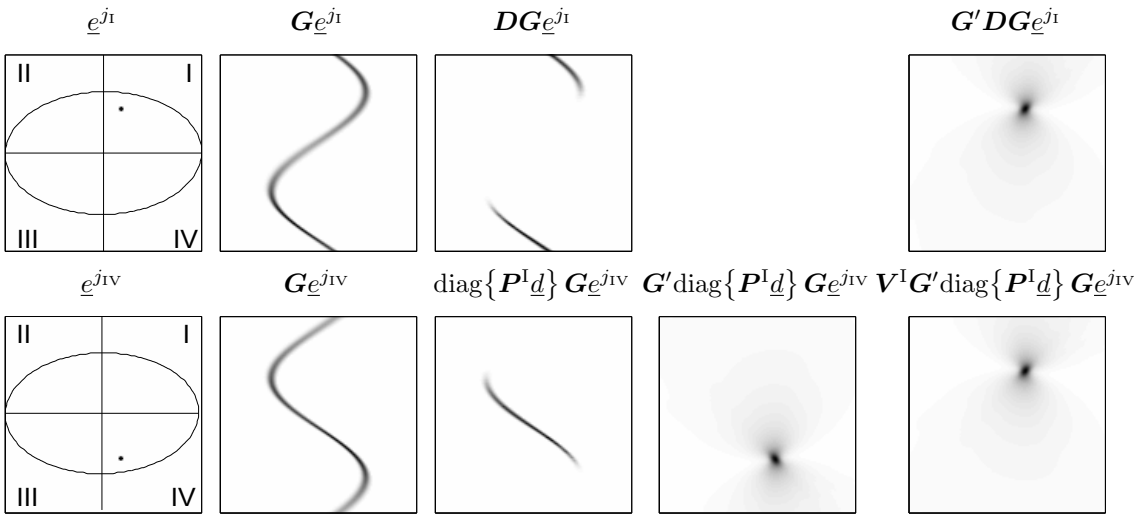


Figure 5.3: Using symmetries in elliptical orbit SPECT to compute weighted responses. Specifically, an illustration showing the calculation of a weighted response in quadrant I from calculations based on a point in quadrant IV. The top row shows a straightforward calculation of the weighted projection-backprojection for a point in quadrant I. The bottom row shows an approximation of the same weighted projection-backprojection, using a point in quadrant IV in conjunction with the permutation operator, $\boldsymbol{P}^{\mathrm{I}}$, and the flip operator, $\boldsymbol{V}^{\mathrm{I}}$.

using simple flips about the axes. Thus, we need to precompute only a single quadrant of linear operators. For circular system geometries, only a single radial line of operators is required.

We illustrate the application of these symmetries to the calculation of weighted projection-backprojections in Figure 5.3. In the top row of this figure, we show a straightforward calculation of a weighted response for a point in quadrant I. In the bottom row we show a sequence of images representing the calculation of the same response using projections and backprojections in quadrant IV. In order to use the quadrant IV calculations, we must appropriately permute elements of the diagonal weighting matrix. We represent this reordering by multiplying diagonal elements, $\underline{d}$, by the matrix $\boldsymbol{P}^{\mathrm{I}}$, which reorganizes projection weightings by reversing the projection order and shifting the weightings by 180°. Additionally, we require an image-domain transformation that flips the image about the x-axis. This operation

is equivalent to a change of variables with $y_{\text{new}} = -y$ and is represented by the matrix $\boldsymbol{V}^{\text{I}}$. The resulting weighted responses shown in the rightmost images are nearly indistinguishable for the two calculations.

Recalling (5.18), since our linear operators involve only weighted responses using the geometric model, $\boldsymbol{G}$, we need to calculate only a single quadrant of operators. Responses in the other quadrants may be obtained using simple permutation and flipping operations. The permutations and flips are quadrant-dependent (as indicated by the superscripts). Therefore, we define generic permutation, $\boldsymbol{P}^j$, and transformation operators, $\boldsymbol{V}^j$, that are position-dependent. If one stores operators only for quadrant IV, these operators are defined as follows:

Quadrant	$\boldsymbol{V}^j$	Action	$\boldsymbol{P}^j$	Action
I	$\boldsymbol{V}^{\text{I}}$	x-axis flip	$\boldsymbol{P}^{\text{I}}$	reverse + 180°
II	$\boldsymbol{V}^{\text{II}}$	x-axis + y-axis flip	$\boldsymbol{P}^{\text{II}}$	+ 180°
III	$\boldsymbol{V}^{\text{III}}$	y-axis flip	$\boldsymbol{P}^{\text{III}}$	reverse
IV	$\boldsymbol{V}^{\text{IV}}$	no action	$\boldsymbol{P}^{\text{IV}}$	no action.

Incorporating this calculation technique into (5.19), we write the approximation of the weighted response as

$$\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{V}^j\boldsymbol{S}^{z_j}\boldsymbol{M}^{(x_j,y_j)}[\boldsymbol{S}_P^{z_j}]^{-1}\boldsymbol{P}^j\underline{d}^j, \tag{5.20}$$

where the $\boldsymbol{M}^{(x_j,y_j)}$ operators are still calculated via (5.18), but only over a subset of locations appropriate for the specific system symmetries.

For circular orbit SPECT, or PET systems with ring geometries, one might choose to store a single radial line[4] of linear operators. In this case, $\boldsymbol{P}^j$ still represents a simple permutation, but $\boldsymbol{V}^j$ must be defined as a position-dependent rotation

[4]We assume that only a single axial slice of operators is being stored. For 3D PET systems, it may be appropriate to store many slices due to the truncated projections. Thus, instead of a radial line of operators, an angular slice is considered.

operation, which can take significantly more time than simple flip operations. Thus, even if circular symmetries are being used, one might choose to precalculate more operators (*e.g.*, over a quadrant), trading off more storage for faster evaluation.

Small Volume of Support

Because $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ is fairly concentrated about voxel j, many calculations involving $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ are also very concentrated about j. For example, in the resolution calculation in (5.2) and the covariance formulation in (5.3), the results of these evaluations typically trend to zero farther from the point of interest. In other words, resolution and covariance functions generally go to zero far from j.

For many applications it is not necessary to estimate far from j. For example, for resolution calculations, one may require only a sampling of the local impulse response that covers two or three times the full-width half-maximum resolution. Beyond this region, the local response is essentially zero.

We have found one can use relatively small regions of support to obtain very good resolution approximations[118]. Thus, it is not necessary to store all the rows of $\boldsymbol{M}^j$. Instead, we choose to store a relatively small volume of support centered around voxel j. If one chooses a small $\eta \times \eta \times \eta$ volume,[5] each $\boldsymbol{M}^j$ is $\eta^3 \times N$. For a typical SPECT system where $P = 128^2 \cdot 64$, a choice of $\eta = 30$ represents a decrease in storage by a factor of almost 40.

Thus, (5.18) and (5.20) become

$$\underline{m}_i^{(x_j,y_j)} = \boldsymbol{T}^j\boldsymbol{G}'\mathrm{diag}\left\{\underline{e}^i\right\}\boldsymbol{G}\underline{e}^{(x_j,y_j,z_0)}, \tag{5.21}$$

and

$$\boldsymbol{T}^j\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \approx \boldsymbol{V}^j\boldsymbol{M}^{(x_j,y_j)}[\boldsymbol{S}_P^{z_j}]^{-1}\boldsymbol{P}^j\underline{d}^j, \tag{5.22}$$

[5]There is no fundamental reason why the subvolumes must be cubical. We choose a cubical subvolume for simplicity.

where $\boldsymbol{T}^j$ represents a position-dependent $\eta^3 \times P$ matrix that represents a truncation function that selects a small volume about pixel j. The image-domain shift operation in (5.19) is no longer necessary due to the truncation function $\boldsymbol{T}^j$, since there is an implicit "centering" of the subvolume. In fact, as we will discuss in Section 5.4.1, this also allows one to eliminate the multiplications involving the $[\mathfrak{F}\{\underline{e}^j\}]$ terms in (5.2) and (5.3).

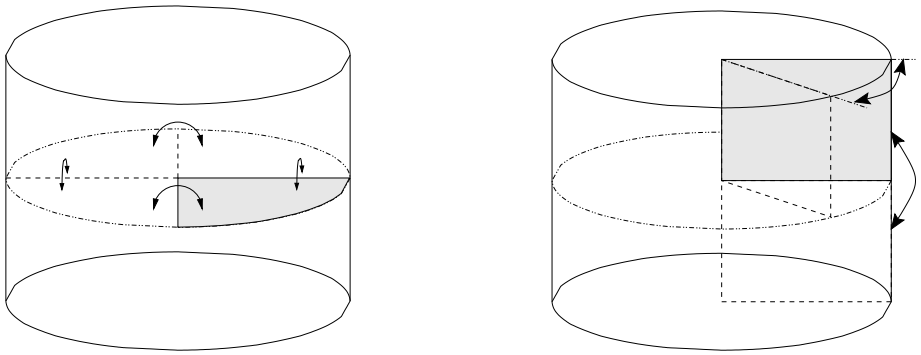
One could generalize the truncation function, $\boldsymbol{T}^j$, to include an invertible transformation that reduces storage requirements. For example, if $\boldsymbol{T}^j$ were a transformation that allows the storage of $\boldsymbol{M}^j$ with fewer coefficients, one could store "compressed" operators. Generally this would require additional computation for the "decompression" step when the operators are applied. We have found that simple truncation can perform quite well without the need for additional transformations. However, it is possible that some system geometries may require larger volumes of support for good noise and resolution predictions.

Spatial Subsampling

The weighted responses, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$, typically vary smoothly with position. Because this is the case, we have found that one can subsample the image-domain and evaluate $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ over a subset of positions and find the remaining positions using interpolation[118].

Using the approximations described earlier in this section, we may subsample the image slice or smaller orbital section (depending on symmetries), calculating linear operators over a grid of every n_dth voxel in both in-plane directions. This reduces storage requirements by $1/n_d^2$.

There are a number of ways to perform the interpolation for the between sample positions. Consider the following two interpolation methods: 1) Bilinear interpola-



(a) 3D SPECT Spatial Subsampling (b) 3D PET Spatial Subsampling

Figure 5.4: Typical regions which are spatially subsampled for 3D PET and SPECT. For elliptical orbit SPECT and 3D PET with truncated projections, one can typically sample geometric responses over the gray regions to fully characterize the system. Responses at other locations may be formed using the transformation operator, $\boldsymbol{V}^j$, which rotates or changes coordinates (as indicated by the arrows).

tion of the nearest four linear operators to obtain an approximate operator, $\hat{\boldsymbol{M}}^{(x_j,y_j)}$, and application of this approximate operator to the vector, $\underline{d}^{(x_j,y_j)}$. 2) Trilinear interpolation of the weighted response from the eight nearest responses. The eight nearest responses are calculated using the sampled operators and the diagonal weights associated with those samples. This second method is appropriate when $\underline{d}^{(x_j,y_j)}$ does not vary too quickly, since both the geometric response and the diagonal weights are interpolated. On the other hand, the first method interpolates only the linear operators and applies the uninterpolated weights, $\underline{d}^{(x_j,y_j)}$. Which method is appropriate depends on a number of factors, including computation time for $\underline{d}^{(x_j,y_j)}$, the size of the truncation function, $\boldsymbol{T}^j$, the degree of space-variance of the diagonal weights, and the acceptable amount of approximation error.

While we have assumed that this sampling takes place in the x-y plane (as is appropriate for SPECT with axially shift-invariant geometric responses), we note that this kind of spatial sampling can be done for any systems with smoothly varying geometric responses. The partial plane which is spatially subsampled for elliptical orbit

SPECT is shown as the gray quadrant in Figure 5.4a. As we mentioned previously, for circular geometries we may sample a single radial line and apply the appropriate permutation, $\boldsymbol{P}^j$, and transformation, $\boldsymbol{V}^j$. However, for 3D PET with truncated projections one should sample many axial slices. Thus, instead of sampling a portion of the x-y plane, one would sample a single angular slice (like the x-z plane) as shown in Figure 5.4b.

5.3.4 Projection-Domain Simplifications

Just as one can approximate $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ using image-domain simplifications, one can make projection-domain approximations that reduce dimensionality, storage requirements, and computation times. Specifically, in the following subsections, we describe approximations that will reduce the number of columns required for the linear operators, $\boldsymbol{M}^j$.

Projection-Constant Weightings

One approximation investigated in [116] relies on the observation that projections of a point are highly localized. That is, for individual projection angles, $\boldsymbol{H}\underline{e}^j$ yields a relatively narrow response. Figure 5.5a shows several projections of a point. The diagonal term, $\boldsymbol{D}$, simply scales each element of the projection and is typically a smoothly varying function over each projection. Recall from (3.18) and (3.19), elements of $\boldsymbol{D}$ are often defined as functions of the mean measurements, which are themselves relatively smooth due to the blur of the projection operator. Because these weightings are relatively smooth for each projection angle and the point projections are highly localized, we can approximate $\boldsymbol{D}$ with a new position-dependent diagonal weighting, $\tilde{\boldsymbol{D}}^j$, which scales projections for individual angles by a single value. In fact, for projection-constant weightings and a shift-invariant geometric re-

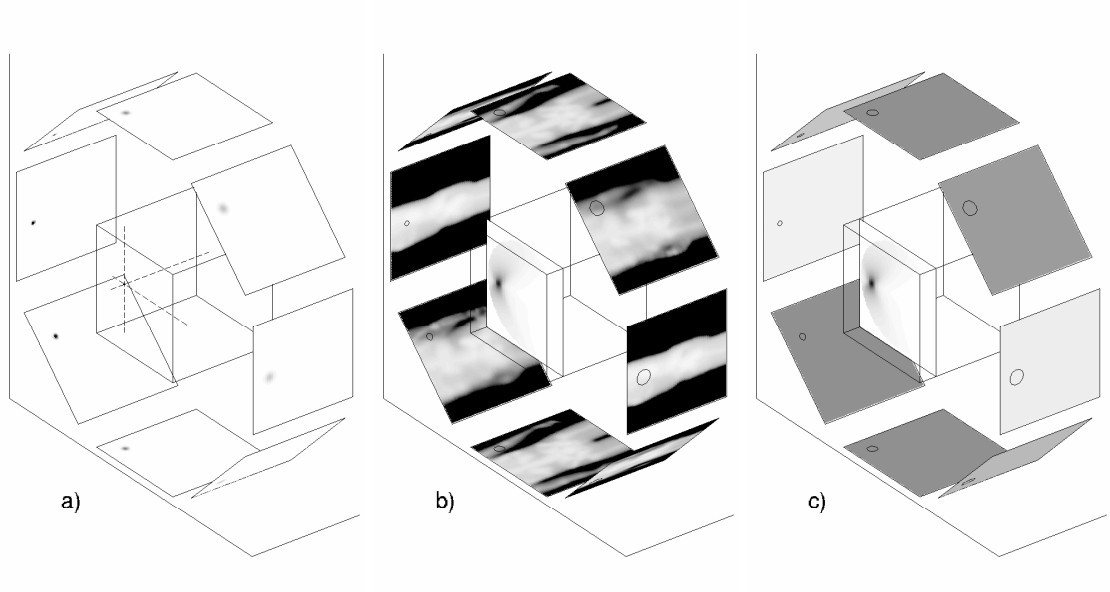


Figure 5.5: Approximation of the weighted point projection-backprojection using a projection-constant weighting.
Figure a) shows a particular point within the imaging volume and a few of its projections, $\boldsymbol{H}\underline{e}^j$. Figure b) shows several projection weightings in the diagonal weighting $\boldsymbol{D}$, and a cross-section of the associated weighted response, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$. We identify the approximate positions of the point projection using a small black circle in each projection. Because the point projections are highly localized, we may approximate the projection weighting using a position-dependent projection-constant weighting, $\tilde{\boldsymbol{D}}^j$, shown in Figure c). The associated weighted response, $\boldsymbol{H}'\tilde{\boldsymbol{D}}^j\boldsymbol{H}\underline{e}^j$, is nearly identical to the unapproximated response, $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$.

sponse, one can show that the weighted projection-backprojection is shift-invariant (see Appendix A for the 2D continuous case). Thus, for shift-invariant penalties and projection-constant weights,[6] the circulant approximation to the local impulse response in (5.2) is exact.

Let $[\boldsymbol{H}]'_{(i,1:P)}$ denote the ith row of $\boldsymbol{H}$, and $\mathcal{P}_a$ denote the set of measurements in the projection at angle a. We make the approximation:

$$\begin{aligned}
\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j &\approx \sum_{a=1}^{n_a}\sum_{i\in\mathcal{P}_a}[\boldsymbol{H}]'_{(i,1:P)}[\tilde{\underline{d}}^j]_a[\boldsymbol{H}]_{(i,1:P)}\underline{e}^j \\
&= \boldsymbol{H}'\mathrm{diag}\left\{\boldsymbol{C}'_{\mathcal{P}}\tilde{\underline{d}}^j\right\}\boldsymbol{H}\underline{e}^j, \\
&= \boldsymbol{H}'\tilde{\boldsymbol{D}}^j\boldsymbol{H}\underline{e}^j, \qquad (5.23)
\end{aligned}$$

where $[\tilde{\underline{d}}^j]_a$ represents the position-dependent, projection-constant weighting for the

[6]The diagonal weights in $\boldsymbol{D}$ are rarely projection-constant. However, it is interesting to note that such weights can lead to a shift-invariant response.

ath angle, and the vector $\underline{\tilde{d}}^j$ denotes the collection of all projection-constant weightings over all n_a angles. The $n_a \times N$ matrix $\boldsymbol{C}_{\mathcal{P}}$ combines measurements within a single projection angle into a single value and is used to form the new diagonal matrix, $\tilde{\boldsymbol{D}}^j$. The combination matrix can be written as:

$$\boldsymbol{C}_{\mathcal{P}} = \left[\begin{array}{ccc} \underline{I}_{\mathcal{P}_1} & \cdots & \underline{I}_{\mathcal{P}_{n_a}} \end{array} \right]', \tag{5.24}$$

where $\underline{I}_{\mathcal{P}_a}$ is an indicator vector where the ith element of the vector is one if the element belongs in the projection at angle a, and is zero otherwise.

While there are many ways to calculate $\underline{\tilde{d}}^j$, elements of this vector can generally be approximated by some form of position-dependent weighted average. For example, one simple technique that weights elements of $\boldsymbol{D}$ by the intensity of a point projection is

$$[\underline{\tilde{d}}^j]_a = \frac{\underline{I}'_{\mathcal{P}_a} \boldsymbol{D} \hat{\boldsymbol{H}} \underline{e}^j}{\underline{I}'_{\mathcal{P}_a} \hat{\boldsymbol{H}} \underline{e}^j}, \tag{5.25}$$

where $\hat{\boldsymbol{H}}$ is some form of the system matrix, $\boldsymbol{H}$. Because we have found that the approximation in (5.23) is relatively insensitive to the exact weightings, it is often sufficient to use an approximate $\hat{\boldsymbol{H}}$. In fact, we find using a simple line integral model without attenuation is often sufficient for $\hat{\boldsymbol{H}}$. Thus, it is straightforward to precompute and store the necessary weightings to compute $\tilde{\boldsymbol{D}}^j$.

Figure 5.5 demonstrates the efficacy of this technique. Figure 5.5a shows a few unweighted projections of a single point. Figure 5.5b shows sample projection weights and a transaxial cross section of the associated weighted response. Approximate positions of the point projection are indicated with small black circles. We find an approximate projection-constant weighting based on (5.25), with $\hat{\boldsymbol{H}}$ equal to a simple line integral model with no attenuation. Thus, (5.25) is simply a bilinear interpolation for each projection. (We suspect that an even simple nearest-neighborhood

interpolation would also be adequate.) Figure 5.5c shows the projection-constant weights and a cross section of the weighted response. The two transaxial cross sections are nearly indistinguishable.

Before we discuss the resulting linear operator form of the approximation, we discuss one additional approximation that further reduces the size of the diagonal weighting.

Angular Subsampling

Rather than computing the projections, $\boldsymbol{H}\underline{e}^j$, over all angles, we further approximate the projection (and backprojection) by reducing the number of projection angles involved. We will divide projection angles into K contiguous blocks, where a single block combines a neighborhood of n_s angles. Letting $\mathcal{S}_k$ denote the set of angles belonging to the kth block, we write

$$\begin{aligned}
\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j &\approx \sum_{k=1}^{K} \sum_{a\in\mathcal{S}_k} \sum_{i\in\mathcal{P}_a} [\boldsymbol{H}]'_{(i,1:P)} [\check{\underline{d}}^j]_k [\boldsymbol{H}]_{(i,1:P)} \underline{e}^j \\
&= \boldsymbol{H}\,\mathrm{diag}\left\{ \boldsymbol{C}'_{\mathcal{S}}\boldsymbol{C}'_{\mathcal{P}}\check{\underline{d}}^j \right\} \boldsymbol{H}\underline{e}^j. \\
&= \boldsymbol{H}\,\mathrm{diag}\left\{ \boldsymbol{C}'_{\mathcal{P}\mathcal{S}}\check{\underline{d}}^j \right\} \boldsymbol{H}\underline{e}^j, \\
&= \boldsymbol{H}\check{\boldsymbol{D}}^j\boldsymbol{H}\underline{e}^j, \qquad (5.26)
\end{aligned}$$

where the combination matrix is defined as

$$\boldsymbol{C}_{\mathcal{S}} = [\underline{I}_{\mathcal{S}_1} \cdots \underline{I}_{\mathcal{S}_K}]', \qquad (5.27)$$

where the indicator vector, $\underline{I}_{\mathcal{S}_k}$, indicates membership of an angle in the set $\mathcal{S}_k$. We also define $\boldsymbol{C}_{\mathcal{P}\mathcal{S}} \triangleq \boldsymbol{C}_{\mathcal{P}}\boldsymbol{C}_{\mathcal{S}}$.

Again, while there are many ways to approximate the position-dependent weighting vector, $\check{\underline{d}}^j$, we choose approximate weights by simply averaging over angles in

each set, $\mathcal{S}_k$. Specifically,

$$[\check{\underline{d}}^j]_k = \frac{1}{n_s} \sum_{a \in \mathcal{S}_k} \frac{\underline{I}'_{\mathcal{P}_a} \boldsymbol{D} \hat{\boldsymbol{H}} \underline{e}^j}{\underline{I}'_{\mathcal{P}_a} \hat{\boldsymbol{H}} \underline{e}^j}. \tag{5.28}$$

The vector, $\check{\underline{d}}^j$, represents a significant decrease in the dimension from the original weighting, $\boldsymbol{D}$. Recall that the diagonal matrix $\boldsymbol{D}$ is $N \times N$, where N is the product of the number of measurements per projection (*i.e.*: the number of pixels in each projection), and the number of projection angles, n_a. In comparison, $\check{\underline{d}}^j$ is contains only K values, where K is the number of projection angles, n_a, divided by the number of angles in each subset, n_s.

5.3.5 Simplified Linear Operators

We now combine the simplifications discussed in the previous sections to obtain a set of linear operators that is practical to implement and store.

Section 5.3.4 discussed two approximations that reduce the dimension of $\boldsymbol{D}$ from N to $K = n_a/n_s$. We may calculate the reduced dimension linear operator by applying the approximations in Section 5.3.4 to (5.21) to obtain

$$\underline{m}_k^{(x_j,y_j)} = \boldsymbol{T}^j \boldsymbol{G}' \mathrm{diag}\left\{ \boldsymbol{C}'_{\mathcal{PS}} \underline{e}^k \right\} \boldsymbol{G} \underline{e}^{(x_j,y_j,z_0)}. \tag{5.29}$$

The projection-constant weighting discussed in Section 5.3.4 eliminates the need for the projection-domain shift operation introduced in (5.19). Thus, we may now write[7]

$$\boldsymbol{T}^j \boldsymbol{H}' \boldsymbol{D} \boldsymbol{H} \underline{e}^j \approx \boldsymbol{V}^j \boldsymbol{M}^{(x_j,y_j)} \boldsymbol{P}^j \check{\underline{d}}^j. \tag{5.30}$$

The vector $\check{\underline{d}}^j$ is formed from joining (5.28) with the attenuation approximation in

[7] Note that the $\boldsymbol{P}^j$ operator has the same function as was described in Section 5.3.3, but now operates on the smaller vector, $\check{\underline{d}}^j$, which contains projection weights for blocks of angles.

(5.15). Specifically, the kth element of $\underline{\check{d}}^j$ is

$$\left[\underline{\check{d}}^j\right]_k = \frac{1}{n_s}\sum_{a\in\mathcal{S}_k}\frac{\underline{I}'_{\mathcal{P}_a}\left[\boldsymbol{D}_c^2\text{diag}\{\boldsymbol{A}\underline{e}^j\}^2\boldsymbol{D}\right]\hat{\boldsymbol{H}}\underline{e}^j}{\underline{I}'_{\mathcal{P}_a}\hat{\boldsymbol{H}}\underline{e}^j}. \tag{5.31}$$

In terms of storage, we now have matrices, $\boldsymbol{M}^{(x_j,y_j)}$, that are $\eta^3 \times K$. From Section 5.3.3, we need to store these matrices within only a single slice, or a single-quadrant of a single slice for orbits with two-fold symmetries. We may further subsample this quadrant to reduce computational costs. Thus, for elliptical orbit SPECT, using all these simplifications in conjunction means we must store

$$\frac{1}{4}\frac{P_xP_y}{n_d^2}\eta^3\frac{n_a}{n_s} \tag{5.32}$$

floating point numbers. Consider a sample SPECT system that incorporates a $128\times 128\times 64$ image volume and projections over 110 angles. For a sampling of every 4th image pixel in x and y, a subvolume of $30\times 30\times 30$, and blocks of 10 angles, we must store about 76 million floating point numbers. If stored as standard single precision floating point numbers, this represents about 290 Mb of storage space.

Equations (5.29), (5.30), and (5.31) represent a set of precomputations and the necessary operations for approximating $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$. While this weighted projection-backprojection may be of interest for some applications, additional simplifications can be made when resolution or covariance prediction is the goal. The following section discusses such simplifications.

5.4 Novel Fast Resolution and Covariance Predictors

5.4.1 Additional Simplifications

To predict resolution or covariance, one can plug the approximation of the weighted projection-backprojection in (5.30) directly into the resolution or covariance predictors in (5.2) and (5.3). However, further investigation allows us to make additional

simplifications that reduce both storage and computation time.

Both (5.2) and (5.3) are based on using a circulant approximation to $\boldsymbol{H'DH}$. Because circulant matrices can be diagonalized using Fourier bases, we may find the eigenvalues of the circulant approximation using Fourier transforms, which allows one to avoid the full matrix inverse computations in (3.15) and (5.1). Because $\boldsymbol{D}$ is a diagonal matrix composed of nonnegative elements, the eigenvalues of $\boldsymbol{H'DH}$ are necessarily real and nonnegative. It is common to enforce these constraints when Fourier transforming the (appropriately shifted) weighted response $\boldsymbol{H'DH}\underline{e}^j$. The real constraint is typically enforced by ensuring point symmetry through the center of the response (*i.e.*, voxel j). An equivalent approach is to only use the real part of the Fourier transformed image. The nonnegativity constraint is often enforced simply by zeroing any negative components. The same constraints are applied to the penalty terms in (5.2) and (5.3). Thus, the resolution and covariance predictors may be written as follows:

$$\underline{l}^j_{\text{circ}} = \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\left\{\underline{e}^j\right\} \odot \underline{\tilde{f}}^j_2}{\underline{\tilde{f}}^j_1 + \underline{\tilde{r}}^j}\right\}, \quad \text{Cov}^j_{\text{circ}} = \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\left\{\underline{e}^j\right\} \odot \underline{\tilde{f}}^j_3}{\left[\underline{\tilde{f}}^j_1 + \underline{\tilde{r}}^j\right]^2}\right\}, \tag{5.33}$$

where

$$\underline{\tilde{f}}^j_n \triangleq \max\left\{\text{re}\left\{\frac{\mathfrak{F}\left\{\boldsymbol{H'D}_n\boldsymbol{H}\underline{e}^j\right\}}{\mathfrak{F}\left\{\underline{e}^j\right\}}\right\}, 0\right\} \tag{5.34}$$

$$\underline{\tilde{r}}^j \triangleq \max\left\{\text{re}\left\{\frac{\mathfrak{F}\left\{\boldsymbol{R}(\breve{\underline{\theta}})\underline{e}^j\right\}}{\mathfrak{F}\left\{\underline{e}^j\right\}}\right\}, 0\right\}, \tag{5.35}$$

and the $\mathfrak{F}\left\{\underline{e}^j\right\}$ are applied to shift the local impulse response or covariance measurement to the jth voxel. (Equivalently, this may be applied as an image-domain shifting operation.)

Since the approximation to $\boldsymbol{H'DH}\underline{e}^j$ discussed in Section 5.3 is eventually plugged into the above expressions, it would be advantageous to include as many of the op-

erations in (5.34) in the precomputation step as possible. Because the Fourier transform is a linear operation, it is natural to incorporate these operations in $\boldsymbol{M}^j$ as well. Specifically, we may now redefine the operators specified in (5.29) as

$$\underline{m}_k^{(x_j,y_j)} = \mathrm{re}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{G}'\mathrm{diag}\left\{\boldsymbol{C}'_{\mathcal{PS}}\underline{e}^k\right\}\boldsymbol{G}\underline{e}^{(x_j,y_j,z_0)}\right\}\right\}. \tag{5.36}$$

Noting that the change of coordinates represented by $\boldsymbol{V}^j$ is invertible, approximation (5.30) becomes

$$\mathrm{re}\left\{\mathfrak{F}\left\{(\boldsymbol{V}^j)^{-1}\boldsymbol{T}^j\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j\right\}\right\} \approx \boldsymbol{M}^{(x_j,y_j)}\boldsymbol{P}^j\check{\underline{d}}^j. \tag{5.37}$$

The transformation $(\boldsymbol{V}^j)^{-1}$ appears inside the Fourier transform, which seems to complicate our task. Fortunately, because the transformation $\boldsymbol{V}^j$ is only a renaming of image coordinates, we may apply the transformation in either image-domain. That is, it may be applied either before the $\mathfrak{F}\{\cdot\}$ operation or after the $\mathfrak{F}^{-1}\{\cdot\}$ operation. Therefore, we may rewrite the circulant approximation to the predictors as

$$\underline{l}^j_{\mathrm{circ}} \approx \boldsymbol{V}^j\mathfrak{F}^{-1}\left\{\frac{\check{\underline{f}}^j_2}{\check{\underline{f}}^j_1 + \check{\underline{r}}^j}\right\}, \tag{5.38}$$

$$\mathrm{Cov}^j_{\mathrm{circ}} \approx \boldsymbol{V}^j\mathfrak{F}^{-1}\left\{\frac{\check{\underline{f}}^j_3}{\left[\check{\underline{f}}^j_1 + \check{\underline{r}}^j\right]^2}\right\}, \tag{5.39}$$

with

$$\check{\underline{f}}^j_n \triangleq \max\left\{\boldsymbol{M}^{(x_j,y_j)}\boldsymbol{P}^j\check{\underline{d}}^j, 0\right\}, \tag{5.40}$$

$$\check{\underline{r}}^j \triangleq \max\left\{\mathrm{re}\left\{\mathfrak{F}\left\{(\boldsymbol{V}^j)^{-1}\boldsymbol{T}^j\boldsymbol{R}(\check{\underline{\theta}})\underline{e}^j\right\}\right\}, 0\right\}. \tag{5.41}$$

Because of the truncation operations, $\boldsymbol{T}^j$, in (5.36) and (5.41), there is an implicit "centering" about location j and the $\mathfrak{F}\{\underline{e}^j\}$ terms of (5.33) are no longer needed. Consequently, the predicted local impulse response in (5.38) and the covariance prediction in (5.39) are evaluated over a smaller support defined by $\boldsymbol{T}^j$. Thus, in order

to form even an (approximate) equality with (5.2) and (5.3), these small support approximations must be embedded into the larger image space. (We have ignored this embedding in (5.38) and (5.39).)

