MATHEMATICAL METHODS FOR ENHANCED INFORMATION SECURITY IN TREATY VERIFICATION

by

Christopher J. MacGahan

BY: =

A Dissertation Submitted to the Faculty of the

COLLEGE OF OPTICAL SCIENCES

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Christopher J. MacGahan
entitled Mathematical Methods for Enhanced Information Security in Treaty Verification
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 12 May 2016
Matthew A. Kupinski

Date: 12 May 2016
Eric W. Clarkson

Date: 12 May 2016
Amit Ashok

Date: 12 May 2016
Erik M. Brubaker

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 12 May 2016
Dissertation Director: Matthew A. Kupinski

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. This work is licensed under the Creative Commons Attribution-No Derivative Works 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

SIGNED: Christopher J. MacGahan

ACKNOWLEDGEMENTS

I want to thank my wife Krista for her love, support and patience through the last 7 years. My parents, Allison and Jon, and my brother, Sean, have always been there whenever I needed them, even though they are 2,000 miles away. Mary, Eddie, Jay and Tyler Irwin—my "Tucson family"—thank you for all of the encouragement, and especially over the last 2 years, a home. The weekly family dinners were always a source of comfort for me, even while seemingly living in three locations at once. This accomplishment is a credit to all of you, whether you helped to foster my love for math and science or supported me when times were tough.

I would like to thank my advisor, Dr. Kupinski, for all of the help and guidance he has given me over the years. This project was certainly a different and challenging one, and his insight was always appreciated and welcome. When funding was limited in my first few years, he was on the lookout for alternative sources and understanding when these commitments ate into my time. I was supported by the Graduate Students and Teachers Engaging in Mathematical Sciences (G-TEAMS) fellowship through 2011-2012 and the Technology Research Initiative Fund in the 2013-2014 academic year. Without this funding, it would have been difficult to complete my studies. I also want to acknowledge the role senior graduate students Jae Hoon Lee, Enrique Montano, and Abhinav Jha have played in my time at Arizona. I appreciated the time to learn from all of you.

Finally, I want to acknowledge my close friends—Phillip Poon, Ricky Gibson, Michael Gehl, and Kristi Behnke—for the interesting discussions (science or not), late nights and all of the memories.

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

LIST OF FIGURES

LIST OF FIGURES – *Continued*

LIST OF FIGURES – *Continued*

LIST OF TABLES

ABSTRACT

Mathematical methods have been developed to perform arms-control-treaty verification tasks for enhanced information security. The purpose of these methods is to verify and classify inspected items while shielding the monitoring party from confidential aspects of the objects that the host country does not wish to reveal. Advanced medical-imaging methods used for detection and classification tasks have been adapted for list-mode processing, useful for discriminating projection data without aggregating sensitive information. These models make decisions off of varying amounts of stored information, and their task performance scales with that information.

Development has focused on the Bayesian ideal observer, which assumes complete probabilistic knowledge of the detector data, and Hotelling observer, which assumes a multivariate Gaussian distribution on the detector data. The models can effectively discriminate sources in the presence of nuisance parameters. The channelized Hotelling observer has proven particularly useful in that quality performance can be achieved while reducing the size of the projection data set. The inclusion of additional penalty terms into the channelizing-matrix optimization offers a great benefit for treaty-verification tasks. Penalty terms can be used to generate non-sensitive channels or to penalize the model's ability to discriminate objects based on confidential information. The end result is a mathematical model that could be shared openly with the monitor. Similarly, observers based on the likelihood probabilities have been developed to perform null-hypothesis tasks.

To test these models, neutron and gamma-ray data was simulated with the GEANT4 toolkit. Tasks were performed on various uranium and plutonium inspection objects. A fast-neutron coded-aperture detector was simulated to image the particles.

CHAPTER 1

A History of Nuclear Nonproliferation and an Approach to Arms-Control-Treaty-Verification Tasks

This chapter serves as an introduction to the reader, giving all of the background information necessary to understand the method development, which is the novel contribution of this thesis. First, in Section 1.1, a history of nuclear weapons is summarized, giving the motivation for the methods developed in this thesis. Next, an introduction to image science gives a description of each of the elements of the imaging equation (Section 1.2). This section also introduces the concept of task-based assessment and stresses the importance that statistics play in imaging. Section 1.3 features a discussion on the relevant tasks to arms-control-treaty verification, such as null-hypothesis and binary-classification tasks. Some background is given on these tasks as well as the different possible applications relevant to the nuclear-security mission.