We have found that $\boldsymbol{M}^{(x_j,y_j)}$ generally contains negative values that are important for prediction. Thus, we cannot apply the negative thresholding in the precomputation step. It must be applied after the operator is applied to $\boldsymbol{P}^j\underline{\check{d}}^j$, as shown in (5.40).

5.4.2 Summary of Computational Burden and Storage

Shift-Variant SPECT

Equations (5.38-5.41) represent the final form of the approximate predictors developed in this paper. For a typical shift-variant SPECT system,[8] these predictors require storage of a set of matrices, $\{\boldsymbol{M}^{(x_j,y_j)}\}$, which consist of

$$\frac{1}{4}\frac{P_x P_y}{n_d^2}(\eta/2+1)\eta^2\frac{n_a}{n_s} \tag{5.42}$$

floating point numbers. The storage requirements are roughly one-half of that which is stated in (5.32) since the Fourier transform of a real signal results in coefficients whose real part is symmetric.

Once the linear operators have been precomputed, the following set of calculations is required for resolution and covariance prediction: 1) The $\underline{\check{d}}^j$ term is calculated via (5.31). Using a simple line integral model requires approximately $25n_a$ floating point operations (flops). 2) One must calculate (5.40), which takes about $2\eta^3 n_a/n_s$ flops, due to the application of the linear operator. (We concentrate on the case when $\underline{\check{f}}_1^j = \underline{\check{f}}_2^j = \underline{\check{f}}_3^j$, which is a realistic assumption for most SPECT systems.) 3) Lastly, one must compute the resolution or covariance prediction using (5.38) or (5.39),

[8] Shift-variant PET systems will have similar storage requirements if the sampling shown in Figure 5.4b is used. Computational requirements will be slightly larger due to the required rotation operations. However, these may still be computed relatively quickly since the truncation operator, $\boldsymbol{T}^j$, reduces the support size.

respectively. This entails a single inverse Fourier transform plus roughly $2\eta^3$ flops for a local resolution prediction and $3\eta^3$ flops for a local covariance estimate.

In many cases (5.41) can be computed once, such as elliptical orbit systems with penalties for which $\boldsymbol{T}^j \boldsymbol{R}(\breve{\underline{\theta}})\underline{e}^j$ exhibits three-fold planar symmetry[9] across each coordinate axis. For example, such is the case if the penalty is isotropic. For anisotropic penalties, one can decompose the penalty into symmetric and asymmetric portions, which can be formed from a small set of bases precomputed from $\breve{\underline{r}}^j$ terms. Thus, (5.41) generally involves relatively little computation.

The remaining computation is in applying a linear operator and a single $\eta \times \eta \times \eta$ inverse Fourier transform for each position j of interest. In comparison, recall the original expressions for the predictors in Section 5.2, which require multiple $P_x \times P_y \times P_z$ Fourier transform operations, a point projection, and a full backprojection for every position.

For some prediction tasks, even the single inverse Fourier transform may be eliminated. For example, for variance prediction one needs only to calculate the peak of the covariance function. Thus, one can eliminate both leading transform operations in (5.39) and simply sum over the η^3 Fourier coefficients and perform an appropriate normalization. Similar simplifications can be made to (5.38) for the contrast recovery coefficient studied in [95].

Shift-Invariant PET

For a PET system with a shift-invariant geometric response, the storage requirements are greatly reduced. Recall that for shift-variant PET systems the system matrix may be factored such that $\boldsymbol{H} = \text{diag}\{c_i\}\,\boldsymbol{G}$, where $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$ is shift-invariant. Also recall that the precomputed operators, such as the one described in (5.36), are

[9] It is important not to confuse the orbital symmetries associated with the $\boldsymbol{V}^j$ transformations with the point symmetry through the origin, which is imposed by the real constraint on the Fourier coefficients.

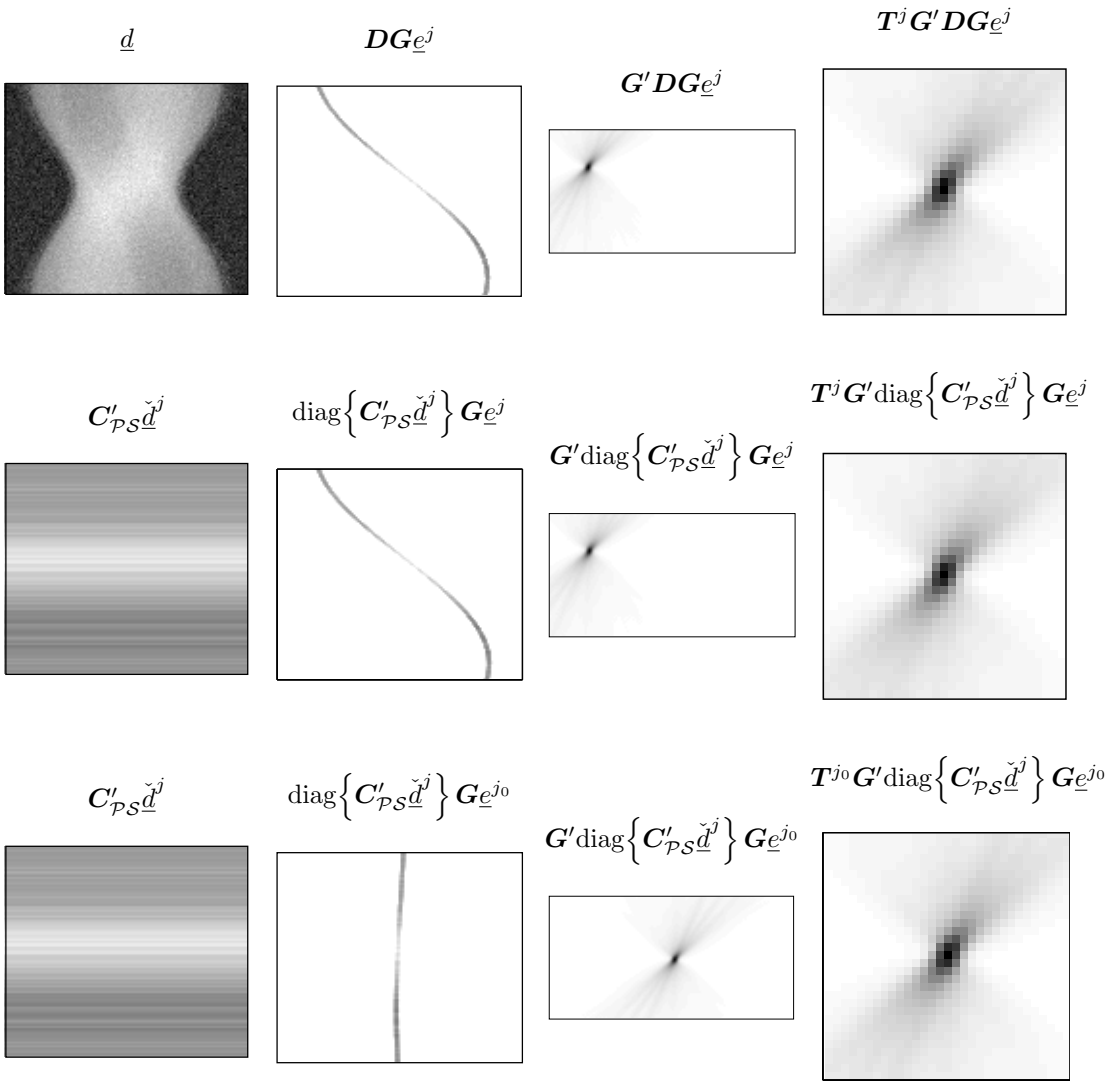


Figure 5.6: Application of approximations to a shift-invariant PET system.
The top row of images shows the direct calculation of the weighted response, $\boldsymbol{G}'\boldsymbol{DG}\underline{e}^j$. The center row shows the calculation of the same weighted response using projection-constant weights (radially constant weights for the 2D case). Since the PET model has a shift-invariant geometric response, we may calculate the same weighted response using only projections and backprojections of the center pixel (shown in the bottom row).

based only on the geometric model, $\boldsymbol{G}$. Consider the case of a weighted response for a 2D PET system and radially constant weights (*i.e.*, projection-constant weights). In Appendix A we show that such radially constant weights lead to a shift-invariant response under a continuous model. Thus, for a system with approximately shift-

invariant $\boldsymbol{G}'\boldsymbol{G}\underline{e}^j$, the weighted response, $\boldsymbol{G}'\boldsymbol{D}\boldsymbol{G}\underline{e}^j$ will also be approximately shift-invariant if $\boldsymbol{D}$ represents a radially constant weighting. Since our linear operators, like the one in (5.36), are functions of a weighted response with radially constant (projection-constant for the 3D case) weights, they too will be shift-invariant. Therefore, for an intrinsically shift-invariant system, only one linear operator needs to be computed.

This important property of weighted responses is illustrated in Figure 5.6. The direct calculation of a weighted response for PET is shown in the top row of Figure 5.6. Using the position-dependent projection-constant approximations of Section 5.3.4, we may calculate an approximate response (which is shown in the middle row of the figure). Because we have used a radially constant weighting in this approximation, we may alternately calculate the same (shifted) response using projections and back-projections at the center pixel. These calculations are shown in the bottom row of Figure 5.6. The truncated responses in the rightmost column of this figure are nearly indistinguishable. Since the linear operator technique is simply an efficient way of calculating the same weighted responses, only the linear operator precalculated for the center pixel is required.

While the storage considerations for the shift-invariant PET case are greatly reduced, the actual penalty design calculation times are typically the same as in the shift-variant case. This is because the linear operator and the position-dependent diagonal weighting must still be computed once for each location of interest.

5.5 Fast Penalty Design

In the previous section we discussed fast techniques for evaluating local impulse responses and covariance estimates. In this section, we discuss methods for perform-

ing fast penalty design for resolution control.

5.5.1 Practical Linearized Penalty Design

Recall the constrained linearized penalty design stated in (4.36). This least-squares design is a function of the $B \times P$ matrix, $\boldsymbol{\Phi}^j$, and the vector, $\underline{\alpha}^j$, that is length P. It is the $\underline{\alpha}^j$ term, that requires the most computation due to the weighted projection-backprojections. Therefore, the methods discussed in Section 5.3 can also be applied to penalty design.

Taking full advantage of the methods in Section 5.3 requires that the approximated responses are highly localized in the image-domain. Thus, recognizing that we have allowed for an arbitrary least-squares weighting in (4.37) and (4.38), we may choose $\boldsymbol{W}^j = \mathfrak{F}^{-1}\{\cdot\}$, so that (4.37) and (4.38) become:

$$\begin{aligned} \boldsymbol{\Phi}^j &\triangleq \mathfrak{F}^{-1}\left\{\text{diag}\left\{\underline{L}_0^j\right\}\mathfrak{F}\left\{\boldsymbol{B}^j\right\}\right\} && (5.43) \\ &= \left[\begin{array}{c|c|c} \text{vec}\left\{l_0^j(m-m_j,n-n_j) ** b_1(m-m_j,n-n_j)\right\} & \ldots & \\ & & \text{vec}\left\{l_0^j(m-m_j,n-n_j) ** b_B(m-m_j,n-n_j)\right\} \end{array}\right] && (5.44) \end{aligned}$$

and

$$\begin{aligned} \underline{\alpha}^j &\triangleq \mathfrak{F}^{-1}\left\{\text{diag}\left\{\mathfrak{F}\left\{\underline{e}^j\right\}\right\}\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\}\right\} - \\ &\qquad \mathfrak{F}^{-1}\left\{\text{diag}\left\{\underline{L}_0^j\right\}\mathfrak{F}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\}\right\} && (5.45) \\ &= \text{vec}\left\{\delta(m-m_j,n-n_j) ** \text{image}\left\{\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j}\right\}\right\} - \\ &\qquad \text{vec}\left\{l_0^j(m-m_j,n-n_j) ** \text{image}\left\{\boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j\right\}\right\}, && (5.46) \end{aligned}$$

where $\text{image}\{\cdot\}$ an operation that is the opposite of $\text{vec}\{\cdot\}$, reorganizing a elements of a vector back into an image. Thus, instead of performing the frequency-domain design that was described in Section 4.3, we may now perform an image-domain design. (As in Section 4.2, we have adopted notation for a 2D reconstruction with

a 2D desired response function, $l_0^j(m,n)$, and 2D penalty bases, $b_b(m,n)$. These 2D functions can easily be replaced with 3D functions for the 3D reconstruction problem.) Similarly, recognizing that image-domain shifts of both components ($\boldsymbol{\Phi}^j$ and $\underline{\alpha}^j$) does not change the least-squares design, we may redefine (5.44) and (5.46) as

$$\boldsymbol{\Phi}^j \triangleq \left[\text{vec}\left\{ l_0^j(m,n) ** b_1(m,n) \right\} \middle| \ldots \middle| \text{vec}\left\{ l_0^j(m,n) ** b_B(m,n) \right\} \right] \quad (5.47)$$

$$\underline{\alpha}^j \triangleq \text{vec}\left\{ \delta(m+m_j, n+n_j) ** \text{image}\left\{ \boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j} \right\} \right\} - \text{vec}\left\{ l_0^j(m+m_j, n+n_j) ** \text{image}\left\{ \boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j \right\} \right\}. \quad (5.48)$$

Thus, if the desired response l_0^j is shift-invariant (which means it is no longer a function of position j), the "centered" design represented by (5.47) and (5.48), requires the calculation of $\boldsymbol{\Phi}^{j_0}$ at a single voxel (*i.e.*, the center voxel, j_0). Moreover, if $\boldsymbol{H}'\boldsymbol{D}_2\mathcal{H}\delta_{\underline{x}_j} \approx \boldsymbol{H}'\boldsymbol{D}_1\boldsymbol{H}\underline{e}^j$ (as is often the case for good system models and typical noise models), we may write a simplified $\underline{\alpha}^j$ as

$$\underline{\alpha}^j \triangleq \text{vec}\left\{ \left(\delta(m+m_j, n+n_j) - l_0^j(m+m_j, n+n_j) \right) ** \text{image}\left\{ \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j \right\} \right\}. \quad (5.49)$$

Let us consider what typical evaluations of the columns of $\boldsymbol{\Phi}^j$ in (5.47) and $\underline{\alpha}^j$ in (5.49) look like. Using the shift-variant 2D PET system we will discuss in Section 6.2.2, Figure 5.7 shows a typical evaluation of a column of $\boldsymbol{\Phi}^j$ and $\underline{\alpha}^j$. The calculation of a single column of $\boldsymbol{\Phi}^j$ is represented in the upper row of images, and a typical $\underline{\alpha}^j$ evaluation is shown in the bottom row. Because we are typically designing penalties with small order neighborhoods, the columns of $\boldsymbol{\Phi}$ (of which the rightmost upper image is an example) generally will be localized in a region similar in size to the desired response, $l_0^j(m,n)$. Similarly, though $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ can have a modest support region, $\underline{\alpha}^j$ is often very localized. This is because $\delta(m,n) - l_0^j(m,n)$ typically takes

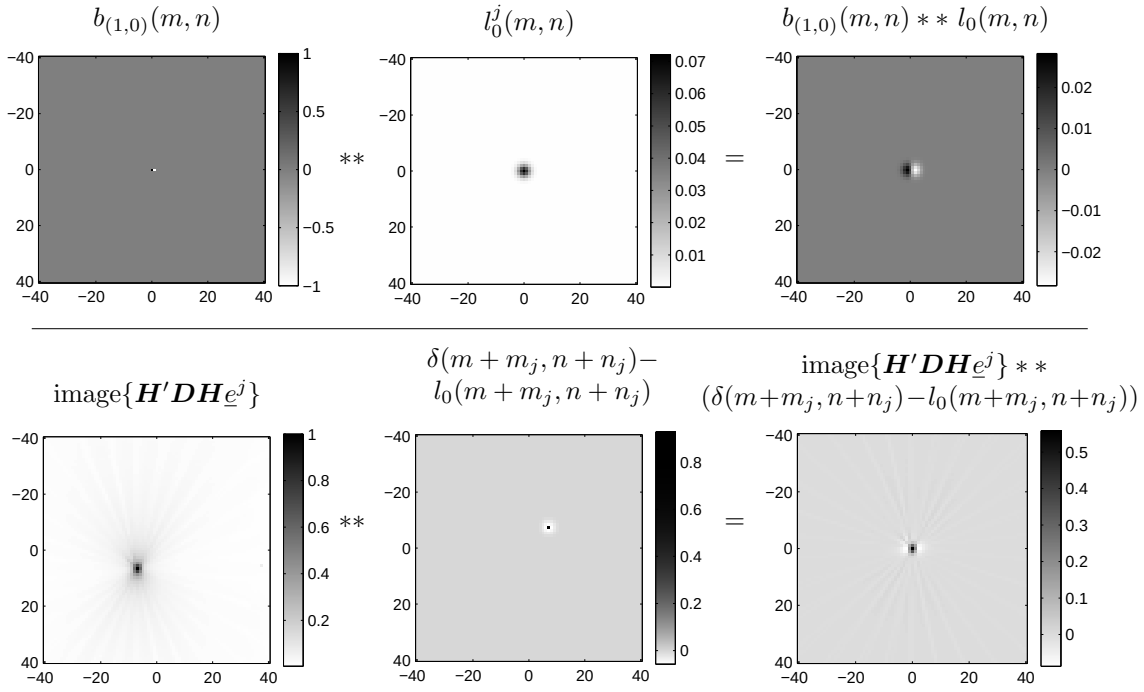


Figure 5.7: An example calculation of least-squares penalty design components. The top row shows the convolution operation for one column of $\boldsymbol{\Phi}$, and the bottom row shows a typical convolution operation for calculation of the $\underline{d}^j$ term. While the image size for this example is 170×170 pixels, the above images are presented zoomed-in for clarity. All images have a linear colormap except for $\delta(m+m_j, n+n_j) - l_0(m+m_j, n+n_j)$ which has been windowed to show details.

the form of a high pass filter (for standard choices of l_0^j) and $\boldsymbol{H'DH}\underline{e}^j$ is generally smoothly varying.

Similarly, even though there may be some structure in $\underline{\alpha}^j$ far away from the origin, these regions are arguably less important for the design. As shown in the top right image in Figure 5.7, far away from the origin, (small neighborhood) penalties have little influence. Similarly, we expect the approximation in (5.2) to be less accurate far from position j, which is equivalent to being far from the origin in $\boldsymbol{\Phi}^j$ and $\underline{\alpha}^j$. Therefore, only a small region near the origin need be evaluated.

Because these components of the least-squares design are highly localized it is natural to incorporate the truncation operator introduced in Section 5.3.3 into an approximation of these terms. Specifically, if the convolution operations are imple-

mented using Fourier transforms,

$$\boldsymbol{\Phi}^j = \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j \underline{l}_0^j\right\} \odot \mathfrak{F}\left\{\boldsymbol{T}^j \boldsymbol{B}^j\right\}\right\} \tag{5.50}$$

$$\underline{\alpha}^j = \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j)\right\} \odot \mathfrak{F}\left\{\boldsymbol{T}^j \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j\right\}\right\}. \tag{5.51}$$

Since the truncation operator implicitly "centers" the design, there is no need for shifting operations or offsets.

In some cases, such as when the system matrix has been precomputed and stored, $\boldsymbol{T}^j\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ may be computed by evaluating partial projections and backprojections. That is, $\boldsymbol{H}\underline{e}^j$ is "evaluated" by simply reading out a column of $\boldsymbol{H}$. Similarly, the partial backprojection, $\boldsymbol{T}^j\boldsymbol{H}'$, uses only a small subset of columns representing the neighborhood around location j. Thus, the penalty design represented by (4.36) using (5.50) and (5.51) can be performed relatively quickly when $\boldsymbol{H}$ has directly accessible columns. This method for fast design was discussed and applied to shift-variant PET and SPECT systems in [118].

Moreover, when the penalty design is for uniform resolution with a desired shift-invariant response given by $\underline{l}_0$, one may use

$$\boldsymbol{\Phi}^{j_0} = \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^{j_0} \underline{l}_0\right\} \odot \mathfrak{F}\left\{\boldsymbol{T}^{j_0} \boldsymbol{B}^{j_0}\right\}\right\} \tag{5.52}$$

$$\underline{\alpha}^j = \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^{j_0}(\underline{e}^{j_0} - \underline{l}_0)\right\} \odot \mathfrak{F}\left\{\boldsymbol{T}^j \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j\right\}\right\}. \tag{5.53}$$

Additionally, for the NNLS algorithm or the greedy algorithm discussed in Table 4.1, which is used to solve the constrained least-squares objective, one needs only $[\boldsymbol{\Phi}^{j_0}]'\boldsymbol{\Phi}^{j_0}$ and $[\boldsymbol{\Phi}^{j_0}]'\underline{\alpha}^j$. Thus, the relatively small $B \times B$ matrix, $[\boldsymbol{\Phi}^{j_0}]'\boldsymbol{\Phi}^{j_0}$, needs to be computed only once and takes little space to store. Actual computation times for this method are discussed in Section 6.3.1 in the application to a 2D SPECT system.

5.5.2 Evaluating Penalty Design Components using Linear Operators

In many cases, one cannot precompute and store the system matrix. This is typically true for systems like 3D SPECT, where the size of the system matrix and its sparsity lead to prohibitively large storage sizes. Thus, these system models are often implemented as "on-the-fly" routines and system matrix is "stored" as a procedure. Because these routines do not necessarily admit to fast column access, the fast penalty design methods from the previous section are not always applicable. In fact, even in cases where the techniques of Section 5.5.1 may be applied, sometimes they are still too slow for some applications. Thus, in this section we describe how one can apply the linear operator methods of Section 5.3 for fast penalty design.

Since the dominant source of computation remains in the evaluation of the weighted projection-backprojection operators it is natural to incorporate the simplified linear operator of Section 5.3.5 into the penalty design. Specifically, substituting $\boldsymbol{T}^j \boldsymbol{H}' \boldsymbol{D} \boldsymbol{H} \underline{e}^j \approx \boldsymbol{V}^j \boldsymbol{M}^{(x_j,y_j)} \boldsymbol{P}^j \check{\underline{d}}^j$ from (5.30) into (5.51) yields:

$$\underline{\alpha}^j \;\approx\; \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j)\right\} \odot \mathfrak{F}\left\{\boldsymbol{V}^j \boldsymbol{M}^{(x_j,y_j)} \boldsymbol{P}^j \check{\underline{d}}^j\right\}\right\}, \tag{5.54}$$

where the linear operator is defined with the columns given in (5.29) and the vector, $\check{\underline{d}}^j$, is given in (5.31). In many cases (5.54) may be used for fast and practical penalty design.[10] However, notice that all the terms to the left of $\boldsymbol{M}$ in (5.54) amount to another linear operation (which is essentially a linear blurring operation). This suggests that these operations may be stored in a precomputed operator as well.

Recalling that $\boldsymbol{V}^j$ may be applied in either the pre- or post-Fourier transform image-domain, we may rewrite (5.54) as

$$\underline{\alpha}^j \;\approx\; \boldsymbol{V}^j \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{[\boldsymbol{V}^j]^{-1}\boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j)\right\} \odot \mathfrak{F}\left\{\boldsymbol{M}^{(x_j,y_j)} \boldsymbol{P}^j \check{\underline{d}}^j\right\}\right\}. \tag{5.55}$$

[10] We also recognize for the case where the discrete operator, $\boldsymbol{H}$, is not so well matched with the continuous operator, $\mathcal{H}$, one may choose to use two separate linear operators for each of the weighted responses in (5.48).

Thus, one main complication to precomputing a linear operator to approximate $\underline{\alpha}^j$ is the $[\boldsymbol{V}^j]^{-1}\boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j)$ term. One alternative is to simply eliminate $\boldsymbol{V}^j$ by eliminating the partial orbital sampling of Section 5.3.3, and sample the orbit completely. (With the appropriate spatial subsampling, this may be practical. We will discuss this later in Section 5.5.3.) However, we note that for many desired responses $[\boldsymbol{V}^j]^{-1}\boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j) = \boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j)$. This, of course, depends on the exact nature of $\boldsymbol{V}^j$ and the desired response. However, $\boldsymbol{V}^j$ typically denotes a rotation or flip operation. Thus, for isotropic desired responses, or other responses with a high degree of symmetry, the transformation $\boldsymbol{V}^j$ has no effect. Therefore, we may approximate $\underline{\alpha}^j$ as

$$\underline{\alpha}^j \approx \boldsymbol{V}^j \breve{\boldsymbol{M}}^{(x_j,y_j)} \boldsymbol{P}^j \breve{\underline{d}}^j, \tag{5.56}$$

where the $\eta^3 \times K$ linear operator $\breve{\boldsymbol{M}}^{(x_j,y_j)}$ is defined with columns given by

$$\underline{\breve{m}}_k^{(x_j,y_j)} = \mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j(\underline{e}^j - \underline{l}_0^j)\right\} \odot \mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{G}'\mathrm{diag}\left\{\boldsymbol{C}'_{\mathcal{PS}}\underline{e}^k\right\}\boldsymbol{G}\underline{e}^{(x_j,y_j,z_0)}\right\}\right\} \tag{5.57}$$

Thus, the $\underline{\alpha}^j$ component of the constrained least-squares design may be almost entirely computed, except for the dependence on the diagonal weighting, $\boldsymbol{D}$, which enters (5.56) through the $\breve{\underline{d}}^j$ term, which is computed via (5.31).

5.5.3 Smaller Design Components and Linear Operators

As mentioned previously, for many algorithms that are used to solve the constrained least-squares problem in (4.36), it is sufficient to provide the algorithm with $[\boldsymbol{\Phi}^j]'\boldsymbol{\Phi}^j$ and $[\boldsymbol{\Phi}^j]'\underline{\alpha}^j$. Because $[\boldsymbol{\Phi}^j]'$ also represents a linear operation on $\underline{\alpha}^j$, this gives us an opportunity to further precompute design components. Similarly, since $\boldsymbol{\Phi}^j$ only has B columns (representing the number of penalty basis functions and consequently the neighborhood size), this should also greatly reduce the storage requirements on the linear operators.

Recalling (5.50) and (5.56), we may write

$$\begin{aligned}\left[\boldsymbol{\Phi}^j\right]'\underline{\alpha}^j &\approx \left[\mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j\underline{l}_0^j\right\}\odot\mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{B}^j\right\}\right\}\right]'\boldsymbol{V}^j\breve{\boldsymbol{M}}^{(x_j,y_j)}\boldsymbol{P}^j\underline{\breve{d}}^j &(5.58)\\ &= \left[\mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j\underline{l}_0^j\right\}\odot\mathfrak{F}\left\{\left[\boldsymbol{V}^j\right]^{-1}\boldsymbol{T}^j\boldsymbol{B}^j\right\}\right\}\right]'\breve{\boldsymbol{M}}^{(x_j,y_j)}\boldsymbol{P}^j\underline{\breve{d}}^j, &(5.59)\end{aligned}$$

where we have again assumed that the desired response is sufficiently symmetric to be insensitive to the transformation operation, $[\boldsymbol{V}^j]^{-1}$. Once again, the $\boldsymbol{V}^j$ term poses some difficulty, since the penalty bases in $\boldsymbol{B}^j$ are generally asymmetric. However, recalling the parameterization of the quadratic penalty discussed in Section 4.2.3, one typically selects penalty basis pairs. For example, in (4.22), the horizontal penalty is represented by both $b_{(-1,0)}$ and $b_{(1,0)}$. Thus, in cases where $\boldsymbol{V}^j$ denotes flip operations, $[\boldsymbol{V}^j]^{-1}\boldsymbol{T}^j\boldsymbol{B}^j$ can be represented by an appropriate column permutation operator, $\boldsymbol{P}_V^j$, such that

$$[\boldsymbol{V}^j]^{-1}\boldsymbol{T}^j\boldsymbol{B}^j = \boldsymbol{T}^j\boldsymbol{B}^j\boldsymbol{P}_V^j. \tag{5.60}$$

Thus, (5.59) becomes

$$\begin{aligned}\left[\boldsymbol{\Phi}^j\right]'\underline{\alpha}^j &\approx \left[\mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j\underline{l}_0^j\right\}\odot\mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{B}^j\boldsymbol{P}_V^j\right\}\right\}\right]'\breve{\boldsymbol{M}}^{(x_j,y_j)}\boldsymbol{P}^j\underline{\breve{d}}^j &(5.61)\\ &= \left[\boldsymbol{P}_V^j\right]'\left[\mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j\underline{l}_0^j\right\}\odot\mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{B}^j\right\}\right\}\right]'\breve{\boldsymbol{M}}^{(x_j,y_j)}\boldsymbol{P}^j\underline{\breve{d}}^j, &(5.62)\end{aligned}$$

which may be calculated using linear operators as

$$\left[\boldsymbol{\Phi}^j\right]'\underline{\alpha}^j \approx \left[\boldsymbol{P}_V^j\right]'\breve{\boldsymbol{M}}^{(x_j,y_j)}\boldsymbol{P}^j\underline{\breve{d}}^j, \tag{5.63}$$

with columns of the $B\times K$ operator, $\breve{\boldsymbol{M}}^{(x_j,y_j)}$, defined as

$$\begin{aligned}\underline{\breve{m}}_k^{(x_j,y_j)} &= \left[\mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j\underline{l}_0^j\right\}\odot\mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{B}^j\right\}\right\}\right]'\cdot\\ &\quad\left[\mathfrak{F}^{-1}\left\{\mathfrak{F}\left\{\boldsymbol{T}^j(\underline{e}^j-\underline{l}_0^j)\right\}\odot\mathfrak{F}\left\{\boldsymbol{T}^j\boldsymbol{G}'\mathrm{diag}\left\{\boldsymbol{C}'_{\mathcal{PS}}\underline{e}^k\right\}\boldsymbol{G}\underline{e}^{(x_j,y_j,z_0)}\right\}\right\}\right]. &(5.64)\end{aligned}$$

As mentioned previously, some complications can be removed by performing a more complete orbital sampling, and eliminating the transformations indicated by

$\boldsymbol{V}^j$ and equivalently $\boldsymbol{P}_V^j$ in (5.63). (This does *not* preclude the type of spatial subsampling where every n_dth voxel is sampled, and the remaining positions are interpolated.) Because we have developed reduced dimension linear operators that are only $B \times K$, it is often feasible to perform such a complete orbital sampling.

Computation and Storage Requirements for Penalty Design

In general, the precomputation phase of the penalty design can be quite computationally expensive. However, since the precomputation step in (5.64) must only be performed once for a given system geometry and desired response, these computations typically do not present a major problem for most applications.

One must store the precomputed design components. For a desired shift-invariant local impulse response, the storage requirements are generally determined by the number and size of the precomputed linear operators used for approximating $[\boldsymbol{\Phi}^{j_0}]'\underline{\alpha}^j$. (There is some additional storage required for the $B \times B$ matrix $[\boldsymbol{\Phi}^{j_0}]'\boldsymbol{\Phi}^{j_0}$.) Thus, for a shift-variant PET system that is sampled over the entire volume (*i.e.*, no attempt to use orbital symmetries) with a 3D subsampling on a grid of every n_dth voxel:

$$\frac{B}{n_d^3} P_x P_y P_z \frac{n_a}{n_s} \tag{5.65}$$

floating point numbers must be stored. For a shift-variant elliptical orbit SPECT system that is subsampled over a single quadrant in one slice:

$$\frac{B}{4} \frac{P_x P_y}{n_d^2} \frac{n_a}{n_s} \tag{5.66}$$

floats are needed. Lastly, in an idealized PET system where the geometric response is shift-invariant only a single $B \times \frac{n_a}{n_s}$ operator is required.

All methods involve application of a $B \times \frac{n_a}{n_s}$ matrix to the position-dependent weights evaluated via (5.31) at each voxel. While there is some computational cost

associated with various permutations and transformations, this is generally small. Additionally, there is a nonnegligible cost in interpolating the matrices for unsampled positions. (One can alternately, interpolate the design weights themselves, although this will yield different results since the penalty design is constrained.) Overall, these methods yield very practical computation times for many applications.

5.5.4 A Scalable Penalty Design

In all of the penalty design methods previously discussed, one must recompute a new penalty matrix $\boldsymbol{R}$ for every desired response, l_0^j. For example, one might want to perform reconstructions with a set of desired responses with different FWHM resolutions. Each resolution requires a separate $\boldsymbol{R}$ calculation, much of which may be precomputed as discussed earlier in this chapter. For further simplification, in this section we present a specific class of desired shift-invariant impulse responses appropriate for idealized PET systems, which require only a single penalty matrix computation.

As mentioned below (2.26), for traditional space-invariant penalties, the w_{jk} terms in (2.26) include the regularization parameter β, which controls the mean global resolution. For shift-invariant penalties where β is a simple multiplication factor we may write $\boldsymbol{R} = \beta \boldsymbol{R}_0$, where $\boldsymbol{R}_0$ specifies the relative penalty strength between pixel pairs and β controls the mean global resolution. Therefore, it is simple to generate new $\boldsymbol{R}$ for different desired resolutions. (One does not have to recompute $\boldsymbol{R}_0$.)

Just as the conventional shift-invariant penalty is a simple function of β, we would like to design the penalty matrix $\boldsymbol{R}$ as a product of a user-selected β and a β-independent $\boldsymbol{R}^\star$, *i.e.* $\boldsymbol{R} = \beta \boldsymbol{R}^\star$, yet still yields uniform resolution properties. In terms of our parameterization of $\boldsymbol{R}$, we would like factorable coefficients such that

$\underline{w}^j = \beta \underline{v}^j$. Making this substitution into (4.36) yields

$$\underline{\hat{w}}^j = \beta \underline{\hat{v}}^j, \qquad \underline{\hat{v}}^j \triangleq \arg\min_{\underline{v}^j \geq \underline{0}} \left\| \beta \boldsymbol{\Phi}^j \underline{v}^j - \underline{\alpha}^j \right\|^2. \tag{5.67}$$

The penalty matrix $\boldsymbol{R}^\star$ is completely specified by $\{\underline{v}^j\}_{j=1}^p$. However, the minimization in (5.67) depends on β. We eliminate this dependence by a particular choice of the target frequency response $\underline{L}_0^j$ and the weighting $\boldsymbol{W}^j$ in (4.37) and (4.38).

Let us consider the idealized PET system where $\boldsymbol{H} = \mathrm{diag}\{c_i\}\,\boldsymbol{G}$ is an appropriate system matrix factorization. In this case, the penalty design components in (4.37) and (4.38) may be written as

$$\boldsymbol{\Phi}^j = \boldsymbol{W}^j \mathrm{diag}\left\{\underline{L}_0^j\right\} \mathfrak{F}\left\{\boldsymbol{B}^j\right\} \tag{5.68}$$

$$\underline{\alpha}^j = \boldsymbol{W}^j \mathrm{diag}\left\{\underline{e}^j - \underline{L}_0^j\right\} \mathfrak{F}\left\{\boldsymbol{G}'\boldsymbol{D}\boldsymbol{G}\underline{e}^j\right\}. \tag{5.69}$$

We will choose the desired response to be equal to the local impulse response of a penalized unweighted least-squares (PULS) estimator with penalty matrix $\boldsymbol{R} = \beta \boldsymbol{R}_0$ and with c_i's are all unity. Not only does this choice for a desired response result in a class of scalable penalties, it closely resembles the responses for conventional PL estimators, and should be relatively easy to achieve via penalty design. (Recall from the discussion in Section 4.3.1, that some desired responses are unachievable, or will require very large penalty neighborhoods.) This response is written:

$$\underline{l}_0^j = [\boldsymbol{G}'\boldsymbol{G} + \beta \boldsymbol{R}_0]^{-1} \boldsymbol{G}'\boldsymbol{G}\underline{e}^j. \tag{5.70}$$

We note that if the c_i terms are not unity, we may obtain the same response by precorrecting the measurement data (by multiplying the data by $\mathrm{diag}\left\{c_i^{-1}\right\}$) and performing a PULS reconstruction using only the geometric model, $\boldsymbol{G}$. If $\boldsymbol{R}_0$ is chosen to be a space-invariant penalty, the response, $\underline{l}_0^j$, is approximately independent of the choice of j since $\boldsymbol{G}'\boldsymbol{G}$ is a nearly shift-invariant operator. This particular choice of

$\underline{l}_0^j$ has a form very similar to the local impulse response in (3.15) and has resolution controlled by the parameter β.

Using the circulant approximation, we express the desired frequency response as

$$\underline{L}_0^j \approx \frac{\mathfrak{F}\left\{\boldsymbol{G}'\boldsymbol{G}\underline{e}^j\right\}}{\mathfrak{F}\left\{\boldsymbol{G}'\boldsymbol{G}\underline{e}^j\right\} + \beta\mathfrak{F}\left\{\boldsymbol{R}_0\underline{e}^j\right\}}. \tag{5.71}$$

Similarly we may write

$$1 - \underline{L}_0^j \approx \frac{\beta\mathfrak{F}\left\{\boldsymbol{R}_0\underline{e}^j\right\}}{\mathfrak{F}\left\{\boldsymbol{G}'\boldsymbol{G}\underline{e}^j\right\} + \beta\mathfrak{F}\left\{\boldsymbol{R}_0\underline{e}^j\right\}} \tag{5.72}$$

For the particular choice (5.70) of $\underline{l}_0^j$, the denominators of (5.71) and (5.72) are identical. Additionally, β is in the numerator of (5.72) and not in the numerator of (5.71). If we choose a least-squares weighting of $\boldsymbol{W}^j = \breve{\boldsymbol{W}}^j \mathrm{diag}\{\mathfrak{F}\{(\boldsymbol{G}'\boldsymbol{G} + \beta\boldsymbol{R}_0)\underline{e}^j\}\}$ the denominators of (5.71) and (5.72) disappear in (5.68) and (5.69), and we can rewrite the penalty design as

$$\begin{aligned}
\hat{\underline{v}}^j &= \arg\min_{\underline{v}^j \geq \underline{0}} \left\| \boldsymbol{\Phi}^j \underline{v}^j - \underline{\alpha}^j \right\|^2 \\
\boldsymbol{\Phi}^j &= \breve{\boldsymbol{W}}^j \mathrm{diag}\left\{\mathfrak{F}\left\{\boldsymbol{G}'\boldsymbol{G}\underline{e}^j\right\}\right\} \mathfrak{F}\left\{\boldsymbol{B}^j\right\} \\
\underline{\alpha}^j &= \breve{\boldsymbol{W}}^j \mathrm{diag}\left\{\mathfrak{F}\left\{\boldsymbol{R}_0\underline{e}^j\right\}\right\} \mathfrak{F}\left\{\boldsymbol{G}'\boldsymbol{D}\boldsymbol{G}\underline{e}^j\right\},
\end{aligned} \tag{5.73}$$

where $\breve{\boldsymbol{W}}^j$ is an additional (optional) least-squares weighting. Note that the design (5.73) is independent of β, as desired.

Once we have calculated the parameters $\{\hat{\underline{v}}^j\}_{j=1}^p$ using (5.73), we construct the penalty matrix $\boldsymbol{R}^\star$ using (4.19) with $\underline{w}^j = \beta\hat{\underline{v}}^j$. This $\boldsymbol{R}^\star$ has been designed to provide global isotropic resolution properties and, because of the least-squares weighting leading to (5.73), $\boldsymbol{R}^\star$ is independent of the choice of the regularization parameter β. Therefore, once $\boldsymbol{R}^\star$ is calculated one may specify a desired global resolution through β. The penalty matrix is given by the simple relation $\boldsymbol{R} = \beta\boldsymbol{R}^\star$. (A method relating β to the FWHM resolution is discussed in [32].)

The computational speed-ups discussed in Sections 5.5.1, 5.5.2, and 5.5.3 can also be applied to (5.73) by using $\breve{\boldsymbol{W}}^j = \mathfrak{F}\{\cdot\}$ and following the same series of developments.

5.6 Summary

In this chapter we have discussed approximations and observations that allow one to calculate weighted projection-backprojections with greatly reduced computation. The key to these fast methods involves factoring the system model into object-independent and object-dependent portions, and applying a precomputed linear operator to the object-dependent component. Once these operators have been precomputed for the object-independent system factors, they may be applied repeatedly using far fewer computation, than if the weighted projection-backprojections were computed directly.

Because the precomputed operators are linear, we may precompute any additional linear operations (like the Fourier transform) that must be applied to the weighted projection-backprojections. This leads to fast computation of local impulse responses and covariance estimates. Similarly, since we have developed a linearized penalty design technique in Section 4.3, one can precompute relatively small operators that may be applied very quickly to generate a space-variant penalty. In the next chapter we demonstrate the use of these operators for resolution and covariance prediction, and for penalty design for uniform resolution.

CHAPTER VI

Application to Emission Tomography

In this chapter we investigate the performance of the fast predictors and resolution control methods discussed for penalized-likelihood estimators in Chapters IV and V. In Section 6.1, we demonstrate the accuracy of our fast resolution and covariance predictors for a fully 3D SPECT system and a simulated anthropomorphic phantom with realistic emission rates and nonuniform attenuation coefficients.

In Section 6.2, we demonstrate the application of the penalty design for resolution control in PET systems. This study includes 2D and 3D systems, systems with shift-invariant and shift-variant geometric responses, test phantoms and anthropomorphic phantoms, simulated data and real data, and various versions of the penalty design techniques discussed in Sections 4.3, 4.4, and 5.5.

In Section 6.3, we demonstrate the application of the penalty design for resolution control in SPECT systems. In this section we are particularly interested in finding methods with exactly matched resolution properties, and methods that can provide uniform resolution across a wide range of desired responses.

6.1 Validation of the Fast Resolution and Covariance Predictors

In this section, we present a study on fast resolution and covariance predictors for penalized-likelihood image reconstructions. Predictors based on the circulant approx-

imation in Section 5.2.1 have been used extensively. Thus, the goal is not to validate that approximation, but to investigate the various other approximations that were discussed in Chapter V. Specifically, we are interested in the SPECT attenuation approximation in (5.14) and the truncation and reduced angle approximations.

We have seen previously in Figure 5.6 that the approximated weighted projection-backprojections for PET are nearly indistinguishable from the direct evaluations. In fact, under certain conditions (projection-constant weights and a shift-invariant geometric response), we know that the PET approximations will be exact. Because we can observe slight differences in the approximated weighted responses for SPECT systems (recall Figure 5.1), it seems more likely that the local impulse response and covariance predictions could be inaccurate in that case. For that reason we present a SPECT investigation on the validity of the resolution and covariance predictors that use the fast techniques of Chapter V. We demonstrate that our fast predictions are highly accurate and we compare the performance of our resolution and covariance predictors versus more traditional predictors and estimators on a simulated fully 3D SPECT system.

6.1.1 3D SPECT system and object model

The SPECT model includes 128 projection angles and 128×30 pixel projection views with 4.5 mm pixels. The image volume is discretized into $128 \times 128 \times 30$ voxels, where each voxel is a 4.5 mm cube. The SPECT camera follows an elliptical orbit with a 283 mm radius on the x-axis and a 220 mm radius on the y-axis. The SPECT detector model includes a depth-dependent Gaussian response that is 1.75 mm FWHM at the face of the collimator and increases linearly with a slope of 0.044 as the distance to the collimator is increased. When the camera aims along the x-axis, this slope corresponds to a FWHM of about 14.2 mm at the center of the

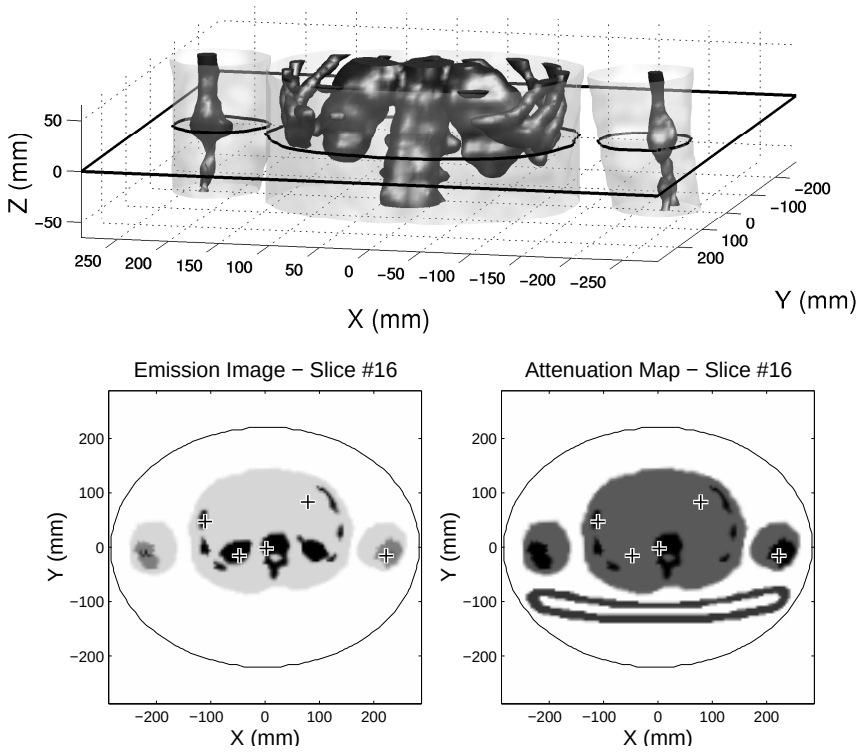


Figure 6.1: 3D digital anthropomorphic phantom used in the 3D SPECT simulation studies. The anthropomorphic phantom simulates a ^{99m}Tc bone scan, with high activity in the bones and kidneys. A central transverse slice of the emission image and the attenuation map is also shown. The orbit of the SPECT camera is indicated by the black ellipse. Additionally, five positions are indicated with + marks for the investigation of resolution and covariance predictors.

field of view.

We chose to simulate a human abdomen ^{99m}Tc bone scan using the Zubal phantom[2, 144]. We modified this phantom to include an attenuating table and resampled the data onto a 4.5 mm grid. Figure 6.1 displays this phantom data. We assigned relative emission rates of 3.0 to the spine, rib cage, and kidneys, 1.5 to the long bones in the arms, 3.0 to the long bone marrow, and 0.5 to the remaining soft tissue background. The attenuation map used attenuation coefficients appropriate for 140 keV photons with 0.23 cm^{-1} for bone, 0.15 cm^{-1} for all soft tissues, and 0.18 cm^{-1} for the table.

We generated simulated SPECT measurements from the above phantom and system model. All studies used pseudo-random Poisson measurement data with a mean of 500,000 counts per slice, including a 20% known uniform background level (the r_i

terms in (2.7)) to approximate the effects of scatter.

6.1.2 Reconstruction

We applied the penalized-likelihood estimator in (2.24) for reconstructing the emission images from the measurement data. The penalized-likelihood objective was maximized using an ordered-subsets paraboloidal surrogates iterative approach [29, 27, 28]. The algorithm was initialized with a filtered backprojection reconstruction. Following many iterations using 16 subsets, we applied convergent single subset iterations, to ensure a nearly converged solution. For the penalty function we use a shift-invariant first-order quadratic penalty with the regularization parameter chosen to yield a spatial resolution of about 2 cm at the center of the field of view. For this penalty, the resolution at the edge of the object was about 4.5 mm. The reconstruction model matches the projection model exactly and used the true attenuation map.

6.1.3 Resolution Prediction

For the above SPECT system with Poisson measurements, the local impulse response of the penalized-likelihood estimator is given in (3.15) with diagonal components,[1]

$$\boldsymbol{D}_1 = \boldsymbol{D}_2 = \operatorname{diag}\left\{ \frac{1}{\bar{Y}_i(\underline{\theta})} \right\}. \tag{6.1}$$

The "traditional" slow approach to computing the local impulse response is to evaluate (3.15) iteratively. We initialized iterations with an impulse at the response position and used 500 conjugate gradient iterations to estimate the response. This yields a well-converged estimate. We compare this approach to the fast predictions described in Section 5.4. For all fast predictions, we used the precomputed linear

[1] For the diagonal in (6.1), we have assumed that the blur due to the system model is much greater than the blur induced by regularization of the estimator. Thus, $\underline{\bar{Y}}(\underline{\theta}) \approx \underline{\bar{Y}}(\underline{\check{\theta}})$.

operators given in (5.36). The predictors were applied using the modified diagonal elements in (5.31).

Because the resolution properties of SPECT systems are space-variant, we investigated the resolution at several positions in the object. These positions are identified with + marks in the left two central slice images in Figure 6.1. From left to right, we label these positions: "Rib," "Left kidney," "Center," "Soft tissue," and "Right elbow."

For the first resolution investigation, we used precomputed operators with a $30 \times 30 \times 30$ subvolume (*i.e.*, $\eta = 30$ in (5.42)) and 32 blocks of 4 angles. We stored operators within a single quadrant of the elliptical orbit and used a spatial subsampling with $n_d = 6$. Operators for unsampled positions are formed using bilinear interpolation. Thus, the precomputed and stored operators may be stored as single precision floating point numbers in approximately 125 Mb.

Figure 6.2 compares the local impulse responses at four different locations. The left set of figures compares local impulse responses calculated at the "Rib" position. Transverse, sagittal, and coronal slices of the 3D response are shown for the iteratively calculated response (top row) and for our fast prediction (middle row). The bottom row shows profiles through each axis of the iteratively calculated response (dashed line) and the fast prediction (solid line). The right set of figures shows axial profiles for three more points. (None of these locations coincide with operator sampling positions. Thus, all fast predictions are based on interpolated operators.) The local impulse responses are space-variant and anisotropic with coarser resolution near the center of the field of view. Despite the multiple approximations and subsampling, our predictions are very close to the iteratively calculated responses. This is true even for the "Rib" position where the attenuation map changes rapidly near the

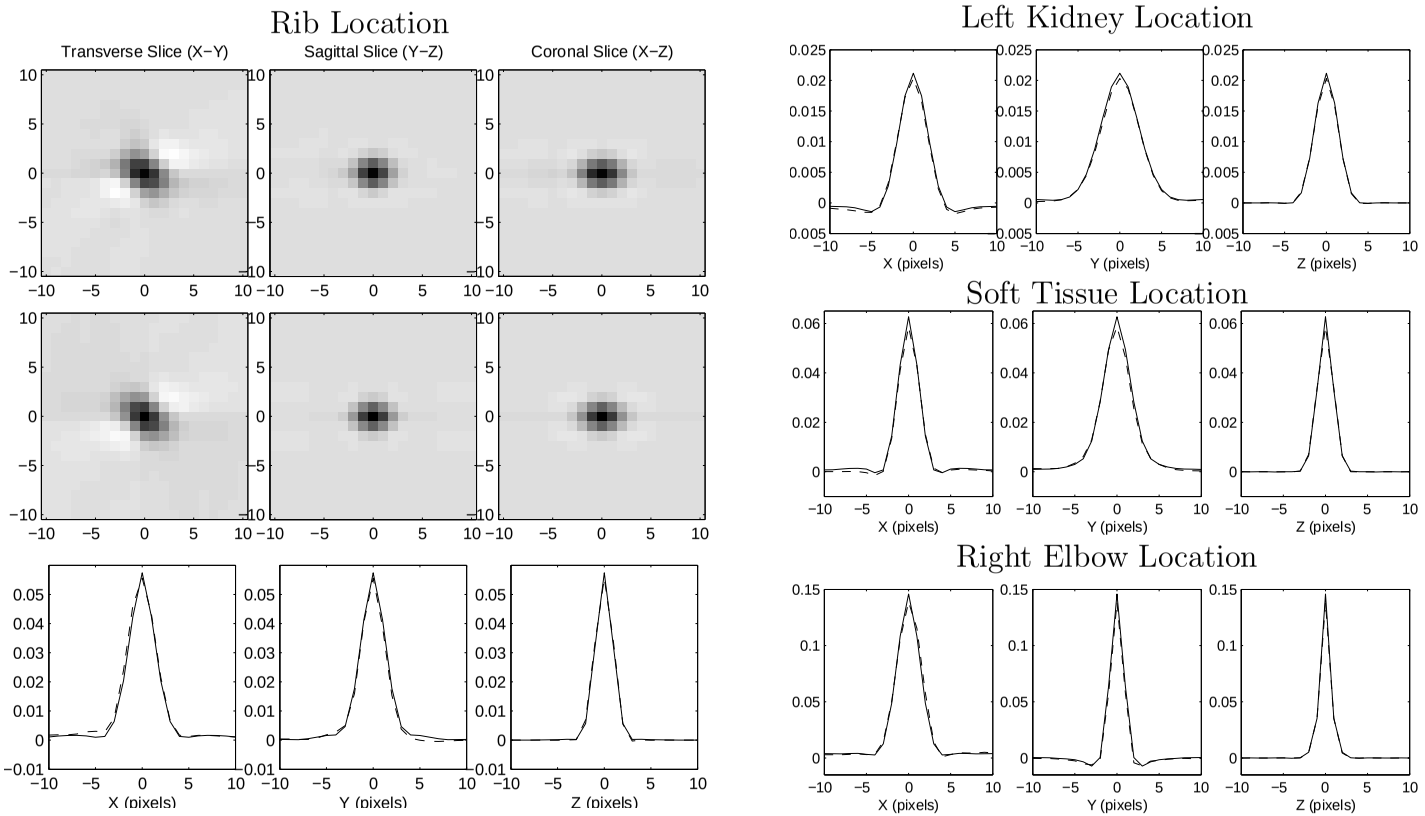


Figure 6.2: A comparison of local impulse responses using our fast predictors versus traditional iterative evaluation. We compare resolution prediction for the two methods at four locations in the image volume identified in Figure 6.1. The left column compares transverse, sagittal, and coronal images of the 3D local impulse response at the rib location using the iterative method (top row) versus the fast predictor (middle row). The bottom row shows profiles through each axis of the response for the iterative method (dashed line) and the fast predictor (solid line). The right column shows the axial profiles for three more image locations.

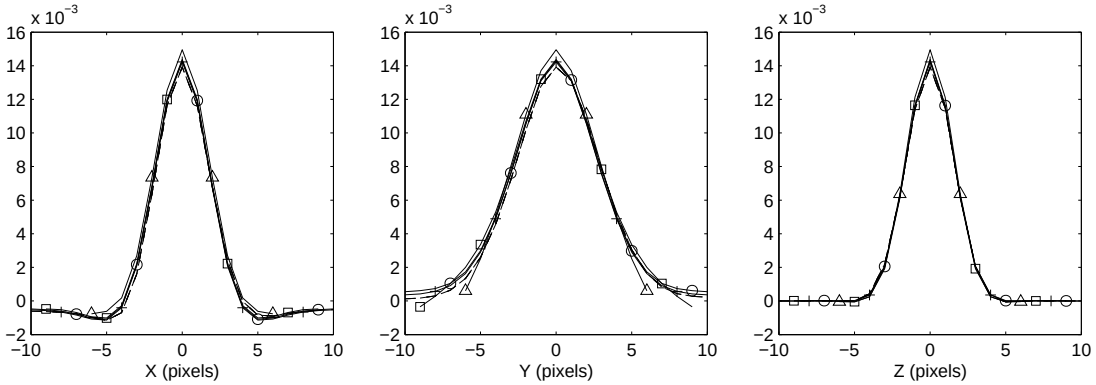
evaluation position.

Assuming the $\boldsymbol{A}$ and $\hat{\boldsymbol{H}}$ matrices in (5.31) have been precomputed and loaded, the Matlab implementation of the resolution predictor used to compute the above predictions, takes roughly 1/15 of a second to compute a single local impulse response on an 800 Mhz Pentium III computer. For comparison with the "traditional" slow iterative approach, we note that a single projection operation, $\boldsymbol{H}\underline{\theta}$, implemented as an "on-the-fly" procedure in an efficient compiled C program takes more than a minute on the same computer.

The required size of the precomputed operators depends on a number of factors including the desired accuracy of the approximation, available storage, desired computation speed, the space-variance of the system, and the space-variance due to the object. We present two studies where the size of the operators are varied and briefly discuss the associated trade-offs.

We first studied the local impulse responses at the five positions shown in Figure 6.1 using operators computed with a range of support sizes. Specifically, cases where 60^3, 30^3, 20^3, and 14^3 voxels are stored. All angles are stored (*i.e.*, 128 blocks) and the locations are sampled positions (therefore no interpolation of operators is performed). The results of this investigation are presented in Figure 6.3.

Most support sizes give remarkably similar predictions across the supported pixels, even for the smaller support sizes where there is significant truncation of the local impulse response function. However, there are some noticeable differences for the smaller support sizes. Specifically, with additional truncation there are growing mismatches in the sidelobe behavior shown in the profiles for the center pixel's response. Similarly, for the smallest subvolume, a mismatch in the peak value of the local impulse response begins to be evident. We quantify this local impulse response



Summary of Normalized Errors

Support Size	Comp. Time	Center	Left Kidney	Right Elbow	Rib	Soft Tissue
60^3	871 ms	4.3%	3.6%	5.0%	7.4%	4.7%
30^3	109 ms	4.8%	4.1%	5.0%	7.8%	6.5%
20^3	32 ms	7.1%	4.2%	5.0%	12%	8.1%
14^3	13 ms	8.0%	7.1%	5.4%	13%	9.5%

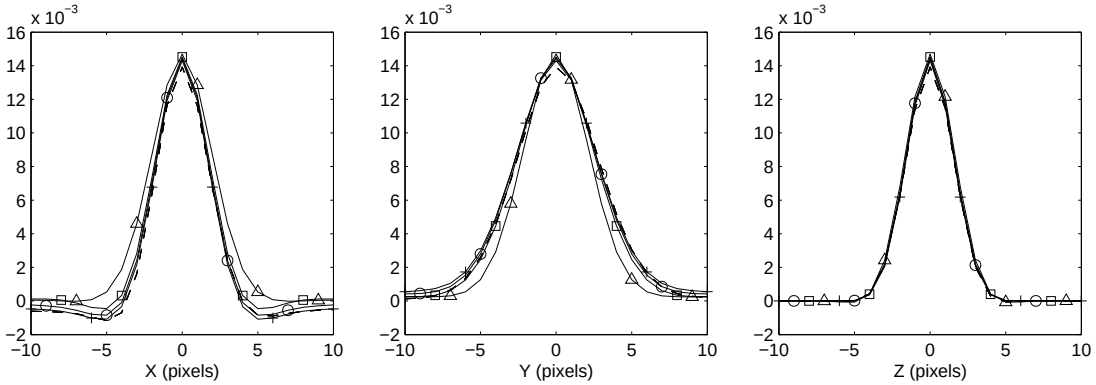
Figure 6.3: Resolution prediction with varying support size.
The plots above show profiles through the X, Y, and Z axes of the 3D local impulse response at the center voxel with a support size of 60^3 voxels (+), 30^3 voxels (○), 20^3 voxels (□), and 14^3 voxels (△). The iteratively computed response is also shown (dashed line). The table summarizes normalized error and computation time for resolution predictions at various locations and support sizes.

mismatch for the five locations in the table in Figure 6.3, where we have defined the normalized error as

$$\frac{\max_k \left| \hat{l}_k^j - l_k^j \right|}{\max_k \left| l_k^j \right|} \cdot 100\%, \tag{6.2}$$

where l_k^j denotes elements of the local impulse response at the jth location, as calculated by the "traditional" iterative approach, and $\hat{l}_k^j$ denotes elements of the response as calculated by the fast approach. We also list the computation time for a single local impulse response evaluation for each support size. Since it appears that relatively good approximations can be made within the stored support, one may only need to store voxels over a region slightly larger than the desired portion of the response. This not only saves storage space for the precomputed operators, but also decreases prediction computation time by greatly reducing the dimension of the matrix multiplications.

We performed a second study, where the support size is held constant using 30^3 voxels and the angular subsampling is varied with 128 blocks, 16 blocks, 8 blocks, and a single block. Figure 6.4 summarizes these results. For the coarser angular sampling, there are significant differences in the sidelobe behavior. These differences are most noticeable in the negative sidelobes in the X profile for the two coarsest samplings. These mismatches should be most pronounced in locations that differ from the geometric response in a very anisotropic fashion. The degree of mismatch will of course depend on the particular angular sampling and the properties of the object and system geometry. For this particular object and geometry, using only 8 blocks still yields approximations with less than 10% normalized error. We note that the "Rib" location generally has higher errors than the other locations. This is most likely due to the rapid local changes in attenuation, which are less likely to fit the approximation made in (5.14).



Summary of Normalized Errors

Angular Blocks	Comp. Time	Center	Left Kidney	Right Elbow	Rib	Soft Tissue
128	109 ms	4.8%	4.1%	5.0%	7.8%	6.5%
16	55 ms	5.7%	4.8%	5.0%	7.6%	7.0%
8	52 ms	8.8%	7.0%	5.2%	8.3%	6.5%
1	48 ms	21%	15%	15%	19%	20%

Figure 6.4: Resolution prediction with varying angular sampling.
The plots above show profiles through the X, Y, and Z axes of the 3D local impulse response at the center voxel with 128 blocks (+), 16 blocks (○), 8 blocks (□), and a single block (△). The iteratively computed response is also shown (dashed line). The table summarizes normalized error and computation time for resolution predictions at various locations and angular samplings.

One other adjustable value is the coarseness of the operator position sampling represented by n_d. We have found that one can use a fairly coarse sampling ($n_d = 6$), since the (unweighted) geometric response varies very smoothly. Finer sampling helps reduce interpolation computations. However, the required sampling is quite coarse, and ultimately depends on the particular system geometry.

6.1.4 Covariance Prediction

We also investigated local covariance predictions. We compared the fast predicted covariance functions versus empirical covariance functions estimated from 500 noisy reconstructions. As with the resolution predictors, we use the precomputed operators given in (5.36) in conjunction with the modified diagonal elements stated in (5.31). We use the covariance equation given in (5.39) and the diagonal weighting $\boldsymbol{D}_3 = \boldsymbol{D}_1$, as in (6.1). We used the same operator dimensions and subsamplings as in the initial resolution investigation.

Figure 6.5 presents the empirical covariance functions and the predicted covariance functions for four positions in the digital phantom. The variation in the sample covariances is quite evident in the image slices and the profiles. Thus, we have included error bars on the sample covariance estimates (based on an assumption of a Gaussian distribution of the reconstructed image values). These error bars indicate plus and minus one standard deviation of the covariance estimate. The covariance predictions appear quite accurate over these four positions, lying within the error bars for most locations. It seems likely that these predictions would be sufficiently accurate for typical applications such as making variance images or evaluating computer observer performance.