The significant contribution of this work to the subject of treaty verification is the development of mathematical methods that can perform different tasks without requiring an information barrier (IB). An IB is a software or hardware barrier that prevents the dissemination of sensitive information on treaty-accountable items (TAIs) to unauthorized individuals. Further information on IBs as well as an explanation of the different approaches to treaty verification can be found in Section 1.5. Section 1.4 describes how similar models have been applied to medical imaging problems in the past. Section 1.6 contains a discussion on the approach this thesis takes to the development of models that process list-mode (LM) events and make decisions based on non-sensitive data.

1.1 A History of Nuclear Weapons and Nuclear Nonproliferation

This section explores a brief history of the development of nuclear weapons—from their development and first use in 1945 to the buildup of the arms race (Section 1.1.1). This culminated in the Cuban Missile Crisis (Section 1.1.2), which is considered the height of the Cold War. A summary is presented of important treaties signed between the United States (U.S.), Soviet Union (U.S.S.R.) and other

nuclear and non-nuclear states over the past several decades in Section 1.1.3. To verify compliance with future treaties, new technologies will be needed. The challenges associated with these different problems are summarized in Section 1.1.4. More details, and summaries of interesting historical events, can be found in Charles Loeber's book (Loeber, 2005).

While I have attempted to partition history into three periods—the arms race and escalation of political conflict between the U.S. and U.S.S.R., the peak of the Cold War in 1962, and the movement towards nonproliferation—reality is not as clean cut (United States Department of State, 2009). The Vietnam war was fought in the 1960s between North and South Vietnam, but heavily backed in money and manpower by the U.S.S.R. and U.S., respectively. In 1968, the U.S.S.R. invaded Czechoslovakia and in 1979 Afghanistan. However, none of these events carried significant risk of nuclear war, and the two sides became more reserved in their tactics after the Cuban Missile Crisis, assisting other countries rather than directly confronting each other.

1.1.1 From the Dropping of the Bombs to the Start of the Cold War

In 1945 the U.S. dropped two nuclear bombs on Japan; one was dropped on August 6 on Hiroshima, killing 140,000 people within a year (Gibson, 1997), and the second was dropped on August 9, killing 70,000 within a year in Nagasaki. These acts hastened the end of World War II as Japan surrendered less than a week later. At that time, the U.S. only had one remaining warhead in their stockpile and was the only nation in the world with the technology. Shortly after Japan's surrender, optimism about the future of nuclear energy was at a high. The Atomic Energy Act (Int, 1946) was signed in 1946 and transferred development and construction of the weapons to the Atomic Energy Commission, which was led by civilians. This hope for a peaceful future is best emphasized by the following quote in the act:

> It is hereby declared to be the policy of the people of the United States that, subject at all times to the paramount objective of assuring the common defense and security, the development and utilization of atomic energy shall, so far as practicable, be directed towards improving the public welfare, increasing the standard of living, strengthening free competition in private enterprise, and promoting world peace.

History, of course, turned out differently. Through a dedicated effort, including the

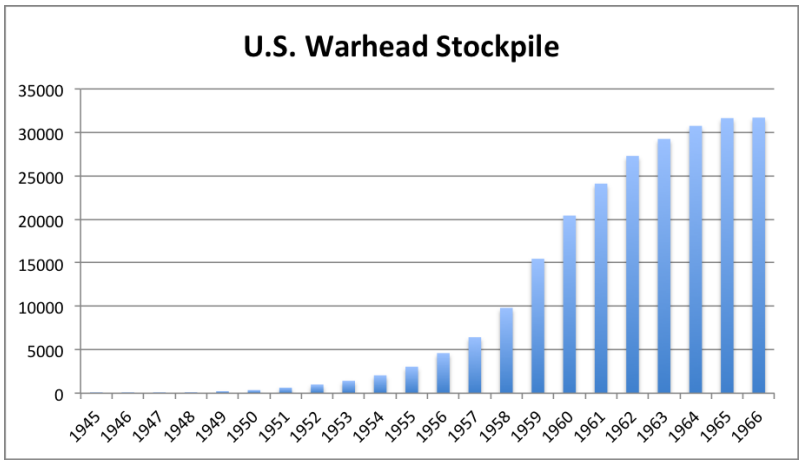


Figure 1.1: The U.S. stockpile increased dramatically from 1946 up to 1966, with peak production occuring in 1957 through 1961

work of spies in the Manhattan Project (Groves, 1983) such as Klaus Fuchs and Theodore Hall (Haynes and Vassiliev, 2009; Williams, 1989), the U.S.S.R. developed an atomic bomb and tested it in 1949. This began the Cold War—the decades long conflict between the U.S. and U.S.S.R. that was mostly fought through proxy wars. This state of tension caused the arms race to explode in an exponential manner (National Resources Defense Council, 2002). The growth of the U.S. stockpile over the next two decades is shown in Figure 1.1.