We performed one final investigation of the accuracy and speed of the predictions. We calculated a variance image for the central slice of the 3D phantom. We used

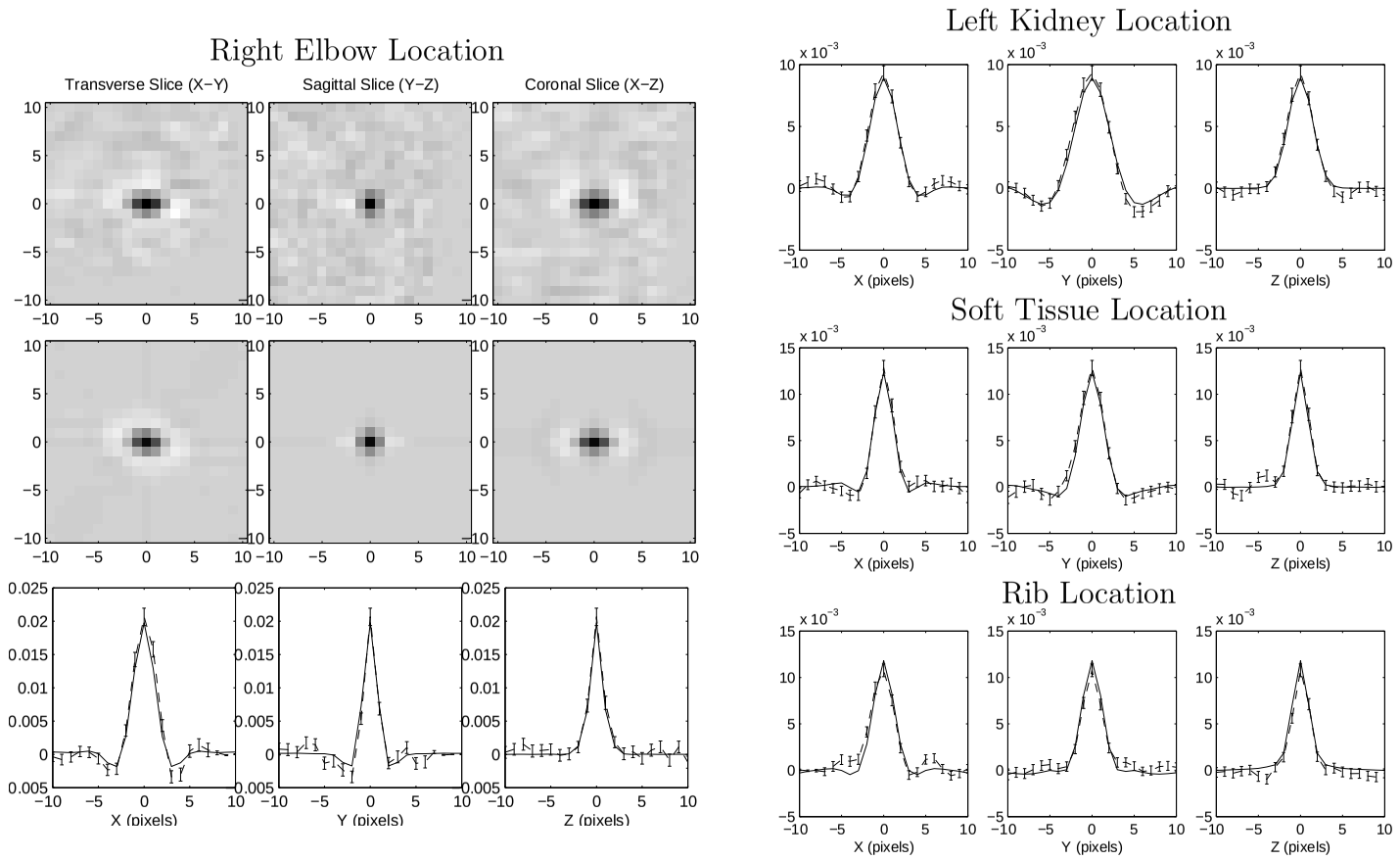


Figure 6.5: A comparison of covariance functions calculated using fast predictors and an empirical sample covariance estimate. Covariance predictions are shown for the four locations in the image volume identified in Figure 6.1. The left column compares transverse, sagittal, and coronal images of the 3D covariance function at the elbow location calculated from the 500 reconstructions (top row) versus the fast predictor (middle row). The bottom row shows profiles through each axis of the response for sample covariance (dashed line) and the fast predictor (solid line). The right column shows the axial profiles for three more image locations.

precomputed operators with 12^3 voxels and 16 blocks of 8 pixels. We stored operators with $n_d = 1$ over a single quadrant (within the elliptical orbit). This takes approximately 160 Mb of storage space.

We used the variance predictor discussed in Section 5.4.2, which eliminates the inverse Fourier transforms. We applied the scaling technique developed by Qi in [96] in an attempt to account for the effects of nonnegativity constraint on the reconstructed images. Figure 6.6 shows the predicted and empirical standard deviation images. Sample standard deviations were calculated using the 500 noisy reconstructions (left image) and the fast predictors (center). We also show a central horizontal profile of the standard deviations, which have been normalized to be a percentage of the warm background in the phantom. Plus and minus single standard deviation error bars on the sample variance estimates are also shown.

The predictions agree very well with the sample variance estimates. The regions with the greatest disagreement appear to be the regions where the nonnegativity constraint is active. This suggests that the post-computation correction factor developed in [96] does not fully model the nonnegativity constraint.

Given the precomputed matrices specified by (5.36) and the precomputed bilinear interpolator, $\hat{\boldsymbol{H}}$ in (5.31), the entire (single slice) standard deviation image was computed in less than 20 seconds using a Matlab implementation on an 800 MHz Pentium III processor. Thus, the variance of the entire volume can be predicted in less than 10 minutes. We expect that efficient routines written in a compiled C program would be significantly faster.

The prediction speed is a function of the size and sampling of the precomputed operators, $\boldsymbol{M}^j$. Thus, the speed is directly related to how many precomputations one is willing to store. We have demonstrated that accurate predictions can be made

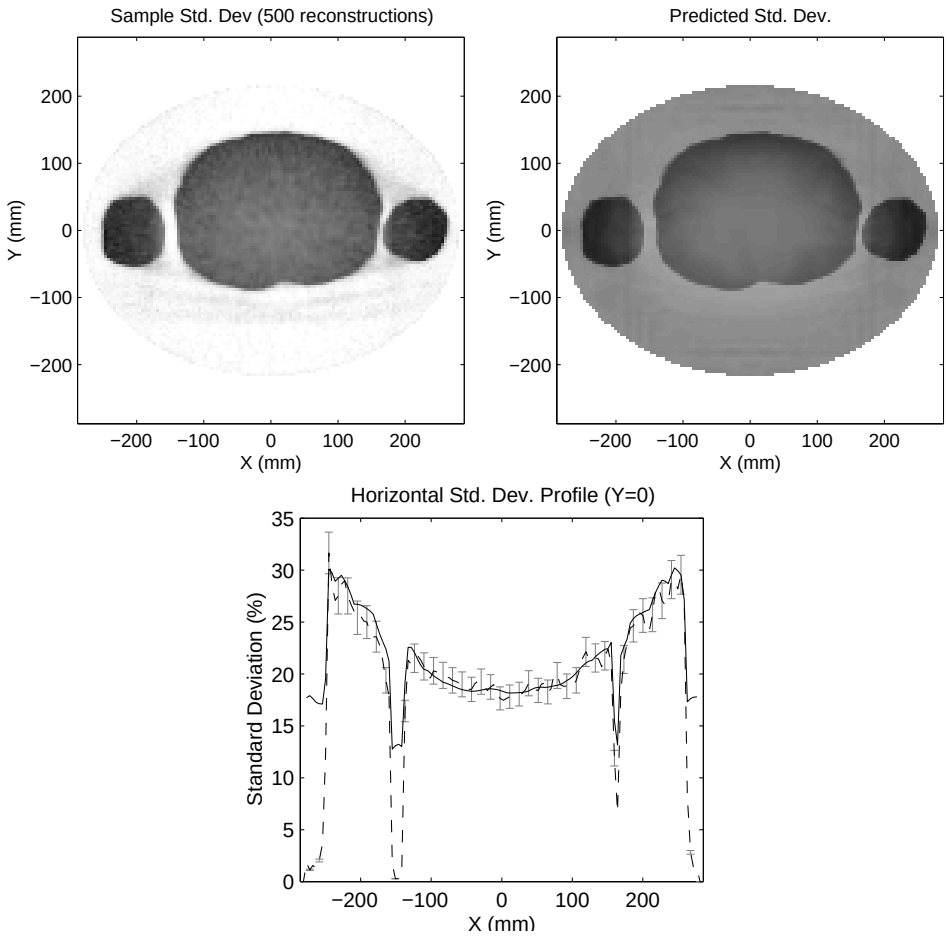


Figure 6.6: A comparison of standard deviations predicted for the 3D anthropomorphic phantom. This figure shows the central slice standard deviation images created from calculating the sample standard deviation of 500 3D reconstructions, and from our fast variance predictor. A horizontal profile shows the sample standard deviation values (dashed line) with error bars and the predicted values (solid line).

for practical storage sizes (*e.g.*, better than 10% error with 125 Mb of storage for the sample SPECT system we have investigated), but the exact trade-off must be specified by the user's requirements on the accuracy and speed of the predictions.

The fast predictors we have developed are most appropriate for situations that require repeated predictions for a static system geometry. Such situations include object-dependent penalty design like that discussed in Chapter IV and in [95], where predictions are required for every voxel. Without fast techniques for resolution and noise prediction, these penalty designs methods would be too slow for practical implementations. Similarly, such fast predictors are also important for the study of computer observers [11], where repeated covariance estimates may be required. While these fast techniques have focused on emission tomography, many of the general ideas may be able to be extended to other imaging systems.

6.2 Resolution Control for PET Systems

In this section, we apply the penalty design methods of Chapter IV to PET systems. Specifically, we concentrate on the goal of uniform resolution. We also use many of the practical techniques discussed in Chapter V to design penalties with reasonable computation times.

6.2.1 2D PET with Shift-Invariant Geometric Response

We first apply our penalty design with the goal of uniform resolution to an idealized PET system that has a shift-invariant geometric response. After calculating the penalty, we performed a local impulse response investigation comparing the relative resolution uniformity of different estimators and regularization schemes.

We return to the system model discussed at the end of Section 3.4.1 with the digital phantom shown in Figure 3.5. We have already shown local impulse response

maps for a conventional first-order shift-invariant penalty in Figure 3.6 and for the certainty-based penalty of [41] in Figure 4.1. To obtain these local impulse responses, we approximated the solution of (3.15) using 40 iterations of a coordinate ascent algorithm initialized with a circulant approximation of the solution. In this section we do the same for our proposed penalty. In addition, we show local impulse response maps for FBP and PULS to demonstrate the relative uniformity. We have implemented the PULS estimator using measurement data that is precorrected for nonuniform c_i terms, so that the estimator has a global response given by (5.70).

Local Impulse Responses

We present two versions of our penalty: (1) the CNLLS penalty of (4.32) computed using a BFGS quasi-Newton method[89], and (2) the computationally efficient linear operator approach in conjunction with the β-independent design discussed in Section 5.5.4 and stated in (5.73). Moreover, for the second approach, we found constrained penalty coefficients using the greedy approach outlined in Table 4.1. The desired response is chosen to be the PULS response given in (5.70) where $\boldsymbol{R}_0$ represents the conventional first-order shift-invariant penalty represented by the filter in (4.21). Both of our proposed designs used a second order neighborhood with eight basis functions (for the eight nearest pixel neighbors[2]). FBP used the apodizing window discussed in [32] to match the PULS target response. All reconstruction methods and penalties were designed with a target resolution of 4.0 pixels (1.2 cm) FWHM resolution. (The relationship between global FWHM resolution and β, and how to calculate β is discussed in [32].)

Figure 6.7 shows the local impulse response map for the penalized-likelihood es-

[2]We have not included the 0th-order magnitude penalty in this penalty design. We have included a basis function for the magnitude penalty in some studies; however, the penalty design almost always results in a zero coefficient for this basis. Thus, in most circumstances this basis can be eliminated.

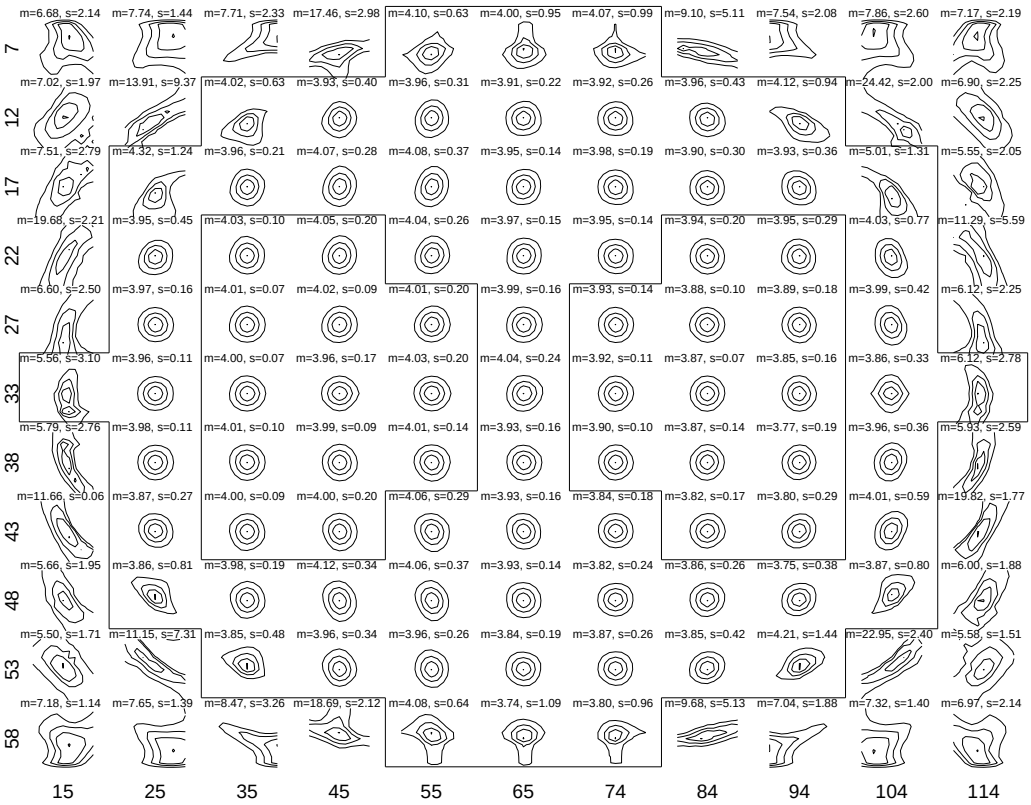


Figure 6.7: Local impulse response map for a PLE with the CNLLS penalty.

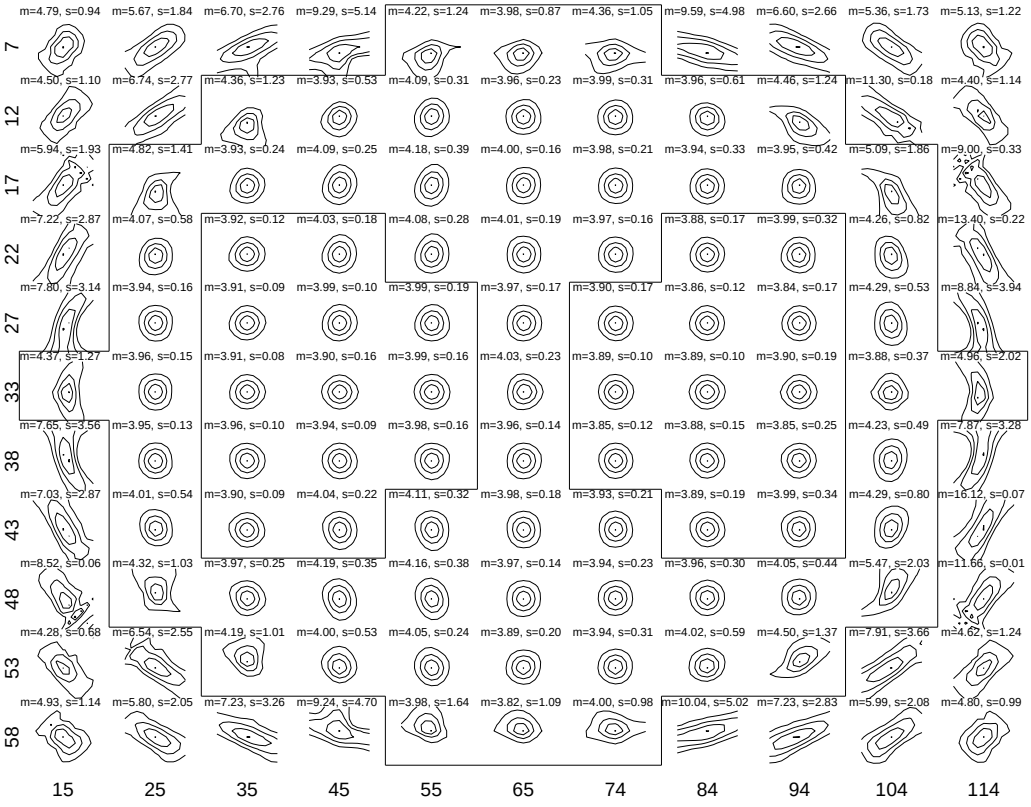


Figure 6.8: Local impulse response map for a PLE with the fast proposed penalty.

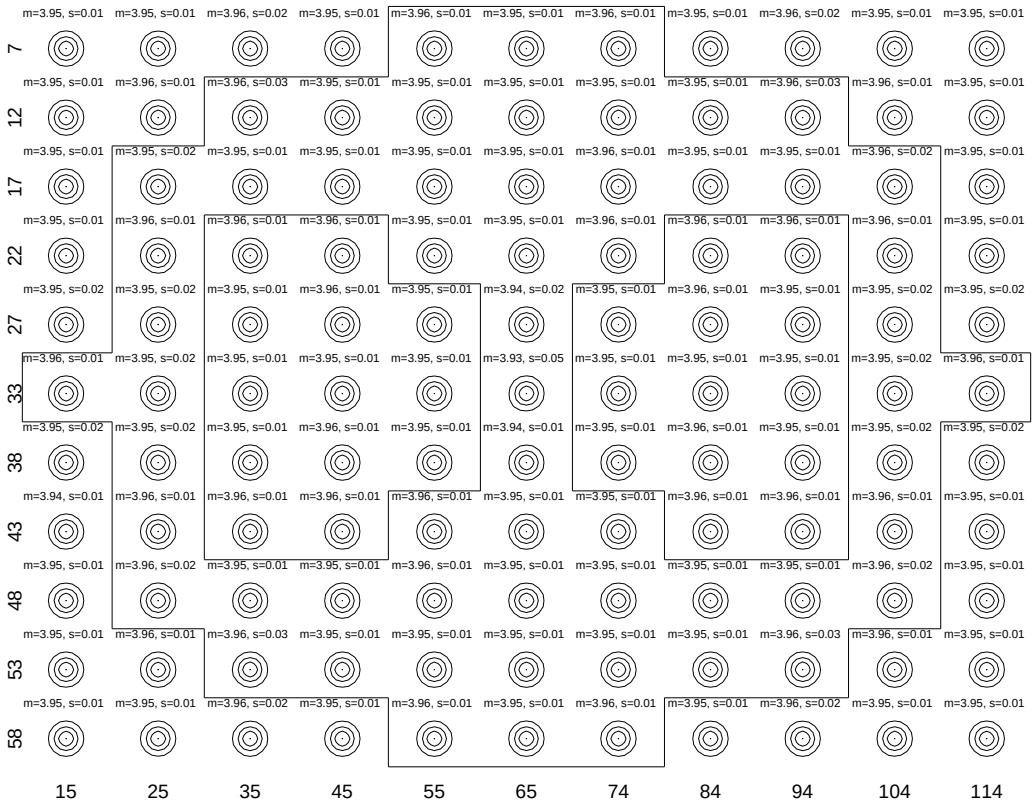


Figure 6.9: Local impulse response map for FBP.

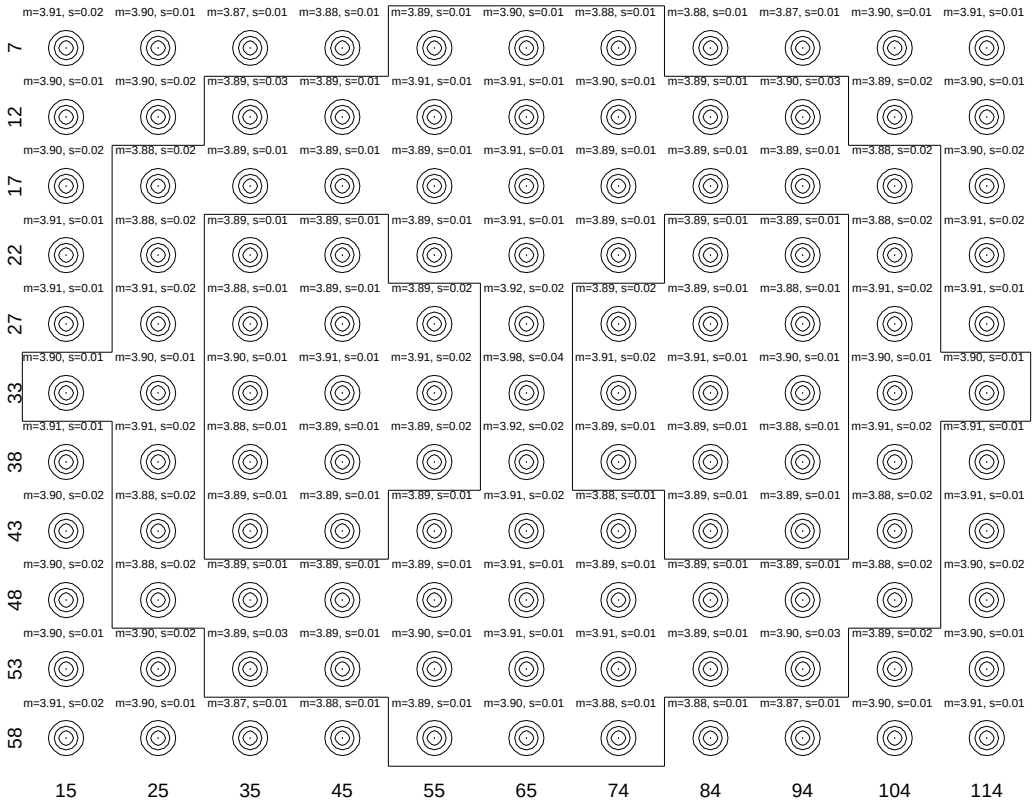
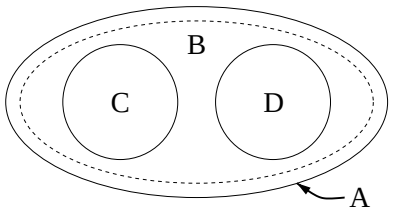


Figure 6.10: Local impulse response map for PULS with a conventional first-order penalty.

timator using the CNLLS penalty. The contours for these responses are nearly radially symmetric and close to the 4.0 pixel FWHM target resolution for pixel locations within the phantom. Outside the phantom the local impulse responses are more irregular. However, these responses are arguably less important since there are few counts outside the image and there is "nothing to smooth." Figure 6.8 presents the responses for the proposed computationally efficient method. The local impulse response contours are also quite symmetric and the average FWHM resolution is very close to the target resolution of 4.0 pixels. Again, the responses outside the object are more irregular. Additionally, note the close agreement between the CNLLS and proposed penalty's responses. The responses are quite similar (particularly inside the phantom). Both these techniques provide much improved spatial uniformity inside the phantom over the conventional shift-invariant penalty shown in Figure 3.6 and the certainty-based penalty in Figure 4.1.

The local impulse responses for FBP are shown in Figure 6.9 and for the precorrected PULS in Figure 6.10. These responses are nearly perfectly symmetric. Recall, our proposed penalty is designed with a target response given by the PULS response (5.70). We see that this response is indeed nearly shift-invariant. Additionally, the responses for FBP and PULS are nearly identical, which is expected since we have designed the FBP window to match the FBP response to the PULS response. Inside the phantom the proposed responses nearly match the PULS responses.

As a quantitative assessment of the resolution uniformity, we calculated the mean absolute radial deviation of the 50% contour from the 2.0 pixel half-maximum target radius. Then we calculated the average value of this deviation over a set of sample locations within the phantom. We performed these calculations over four pixel sets: Set A consists of all pixels within the digital phantom object; Set B contains roughly



Estimator	Pixel Set			
	A	B	C	D
PLE: Conventional	0.29	0.26	0.27	0.23
PLE: Certainty-Based	0.33	0.26	0.31	0.20
PLE: CNLLS Penalty	0.19	0.10	0.06	0.08
PLE: Proposed Penalty	0.24	0.12	0.07	0.09
Filtered Backprojection	0.02	0.02	0.02	0.02
PULS	0.05	0.05	0.05	0.05

Figure 6.11: Summary of resolution uniformity in shift-invariant PET for different estimators. The mean absolute radial deviations in the FWHM contours of the local impulses responses for various methods applied to the PET reconstruction of the object in Figure 3.5. Uniformity over the following regions are presented: (A) All Phantom Pixels, (B) Interior Pixels, (C) Cold Disc Pixels, and (D) Hot Disc Pixels.

80% of the interior pixels of the phantom excluding the outer edge pixels; Set C contains all pixels in the cold disc; and Set D contains all pixels in the hot disc. These results are summarized in Figure 6.11. All values are in pixels. The certainty-based penalty and the conventional penalty have the greatest deviation, while the CNLLS penalty and the proposed penalty are more uniform. The improvement in uniformity with these penalties is more dramatic for the interior pixels (sets B, C, and D), indicating that these penalties provide less uniform resolution at the edges of the phantom. FBP and PULS have the lowest deviations with no variation between sets.

Comparing the Proposed Penalties

In this local impulse response investigation we have seen that the CNLLS penalty, and the penalty calculated using linear operators with the greedy constraint approach yield very similar results. We present the actual penalty coefficients for these two penalty designs in Figure 6.12. These two designs produce very similar coefficients; however, there are some discrepancies between the two sets of coefficients. However, despite the differences between the two designs, we have seen that the local impulse responses for the two methods are very similar. This indicates that the fast penalty

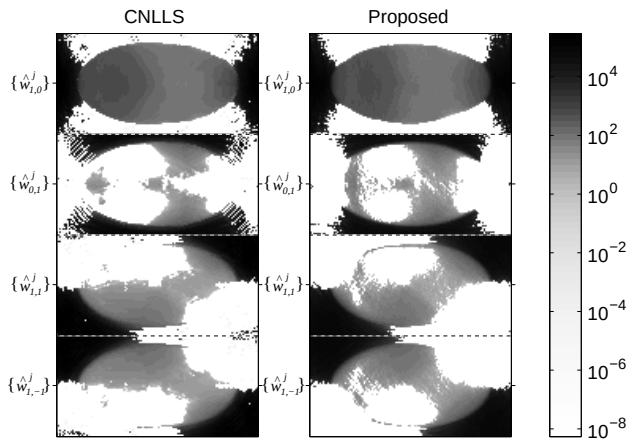


Figure 6.12: Comparison of calculated penalty weights for the CNLLS penalty and the penalty computed using the fast linear operator approach.
The four interpixel weightings associated with a second-order penalty are shown for both methods. Note the logarithmic color scale. White regions indicate a value of zero.

design using linear operators is a practical method. In fact, while the computation of the CNLLS design for this problem took about two hours, the application of the design using precomputed linear operators took less than two seconds (on a 266 MHz Pentium II processor using an efficient C program). For comparison, a single reconstruction on the same machine using 30 iterations of the SAGE algorithm[38] took about 20 seconds. Because the precomputation step requires only a single linear operator for this intrinsically shift-invariant system, the precomputations took only 23 seconds.

While the precomputed operator technique represents a method that is fast enough to use in practical situations, both this method and the CNLLS penalty yield local impulse responses that still have residual anisotropy. Returning to the penalty coefficients shown in Figure 6.12, we see that the nonnegativity constraints are quite active. It appears that the nonnegativity constraints may be preventing full resolution control. Thus, we have attempted to use the relaxed constraints of Section 4.4 to provide increased resolution uniformity.

Relaxed Design Constraints

To use the relaxed design constraints from Section 4.4, we have applied the heuristics following (4.41) to the same shift-invariant PET system and test phantom in Figure 3.5 and the same PULS target response discussed previously. Figure 6.13 shows the resulting "constraint map" over a large portion of the image area for this particular study. The figure shows all the loop-type constraints of (4.40) using triangles to connect the constrained penalty coefficients. The remaining coefficients which are not shown are constrained using the simple nonnegativity constraint. We have used a larger penalty neighborhood than the previous studies. Specifically, we have used a pixel neighborhoods including the 20 nearest neighbors. (Recall that the Fourier constrained "toy" problem in Figure 4.3 required a fairly large neighborhood to show significant improvement.)

Once the constraint map has been calculated, we perform the penalty design represented by (4.41) using a sequential quadratic programming algorithm[111, 110]. Figure 6.14 shows the penalty coefficients[3] for one particular direction (*i.e.*, the horizontal penalty) for both the traditional nonnegativity constraints and the proposed relaxed constraints. (Both methods used a 20 pixel neighborhood.)

The upper half of each image shows the positive weight values and the lower half shows the negative values. For the individual nonnegativity method, there are no negative values and the lower half is blank. All those positions colored white indicate a zero weight. In comparison, the design with the proposed relaxed constraints does include negative weights. Additionally, if one visually combines the top and bottom halves of the right image, there are relatively few positions that are zero (*i.e.*neither

[3]We note that these penalty coefficients are slightly different than those shown in Figure 6.12. This is due not only to the larger penalty neighborhood, but also because the system model is slightly different. Specifically, the model used here does not include random detector efficiencies. However, other than that, the models are identical.

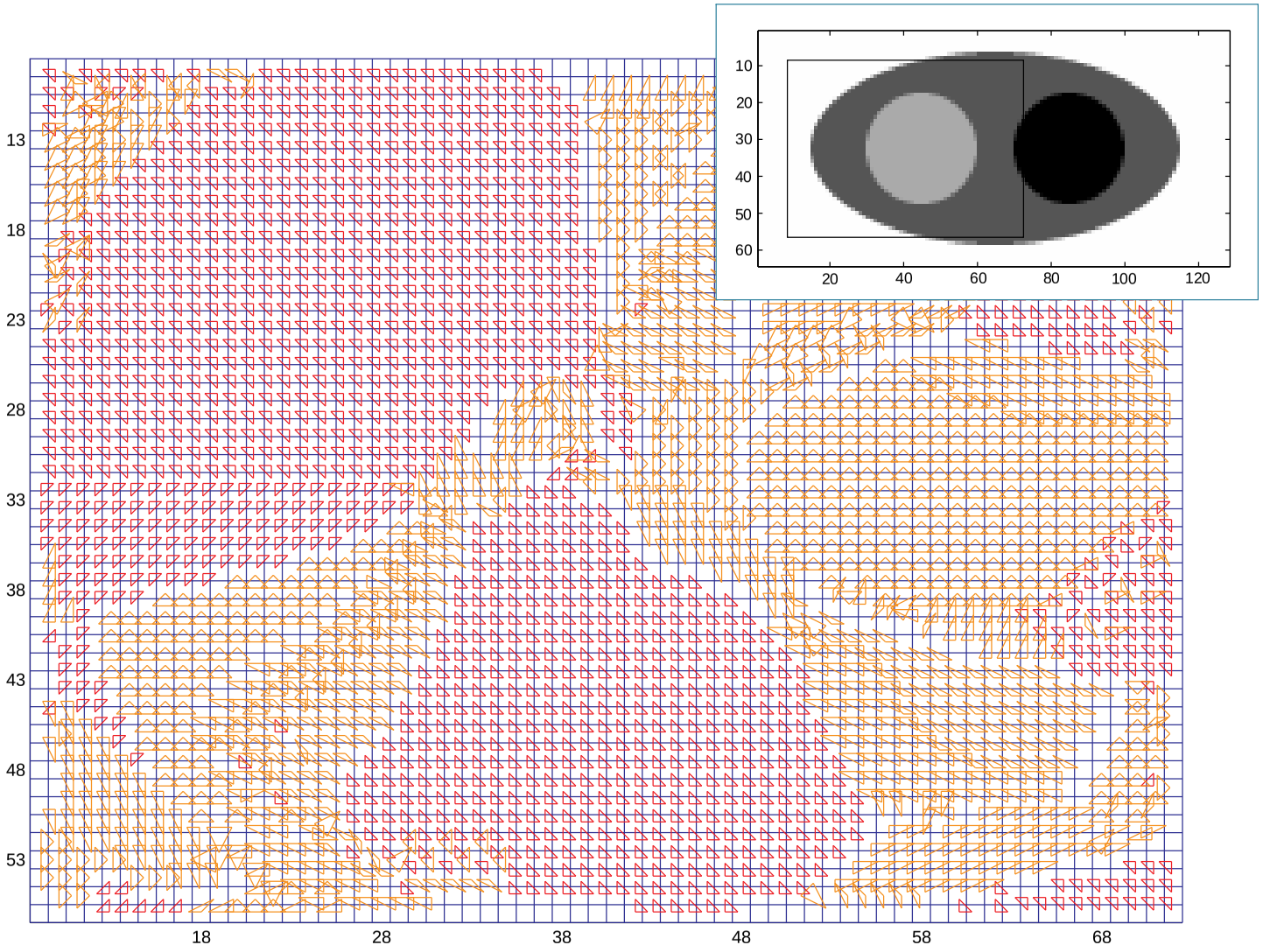


Figure 6.13: An illustration of the relaxed constraints used in PET penalty design.
The above image shows the loop-type constraints of (4.40) for a large portion of the PET test phantom (inset above right). These constraints are applied for a penalty with a neighborhood of 20 pixels. Loops that use only adjacent pixels are shown in red and loops that include pixels up to two pixels away are shown in orange. The remaining weights (not shown) are nonnegatively constrained.