The arms race was not solely limited to the production of more weapons between the U.S. and the U.S.S.R. The United Kingdom (U.K.) developed its first weapon in 1952, France in 1960, and China in 1964 (Gibson, 1997). In addition to an increase in nuclear proliferation, the arms race also caused a dramatic increase in yield. The Fat Man bomb dropped on Japan had a yield of 21kT. In the late 1940s, significant investment took place in the development of thermonuclear fusion bombs. In May of 1951, George, with a yield of 225 kT, was tested around the Marshall Islands. In October 1952, the first staged-thermonuclear device was tested, with a yield of 10.4Mt.

1.1.2 The Cuban Missile Crisis

The height of the Cold War was arguably the Cuban Missile Crisis. Fidel Castro's communist regime gained power through a revolution that ended in January, 1959 (Pérez-Stable, 1999). Castro regularly made anti-American remarks, and fear of a nearby government politically similar to the U.S.S.R. led the U.S. to cancel sugar and oil imports, beginning a dramatic escalation of tension between the two

countries. Cuba sought other trade partners, turning to Nikita Khrushchev and the U.S.S.R. to fulfill their economic needs, raising the ire of the U.S. In response to Cuban policies such as the nationalization of American-owned Cuban oil refineries, President Eisenhower levied an embargo on exports to Cuba (Young, 1960) on October 19, 1960. In 1962, this was further extended to cover almost all imports.

The U.S. attempted to overthrow Castro in April of 1961 (Gilman, 2004). The CIA trained roughly 1,500 rebels in Florida, Guatemala and other countries countries beginning in 1959. It was intended to be a covert operation, appearing to be an independent rebellion, but by the time the invasion occurred, the plan had been well publicized. The *New York Times* reported on the U.S.'s plans before the operation was underway. The attempt to overthrow Castro failed, serving only to further cement Castro's power, as the Cuban public was outraged. Furthermore, the U.S.S.R. used the crisis as a political opportunity, obtaining the Cuban government's agreement to place ballistic-missile launch sites on Cuban soil.

The Cuban Missile Crisis covered a 13 day period in October, 1962 (Gale, 2008). On the morning of October 14, 1962, the US received intelligence that Soviet ballistic missiles were stationed in Cuba, some of which were believed to have a range of 2,200 miles. Rather than strike the launch sites or invade Cuba, President Kennedy opted to create a blockade, quarantining all offensive military equipment being shipped to Cuba. On October 24th, the quarantine went into effect, and while most Soviet ships reversed course, three still moved toward the quarantine line. These moments created the most tense moment of the Cold War. Shortly before the U.S. would have been forced to act, the Russian ships turned around. By October 25th, some of the missiles were believed to be operational, and the U.S. considered offensive action to remove them. On the 26th, the Soviets offered to dismantle their missiles in exchange for the US guaranteeing they would not invade Cuba, and after debate, the U.S. agreed. The missiles were dismantled on October 28th, ending the episode that was the closest the world has ever came to nuclear war.

1.1.2.1 Positive Impacts of the Cuban Missile Crisis

The impact this event had on U.S. and Soviet relations was largely positive. It led to the institution of the "hot line", allowing direct communication between the leaders of the two countries. It also served as an impetus for the Test-Ban Treaty. Perhaps the most interesting perspectives were from the two men at the heart of the incident. In November of 1962, shortly after the end of Cuban Missile Crisis, Kruschev said,

> They talk about who won and who lost. Human reason won. Mankind won.

Similarly, Kennedy, speaking at an American University Commencement address in 1963 (Kennedy, 1963) stated,

> ...even in the cold war—which brings burdens and dangers to so many countries, including this nationś closest allies—our two countries bear the heaviest burdens. For we are both devoting massive sums of money to weapons that could better be devoted to combat ignorance, poverty, and disease.
>
> We are both caught up in a vicious and dangerous cycle with suspicion on one side breeding suspicion on the other, and new weapons begetting counter-weapons.
>
> In short, both the United States and its allies, and the Soviet Union and its allies, have a mutually deep interest in a just and genuine peace and in halting the arms race. Agreements to this end are in the interests of the Soviet Union as well as ours—and even the most hostile nations can be relied upon to accept and keep those treaty obligations and only those treaty obligations which are in their own interest.

This speech was not only remarkable for its message of peace just eight months after the world was on the brink of nuclear war, but for an important policy proposal as well. Negotiations on a test-ban treaty had begun in 1958, four years before the Cuban Missile Crisis, between the U.S., U.S.S.R. and U.K. However, there were still disagreements over verification of compliance with the treaty. The U.S. and U.K. wanted the capability to inspect Soviet missile sites in the event that an explosion was detected inside the U.S.S.R.'s borders, a stipulation that the U.S.S.R. did not acquiesce to (Office of the Historian, 1963). In his commencement speech, Kennedy also included a policy declaration, leading to the formation of the Limited Test Ban Treaty:

> I now declare that the United States does not propose to conduct nuclear tests in the atmosphere so long as other states do not do so. We will not be the first to resume. Such a declaration is no substitute for a formal binding treaty—but I hope it will help us achieve one. Nor would such

> a treaty be a substitute for disarmament—but I hope it will help us achieve it.

1.1.3 International Progress Towards Nuclear Nonproliferation and Arms Reduction

This section summarizes the various treaties that have been signed into force.

1.1.3.1 Limited Test Ban Treaty

The Limited Test Ban Treaty (United States and Union of Soviet Socialist Republics, 1963) was signed in 1963. After years of tests over land and at sea and concerns over nuclear fallout from those tests, the U.S., U.S.S.R. and U.K. agreed on a treaty to ban tests that caused fallout over adjacent countries, essentially forcing all future tests to be conducted underground. Verification for this treaty was done by national means, using radionuclide testing of the atmospheric particles. To detect a possible explosion, the concentration of decay chain products such as Americium 241, Cesium 137, Iodine 131 and Strontium 90 were measured (CTBTO Preparatory Commission, 2012). The test ban was expanded upon in 1974, when the countries signed the Threshold Test Ban Treaty, preventing detonation of any nuclear devices over 150kT underground. Though this treaty was signed in 1974, it did not enter into force until 1990 (United States and Union of Soviet Socialist Republics, 1974).

1.1.3.2 Nuclear Non-Proliferation Treaty

The incredible size of the U.S. and U.S.S.R. stockpile certainly served as an impetus for disarmament treaties between the countries. While the exact cost is not known, the Brookings Institute (Schwartz, 2011) has arrived at a minimum estimate (shown in Figure 1.2). They found that by 1996, the U.S. had spent $409 billion in weapon development, $3.2 trillion in deployment costs, $830 billion in targeting and control costs, and $937 billion on nuclear defense (costs in 1996 dollars). Due in part to the above costs, as well as the existential threat that nuclear weapons posed to humanity, there was momentum in the late 1960s for an international nonproliferation treaty. In 1968, the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) was signed (United Nations Office of Disarmament Affairs, 1968) by the five nuclear states—the U.S., the U.S.S.R., the U.K., France, and China. There were three central tenets behind this treaty:

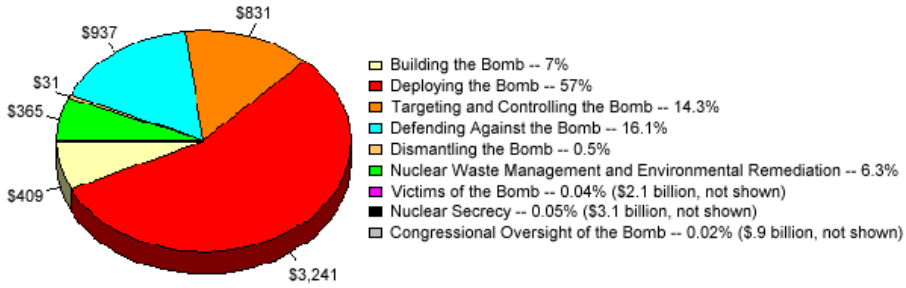


Figure 1.2: Pie chart of the estimated cost to keep the U.S. stockpile up to date between 1940 and 1996 in 1996 U.S. dollars (Schwartz, 2011).