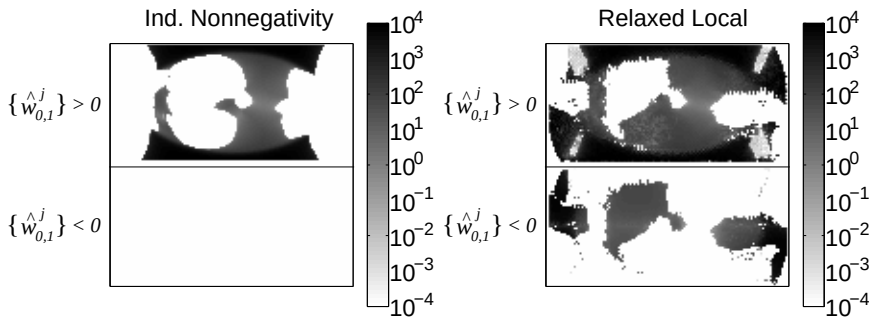


Figure 6.14: Comparison of calculated penalty weights for the nonnegatively constrained penalty and the penalty using relaxed constraints.
The vertical interpixel weightings for each method are shown. Positive values are indicated in the upper half of each image, and negative values in the lower half. The color scale is logarithmic and values that are zero are shown as white in both halves.

positive nor negative). Other weighting "directions" show similar results.

After designing quadratic penalties using the two different constraint choices, we found the resulting local impulse responses. Local impulse responses for one particular location are shown in Figure 6.15. This position, is the same position shown in Figure 4.3 that was used for the "toy" design problem, where resolution anisotropy persisted even after the application of a penalty designed using the individual nonnegativity constraints. Local impulse responses for both the old and new constraint choices are shown in Figure 6.15. Specifically, we present the 20%, 40%, 60%, 80%, and 99% contours of the local impulse response for not only the nonnegatively constrained design and design with relaxed constraints, but also the responses for the conventional shift-invariant penalty and the certainty-based penalty of Section 4.1.1.

The local impulse response using the proposed relaxed constraints shows contours (particularly the innermost contour) that are closer to the desired response for which the penalty was designed (indicated by the dashed contours). While the nonnegatively constrained design shows increased blur in a slightly off-vertical direction, the relaxed design shows improved isotropy of the response. Thus, the increased design flexibility of the proposed constraints yields improved resolution uniformity.

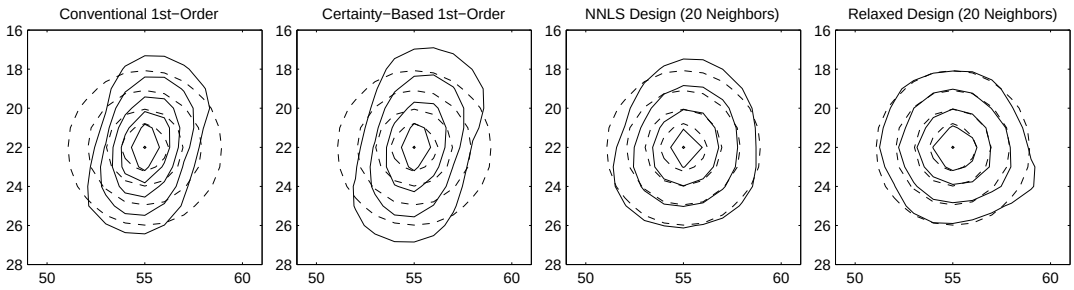


Figure 6.15: Illustration of the relative resolution uniformity using the relaxed design constraints.

The above images show contours of the desired local impulse response (dashed) and actual local impulse response (solid) for various penalties including the proposed design using the individual nonnegativity constraints and the relaxed design constraints.

Most local impulses throughout the image show either similar improvements or performance as good as the individual nonnegativity constraints. Indeed, we expect that the proposed constraints should yield a design no worse than the nonnegatively constrained design, since the nonnegative solution is in the feasible region of the proposed constraints. However, there are a few locations where there are slight degradations in the uniformity. We suspect that such degradations are the result of the suboptimal greedy iterative optimization approach used to calculate the weights. Such results may be due to incomplete convergence of the design optimization, or limit cycles. It is possible that some kind of regularization (*i.e.*: smoothing of the weight values between iterations) might decrease these effects. It may also be possible to develop an alternative optimization approach that is less susceptible to such problems. However, it should also be noted that these effects are generally relatively small, and that the uniformity improvements outweigh the degradations.

While the relaxed constraints lead to a penalty design with increased design flexibility, the uniformity improvements are somewhat marginal, especially in light of the increased complication of the penalty design. Because the relaxed design requires iterative solution, it takes a few hours to find a constraint map and to perform the

subsequent penalty design. Moreover, while the relaxed design yields improvement in some regions, the relatively large anisotropy found at the edges of the phantom do not change appreciably from the nonnegatively constrained design. Thus, we suspect the relaxed design method is inappropriate for many practical applications.

The relatively large anisotropy at the edges of the phantom appears to be a very difficult nonuniformity to correct. We suspect that designing for uniform local impulse responses in these regions requires either a combination of very large penalty neighborhoods and relaxed constraints, or such responses are fundamentally unattainable with this kind of penalized-likelihood reconstruction (recall that from the discussion in Section 4.3.1, some responses may not be achievable). While reconstructions can be visually improved by truncating penalty coefficients at the edges, or by designing for a finer resolution at the edges, the uniform resolution goal appears difficult or impossible to attain using this kind of penalized-likelihood estimation.

Sample reconstructions of an Anthropomorphic Phantom

While the local impulse responses shown previously may be used to quantitatively identify the degree of resolution uniformity, it is often the images themselves that are the point of interest. Thus, we return to the anthropomorphic torso phantom that was shown in Figure 3.2 and discussed in Section 3.2.2

We reconstructed the noiseless emission measurements using FBP, penalized unweighted least-squares, and penalized-likelihood estimators with the conventional, certainty-based, and (linearized) proposed penalties. All statistical methods enforced nonnegativity of the image and negatives in the image reconstructed via FBP were set to zero. All methods used a target FWHM resolution of 3.0 pixels (1.25 cm). For PULS and PL with conventional regularization, the penalties were chosen so that $\boldsymbol{R}_0$ corresponds to the shift-invariant first-order penalty represented by the filter

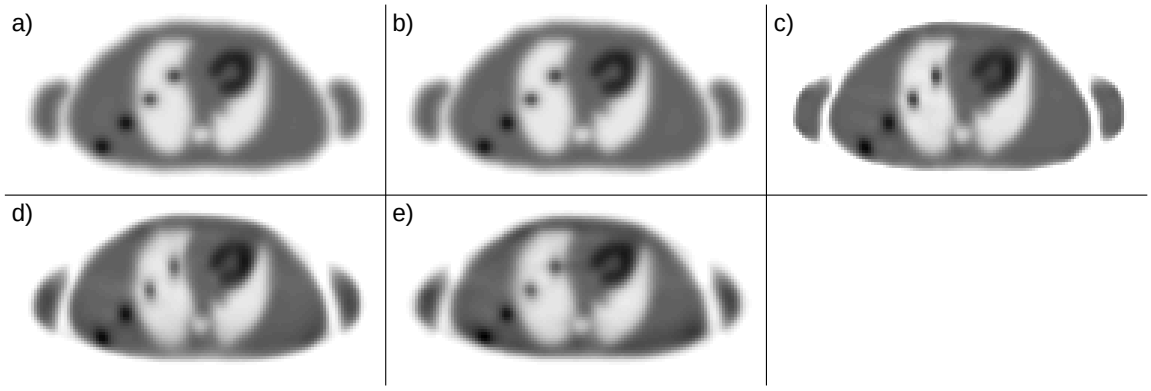


Figure 6.16: Reconstruction of a 2D PET thorax phantom using various reconstruction methods. We reconstruct simulated data from the phantom in Figure 3.2 using a) filtered backprojection, b) Penalized unweighted least-squares, c) Penalized-likelihood with conventional regularization, d) Penalized-likelihood with certainty-based penalty, and e) Penalized-likelihood with the proposed penalty.

in (4.21). The proposed penalty uses the β-independent design (5.73) with second-order bases, and the same target $\boldsymbol{R}_0$ as PULS, and is applied using the (single) linear operator approach.

The reconstructions using these methods are presented in Figure 6.16. The FBP reconstruction in Figure 6.16a has uniform resolution properties. This is evident from the uniformly smooth edges and radially symmetric tumors. Similarly, the PULS reconstruction in Figure 6.16b shows the expected nearly identical results. (Recall the nearly identical local impulse responses of FBP and PULS in Figures 6.9 and 6.10.) The PL reconstruction using conventional regularization is shown in Figure 6.16c. There is significant distortion of the four round "tumors" (particularly in the lungs) in this reconstruction. These hot spots are preferentially blurred vertically and appear elliptical. Another indication of resolution nonuniformity is evident at the outer boundaries of the arms. These boundaries are sharper than those in FBP and PULS. The reconstruction with certainty-based penalty in Figure 6.16d shows some improvement. Most notably, the outer edges of the arms are smoothed

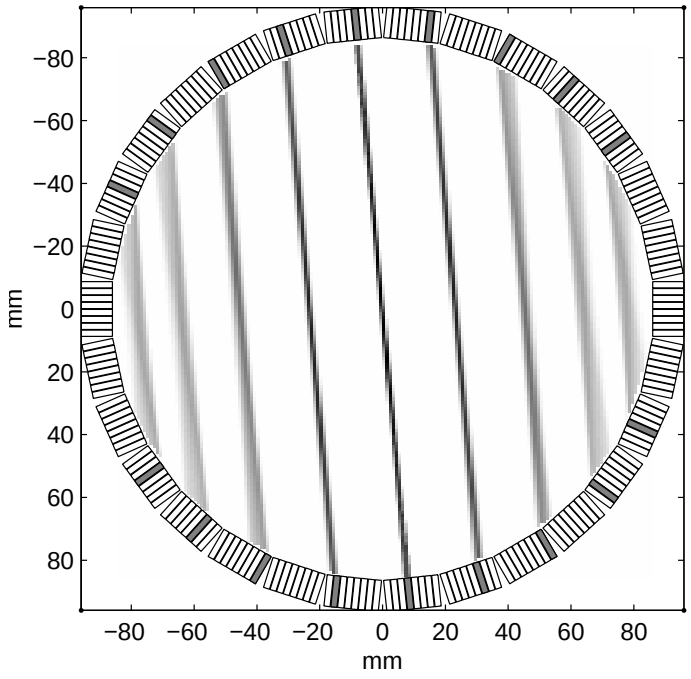


Figure 6.17: A space-variant small animal PET system.
The above PET system model has physical characteristics chosen to simulate the MicroPET rodent scanner. In the above image, rays connecting every 10th detector are shown. Both nonuniform sampling and varying detector response due to crystal penetration are incorporated in this model.

in a more uniform fashion. However, the tumors are still smoothed preferentially in the vertical direction. Figure 6.16d shows the PL reconstruction with our proposed penalty. The resolution uniformity appears much improved over the other PL methods. The tumors appear nearly radially symmetric and the edges appear more uniformly smoothed, although some anisotropy at the edges remains.

6.2.2 2D PET with Shift-Variant Geometric Response

We have also applied our penalty design technique to a space-variant small animal PET system. Specifically, we have modeled a MicroPET rodent scanner. This system has 2 mm (square) by 10 mm crystals in 30 8×8 blocks. The full field of view of 170×170, 1 mm pixels, is modeled using finite integration of over all angles and pixels, and includes crystal penetration effects. Figure 6.17 shows responses for detectors

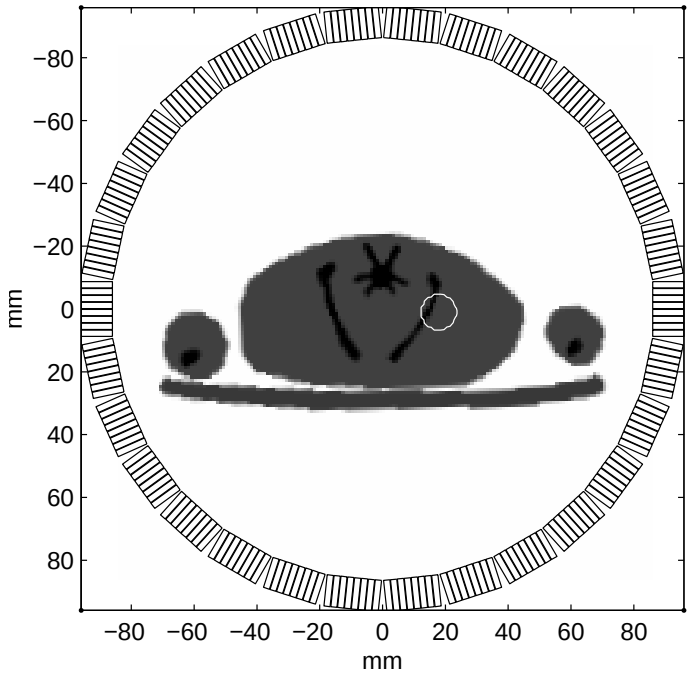


Figure 6.18: The small animal PET system with a simulated rat phantom.
This figure shows the attenuation map for a transverse slice of a rat's lower pelvis and thighs, where the animal takes up a very wide field-of-view. The emission image is uniform, except for a hot lesion in the location indicated by the white circle.

pairs over regular intervals. Both the nonuniform sampling and the space-variant detector responses are evident in this figure.

Figure 6.18 shows a sample digital phantom placed in the scanner. This image shows the attenuation map for a digital rat phantom in a slice at the bottom of the pelvis, where the rat takes up a very large portion of the field of view. This data was obtained by manually segmenting MRI data obtained from [1]. The attenuation values are 0.0096, 0.013, and 0.010 mm^{-1} for the soft tissue, bone, and the table, respectively, and correspond to appropriate values for 512 keV photons. The emission image has a uniform background with emission rate of 1.0, and a single circular lesion in the right half of the phantom with an emission rate of 2.0 (indicated in Figure 6.18 by the white circle). Projections contain 10 million counts with 5% percent random coincidences.

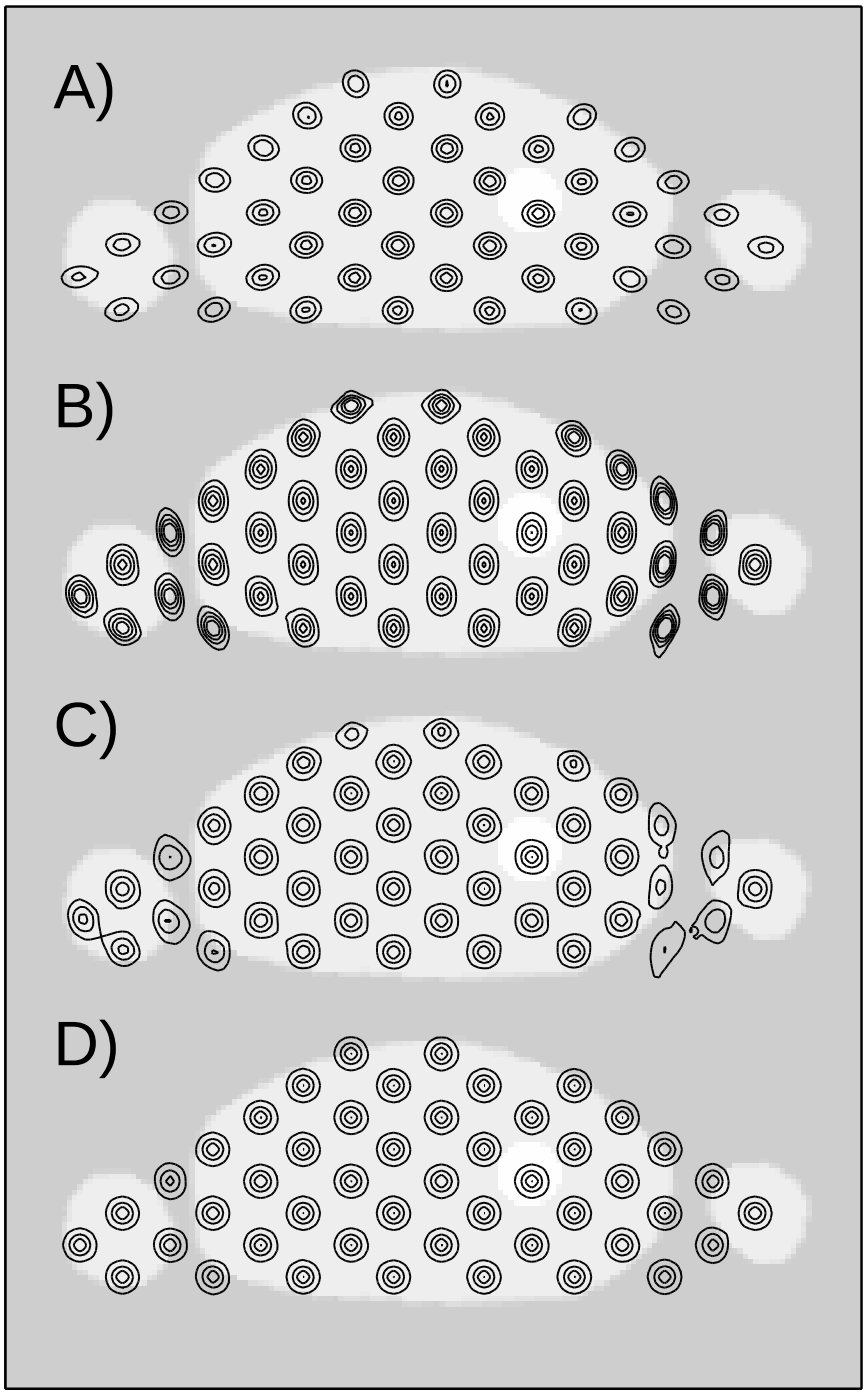


Figure 6.19: Local impulse responses in shift-variant PET reconstructions.
This figure shows contours of the local impulse responses for A) filtered-backprojection, B) PL with space-invariant penalty, C) PL with proposed penalty, and D) post-smoothed ML. Contours are superimposed on the emission image to show position.

Figure 6.19 presents local impulse response contours for four different reconstruction methods. These contours are made at the 25%, 50%, 75%, and 99% levels of the target response. The target response was chosen to be the PULS response at the center of the field of view (recall PULS will not yield shift-invariant responses for a shift-variant system; however, at the center of the field of view, the response is highly symmetric) with β such that the response has a FWHM resolution of 4.0 mm. The local impulse response contours are superimposed on the rat slice emission image so that the position of the responses are apparent. We performed FBP reconstruction by radially resampling the cylindrical projections (arc correction). And although the system is shift-variant, we again used the least-squares filter of [32] in an attempt to match resolution properties. Figure 6.19A shows responses for FBP, which, while relatively well-matched at the center, have reduced peaks toward the edges, indicating coarser resolution properties as expected.

In contrast, the responses arising from conventional PL (shift-invariant first-order penalty) shown in Figure 6.19B are narrower at the edges. There are competing effects in PL reconstruction. While the system model suggests decreased resolution at the edges due to the detector responses, there is actually finer sampling at the edges (in effect better conditioning the reconstruction than if uniformly sampled data were acquired). However, for emission tomography, the FWHM resolution of conventional PL varies inversely with ray certainty. Thus, at the edges, where ones obtains lower count measurements and thus increased certainty (under the Poisson model), one expects decreased (finer) resolution. While these competing effects actually appear to yield more uniform resolution than if the system model were idealized to have uniformly sampled projections, the effects of attenuation are clear in the responses, resulting in greater vertical smoothing.

Figure 6.19C shows contours for PL with the proposed space-variant (second-order) penalty. The responses are very uniform in the interior of the object, but degrade near the edges and outside the object. In general, the proposed technique yields more uniform results than conventional PL. If more uniform results are desired, a larger order penalty neighborhood may be required, or relaxed design constraints of Section 4.4 may need to be applied.

Lastly, we present the contours for the case of post-smoothed ML in Figure 6.19D. These responses are very uniform throughout the image and are very well matched to the target response. We find the greatest uniformity and the ability to match a target response with the post-smoothed ML and proposed PL techniques.

6.2.3 3D PET with Shift-Invariant Geometric Response

We have also used our penalty design methods on a 3D PET system with a shift-invariant geometric response (*i.e.*, no truncation of projections). In order to evaluate the 3D design, we used the anthropomorphic phantom presented in Figure 6.20. This phantom is $86 \times 86 \times 32$ with 5mm pixels. The PET system model used 32 transaxial rotation angles covering 360° and 9 axial "tilt" angles covering -15 to 15 degrees. Attenuation effects are also modeled using the previously mentioned linear attenuation coefficients appropriate for 512 keV photons in the lungs, soft tissue, and bone.

For the desired response, we chose a PULS response. If one chooses a conventional uniform first-order 3D penalty represented by the filter in (4.23), the response generally will not be radially symmetric due to incomplete sampling of the spherical data. Usually not all axial "tilts" from $-90°$ to $90°$ are included, leading to less intrinsic smoothing in the axial direction. The geometric response tends to be isotropic in the transaxial planes, since the image space is uniformly sampled in transaxial angles.

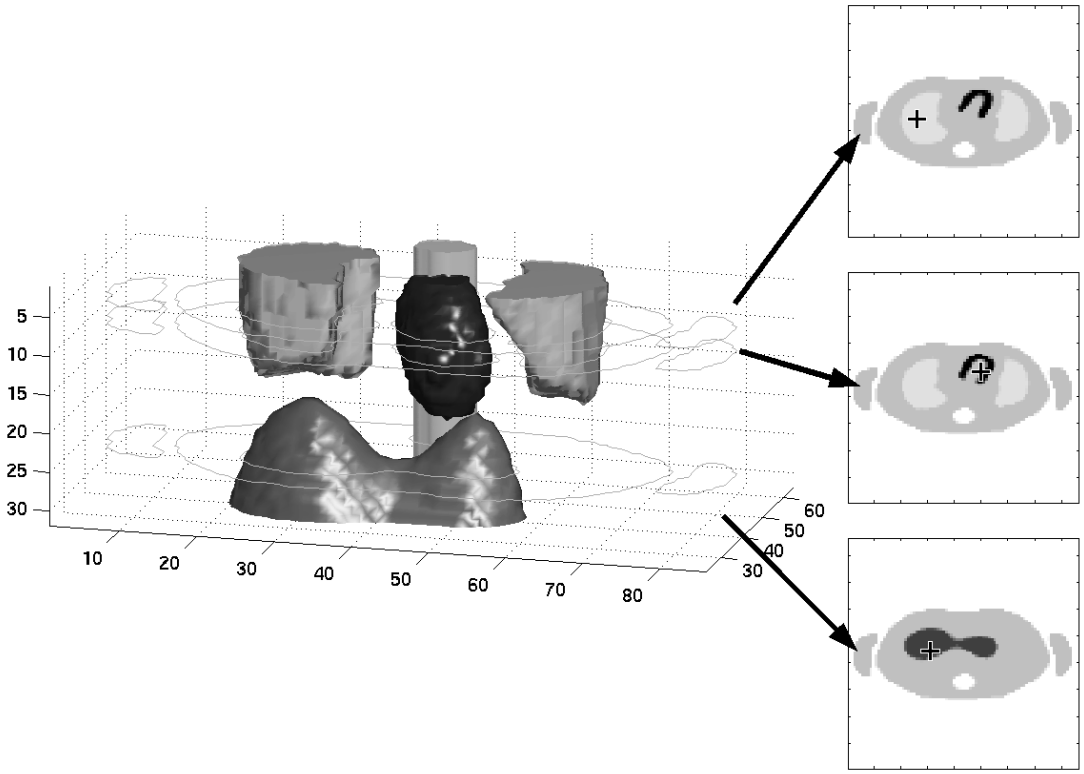


Figure 6.20: 3D PET simulated thorax phantom with sample locations for a local impulse response investigation.

In this case one often splits the penalty into in-plane and cross-plane portions. For example, using the filter representation, the penalty is

$$\beta_{xy} \quad + \beta_z \quad . \tag{6.3}$$

The β_{xy} term controls the in-plane resolution and β_z controls the smoothing between planes. Therefore, we choose β_{xy} and β_z to make the desired PULS response have isotropic smoothing with 3.3 pixel FWHM resolution.

Sample Reconstructions

To compare the conventional penalty versus our proposed penalty we use a modified version of the phantom presented in Figure 6.20, that has a spherical hot spot

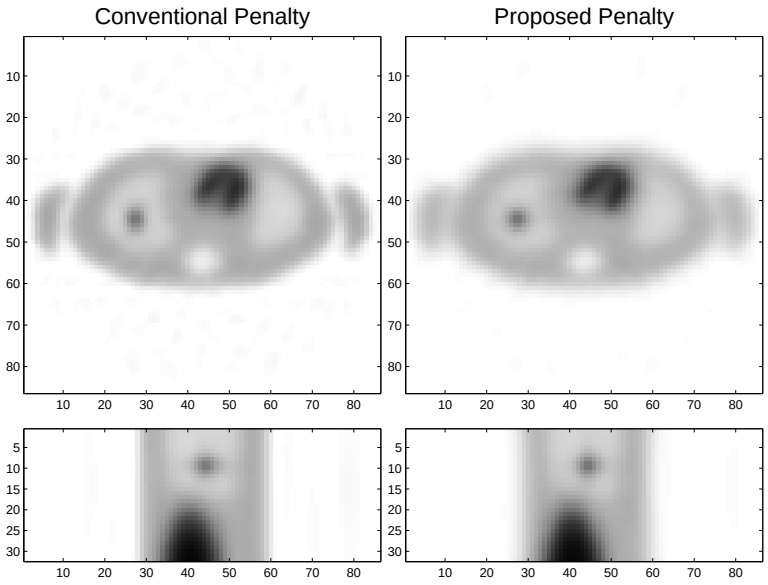
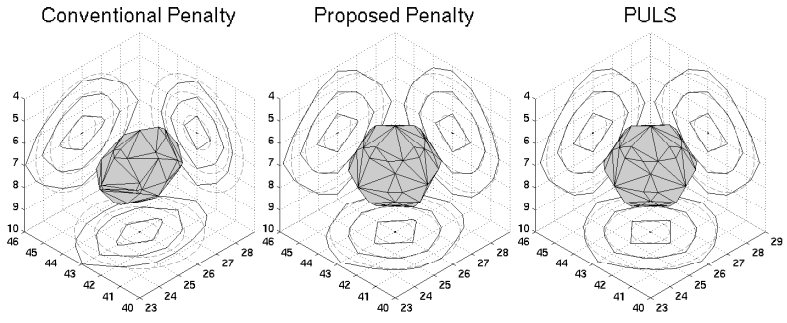


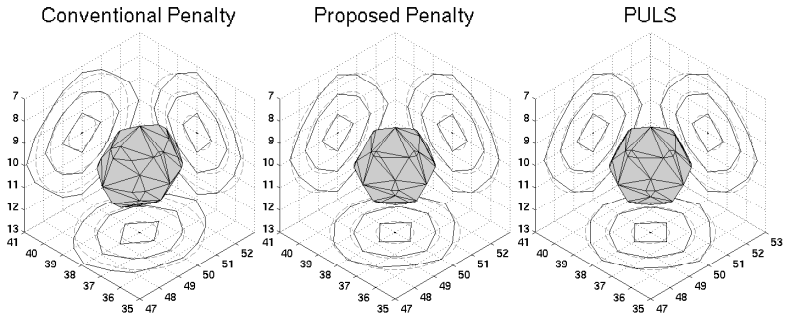
Figure 6.21: Sample reconstruction of the 3D PET thorax phantom using a conventional penalty and the proposed penalty.
These reconstructions use the 3D thorax phantom of Figure 6.20 with the addition of a single hot spot in the lung.

added to the left lung region. From this phantom we obtained noiseless projections and reconstructed the image volume using the following two methods: (1) a conventional penalty with a kernel of the form given in (6.3), where β_{xy} and β_z are chosen to be the same as the PULS values, and (2) the proposed 3D penalty with 26 basis functions filling the 3x3x3 cube and a desired response equal to the PULS response. Slices of this reconstructed volume using the two penalties are presented in Figure 6.21.

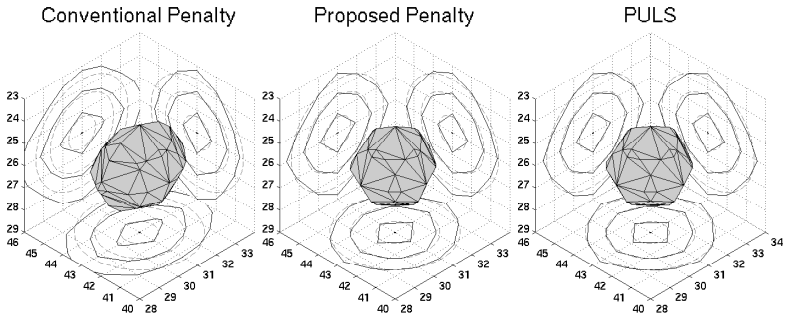
In the reconstruction that uses the conventional penalty, the nonuniform smoothing properties of the estimator are evident. The spherical hot spot has been stretched vertically in the transaxial slice and horizontally in the sagittal slice. Compare this to the reconstruction using the proposed penalty. With the proposed penalty reconstruction, the hot spot is nearly radially symmetric in both the transaxial and sagittal slices.



(a) Local impulse responses in the lung region



(b) Local impulse responses in the heart region



(c) Local impulse responses in the liver region

Figure 6.22: Local impulse response investigation for three points in the 3D PET phantom. These points lie in the (a) lung, (b) heart, and (c) liver. For each location, the half-maximum surface of the 3D local impulse response is plotted for reconstruction using a PWLS with conventional penalty, a PWLS with proposed penalty, and PULS with conventional penalty. Additionally, slice contours are presented on the planes passing through the coordinate axes.

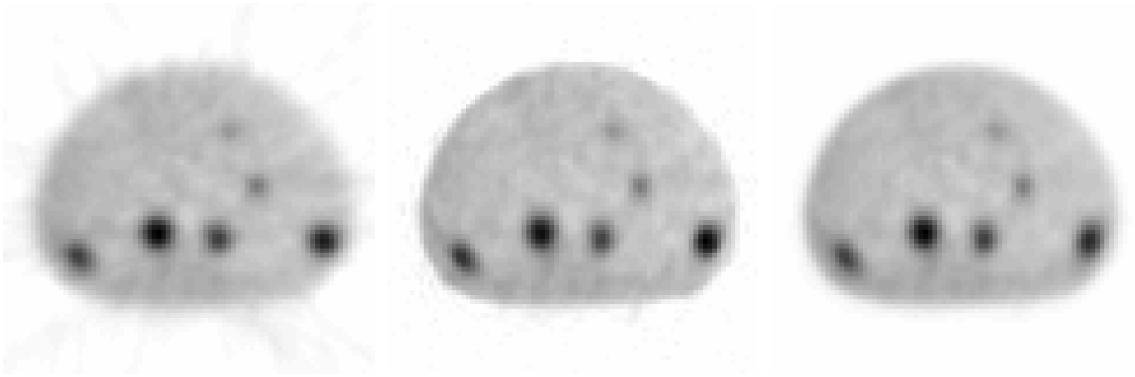
3D Local Impulse Responses

In addition to the noiseless reconstructions, we performed a local impulse response investigation. In Figure 6.20, we have identified three locations for this investigation; (1) a voxel in the right lung, (2) a voxel in the heart, and (3) a voxel in the liver. This investigation compares three different methods: the conventional penalty and the proposed penalty used in 3D sample reconstructions earlier in this section, and the PULS response that was used as a desired response. The results of this investigation are presented in Figure 6.22. As in the 2D local impulse response investigation discussed earlier in this section, for locations in the interior of the phantom, the proposed penalty yields increased resolution uniformity with responses closely matching the PULS objective.