1. Non-Proliferation of Weapons: Nuclear-weapon states were not to share their weapons or knowledge of construction with non-nuclear-weapon states, preventing the proliferation of knowledge required to construct weapons. (Article I). Non-nuclear weapon states were not to receive or seek out information on building nuclear weapons (Article II).

2. Disarmament: Nuclear weapon states were to make an effort to stop the arms race and negotiate towards disarmament (Article VI)

3. Spread of Knowledge for Peaceful Purposes: Non-nuclear states were to submit to standards set by the International Atomic Energy Agency (IAEA). In exchange, the weapon states were to share knowledge about any peaceful uses of nuclear energy (Article V).

The NPT has always been a political balancing act, and is a grand bargain of sorts between the weapon and non-weapon states. Over the past 25 years, significant progress has been made on these fronts. However, non-weapon states argue that disarmament has not gone far enough considering the sacrifice they are making in foregoing a critical means of defense (Bird, 2015). Australia, for example, expressed their disappointment in a 2015 statement to the United Nations:

> ...a growing number of states have expressed frustration at the slow pace of nuclear disarmament. These have led for some to call for a treaty banning nuclear weapons...
>
> ...Australia, like many States, is concerned that 45 years since the NPT entered into force some 16,000 warheads still exist. We (Australia) remain concerned that some states will continue to produce weapons-grade

> uranium and plutonium, although we acknowledge that a number of nuclear weapon states have declared moratoria on the production of fissile material. We are concerned that some States are developing new, small, battlefield scale nuclear weapons.

1.1.3.3 Various Disarmament Treaties

In the initial nonproliferation agreements, nuclear disarmament focused on the number of delivery systems, as these were the most easily verified aspects of the weapons. In 1972, Nixon and Brezhnev signed the Strategic Arms Limitation Treaty (SALT 1) (Bureau of International Security and Nonproliferation, 1972). This treaty stopped further production of land-based intercontinental ballistic missiles (ICBMs). SALT 1 limited the U.S. to 1,054 ICBMS and the U.S.S.R. to 1,618 ICBMs. It also set limits on submarine-launched ballistic missiles (SLBMs)—710 missiles on 44 submarines for the U.S. and 950 missiles on 62 submarines for the U.S.S.R. Verification of these treaties was done by national technical means, and the two states agreed not to conceal information that would prevent verification. This was largely done by satellite—one missile was contained in each silo, and even as the U.S. and U.S.S.R. began to move their silos underground, they were still detectable from satellite. Due to the large size of the missile-launch sites and the manpower required to maintain them, it is difficult to hide them from an observer. SLBM verification was accomplished through monitoring the subs themselves. These submarines are large and require visible production and support infrastructure (OTA Project Staff, 1990).

One aspect that the SALT I treaty did not address was Multiple-Independently-Targeted Re-entry Vehicles (MIRV), which the U.S. was using in their weapons increasingly often. As an example, the LGM-30 Minuteman III (Federation of American Scientists, 2015) has three re-entry vehicles that can be pointed at three different targets. SALT II sought to address this, specifically putting a limit on the number of MIRVed missiles (United States and Union of Soviet Socialist Republics, 1979), and limiting the total number of missiles to 2,250. This agreement was signed in January of 1979 by Carter and Brezhnev, but was not ratified by the U.S. Congress as concerns grew over the U.S.S.R.'s invasion of Afghanistan. While the treaty was never formally ratified, both countries did state they would comply (Arms Control Association, 2014b).

In 1988, the Intermediate-range Nuclear Forces treaty (Arms Control Association, 2014a) eliminated all missiles with a range of 500-5,000km. This treaty effec-

Strategic Nuclear Arms Control Agreements

Strategic Nuclear Arms Control Agreements							
	SALT I	SALT II	START I	START II	START III	SORT	New START
Status	Expired	Never Entered Into Force	Expired	Never Entered Into Force	Never Negotiated	Replaced by New START	In Force
Deployed Warhead Limit	NA	NA	6,000	3,000-3,500	2,000-2,500	1,700-2,200	1,550
Deployed Delivery Vehicle Limit	US: 1,710 ICBMs & SLBMs USSR: 2,347	2,250	1,600	NA	NA	NA	700
Date Signed	May 26, 1972	June 18, 1979	July 31, 1991	Jan. 3, 1993	NA	May 24, 2002	April 8, 2010
Date Ratifed, U.S.	Aug. 3, 1972	NA	Oct. 1, 1992	Jan. 26, 1996	NA	March 6, 2003	Dec. 22, 2010
Ratification Vote, U.S.	88-2	NA	93-6	87-4	NA	95-0	71-26
Date Entered Into Force	Oct. 3, 1972	NA	Dec. 5, 1994	NA	NA	June 1, 2003	Feb. 5, 2011
Implementation Deadline	NA	NA	Dec. 5, 2001	NA	NA	NA	Feb. 5, 2018
Expiration Date	Oct. 3, 1977	NA	Dec. 5, 2009	NA	NA	Feb. 5, 2011	Feb. 5, 2021