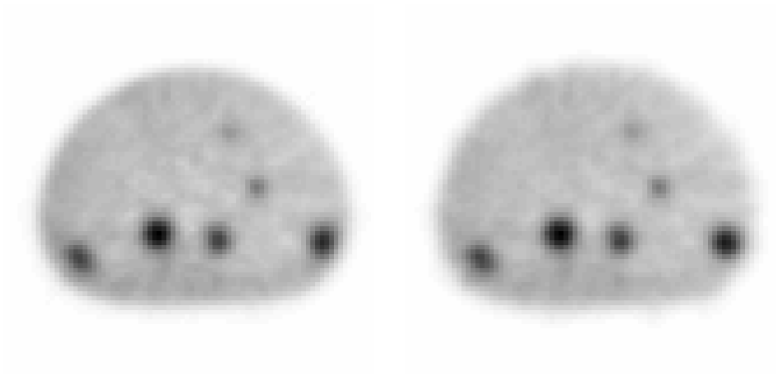
6.2.4 Reconstruction of Real PET Data

All of the reconstructions shown up to this point have been performed on simulated data. In this section we show the reconstruction of PET data acquired from a CTI 921 ECAT EXACT scanner. The CTI PET system was modeled using equally spaced strip integrals. Data were acquired with 1.5 million counts per slice from a phantom prepared with several hot spheres and a warm background.

We performed 2D reconstruction of this data using various methods and a target response equal to the "natural" response of (5.70) with 1.2 cm FWHM resolution. We show these reconstructions in Figure 6.23. Figure 6.23a shows the filtered backprojection reconstruction. While the spheres are generally fairly symmetric, the image appears significantly noisier than the other reconstructions. Figures 6.23b and c show penalized-likelihood estimates using a conventional shift-invariant penalty and the certainty-based penalty of Section 4.1.1. Both of these reconstructions exhibit



(a) Filtered Backprojection (b) PL - Shift-Invariant Penalty (c) PL - Certainty-Based Penalty



(d) PL - Penalty Design (e) Post-Filtered ML

Figure 6.23: Rconstructions of PET data from a CTI 921 ECAT EXACT scanner using various reconstruction methods.

nonuniform resolution properties with the cross-sections of the spheres appearing elliptical. In comparison, consider the penalized-likelihood reconstruction with our penalty design[4] for uniform resolution shown in Figures 6.23d. The resolution in this image is highly uniform and the spheres are very symmetric except for the leftmost sphere, which is not symmetric in any of the reconstructions. (This could possibly be due to noise or model mismatches.) Lastly, a post-filtered maximum-likelihood estimate is shown in Figure 6.23e. This estimate also has very uniform resolution with slightly better uniformity at the edges (see rightmost sphere). Thus, we see that our penalty design can be successfully applied to real data to provide nearly

[4]Specifically, we used the efficient scalable design procedure of Section 5.5.4.

uniform resolution properties.

6.3 Resolution Control for SPECT Systems

In this section we concentrate on applying our penalty design techniques to SPECT systems, which are inherently shift-variant. We pay particular attention to what methods can provide uniform resolution and how to exactly match the resolution properties of two different methods.

6.3.1 2D SPECT

For our SPECT investigation we return to the 2D SPECT model of Section 3.2.3 with a depth-dependent Gaussian response, and the "cold rod" phantom in Figure 3.3. Recall from Section 5.5.1, when the system matrix is precomputed, we have direct access to the columns of $\boldsymbol{H}$. Since this is the case for our simulated 2D SPECT system, we may apply the "truncated" design[5] represented by (5.52) and (5.53) without using linear operators or the associated precomputations.

A Brief Discussion of Penalty Computation Times

Before discussing the resolution properties of various reconstruction techniques, we first demonstrate the feasibility of the proposed "truncated" design in terms of computation time. Table 6.1 lists computation times for the space-variant penalty for this SPECT system using a `gcc`-compiled ANSI C implementation of the design discussed in Section 5.5.1. For comparison, the time to complete a single projection-backprojection, (*i.e.*, $\boldsymbol{H}'\boldsymbol{H}\underline{\theta}$), is approximately 1.5 seconds. We present results for two different support sizes and four different spatial subsamplings (*i.e.*, evaluating at every n_dth pixel and filling in the penalty coefficient gaps by interpolation). Due to

[5] Generally the linear operator approach will yield penalties faster than the "truncated" design approach. However, if one does not wish to perform such precomputations and one has direct access to the columns of $\boldsymbol{H}$, then this method is often still quite practical (as we demonstrate here).

Table 6.1: Calculation times for the proposed penalty on an 800 MHz Pentium-III processor.

Spatial Subsampling	20×20 Support	12×12 Support
1	128 s	60 s
2	33 s	16 s
3	15 s	8 s
5	6 s	4 s

zero padding[6] and the use of radix-2 FFTs, the 20×20 support size uses 32×32 FFTs and the 12×12 support uses 16×16 FFTs. All methods used a second-order penalty, incorporating the eight nearest pixels. The computation times are very reasonable, particularly for the larger subsampling values.

SPECT Reconstructions

For our resolution investigation we would like to compare a number of different methods and to match the resolution properties of those methods as closely as possible. Since different methods have different resolution properties and are generally at least slightly shift-variant, we have attempted to match resolution as closely as possible for the *center pixel* in the image.

Furthermore, we have chosen the following target impulse response,

$$\underline{l}_0 = [\boldsymbol{H}'\boldsymbol{H} + \boldsymbol{R}_0]^{-1}\boldsymbol{H}'\boldsymbol{H}\underline{e}^{j_0}, \tag{6.4}$$

where we have selected a conventional space-invariant penalty and j_0 denotes the center pixel in the image. Equation (6.4) represents the local impulse response for a conventional penalized unweighted least-squares reconstruction. We evaluate this target response (6.4) using iterative techniques. This response is also essentially radially symmetric since the response lies at the center pixel for a SPECT model

[6]Technically the zero padding applied these cases is insufficient to completely eliminate wrap-around effects from periodic convolution. However, because the $\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^j$ responses are fairly smooth (especially at the edges) and the blur operation uses a high pass filter, we accept small amount of wrap-around in the penalty design to reduce computation.

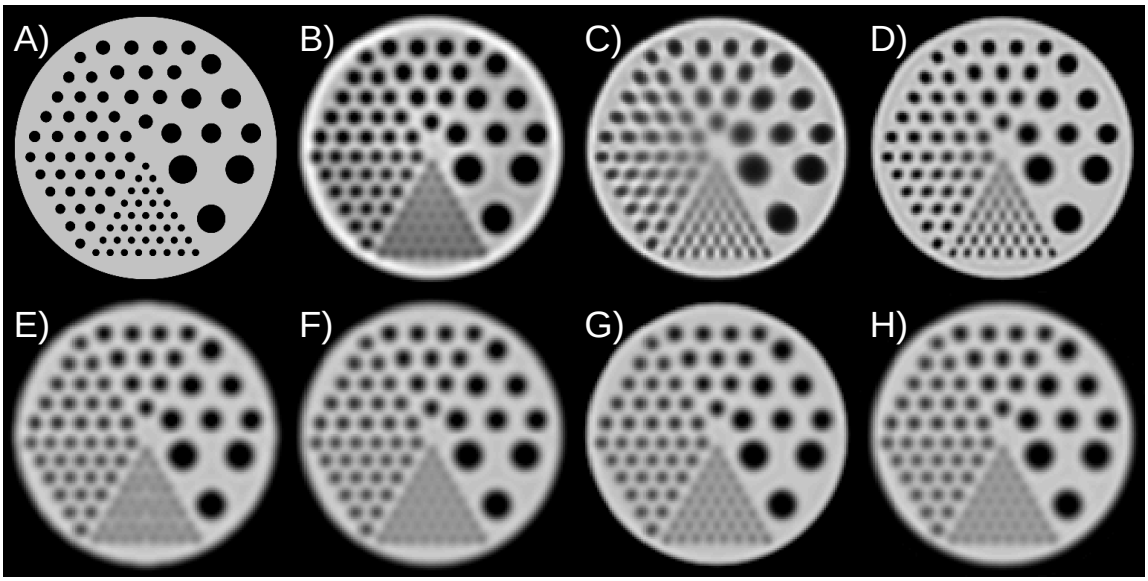


Figure 6.24: Noiseless 2D SPECT reconstructions using various estimators.
A) True emission image, B) FBP with uniformity correction using the frequency-distance principle and attenuation correction, C) Truncated OSEM, D) PL with standard space-invariant penalty, E) (Subsampled) true image smoothed with desired blur, F) Post-smoothed ML, G) PL with modified penalty, and H) the hybrid post-smoothed PL approach of Section 6.3.2.

that incorporates a circular orbit, and since the object is a centered, uniformly attenuating, disc-shaped object. For $\boldsymbol{R}_0$ in (6.4), we chose a standard penalty matrix that uses a first-order neighborhood with four equal penalty weights and a weighting chosen to yield a FWHM resolution of 10 mm.

Choosing the target response (6.4) allows one to match exactly the reconstruction resolutions for many methods since it represents a form achievable by many penalized-likelihood and filtering methods. Figure 6.24E shows the true image downsampled to 128×128 and blurred with the target response (6.4).

Figure 6.24 shows images reconstructed from noiseless projections of the cold rod phantom using a variety of techniques. Figure 6.24B shows a filtered-backprojection (FBP) reconstruction using the frequency-distance principle to correct for nonuniform resolution[135] and Chang-type attenuation correction[19]. Because the nonuniform detector response cannot be completely eliminated by frequency-distance prin-

ciple filtering, we use the following approach to match the resolution properties with the target response in (6.4).

When the response of an estimator, such as FBP, is known, and does not match (6.4) perfectly, one can force a match by applying post-filtering. The overall response is then a combination of the estimator response and the post-filter. Specifically,

$$l_{\text{overall}}(m,n) = l_{\text{est}}(m,n) ** l_{\text{post}}(m,n), \tag{6.5}$$

where $l_{\text{overall}}(m,n)$, $l_{\text{est}}(m,n)$, and $l_{\text{post}}(m,n)$ represent the overall response, the response due to the estimator, and the post-smoothing filter, respectively. Thus, given an overall desired target response and the estimator response, one can find the appropriate post-smoothing filter by

$$l_{\text{post}}(m,n) = \mathfrak{F}^{-1}\left\{\frac{\mathfrak{F}\left\{l_{\text{overall}}(m,n)\right\}}{\mathfrak{F}\left\{l_{\text{est}}(m,n)\right\}}\right\}. \tag{6.6}$$

Depending on the form of $l_{\text{est}}(m,n)$, it may not be possible to obtain any overall desired response because of zeros in the frequency-domain. However, one can find approximate post-filters for a wide range of overall desired responses.

Therefore, even though ramp-filtered FBP with the frequency-distance-based uniformity correction yields an imperfect response, we match the overall target response, (6.4), by using a post-filter calculated from (6.6). Because the ramp-filtered FBP estimator generally yields space-variant results, we match the target response only at the center pixel. That is, we find $l_{\text{est}}(m,n)$ for the center pixel by propagating an impulse through the ramp-filtered FBP estimator, and find a single shift-invariant post-filter using (6.6) to match the target response. The resulting reconstruction, shown in Figure 6.24B, has relatively good resolution uniformity, but suffers from ringing artifacts, most noticeable at the edges of the object.

Figure 6.24C shows a reconstruction using an ordered subsets expectation maximization (OSEM) algorithm with 10 subsets. We initialize the algorithm with a uniform image and perform nine iterations. Starting with a flat image and using only a few iterations is sometimes used as a noise-control technique, since higher spatial frequencies generally take more iterations to appear in the image estimate. The resolution properties are highly nonuniform, and only roughly matched even at the center due to the poor (object-dependent) resolution control available with this method.

Figure 6.24D shows a standard penalized-likelihood reconstruction using a space-invariant penalty. We may write an approximate local impulse response for this estimator at the center pixel as

$$\underline{l}_{\mathrm{PL}}^{j_0} = [\boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} + \beta \boldsymbol{R}_0]^{-1} \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}\underline{e}^{j_0}. \tag{6.7}$$

However, for the center pixel in this particular phantom the diagonal weighting denoted by $\boldsymbol{D}$ is very uniform and the response (6.7) is indistinguishable from

$$\begin{aligned} \underline{l}_{\mathrm{PL}}^{j_0} &\approx [d\boldsymbol{H}'\boldsymbol{H} + \beta \boldsymbol{R}_0]^{-1} d\boldsymbol{H}'\boldsymbol{H}\underline{e}^{j_0} \\ &= [\boldsymbol{H}'\boldsymbol{H} + \frac{\beta}{d}\boldsymbol{R}_0]^{-1} \boldsymbol{H}'\boldsymbol{H}\underline{e}^{j_0}, \end{aligned} \tag{6.8}$$

where d denotes the uniform diagonal weighting. Thus, using the same penalty, $\boldsymbol{R}_0$, as in (6.4) with an appropriate scaling β, we have matched the center pixel's response nearly exactly. We estimate the solution with 200 iterations of an ordered subsets version of De Pierro's algorithm[22] with 10 subsets, initialized with an FBP reconstruction, followed by 20 iterations with one subset. For typical image reconstruction problems, this represents many more iterations than are generally necessary to form a good image estimate. However, we would like a solution that is well-converged so that we may guarantee that any resolution mismatches (or, noise mismatches later in

Section 7.2) are due entirely to the objective function, not to insufficient convergence of the algorithm used to find the estimate. While the resolution properties for the PL estimate in Figure 6.24D are nearly exactly matched at the center, the nonuniform resolution properties away from the center are clearly evident.

Figure 6.24F is a reconstruction using a post-smoothed ML technique, using 200 OSEM iterations (10 subsets) initialized with an FBP image, followed by 20 EM iterations to ensure a nearly converged solution. Since we have post-smoothed with the desired target response in (6.4), the resolution properties are essentially exactly matched, as seen by comparing Figure 6.24E and Figure 6.24F.

Lastly, we applied our proposed space-variant penalty, using 200 iterations of the ordered-subsets De Pierro's algorithm (10 subsets), initialized with an FBP image and followed by 20 iterations using one subset. Figure 6.24G shows the reconstruction resulting from our penalty design using the 20×20 support with no spatial subsampling. The resolution properties are virtually exactly matched at the center since the target response is easily achieved using the space-variant design. That is, because a space-invariant penalty achieves this response, the space-variant design easily achieves the same response. The global resolution properties are mostly very uniform, with some mild nonuniformities at the object edges, where approximation (3.20) is less accurate.

Using the other choices of support size and spatial subsampling shown in Table 6.1 yielded nearly identical results in the interior of the object. Significant nonuniformity was noticeable only at the edges of the object when using coarser spatial subsampling. One could use a region-dependent subsampling of positions in (4.36) to sample more finely at the object edges to provide nearly the same results with fast computation.

Local Impulse Responses

We also investigate the resolution properties various techniques for this SPECT system by evaluating[7] the local impulse response at a variety of locations and compare them to the target response. Such a sampling of responses is shown in Figure 6.25. The local impulse responses are contoured at four levels indicating the 25%, 50%, 75%, and 99% levels of the target response.

The relatively narrow responses of conventional PL are evident away from the center of the object in Figure 6.25B. In contrast, the uniformity-corrected FBP, PL with the space-variant penalty, and post-smoothed ML, shown in Figures 6.25A, 6.25C, and 6.25D, respectively, yield very uniform responses. That is, the responses show a high degree of symmetry and spatial uniformity, and the response peaks and contours are closely matched to the target in (6.4). The response of the center pixel (shown in the lower right corner of each subfigure) is indistinguishable from the target response for all these methods. (We do not present local impulse responses for OSEM with truncated iterations; however, we would expect very nonuniform responses that have mismatch even at the center pixel.) Post-smoothed ML appears to have the best uniformity, whereas our proposed PL method shows very slight asymmetries at the edges of the object.

While post-smoothed ML appears to yield more uniform resolution properties than the proposed PL technique, we find that there are still resolution nonuniformities for the post-smoothed ML techniques. When we investigate the the resolution properties of conventional ML with no filtering through a systematic evaluation of local impulse responses, we find that the FWHM resolution of the responses varies from

[7]For most statistical methods we evaluate (3.15) using iterative techniques (we choose 100 iterations of a coordinate ascent algorithm initialized with the target response). For ML techniques where the invertibility conditions for (3.15) may not hold, we use the techniques described in [133], where the emission image is perturbed with an impulse, and differences in reconstructions with and without the perturbation are obtained. For linear techniques like FBP, we simply propagate an impulse response through the system to find the local response.

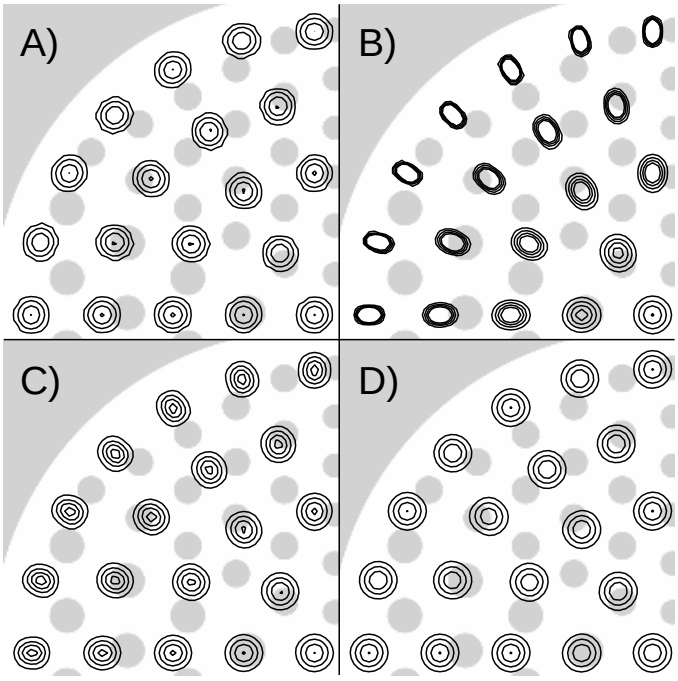


Figure 6.25: Local impulse responses for various methods in shift-variant SPECT. This figure shows responses for A) Uniformity and attenuation-corrected FBP, B) PL with space-invariant penalty, C) PL with proposed penalty, and D) post-smoothed ML. Each estimator tries to match a 10.0 mm FWHM target at the center of the field of view. All responses are superimposed on the upper left quadrant of the phantom to illustrate the sample locations for these impulse response.

about 3 mm at the edges of the phantom to nearly 7 mm at the center. This is an indication that the system matrix, $\boldsymbol{H}$, is rank-deficient, and the ML estimator cannot resolve single pixels. Thus, the post-smoothed estimates must also have nonuniform resolution properties. For relatively large target responses, the post-smoothing blur dominates and these nonuniformities are very small (as we have seen for the 10.0 mm target). However, for smaller desired responses, simple post-smoothing will not yield the desired target. However, we can adopt a post-filter approach that compensates for the intrinsic blur of the ML estimator by applying (6.6).

We use (6.6) to find a post-smoothing filter for ML for a target response of the form in (6.4) with a FWHM resolution of 7.7 mm. Figure 6.25 shows local impulse responses for the 7.7 mm target for the proposed PL estimator and for the

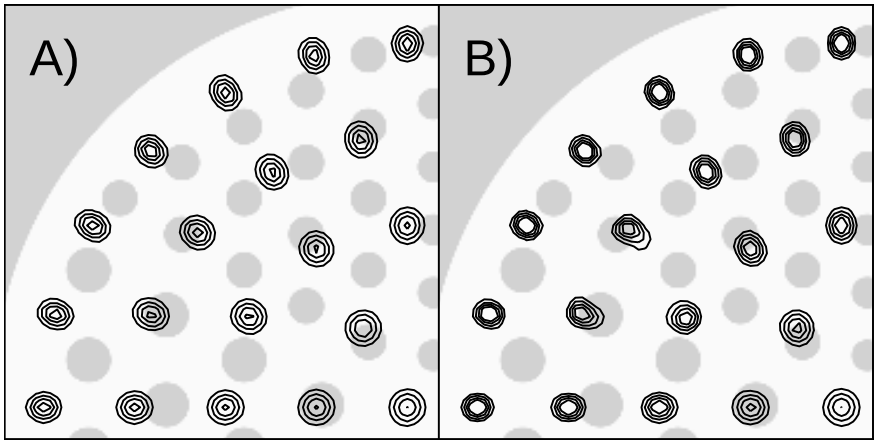


Figure 6.26: SPECT local impulse responses for a 7.7 mm FWHM target. A) PL with proposed penalty and B) post-smoothed ML.

post-smoothed ML approach. Despite matching the target response at the center pixel (lower right corners in each subfigure in Figure 6.26), the ML approach clearly yields nonuniform resolution properties with narrower responses toward the edges. In comparison, the PL approach yields more uniform results.

In summary, the only 2D SPECT reconstruction methods presented here that yield nearly uniform resolution properties are post-smoothed ML (for larger FWHM targets), the proposed PL approach, and FBP with frequency-distance corrections. We compare the noise properties of these methods in Section 7.2.

6.3.2 A Hybrid Regularization Approach

An interesting alternative to choosing between post-smoothed ML and the space-variant PL approaches is to use both! One can use a hybrid method that includes a degree of regularization that keeps responses fairly uniform and increases convergence rates for iterative algorithms, and then apply a post-smoothing filter to set the overall target resolution. This approach is attractive for a number of reasons. Using the PL approach keeps the responses uniform even for fairly small target responses. Post-

smoothing will generally reduce any of the remaining resolution nonuniformities, and can be applied quickly for a number of desired FWHM resolutions or responses without additional iterative reconstructions. And, convergence rates are increased over the unregularized ML approach, reducing computation.

This hybrid post-smoothed PL approach can implemented easily, using (6.6) to find the appropriate post-smoothing filter for a desired overall response. In this case, the l_{est} term represents the "first-pass" resolution induced by the penalized like-likelihood objective. For our proposed space-variant penalty, l_{est} is equal to the "first-pass" target response. Figure 6.24H shows a sample reconstruction using this technique for the SPECT problem. For this hybrid estimator, we apply our PL approach with a target of the form in (6.4) with a 7.7 mm FWHM, followed post-filtering via (6.6) using same overall target as the other methods shown in that figure. One can see the increased uniformity as compared with the non-hybrid PL approach in Figure 6.24G.

6.4 Summary

In this chaper we have demonstrated the fast resolution and covariance prediction methods discussed in Chapter V. These methods are fast and accurate, and can be applied to large shift-invariant tomographic systems like 3D SPECT. Similarly, we have demonstrated the fast penalty design methods of Chapter V produce highly uniform resolution properties comparable to the slower design methods (*i.e.*, the CNLLS penalty) of Section 4.3.3. These methods are practical for both shift-invariant and shift-variant imaging systems, including 2D and 3D, PET and SPECT systems.

CHAPTER VII

Noise Performance of Uniform Resolution Estimators

In this chapter we discuss the relative performance of different reconstruction methods. In particular, we are interested in the noise properties of the images produced by different methods. In Section 7.1 we investigate the noise properties of a shift-invariant PET system, by looking at standard deviation images, bias-variance curves, and correlation images. Because we have found that the noise properties of an image are very closely tied to its resolution properties, we perform noise studies in Section 7.2 on SPECT estimators that are very carefully resolution matched. This study includes a discussion of the performance and the advantages and disadvantages of various uniform resolution reconstruction methods.

7.1 PET Studies

We begin our noise investigation with a study of the noise properties of images from the shift-invariant PET system and digital phantom in Figure 3.5.

7.1.1 Variance in Reconstructed Images

To form sample standard deviation images, we simulated 400 noisy measurement realizations for the digital phantom in Figure 3.5. The PET model included 10% random coincidences and averaged 1 million counts per realization.

We reconstructed each of these 400 realizations using 30 iterations of the SAGE algorithm [39] with the same regularization methods used in the resolution properties investigation in Section 6.2.1. For all of the statistical methods except the CNLLS penalty, we use the noisy measurements, $\underline{Y}$, for calculation of $\boldsymbol{R}$. Because of the extensive computation time associated with calculation of the CNLLS penalty, the noiseless, $\underline{\bar{Y}}$ were used. (*i.e.*: The same penalty based on the noiseless measurements was used for all realizations.)

The results of this noise investigation are presented in Figure 7.1. The sample standard deviation images are shown on the left side of the figure. Horizontal and vertical profiles of these images are shown in the remaining plots. The horizontal profile is taken through the image center and the vertical profile is taken through the center of the cold disc. These profiles are represented by dotted lines in the images. Pixel standard deviations in these plots are expressed in terms of a percentage of the background ellipse intensity. If one included error bars on these plots, the error bars would be smaller than the plot markers. Therefore, we have eliminated the error bars for clarity. For conventional regularization, the standard deviation estimate is nearly uniform. FBP and PULS generally have the highest standard deviation and the certainty-based penalty have the lowest standard deviation. Not only do FBP and PULS share similar resolution properties, but also similar noise properties. The close agreement in standard deviation between the proposed method and the CNLLS penalty further justifies our computationally efficient design technique.

At first glance, it appears that uniform resolution properties come at the price of a variance increase as compared with the certainty-based penalty. However, the certainty-based penalty and the proposed penalty have *different* resolution properties. The certainty-based reconstruction often has a greater maximum diameter of the

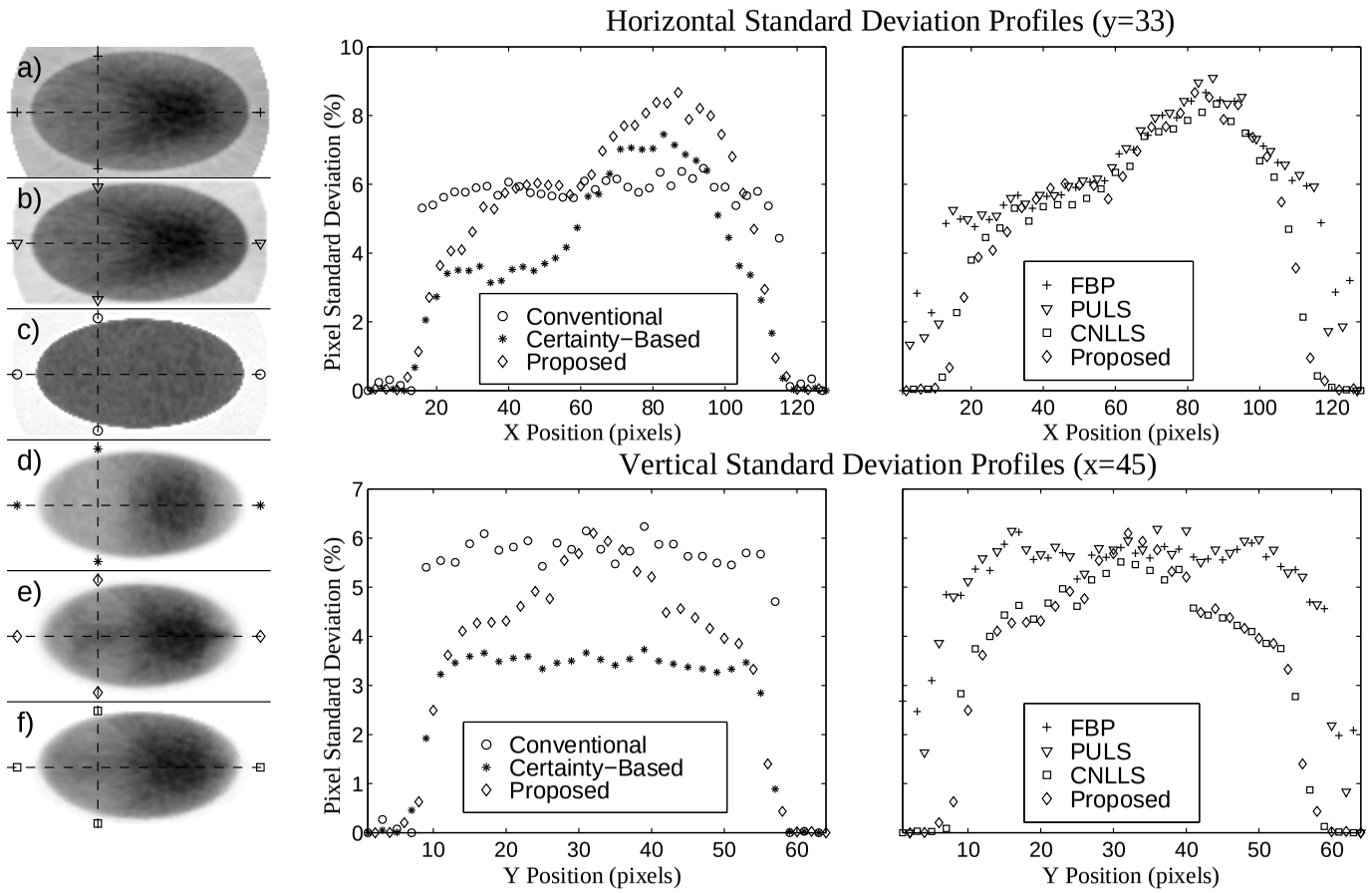


Figure 7.1: Sample standard deviation images and profiles for 2D PET reconstructions.
This figure shows standard deviation images and profiles for the following estimators: a) Filtered backprojection (+), b) Penalized unweighted least-squares PULS ($\triangledown$), c) PLE with conventional regularization ($\circ$), d) PLE with certainty-based penalty ($*$), e) PLE with proposed penalty ($\lozenge$), and f) PLE with CNLLS penalty ($\square$).

local impulse response (compare Figure 4.1 and Figure 6.8). This can be interpreted as increased smoothing, and therefore yields reconstructions with lower variance.

7.1.2 Bias-Variance Curves

We would like to produce a resolution-noise curve comparing the relative performance of these two methods over a range of target resolutions, but this is difficult because they have different resolution properties. Using the angularly averaged FWHM as a resolution metric (cf [41]) unfairly handicaps estimators with isotropic resolution properties.[1] Estimators with anisotropic responses can reduce noise by smoothing "optimally" in each direction while maintaining the same average FWHM as an estimator with isotropic responses. Rather than creating resolution-noise curves where each point on the curve corresponds to a single resolution value and a single standard deviation, we created "banded curves" as follows. For the ordinate, we used the sample standard deviations of pixel values in images reconstructed from 400 noisy sinogram realizations, for each of several target spatial resolutions. For each target resolution we also computed the local impulse response and found the smallest and largest diameters of its half-maximum contour. We specified the abscissae in the banded plot as the interval between the minimum and maximum diameters. For each pixel location and target resolution, these plots describe the (single) pixel standard deviation value as well as the *range* of spatial resolutions spanned by the local impulse response. A method with isotropic resolution properties would appear as a single line in such plots, whereas a method with a highly anisotropic response appears as a thick band.

We calculated such trade-off curves for four pixel positions. The curves for the conventional and proposed penalties are shown in Figure 7.2. The lighter band

[1] The angularly averaged FWHM resolution also ignores the tails of the response which can have a large effect on the bias.

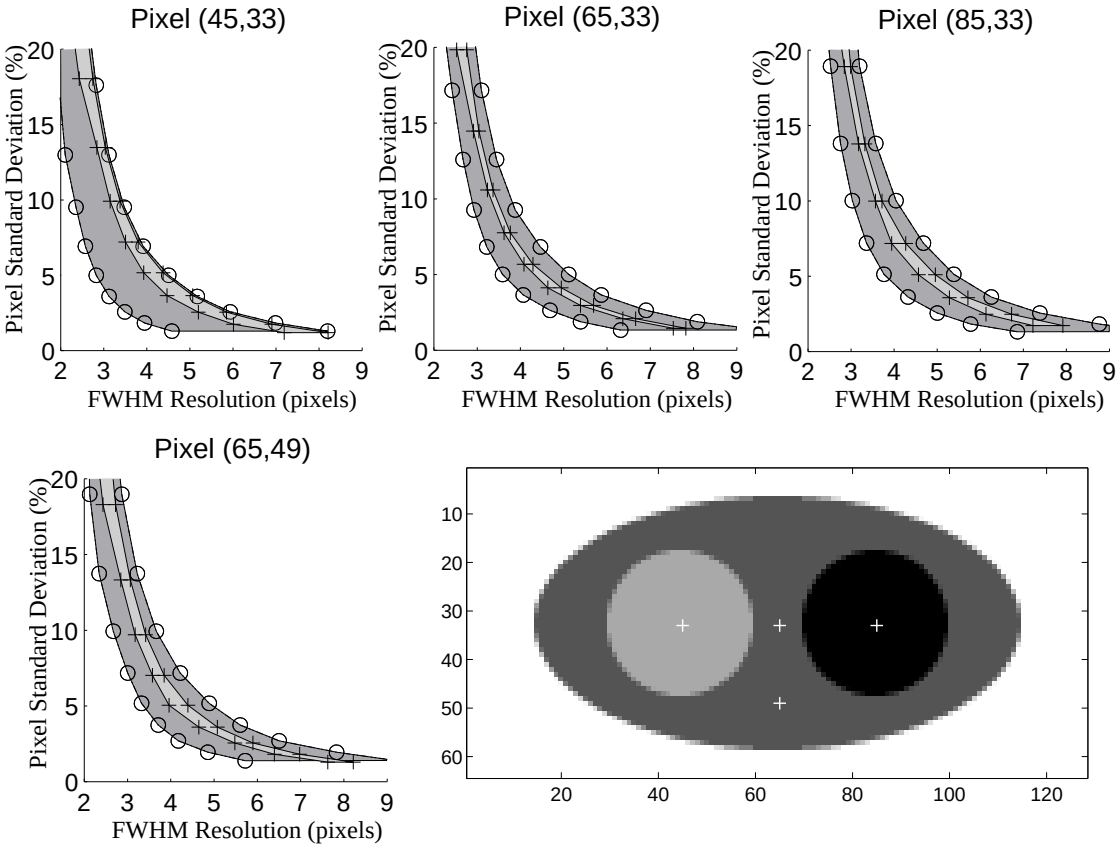


Figure 7.2: Banded bias-variance curves for a conventional penalty and our proposed penalty. This figure shows the resolution/noise trade-off for penalized-likelihood emission image reconstruction with conventional (∘/dark) and proposed penalties (+/light). The four locations for which these curves are plotted are shown in the lower right figure.

with "+" symbols on the border represents the resolution/noise trade-off curve for the proposed regularization, while the darker band with "∘" symbols on the border is the curve for reconstruction with conventional regularization. (The light band partially obscures the dark band; however, the borders are marked by symbols and lines so that the degree of overlap is visible.)

We also produced a banded resolution/noise trade-off plot using the certainty-based regularization of [41]. Since the certainty-based technique produced a curve nearly identical to the conventional regularization, we have omitted the plot. Similar behavior was observed in [41] using a mean FWHM resolution criterion. Essentially this means each pixel simply moves up or down its resolution/noise curve to the

specified resolution. This is another indication that the certainty-based method does not yield isotropic resolution properties. While the average FWHM resolution may be improved, the local impulse responses are still anisotropic yielding a wide resolution band in our banded resolution/noise trade-off curves.

In Figure 7.2 the banded curves for the proposed penalty span a small resolution range (*i.e.*, the curve is thin horizontally), indicating isotropic smoothing properties relative to the conventional penalty. If our design were ideal, minimum and maximum FWHM resolution would be identical and we would have a line instead of a band. Note that the proposed penalty band lies inside the conventional penalty band. If the proposed penalty band laid above the conventional penalty band over the same resolution interval, then the proposed penalty would arguably have worse noise properties. The proposed penalty band generally lies in the center of the conventional penalty band. However, this is not the case for the pixel (45,33) in the cold disc. Note that the local impulse response for the conventional penalty at this pixel is especially asymmetric (see Figure 3.6) having the largest difference between the min and max FWHM resolutions. If this local impulse response yields an "optimal" kind of smoothing (with its predominantly vertical orientation), it is logical that an isotropic response would decrease the variance little with additional horizontal smoothing (note that max resolution for the conventional local impulse response is very close to the 4.0 target). Using this rationale the proposed penalty bands for the other pixel locations lie roughly in the middle of the conventional penalty's band since the local responses for these points are less asymmetric (with the max resolution greater than 4.0 and min less than 4.0 pixels). The isotropic response reduces the max resolution and increases the min resolution as compared with the conventional response. The "optimal" smoothing of the conventional response is arguably not so

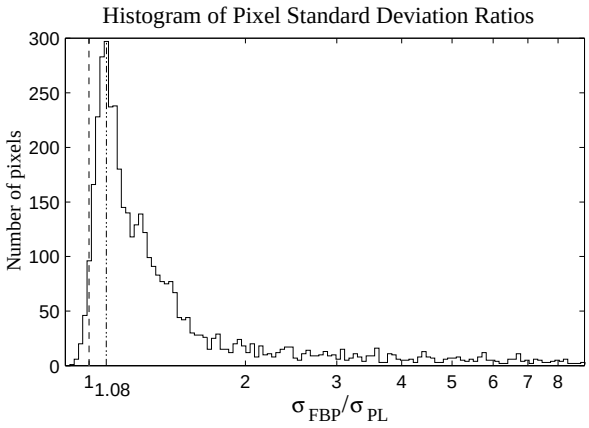


Figure 7.3: Relative noise performance of FBP and PL estimators.
This figure shows a histogram demonstrating the distribution of the ratio of the pixel standard deviation using filtered backprojection (σ_{FBP}) to the pixel standard deviation using a PLE with the proposed penalty (σ_{PL}).

directionally dependent in this case and an isotropic response can provide roughly the same variance. While these two methods have different resolution properties, it appears that our penalty design has not adversely affected the noise properties of the estimator.

7.1.3 Comparing Estimators with Approximately Matched Resolutions

It is difficult to compare *globally* our proposed penalty with the conventional and certainty-based methods for an entire image reconstruction because they possess different resolution properties for every pixel. On the other hand, FBP and the proposed penalty both yield nearly the same local impulse responses, so a comparison seems more appropriate. Since these methods have nearly the same resolution properties, we should be able to identify which provides better global noise properties. Note, particularly in the vertical profile in Figure 7.1, that reconstructions based on the proposed penalty often have lower variance than FBP.

There are a few points in Figure 7.1 where the standard deviation estimate is slightly greater for the proposed penalty. To illustrate the relative global noise prop-

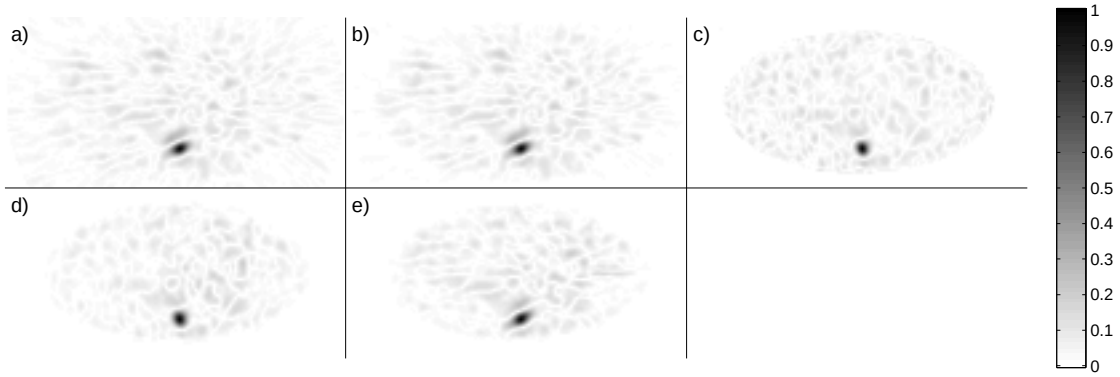


Figure 7.4: Correlation maps for various estimators for 2D PET.
This figure shows sample absolute correlation maps for pixel (65,49) using the following estimators: a) Filtered backprojection (FBP), b) Penalized unweighted least-squares (PULS), c) PLE with conventional regularization, d) PLE with certainty-based penalty, e) PLE with proposed penalty.

erties of FBP and the PLE with the proposed regularization, we generated a histogram of the relative variances. For each pixel in the object, we calculated the ratio of the sample standard deviation at that pixel using filtered backprojection (σ_{FBP}) to the sample standard deviation at that pixel using the PLE with the proposed regularization (σ_{PL}). For pixels where $\sigma_{\mathrm{FBP}}/\sigma_{\mathrm{PL}}$ is greater than one, filtered backprojection has higher standard deviation. This histogram is shown in Figure 7.3. The vertical dashed line indicates the position where this ratio equals one. For nearly every pixel the PLE with the proposed regularization produces lower variance estimates and, for those pixels that have higher variances the difference is only slight. More than 50% of the pixels have over a 20% reduction in reconstructed pixel standard deviation.

Although it appears that the PL approach has better noise performance than FBP, we know these methods are not perfectly resolution-matched. Therefore, particular care should be taken in interpreting these results. In Section 7.2 we concentrate on SPECT reconstructions that are nearly exactly resolution matched.

7.1.4 Correlation Investigation

In addition to the variance investigation, we present a brief correlation investigation. We have included a set of typical correlation maps in Figure 7.4 for FBP, PULS, and the PLEs with conventional, certainty-based, and proposed penalties. These maps represent the absolute value of the correlation between each pixel and pixel (65,49). FBP and PULS have nearly identical correlation maps (particularly inside the object). The PLEs with conventional and certainty-based penalties have similar correlation maps, but are noticeably different due to the different resolutions. The proposed method is shown in Figure 7.4e. The structure of the correlation immediately surrounding (65,49) is quite similar to FBP and PULS, having lost the nearly isotropic effect of the other PLEs. This behavior is somewhat counter-intuitive since PLEs usually have much narrower correlation sidelobes than FBP and PULS.

We have seen that the PL approach with our proposed penalty has noise properties that are similar to methods like PULS and FBP. The variance properties of our PL approach appeared only marginally better than FBP and PULS for this particular shift-invariant PET system and object. We believe more significant differences can be found in shift-variant systems, where it is often more difficult to reconstruct images with uniform resolution properties. The next section addresses such a system.

7.2 SPECT Studies

In this section we consider the performance of various estimators that are applied to an intrinsically shift-variant SPECT system. We first must identify estimators with nearly exactly matched resolution properties. We have found that simply matching FWHM resolution is insufficient for comparison, as the sidelobe behavior and overall shape of the response can greatly affect the noise performance.

Recalling the investigations in the Section 6.3.1, the only uniform resolution methods we have investigated with well matched responses are uniformity-corrected FBP, post-smoothed ML, and our proposed PL technique. Additionally, we know that these methods are not globally exactly matched. However, in practice we can match these methods at (at least) one pixel by choosing to post-smooth the ML and FBP approaches using (6.6) and a target response equal to the estimated PL response. Thus, we can very nearly exactly match the local impulse response at, for example, the center pixel. Other pixel positions will generally be only approximately matched.

7.2.1 Variance of Estimators with Exactly Matched Resolution

Returning to the SPECT model of Section 3.2.3 with the "cold rod" phantom of Figure 3.3, we performed 400 noisy reconstructions for the uniformity-corrected FBP techniques (using the frequency-distance principle filtering method), our PL approach, and the post-smoothed ML method. For these techniques we applied the same reconstruction algorithms as were applied in Section 6.3.1. This was performed over a range of target resolutions with FWHM from 7.7 mm to 17 mm, using the target response of (6.4). No targets below 7.7 mm were calculated because even unpenalized ML yields a response of about 6.9 mm at the center pixel. This minimum resolution represents a barrier for both methods since the PL method approaches the ML estimate for small target resolutions. We chose the post-filters for the FBP and ML techniques using (6.6) over the entire range of targets. Thus, the resolution properties are essentially exactly matched for all methods at the center even for the smaller target responses.

Figure 7.5 shows standard deviations for the center pixel for these methods. One standard deviation error bars are shown for each estimate. The plots for the proposed PL approach and the post-smoothed ML estimates are nearly identical with

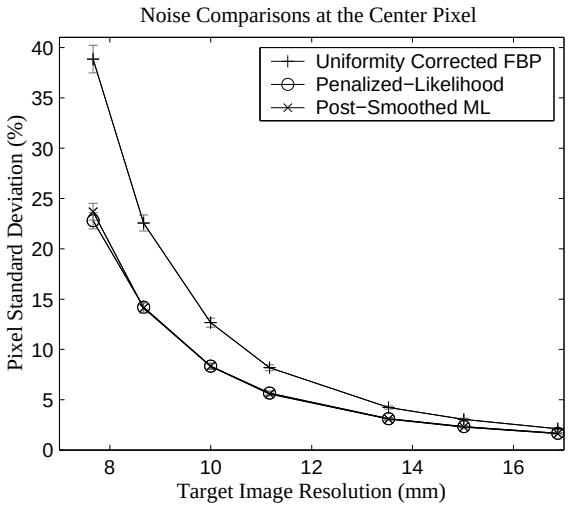


Figure 7.5: Noise/resolution trade-off for exactly matched SPECT estimators.
This figure shows standard deviations for the uniformity-corrected FBP (+), the penalized-likelihood (∘), and the post-smoothed maximum-likelihood (×) techniques for the center pixel where the local impulse responses are exactly matched over a range of target FWHM resolutions.

small differences well within the error bars. Thus, in terms of variance the methods appear to have the same noise performance when the spatial resolutions are carefully matched. In contrast, the FBP approach suffers from increased noise in the reconstructions.

Covariance Study

We also study the covariances in the reconstructions. Covariance functions are arguably a more important feature than variances for evaluating different methods with specific tasks in mind. (For example, many computer observer models require the covariance functions to assess performance.) We calculated the sample covariance function for the center pixel for the uniform resolutions methods using the 400 reconstructions. These covariances are shown in Figure 7.6 for 7.7 mm, 10.0 mm, and 15.0 mm targets. The plots for the post-smoothed ML and our proposed PL approaches are nearly indistinguishable. Thus, for this system and target, the post-smoothed ML and PL approaches have essentially the same noise perfor-

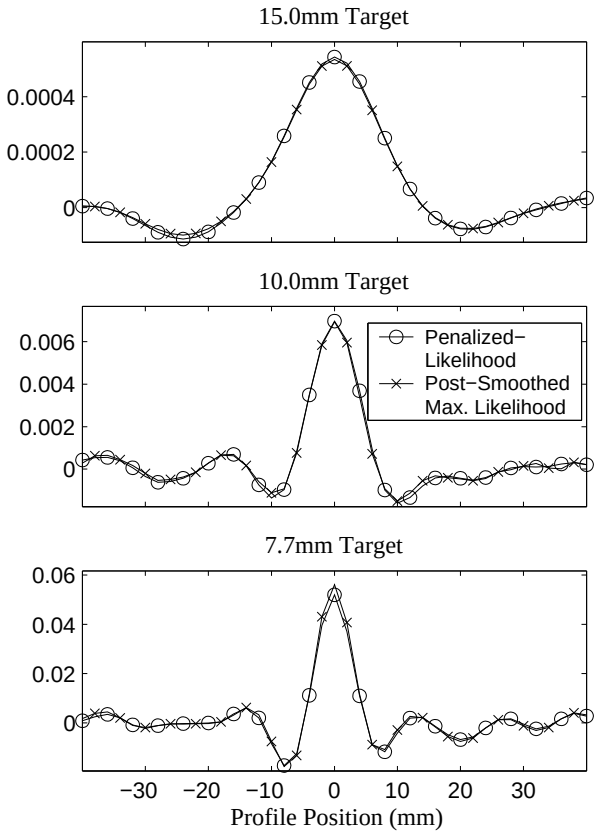


Figure 7.6: Comparison of covariance functions for PL and PSML.
These plots show sample covariance functions for PL (∘) and PSML (×) with exactly matched resolutions for three different target resolutions.

mance. Neither method appears to have an advantage over a wide range of practical reconstruction resolutions. This result is not entirely unexpected. In Section 7.2.3 we present an analysis for a linear measurement model and a Gaussian noise model, and argue that the post-smoothed ML and exactly matched PL methods should yield identical covariance properties. Thus, for cases where the Poisson statistics are modeled well by a Gaussian approximation, it is not surprising that the same conclusions hold. In contrast, the uniformity-corrected FBP yields different covariance functions. However, whether or not FBP's covariance is desirable will depend on the task for which the images are made and if the associated reconstruction artifacts are

tolerable.

We also explored the noise performance at other pixel locations found similar equivalence of post-smoothed ML and PL for the target in (6.4). However, for other pixel positions and other targets at particularly fine matched resolutions sometimes one or the other algorithm would have lower standard deviation depending on the particular location and target response. Rather than attempting to draw general conclusions about the relative merits of the two approaches for all conditions it seems advisable for algorithm designers to compare the two for the given system model and target resolutions of interest.

7.2.2 Convergence Rates

Since the noise performance for PL and post-smoothed ML are indistinguishable for the investigations in the previous section, other considerations such as computation time may be more important. It is popularly held that unregularized methods converge more slowly than regularized methods due to the conditioning of the problems. However, unregularized algorithms converge to different limits than the regularized algorithms making analytical comparisons difficult.

We performed a simple investigation of the convergence rates of matched post-smoothed ML and PL approaches. We compared the normalized mean squared difference between the image estimate at the nth iteration, $\underline{\hat{\theta}}_n$, and the fully converged solution, $\underline{\hat{\theta}}_\infty$. For the PL approach, $\underline{\hat{\theta}}_n$ is simply the estimate at the nth iteration. For the post-smoothed ML technique, $\underline{\hat{\theta}}_n$ is the ML estimate at the nth iteration, with a post-smoothing filter applied.

We initialized with an FBP image and used the same ordered-subsets techniques and the same 10.0 mm target response mentioned in Section 6.3.1. Estimates of $\underline{\hat{\theta}}_\infty$, were calculated using 500 ordered-subsets iterations, followed by 100 single subset

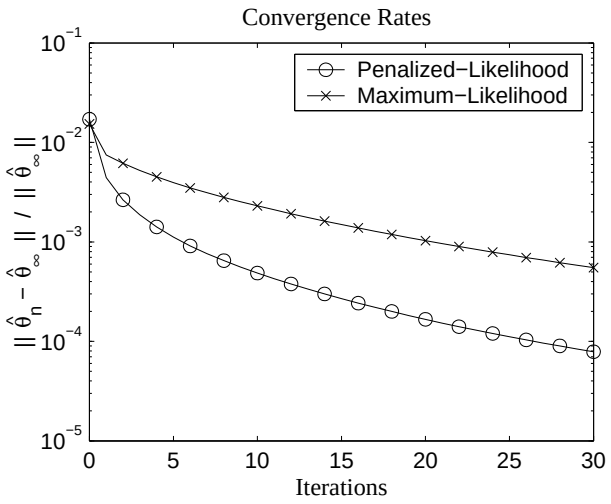


Figure 7.7: Convergence rates of PL and ML.
This figure shows the relative convergence rates of the PL ($\circ$) and post-smoothed ML ($\times$) approaches for a 10 mm target FWHM resolution.

iterations.

Figure 7.7 shows that the PL approach converges more quickly than the ML approach. For a similar level of convergence, it appears that the ML technique takes roughly three times the number of iterations. Such speed-ups depend on the target resolution, since increased regularization leads to better conditioning. However, we expect similar rank performance for the two methods for different target resolutions.

7.2.3 Theoretical Analysis of PL and PSML with Exactly Matched Resolution

Following [40], in this section we describe conditions under which a post-filtered weighted least-squares reconstruction is identical to a penalized weighted least-squares reconstruction. This mathematical equivalence corroborates our empirical findings for post-filtered ML and penalized-likelihood reconstructions in Section 7.2. Recall that for a linear measurement model, the PWLS estimate maximizes an objective function of the following form:

$$\boldsymbol{\Phi}(\underline{\theta}, \underline{Y}) = -\left(\underline{Y} - \boldsymbol{H}\underline{\theta}\right)' \boldsymbol{K}^{-1} \left(\underline{Y} - \boldsymbol{H}\underline{\theta}\right) - \underline{\theta}' \boldsymbol{R} \underline{\theta}, \tag{7.1}$$

where $\boldsymbol{K} = \text{cov}\{\underline{Y}\}$. Assuming appropriate invertibility conditions hold, the minimizer has the following closed-form solution

$$\hat{\underline{\theta}} = [\boldsymbol{H}'\boldsymbol{K}^{-1}\boldsymbol{H} + \boldsymbol{R}_{\text{sym}}]^{-1}\boldsymbol{H}'\boldsymbol{K}^{-1}\underline{Y}, \tag{7.2}$$

where $\boldsymbol{R}_{\text{sym}}$ was defined in (4.16). We will also assume that the actual system model is exactly matched to the reconstruction model with measurements related to the object through the system matrix, $\boldsymbol{H}$. From (7.2), it is straightforward to write the covariance matrix for $\hat{\underline{\theta}}$ as

$$[\boldsymbol{F} + \boldsymbol{R}_{\text{sym}}]^{-1}\boldsymbol{F}[\boldsymbol{F} + \boldsymbol{R}_{\text{sym}}]^{-1}. \tag{7.3}$$

Thus, for a given penalty matrix, we can write out the covariance of the reconstructed images.

Recall the explicit solution for a penalty matrix presented in Section 4.3.1. While this is not a convenient form for penalty design, it is convenient for theoretical investigation. For the specific collection of desired impulse responses represented by $\boldsymbol{L}_0$, and the PWLS objective, (4.27) may be written[2] as

$$\boldsymbol{R}^{\star}_{\text{sym}} = \boldsymbol{F}[\boldsymbol{L}_0^{-1} - \boldsymbol{I}]. \tag{7.4}$$

Thus, for a specific set of desired responses, we may plug (7.4) into (7.3) to obtain the covariance matrix for penalized-likelihood reconstruction:

$$\begin{aligned} \text{Cov}\{\hat{\underline{\theta}}_{\text{PL}}\} &= [\boldsymbol{F}\boldsymbol{L}_0^{-1}]^{-1}\boldsymbol{F}[\boldsymbol{F}\boldsymbol{L}_0^{-1}]^{-1} \\ &= \boldsymbol{L}_0\boldsymbol{F}^{-1}\boldsymbol{L}_0. \end{aligned} \tag{7.5}$$

Similarly, because post-smoothing is a linear operation: $\hat{\underline{\theta}}_{\text{PSML}} = \boldsymbol{L}_0\hat{\underline{\theta}}_{\text{ML}}$, we may find the covariance for post-smoothed ML reconstruction by first finding the covariance

[2] Recall that this solution only exists under particular circumstances discussed in Section 4.3.1.

for the ML approach by setting $\boldsymbol{R}_{\text{sym}} = \mathbf{0}$ in (7.3). That is,

$$\begin{aligned} \text{Cov}\{\hat{\underline{\theta}}_{\text{PSML}}\} &= \boldsymbol{L}_0 \text{Cov}\{\hat{\underline{\theta}}_{\text{ML}}\} \boldsymbol{L}_0 \\ &= \boldsymbol{L}_0 \boldsymbol{F}^{-1} \boldsymbol{L}_0. \end{aligned} \tag{7.6}$$

Thus, (7.5) and (7.6) are identical. Therefore, when resolution properties are exactly matched under this system model, the penalized weighted least-squares and post-smoothed weighted least-squares approaches will yield the exact same noise performance. In the empirical investigations of Section 7.2 we observed similar results under the Poisson model.

It is plausible that under other noise models, where the noise cannot be well approximated by a Gaussian model, or when (7.4) cannot be solved, the noise performance will be significantly different in the post-smoothed ML and PL cases.

CHAPTER VIII

Conclusion

8.1 Summary

In this dissertation we have developed practical methods for controlling the resolution properties of images produced by penalized-likelihood estimators. The methods described apply generally to a broad class of imaging systems. We have concentrated on (emission) tomography systems that can possibly be shift-variant, and have developed fast methods to compute the shift-variant penalty design, which is necessary for resolution control.

We also have derived a new formulation for the local impulse response (a resolution predictor) appropriate for systems where the discrete reconstruction model is mismatched with the "real world" continuous object and measurement model. The same fast methods used for penalty design may be applied to both resolution and covariance prediction. We have shown that such predictions can be made for both PET and SPECT system models with increased speed and accuracy over traditional prediction techniques.

We have used our resolution control methods for many different tomographic models to provide images with nearly uniform (shift-invariant and isotropic) resolution properties. Uniform resolution appears to be important for tasks such as image

registration where the shape of features in the image can potentially be distorted by nonuniform resolution properties. Previously, conventional penalized-likelihood techniques were not able to provide uniform resolution even in the case of a shift-invariant system model. Our methods provide nearly uniform resolution even in the case of intrinsically shift-variant systems like SPECT. Whereas, very few estimators, including conventional penalized-likelihood methods, are able to achieve good resolution uniformity.

While it is unclear whether uniform resolution is important for many tasks, it seems that the issue of resolution controllability is very important. Rather than letting the resolution properties be controlled by the specifics of the system geometry and measurement noise, using the methods developed in this work, one can now specify user-defined resolution properties. The methods we have presented are general and may be applied to achieve user-specified shift-variant resolution properties (*e.g.*, creating uniform resolution in regions while preserving edges).

Because we have the ability to match the resolution properties of different estimators, we now have a foundation for the fair comparison of estimators. We have performed such an investigation for a SPECT system using nearly exactly matched FBP, post-smoothed ML, and the proposed PL methods. As one might expect the statistical methods that fully incorporate all physical effects and take the noise model into account provide reconstructions with lower noise than FBP. However, we have also found that the noise characteristics of our PL approach and post-smoothed ML are nearly identical. Thus, the decision to use one estimator over another must be made on other considerations such as resolution uniformity or computation time.

Such decisions depend on one's exact reconstruction needs. As we have seen in some cases, methods that depend on fully resolving (*i.e.*, for pixelized images the

local response equals to an impulse function) image parameters, like post-smoothed ML, cannot always provide uniform resolution. Generally this means that the system model and the object parameterization results in an underdetermined (or rank-deficient) problem. Thus, while methods like post-smoothed ML will yield nonuniform resolution properties, we have shown that PL with our proposed penalty can still yield images with very uniform resolution.

In many other cases, the nonuniformities produced by post-smoothed ML are relatively insignificant. Thus, things like computation time and convergence rates are important. While we have shown that our PL approach generally has increased convergence rates, it also often yields images with less resolution uniformity than post-smoothed ML.

Typically all these factors must be weighed when choosing an estimator. We have also discussed a hybrid estimator that includes both penalty design and post-reconstruction filtering. This method provides added flexibility in deciding where various trade-offs are made.

8.2 Future Work

While we have accomplished much for predicting and controlling resolution for penalized-likelihood estimators, there are always improvements or extensions that can be made and other imaging systems to which these methods could be applied.

The prediction and penalty design techniques presented in this paper could also be applied to other imaging modalities like x-ray computed tomography and magnetic resonance imaging. Similarly, our techniques may need to be extended to for more realistic systems. For example, modern SPECT systems use body contouring orbits to improve image quality. This adds an additional kind of object-dependence into

the SPECT system model, which must be taken into account to use the fast linear operator approach discussed in Chapter V.

It is possible that one can find better techniques for relaxing the design constraints on the penalty design. While we have investigated relaxed constraints in some detail, the resulting penalty designs were cumbersome and too slow for practical use. Other relaxed constraints, such as ones that are based on the componentwise distance to the nearest singular matrix[101], might be applied to a technique that updates a penalty matrix.

We also believe that more work need to be done on investigating the properties of images under a wider range of desired responses. We have concentrated on desired responses that are similar to the "natural" responses penalized-likelihood estimators. We know that some responses are difficult or impossible to achieve. A better classification of achievable responses would be most helpful. This should include a deeper investigation of the hybrid approach that mixes both objective function regularization and post-smoothing. Similarly, since the "natural" responses often possess ringing, one might choose to adopt a constraint to reduce ringing, such as the Tchebychev equi-ripple model.

It would also be natural to extend these penalty design approaches to non-quadratic penalties. Specifically, just as a concept of uniform resolution may be important for a "smoothing" penalty, there is a concept of uniform "edgeness" for edge-preserving penalties. For example, one might desire that edges in high count regions are just as likely to be formed for a given edge size, as those in low count regions. Recall from Chapter III that nonquadratic penalties have Hessians that are object-dependent. Thus the local impulse responses (or covariance predictors) can also be highly object-dependent and may require better approximations. Similarly,

one might be able to use the square-root penalty discussed in Section 4.2.1 to achieve more uniform resolution properties in particularly troublesome cases.

Also, we would ultimately like to be able to design the resolution and covariance properties of an estimator to be optimal for a certain task. Specifically, we might want to design the optimal regularization for a tumor detection task using a computer observer. It is not obvious that the naturally induced resolution properties of conventional PL are optimal. We expect that this will require further work in the area of quantifying performance of many tasks like the joint localization and detection task (such as LROC curves). However, we have already developed many important tools for controlling the resolution properties of an image, and predicting the covariance in reconstructed images with practical computation times.

We believe the approximations for SPECT attenuation made in Section 5.3 have particular potential for application in other areas. For example, our approximations could be incorporated into preconditioning methods as in [35] to speed convergence rates. These approximations can also be incorporated into computer observer investigations, various estimation bounds, or other areas where the Fisher information matrix (or the individual weighted responses) must be calculated repeatedly for different measurement data or other geometry-independent parameters.

APPENDICES

APPENDIX A

Space-Invariant Weighted Responses

In this appendix we show that the backprojected weighted projection of an impulse results in a space-invariant blur for radially constant weightings and detector responses that are not depth-dependent and shift-invariant.

Returning to the idealized continuous model in Section 2.2.1, we may write the weighted projection-backprojection of an image, f, as $\mathcal{P}'_{\text{blur}}\mathcal{W}\mathcal{P}_{\text{blur}}f$, where the operator $\mathcal{W}$ represents the application of a projection domain weighting function, $w_\theta(r)$. For non-depth-dependent and shift-invariant detector blur, $b(r)$, we may write the blurred projection of an image, $\mathcal{P}_{\text{blur}}f$, as

$$p_\theta(r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b(r-s)f(l\cos\theta + s\sin\theta, l\sin\theta - s\cos\theta)\,dl\,ds. \tag{A.1}$$

Similarly, when one chooses a radially constant weighting such that $w_\theta(r) = w_\theta$, the weighted backprojection of the blurred projections is written:

$$f_b(x,y) = \int_{-\infty}^{\infty}\int_{0}^{\pi} b(s)w_\theta p_\theta(x\cos\theta + y\sin\theta - s)\,d\theta\,ds. \tag{A.2}$$

The 2D Fourier transform of the weighted projection-backprojection of an image is

written as

$$F_b(\rho,\phi) = \int\int f_b(x,y)e^{i2\pi\rho(x\cos\phi+y\sin\phi)}\,dx\,dy \tag{A.3}$$

$$= \int\int\left(\int_{-\infty}^{\infty}\int_0^{\pi} b(s)w_\theta p_\theta(x\cos\theta+y\sin\theta-s)\,d\theta\,ds\right)e^{i2\pi\rho(x\cos\phi+y\sin\phi)}\,dx\,dy \tag{A.4}$$

Making a change of coordinates such that $\tilde{x} = x\cos\theta+y\sin\theta$ and $\tilde{y} = y\cos\theta-x\sin\theta$, we may rewrite (A.4) as

$$\begin{aligned} F_b(\rho,\phi) &= \int\int\int_{-\infty}^{\infty}\int_0^{\pi} b(s)w_\theta p_\theta(\tilde{x}-s)e^{i2\pi\rho[\tilde{x}\cos(\phi-\theta)+\tilde{y}\sin(\phi-\theta)]}\,d\theta\,ds\,d\tilde{x}\,d\tilde{y} \\ &= \int w_\theta\left\{\int\left(\int_{-\infty}^{\infty} b(s)p_\theta(\tilde{x}-s)\,ds\right)e^{i2\pi\rho\tilde{x}\cos(\phi-\theta)}\,d\tilde{x}\int e^{i2\pi\rho\tilde{y}\sin(\phi-\theta)}\,d\tilde{y}\right\}d\theta \\ &= \int w_\theta B\left(\rho\cos(\phi-\theta)\right)P_\theta\left(\rho\cos(\phi-\theta)\right)\delta\left(\rho\sin(\phi-\theta)\right)\,d\theta, \end{aligned} \tag{A.5}$$

where $B(\cdot)$ denotes the 1D Fourier transform of the blur function $b(r)$ and $P_\theta(\cdot)$ denotes the 1D Fourier transform of the blurred projections along the radial direction. Using properties of delta functions,

$$\delta\left(\rho\sin(\phi-\theta)\right) = \frac{\delta(\theta-\phi)}{\left|\frac{d}{d\theta}\left(\rho\sin(\phi-\theta)\right)\right|_{\theta=\phi}}. \tag{A.6}$$

Thus, we may write (A.5) as

$$\begin{aligned} F_b(\rho,\phi) &= \int w_\theta B\left(\rho\cos(\phi-\theta)\right)P_\theta\frac{1}{\rho}\delta(\theta-\phi)\,d\theta \\ &= \frac{w_\phi}{\rho}B(\rho)P_\phi(\rho). \end{aligned} \tag{A.7}$$

The 1D Fourier transform of the blurred projections is

$$P_\phi(\rho) = \mathfrak{F}_1\left\{\mathcal{P}_{\text{blur}}f\right\} = \mathfrak{F}_1\left\{b(r)\right\}\mathfrak{F}_1\left\{\mathcal{P}f\right\} \tag{A.8}$$

$$= B(\rho)\mathfrak{F}_2\left\{f(x,y)\right\} = B(\rho)F(\rho,\phi), \tag{A.9}$$

where $\mathfrak{F}_2\left\{\cdot\right\}$ denotes the 2D Fourier transform and we have used the Fourier slice theorem[80] which says $\mathfrak{F}_1\left\{\mathcal{P}f\right\} = \mathfrak{F}_2\left\{f(x,y)\right\}$. Thus, combining (A.7) and (A.9)

we find that

$$F_b(\rho,\phi) = \frac{w_\phi}{\rho} B^2(\phi) F(\rho,\phi), \tag{A.10}$$

which implies

$$f_b(x,y) = f(x,y) ** \frac{w_\phi + \frac{\pi}{2}}{r} ** \mathfrak{H}^{-1}\left\{B^2(\rho)\right\}. \tag{A.11}$$

Thus, the blur due to a weighted projection-backprojection with radially constant blur is shift-invariant. However, note that this blur is anisotropic when w_ϕ is nonuniform.

APPENDIX B

Resolution Properties of Filtered Backprojection

Adopting the idealized continuous model in Section 2.2.1, one can show that the resolution properties of filtered backprojection (FBP) are space-invariant. Given ideal projections with no detector blur, $g_\phi(r)$, the windowed FBP reconstruction with window $W(\rho)$ is written:

$$\begin{aligned} \tilde{f}(x_1, x_2) &= \mathcal{P}'\mathfrak{F}_1^{-1}\left\{\mathfrak{F}_1\left\{g_\phi(r)\right\}|u|W(u)\right\} && \text{(B.1)} \\ &= \int_0^\pi \tilde{q}_\phi(x_1\cos\phi + x_2\sin\phi)d\phi && \text{(B.2)} \end{aligned}$$

with

$$\tilde{q}_\phi(\rho) = \int G_\phi(\rho)|\rho|W(\rho)e^{i2\pi r\rho}d\rho, \qquad \text{(B.3)}$$

where $G_\phi(u)$ denotes the 1D Fourier transform of $g_\phi(r)$. The 2D Fourier transform of $\tilde{f}(x_1, x_2)$ is $\tilde{F}(u_1, u_2)$, which may be written as

$$\begin{aligned} & \iint \left(\int_0^\pi \int G_\phi(\rho)|\rho|W(\rho)e^{i2\pi\rho(x_1\cos\phi + x_2\sin\phi)}d\rho d\phi\right) e^{-i2\pi(u_1x_1+u_2x_2)}dx_1dx_2 \\ = & \iint \left(\int_0^\pi \int F(\rho,\phi)|\rho|W(\rho)e^{i2\pi\rho(x_1\cos\phi + x_2\sin\phi)}d\rho d\phi\right) e^{-i2\pi(u_1x_1+u_2x_2)}dx_1dx_2 \\ = & \iint \left(\iint F(v_1,v_2)W(\sqrt{v_1^2+v_2^2})e^{i2\pi(x_1v_1+x_2v_2)}dv_1dv_2\right) e^{-i2\pi(u_1x_1+u_2x_2)}dx_1dx_2 \\ = & \iint F(v_1,v_2)W(\sqrt{v_1^2+v_2^2})\left(\iint e^{i2\pi(x_1v_1+x_2v_2)}e^{-i2\pi(u_1x_1+u_2x_2)}dx_1dx_2\right)dv_1dv_2 \\ = & \iint F(v_1,v_2)W(\sqrt{v_1^2+v_2^2})\delta(u_1-v_1,u_2-v_2)dv_1dv_2 \\ = & F(u_1,u_2)W(\sqrt{u_1^2+u_2^2}), \qquad \text{(B.4)} \end{aligned}$$

where F denotes the Fourier transform of the true image. Thus,

$$\tilde{F}(\rho,\phi) = F(\rho,\phi)W(\rho) \tag{B.5}$$

$$\Longrightarrow \quad \tilde{f}(x_1,x_2) = f(x_1,x_2) * * w(r). \tag{B.6}$$

The first step follows from the central section theorem ($G_\phi(\rho) = F(\rho,\phi)$). The resulting impulse response $w(r)$ in (B.6) can be found by obtaining the inverse Hankel transform of $W(\rho)$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] The Whole Frog Project. Lawrence Berkeley National Laboratory, Office of Science, U.S. Department of Energy. [Online]. Available: `http://www.itg.lbl.gov/Frog`.

[2] The Zubal Phantom. Image Processing and Analysis Group, Dept. of Diagnostic Radiology and Electrical Engineering and the Yale School of Medicine, Yale University. [Online]. Available: `http://noodle.med.yale.edu/zubal/`.

[3] S Ahn and J A Fessler. Globally convergent ordered subsets algorithms: Application to tomography. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 2001.

[4] S Alenius and U Ruotsalainen. Bayesian image reconstruction for emission tomography based on median root prior. *Eur. J. Nuc. Med.*, 24(3):258–65, 1997.

[5] C R Appledorn. An analytical solution to the nonstationary reconstruction problem in single photon emission computed tomography. In D A Ortendahl and J Llacer, editors, *Info. Proc. in Med. Im.*, pages 69–79. Liss, NY, 1991.

[6] R D Badawi, M A Lodge, and P K Marsden. Algorithms for calculating detector efficiency normalization coefficients for true coincidences in 3D PET. *Phys. Med. Biol.*, 43(1):189–205, January 1998.

[7] J R Baker, T F Budinger, and R H Huesman. Generalized approach to inverse problems in tomography: Image reconstruction for spatially variant systems using natural pixels. *Crit. Rev. Biomed. Eng.*, 20:47–71, 1992.

[8] H H Barrett, D W Wilson, and B M W Tsui. Noise properties of the EM algorithm: I. Theory. *Phys. Med. Biol.*, 39:833–846, 1994.

[9] S Bellini, M Piacentini, C Cafforio, and F Rocca. Compensation of tissue absorption in emission tomography. *IEEE Tr. Acoust. Sp. Sig. Proc.*, 27(3):213–8, June 1979.

[10] A Blake and A Zisserman. *Visual reconstruction*. MIT Press, Cambridge, MA, 1987.

[11] Paola Bonetto, Jinyi Qi, and Richard M Leahy. Covariance approximation for fast and accurate computation of channelized Hotelling observer statistics. *IEEE Tr. Nuc. Sci.*, 47(4):1567–72, August 2000.

[12] T Bortfield and U Oelfke. Fast and exact 2D image reconstruction by means of Chebyshev decomposition and backprojection. *Phys. Med. Biol.*, 44(4):1105–20, April 1999.

[13] D Boulfelfel, R M Rangayyan, L J Hahn, R Kloiber, and G R Kuduvalli. Two-dimensional restoration of single photon emission computed tomography images using the Kalman filter. *IEEE Tr. Med. Im.*, 13(1):102–9, March 1994.

[14] C Bouman and K Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Tr. Im. Proc.*, 2(3):296–310, July 1993.

[15] C A Bouman and K Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Tr. Im. Proc.*, 5(3):480–92, March 1996.

[16] J E Bowsher, V E Johnson, T G Turkington, R J Jaszczak, C E Floyd, and R E Coleman. Bayesian reconstruction and use of anatomical a priori information for emission tomography. *IEEE Tr. Med. Im.*, 15(5):673–86, October 1996.

[17] R Bracewell. *The Fourier transform and its applications*. McGraw-Hill, New York, 1978.

[18] R E Carson, Y Yan, B Chodkowski, T K Yap, and M E Daube-Witherspoon. Precision and accuracy of regional radioactivity quantitation using the maximum likelihood EM reconstruction algorithm. *IEEE Tr. Med. Im.*, 13(3):526–537, September 1994.

[19] L T Chang. A method for attenuation correction in radionuclide computed tomography. *IEEE Tr. Nuc. Sci.*, 25(1):638–643, February 1978.

[20] J G Colsher. Fully three dimensional positron emission tomography. *Phys. Med. Biol.*, 25:103–15, 1980.

[21] M E Daube-Witherspoon and G Muehllehner. Treatment of axial data in three-dimensional PET. *J. Nuc. Med.*, 28:1717–24, 1987.

[22] A R De Pierro. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Tr. Med. Im.*, 14(1):132–137, March 1995.

[23] A R De Pierro and M E B Yamagishi. Fast EM-like methods for maximum 'a posteriori' estimates in emission tomography. *IEEE Tr. Med. Im.*, 20(4):280–8, April 2001.

[24] M Defrise, P E Kinahan, D W Townsend, C Michel, M Sibomana, and D F Newport. Exact and approximate rebinning algorithms for 3-D PET data. *IEEE Tr. Med. Im.*, 16(2):145–58, April 1997.

[25] M Defrise, D W Townsend, and R Clack. Three-dimensional image reconstruction from complete projections. *Phys. Med. Biol.*, 34:571–87, May 1989.

[26] E V R Di Bella, A B Barclay, R L Eisner, and R W Schafer. A comparison of rotation-based methods for iterative reconstruction algorithms. *IEEE Tr. Nuc. Sci.*, 43(6):3370–6, December 1996.

[27] H Erdoğan and J A Fessler. Monotonic algorithms for transmission tomography. *IEEE Tr. Med. Im.*, 18(9):801–14, September 1999.

[28] H Erdoğan and J A Fessler. Ordered subsets algorithms for transmission tomography. *Phys. Med. Biol.*, 44(11):2835–51, November 1999.

[29] J A Fessler and H Erdoğan. A paraboloidal surrogates algorithm for convergent penalized-likelihood emission image reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1132–5, 1998.

[30] J A Fessler. Penalized weighted least-squares image reconstruction for positron emission tomography. *IEEE Tr. Med. Im.*, 13(2):290–300, June 1994.

[31] J A Fessler. Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans. *IEEE Tr. Im. Proc.*, 4(10):1439–50, October 1995.

[32] J A Fessler. Resolution properties of regularized image reconstruction methods. Technical Report 297, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, August 1995.

[33] J A Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Tr. Im. Proc.*, 5(3):493–506, March 1996.

[34] J A Fessler. Approximate variance images for penalized-likelihood image reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 949–52, 1997.

[35] J A Fessler and S D Booth. Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction. *IEEE Tr. Im. Proc.*, 8(5):688–99, May 1999.

[36] J A Fessler, N H Clinthorne, and W L Rogers. Regularized emission image reconstruction using imperfect side information. *IEEE Tr. Nuc. Sci.*, 39(5):1464–71, October 1992.

[37] J A Fessler and E P Ficaro. Fully 3D PET image reconstruction using a Fourier preconditioned conjugate-gradient algorithm. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1599–1602, 1996.

[38] J A Fessler and A O Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Tr. Sig. Proc.*, 42(10):2664–77, October 1994.

[39] J A Fessler and A O Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Tr. Im. Proc.*, 4(10):1417–29, October 1995.

[40] J A Fessler and A O Hero. Comparing estimator covariances at matched spatial resolutions for imaging system design. In *Proc. 1999 IEEE Information Theory Workshop on Detection, Estimation, Classification and Imaging (DECI)*, page 15, 1999.

[41] J A Fessler and W L Rogers. Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs. *IEEE Tr. Im. Proc.*, 5(9):1346–58, September 1996.

[42] D Gagnon, N Pouliot, L Laperriére, M Therrien, and P Oliver. Maximum likelihood positioning in the scintillation camera using depth of interaction. *IEEE Tr. Med. Im.*, 12(1):101–107, March 1993.

[43] H C Gifford, R G Wells, and M A King. A comparison of human observer LROC and numerical observer ROC for tumor detection in SPECT images. *IEEE Tr. Nuc. Sci.*, 46(4):1032–7, August 1999.

[44] G Glatting, M Wuchenauer, and S N Reske. Iterative reconstruction for attenuation correction in positron emission tomography: Maximum likelihood for transmission and blank scan. *Med. Phys.*, 26(9):1838–42, September 1999.

[45] S J Glick, M A King, T-S Pan, and E J Soares. An analytical approach for compensation of non-uniform attenuation in cardiac SPECT imaging. *Phys. Med. Biol.*, 40(10):1677–94, October 1995.

[46] S J Glick, B C Penney, and C L Byrne. A fast projector backprojector pair for use in iterative reconstruction of SPECT images. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, 1993.

[47] S J Glick, B C Penney, M A King, and C L Byrne. Noniterative compensation for the distance-dependent detector response and photon attenuation in SPECT. *IEEE Tr. Med. Im.*, 13(2):363–74, June 1994.

[48] R Gordon. A tutorial on ART (algebraic reconstruction techniques). *IEEE Tr. Nuc. Sci.*, 21:78–93, 1974.

[49] R Gordon, R Bender, and G T Herman. Algebraic reconstruction techniques (ART) for the three-dimensional electron microscopy and X-ray photography. *J. Theor. Biol.*, 29:471–81, 1970.

[50] T J Hebert. Statistical stopping criteria for iterative maximum likelihood reconstruction of emission images. *Phys. Med. Biol.*, 35(9):1221–32, 1990.

[51] T J Hebert and R Leahy. Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. *IEEE Tr. Sig. Proc.*, 40(9):2290–303, September 1992.

[52] G T Herman, A Lent, and S W Rowland. ART: mathematics and applications (a report on the mathematical foundations and on the applicability to real data of the algebraic reconstruction techniques). *J. Theor. Biol.*, 42:1–32, 1973.

[53] Flemming Hermansen, Terry J Spinks, Paolo G Camici, and Adriaan A Lammertsma. Calculation of single detector efficiencies and extension of the normalization sinogram in PET. *Phys. Med. Biol.*, 42(6):1143–54, June 1997.

[54] A O Hero, J A Fessler, and M Usman. Exploring estimator bias-variance tradeoffs using the uniform CR bound. *IEEE Tr. Sig. Proc.*, 44(8):2026–41, August 1996.

[55] E J Hoffman, S C Huang, M E Phelps, and D E Kuhl. Quantitation in positron emission computed tomography: 4 Effect of accidental coincidences. *J. Comp. Assisted Tomo.*, 5(3):391–400, 1981.

[56] T J Holmes, D L Snyder, and D C Ficke. A statistical analysis of count normalization methods used in positron-emission tomography. *IEEE Tr. Nuc. Sci.*, 31(1):521–525, February 1984.

[57] P J Huber. *Robust statistics*. Wiley, New York, 1981.

[58] H M Hudson and R S Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Tr. Med. Im.*, 13(4):601–9, December 1994.

[59] R J Jaszczak, K L Greer, C E Floyd, C C Harris, and R E Coleman. Improved SPECT quantification using compensation for scattered photons. *J. Nuc. Med.*, 25(8):893–900, 1984.

[60] V E Johnson. A model for segmentation and analysis of noisy images. *J. Am. Stat. Ass.*, 89(425):230–41, March 1994.

[61] Chien-Min Kao, Xiaochuan Pan, and Chin-Tu Chen. Accurate image reconstruction using DOI information and its implications for the development of compact PET systems. *IEEE Tr. Nuc. Sci.*, 47(4-2):1551–60, August 2000.

[62] B Karuta and R Lecomte. Effect of detector weighting functions on the point spread function of high-resolution PET tomographs. *IEEE Tr. Med. Im.*, 11(3):379–85, September 1992.

[63] P E Kinahan and J G Rogers. Analytic 3D image reconstruction using all detected events. *IEEE Tr. Nuc. Sci.*, 36(1):964–8, February 1989.

[64] Leonid A Kunyansky. A new SPECT reconstruction algorithm based on the Novikov explicit inversion formula. *Inverse Prob.*, 17(2):293–306, April 2001.

[65] P J La Riviere and X Pan. Nonparametric regression sinogram smoothing using a roughness-penalized poisson likelihood objective function. *IEEE Tr. Med. Im.*, 19(8):773–86, August 2000.

[66] S Lakshmanan and H Derin. Parameter space for Gaussian Markov random fields with separable autocorrelations. In *Proc. 15th Ann. Conf. Inf. Sci. Syst.*, pages 226–231, 1990.

[67] S Lakshmanan and H Derin. Valid parameter space for 2-D Gaussian Markov random fields. *IEEE Tr. Patt. Anal. Mach. Int.*, 39(2):703–9, March 1993.

[68] D S Lalush and B M W Tsui. Mean-variance analysis of block-iterative reconstruction algorithms modeling 3D detector response in SPECT. *IEEE Tr. Nuc. Sci.*, 45(3):1280–7, June 1998.

[69] K Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Tr. Med. Im.*, 9(4):439–46, December 1990. Corrections, 10:2(288), June 1991.

[70] K Lange and R Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomo.*, 8(2):306–16, April 1984.

[71] C L Lawson and R J Hanson. *Solving least squares problems*. Prentice-Hall, New Jersey, 1974.

[72] K J Leaf-Lensmire, C W Stearns, and J G Colsher. PET transmission scanning with a variable speed orbiting rod source. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 979–81, 1992.

[73] R Leahy and X H Yan. Statistical models and methods for PET image reconstruction. In *Proc. of Stat. Comp. Sect. of Amer. Stat. Assoc.*, pages 1–10, 1991.

[74] C S Levin and E J Hoffman. Calculation of positron range and its effect on the fundamental limit of positron emission tomography. *Phys. Med. Biol.*, 44(3):781–99, March 1999.

[75] R M Lewitt. Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys. Med. Biol.*, 37(3):705–16, 1992.

[76] R M Lewitt, P R Edholm, and W Xia. Fourier method for correction of depth-dependent collimator blurring. In *Proc. SPIE 1092, Med. Im. III: Im. Proc.*, pages 232–43, 1989.

[77] R M Lewitt, G Muehllehner, and J S Karp. Three-dimensional image reconstruction for PET by multi-slice rebinning and axial image filtering. *Phys. Med. Biol.*, 39(3):321–9, March 1994.

[78] J S Liow and S C Strother. The convergence of object dependent resolution in maximum likelihood based tomographic image reconstruction. *Phys. Med. Biol.*, 38(1):55–70, January 1993.

[79] C A Lowry and D N Taylor. Improvements in image contrast and quantification using scatter subtraction in SPECT. *Br. J. Radiol.*, 59:728, 1986.

[80] A Macovski. *Medical imaging systems*. Prentice-Hall, New Jersey, 1983.

[81] S Matej and R M Lewitt. Practical considerations for 3-D image reconstruction using spherically symmetric volume elements. *IEEE Tr. Med. Im.*, 15(1):68–78, February 1996.

[82] A W McCarthy and M I Miller. Maximum likelihood SPECT in clinical computation times using mesh-connected parallel processors. *IEEE Tr. Med. Im.*, 10(3):426–436, September 1991.

[83] C E Metz and X Pan. A unified analysis of exact methods of inverting the 2-D exponential Radon transform with implications for noise control in SPECT. *IEEE Tr. Med. Im.*, 14(4):643–58, December 1995.

[84] M I Miller and B Roysam. Bayesian image reconstruction for emission tomography incorporating Good's roughness prior on massively parallel processors. *Proc. Natl. Acad. Sci.*, 88:3223–3227, April 1991.

[85] T R Miller and J W Wallis. Clinically important characteristics of maximum-likelihood reconstruction. *J. Nuc. Med.*, 33(9):1678–84, September 1992.

[86] W W Moses, S E Derenzo, C L Melcher, and R A Manente. A room temperature LSO/PIN photodiode PET detector module that measures depth of interaction. *IEEE Tr. Nuc. Sci.*, 42(4):1085–9, August 1995.

[87] E Ü Mumcuoğlu, R M Leahy, and S R Cherry. Bayesian reconstruction of PET images: methodology and performance analysis. *Phys. Med. Biol.*, 41(9):1777–1807, September 1996.

[88] S Mustafovic, K Thielemans, D Hogg, and P Bloomfield. Object dependency of resolution and convergence rate in OSEM with filtering. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, pages 1786–90, 2001.

[89] S H Nash and A Sofer. *Linear and nonlinear programming*. McGraw-Hill, New York, 1996.

[90] M Nikolova, J Idier, and A Mohammad-Djafari. Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF. *IEEE Tr. Im. Proc.*, 7(4):571–85, April 1998.

[91] J Nunez and J Llacer. Variable resolution Bayesian image reconstruction. In *Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, 1995.

[92] J M Ollinger. Model-based scatter correction for fully 3D PET. *Phys. Med. Biol.*, 41(1):153–76, January 1996.

[93] P H Pretorius, M A King, S J Glick, T S Pan, and D S Luo. Reducing the effect of nonstationary resolution on activity quantitation with the frequency distance relationship in SPECT. *IEEE Tr. Nuc. Sci.*, 43(6):3335–41, December 1996.

[94] J Qi and R M Leahy. Fast computation of the covariance of MAP reconstructions of PET images. In *Proc. SPIE 3661, Med. Im. I: Im. Proc.*, pages 344–55, February 1999.

[95] J Qi and R M Leahy. A theoretical study of the contrast recovery and variance of MAP reconstructions with applications to the selection of smoothing parameters. *IEEE Tr. Med. Im.*, 18(4):293–305, April 1999.

[96] Jinyi Qi and Richard M Leahy. Resolution and noise properties MAP reconstruction for fully 3D PET. *IEEE Tr. Med. Im.*, 19(5):493–506, May 2000.

[97] J Radon. On the determination of functions from their integrals along certain manifold. *Berichte Sächs. Akad. Wiss. (Leipzig)*, 69:262–78, 1917. Uber die Bestimmung von Funktionen durch ihre Intergralwerte Langs gewisser Manningfultigkeiten.

[98] S J Reeves. Optimal space-varying regularization in iterative image restoration. *IEEE Tr. Im. Proc.*, 3(3):319–23, May 1994.

[99] H W Reist, O Stadelmann, and W Kleeb. Study on the stability of the calibration and normalization in PET and the influence of drifts on the accuracy of quantification. *Eur. J. Nuc. Med.*, 15:732–735, 1989.

[100] J G Rogers, R Harrop, and P E Kinahan. The theory of three-dimensional image reconstruction for PET. *IEEE Tr. Med. Im.*, 6(3):239–43, September 1987.

[101] S M Rump. Bounds for the componentwise distance to the nearest singular matrix. *SIAM J. Matrix Anal. Appl.*, 18(1):83–103, January 1997.

[102] V V Selivanov, Y Picard, J Cadorette, S Rodrigue, and R Lecomte. Detector response models for statistical iterative image reconstruction in high resolution PET. *IEEE Tr. Nuc. Sci.*, 47(3):1168–75, June 2000.

[103] B W Silverman, C Jennison, J Stander, and T C Brown. The specification of edge penalties for regular and irregular pixel images. *IEEE Tr. Patt. Anal. Mach. Int.*, 12(10):1017–24, October 1990.

[104] D L Snyder and M I Miller. The use of sieves to stabilize images produced with the EM algorithm for emission tomography. *IEEE Tr. Nuc. Sci.*, 32(5):3864–71, October 1985.

[105] D L Snyder, M I Miller, L J Thomas, and D G Politte. Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography. *IEEE Tr. Med. Im.*, 6(3):228–38, September 1987.

[106] E J Soares, C L Byrne, and S J Glick. Noise characterization of block-iterative reconstruction algorithms. I. Theory. *IEEE Tr. Med. Im.*, 19(4):261–70, April 2000.

[107] E J Soares, C L Byrne, S J Glick, C R Appledorn, and M A King. Implementation and evaluation of an analytical solution to the photon attenuation and nonstationary resolution reconstruction problem in SPECT. *IEEE Tr. Nuc. Sci.*, 40(4):1231–1237, August 1993.

[108] James A Sorenson and Michael E Phelps. *Physics in nuclear medicine*. Saunders, Philadelphia, 2 edition, 1987.

[109] Saowapak Sotthivirat and J A Fessler. Image recovery using partitioned-separable paraboloidal surrogate coordinate ascent algorithms. *IEEE Tr. Im. Proc.*, 11(3):306–17, March 2002.

[110] P Spellucci. A new technique for inconsistent problems in the SQP method. *Math. Meth. of Oper. Res.*, 47:355–400, 1998.

[111] P Spellucci. An SQP method for general nonlinear programs using only equality constrained subproblems. *Math. Prog.*, 82:413–48, 1998.

[112] J A Stamos, W L Rogers, N H Clinthorne, and K F Koral. Object-dependent performance comparison of two iterative reconstruction algorithms. *IEEE Tr. Nuc. Sci.*, 35(1):611–614, February 1988.

[113] H Stark and J W Woods. *Probability, random processes, and estimation theory for engineers*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[114] J W Stayman and J A Fessler. Spatially-variant roughness penalty design for uniform resolution in penalized-likelihood image reconstruction. In *Proc. IEEE Intl. Conf. on Image Processing*, volume 2, pages 685–9, 1998.

[115] J W Stayman and J A Fessler. Penalty design for uniform spatial resolution in 3D penalized-likelihood image reconstruction. In *Proc. of the 1999 Intl. Mtg. on Fully 3D Im. Recon. in Rad. Nuc. Med.*, 1999.

[116] J W Stayman and J A Fessler. Regularization for uniform spatial resolution properties in penalized-likelihood image reconstruction. *IEEE Tr. Med. Im.*, 19(6):601–15, June 2000.

[117] J W Stayman and J A Fessler. Nonnegative definite quadratic penalty design for penalized-likelihood reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 2001.

[118] J W Stayman and J A Fessler. Compensation for nonuniform resolution using penalized-likelihood reconstruction in space-variant imaging systems. *IEEE Tr. Med. Im.*, 2002. accepted.

[119] J W Stayman and J A Fessler. Efficient calculation of resolution and covariance for fully 3D SPECT. *IEEE Tr. Med. Im.*, 2002. submitted.

[120] J W Stayman and J A Fessler. Fast methods for approximation of resolution and covariance for spect. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 2002.

[121] R G Swensson. Unified measurement of observer performance in detecting and localizing target objects on images. *Med. Phys.*, 23(10):1709–25, October 1996.

[122] M M Ter-Pogossian, M E Raichle, and B E Sobel. Positron-emission tomography. *Scientific American*, 243:171–181, October 1980.

[123] A Tikhonov and V Arsenin. *Solution of ill-posed problems*. Wiley, New York, 1977.

[124] M Usman, A O Hero, and J A Fessler. Bias-variance tradeoffs analysis using uniform CR bound for image reconstruction. In *Proc. IEEE Intl. Conf. on Image Processing*, volume 2, pages 835–839, 1994.

[125] M Usman, A O Hero, and J A Fessler. Uniform CR bound: implementation issues and applications. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1443–1447, 1994.

[126] L van Elmbt and S Walrand. Simultaneous correction of attenuation and distance-dependent resolution in SPECT: an analytical approach. *Phys. Med. Biol.*, 38(9):1207–17, September 1993.

[127] E Veklerov and J Llacer. Stopping rule for the MLE algorithm based on statistical hypothesis testing. *IEEE Tr. Med. Im.*, 6(4):313–9, December 1987.

[128] P R G Virador, W W Moses, and R H Huesman. Reconstruction in PET cameras with irregular sampling and depth of interaction capability. *IEEE Tr. Nuc. Sci.*, pages 1225–1230, June 1998.

[129] W Wang and G Gindi. Noise analysis of MAP-EM algorithms for emission tomography. *Phys. Med. Biol.*, 42(11):2215–32, November 1997.

[130] C C Watson, D Newport, M E Casey, R A de Kemp, R S Beanlands, and M Schmand. Evaluation of simulation-based scatter correction for 3-D PET cardiac imaging. *IEEE Tr. Nuc. Sci.*, 44(1):90–7, February 1997.

[131] A Welch and G T Gullberg. Implementation of a model-based nonuniform scatter correction scheme for SPECT. *IEEE Tr. Med. Im.*, 16(6):717–26, December 1997.

[132] D W Wilson and B M W Tsui. Noise properties of filtered-backprojection and ML-EM reconstructed emission tomographic images. *IEEE Tr. Nuc. Sci.*, 40(4):1198–1203, August 1993.

[133] D W Wilson and B M W Tsui. Spatial resolution properties of FB and ML-EM reconstruction methods. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1189–1193, 1993.

[134] D W Wilson, B M W Tsui, and H H Barrett. Noise properties of the EM algorithm: II. Monte Carlo simulations. *Phys. Med. Biol.*, 39:847–872, 1994.

[135] W Xia, R M Lewitt, and P R Edholm. Fourier correction for spatially variant collimator blurring in SPECT. *IEEE Tr. Med. Im.*, 14(1):100–15, March 1995.

[136] Y Xing and G Gindi. Rapid calculation of detectability in Bayesian SPECT. In *Proc. IEEE Intl. Symp. Biomedical Imaging*, pages 78–81, July 2002.

[137] Y Xing, I T Hsiao, and G R Gindi. Efficient calculation of resolution and variance in 2D circular-orbit SPECT. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, pages M5B–5, 2001.

[138] B Xu, X Pan, and C T Chen. An innovative method to compensate for distance-dependent blurring in 2D SPECT. *IEEE Tr. Nuc. Sci.*, 45(4):2245–52, August 1998.

[139] J Yao and H H Barrett. Predicting human performance by a channelized Hotelling observer model. *Math. Meth. in Med. Im., SPIE*, 1768:161–8, 1992.

[140] M Yavuz and J A Fessler. Objective functions for tomographic reconstruction from randoms-precorrected PET scans. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1067–71, 1996.

[141] M Yavuz and J A Fessler. Penalized-likelihood estimators and noise analysis for randoms-precorrected PET transmission scans. *IEEE Tr. Med. Im.*, 18(8):665–74, August 1999.

[142] D F Yu and J A Fessler. Mean and variance of photon counting with deadtime. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1470–4, 1999.

[143] R E Ziemer, W H Tranter, and D R Fannin. *Signals and systems: continuous and discrete*. Prentice-Hall, New Jersey, 1998.

[144] G Zubal, G Gindi, M Lee, C Harrell, and E Smith. High resolution anthropomorphic phantom for Monte Carlo analysis of internal radiation sources. In *IEEE Symposium on Computer-Based Medical Systems*, pages 540–7, 1990.

ABSTRACT

Spatial Resolution in Penalized-Likelihood Image Reconstruction

by

Joseph Webster Stayman

Chair: Jeffrey A. Fessler

Penalized-likelihood methods have been used widely in image reconstruction since they can model both the imaging system geometry and measurement noise very well. However, images reconstructed by conventional penalized-likelihood methods are subject to anisotropic and shift-variant spatial resolution properties, which can complicate selection of the regularization parameter and make the analysis of the resulting images more difficult. The local impulse response is a resolution predictor that may be used to quantify these shift-variant spatial resolution properties. We have derived a new formulation of the local impulse response for penalized-likelihood estimators. This formulation is appropriate for a general class of imaging systems that acquire a finite number of measurements from a continuous object and reconstruct that object using a discrete model. We have developed fast techniques for evaluating both spatial resolution and covariance predictors for emission tomography systems even when the geometric system model is inherently shift-variant. We

have also developed practical methods based on these rapid predictions to provide increased resolution control by designing an appropriate penalty function. The penalty function design allows for the specification of user-defined resolution properties like uniform resolution (i.e., both isotropic and shift-invariant). We show that these penalty design techniques can provide nearly uniform resolution even in intrinsically shift-variant imaging systems; whereas many traditional reconstruction techniques cannot fully compensate for the shift-variant effects. We discuss the relative resolution uniformity of different reconstruction methods and examine the relative noise performance of estimators for which the resolution properties are exactly matched. Among these matched estimators we find that the penalized-likelihood approach and the post-filtered maximum-likelihood approach often produce identical noise properties, and both provide reduced noise relative to classical filtered backprojection.