Figure 1.3: Summary of stockpile and delivery-system reduction agreements made by the U.S. and U.S.S.R./Russia.

tively prevented the U.S. and U.S.S.R. from putting weapons in friendly territories close to the opposing countries (such as Cuba). In total over 2,500 missiles were destroyed. It also was the first treaty with intrusive on-site inspections for verification. Verification was done through inspections one to three months before the treaty entered into force. Up to 20 limited notice inspections a year and the monitoring of missile production sites were allowed in the treaty as well. Monitors were welcome to oversee missile destruction. When the U.S.S.R. disbanded, Russia kept the agreement with the U.S. All other U.S.S.R. successor states signed by 2002 and likewise destroyed their intermediate-range missiles.

The Strategic Arms Reduction Treaty (START) (Federation of American Scientists, 1999) was signed in 1991 and went into effect in December of 1994. It stated that within seven years, both countries must limit the total number of warheads to 6,000 and delivery vehicles to 1,600. Unlike the SALT treaty, START incorporated an intrusive verification structure. Baseline readings were taken in 1995. Starting at that time, the U.S. began portal monitoring activities at two Russian missile assembly sites, though Russia chose not to continuously monitor the U.S. portal listed

in the treaty.

Verification of warhead totals is a far more difficult task than delivery system totals. Missile construction is visible from satellites, and on site inspections do not reveal sensitive information, unlike measuring a warhead. The details of warhead construction are closely guarded secrets by each country. Hence, the verification process for the first START treaty was fairly simple. Each missile type was assigned a certain number of warheads, and the inspecting country could come in and verify that the correct number of warheads (or fewer) were loaded up on the missiles (Tre, 1991) by verifying the absence of nuclear material on some of the reentry vehicles.

SORT (Strategic Offensive Reductions Treaty) (United States and Union of Soviet Socialist Republics, 2003) entered into force in 2003. SORT obligated the U.S. and U.S.S.R. to reduce their stockpiles to a maximum of 2,200 warheads (by START counting rules) in 2012. Both countries stated compliance by 2009. In 2011, New START was signed (U.S. Department of State, 2011). This further reduced the number of missiles to 700 total land and sea based missiles, and 1550 total warheads. New START counted re-entry vehicles as warheads, as opposed to the original START which verified a pre-assigned number for each missile (Arms Control Association, 2012). In addition, bombers were counted as a single warhead, though warheads aren't typically deployed on bombers, so there will be less than 1550 total deployed warheads when both countries comply. The treaty allows for 10 yearly on-site inspections at submarine and air bases to count the number of deployed warheads. The US declared compliance with the state warhead limit in October, 2015 (Kristensen, 2015).

1.1.4 Future Steps Towards Arms Control

Of great importance to this thesis, and the cause of this project's existence, is the potential for future limitations on warheads in arms-control treaties. To reduce the stockpile further, the two countries will need the capability to verify missile disarmament and warhead dismantlement. Methods need to be developed that answer the question "Is this measured item a warhead? If so, what type?" This task is inherently difficult. Neither side wants to reveal secrets about its weapons, so a measurement must be made that can verify the tested object is a warhead without gaining access to the intimate details of that object.

The Comprehensive Nuclear Test Ban Treaty (CTBT) (CTBTO Preparatory

Commission and others, 1996) was ready for signing in 1996. It would ban any nuclear tests, even underground, essentially preventing non-weapon states from gaining the experience necessary to build a functioning nuclear weapon, or risk deploying a non-functional weapon. It would also prevent weapon states from testing new weapons. A particular goal of this treaty was to avoid a regional arms-race between India and Pakistan that could have the potential to incite nuclear war. Unfortunately, this goal has not been accomplished—North Korea, India and Pakistan, the states that are arguably three of the most important to international nuclear stability, have not signed the treaty. In fact North Korea has tested weapons in 2006, 2009 and 2015 (BBC News, 2015). In addition, Iran, Israel, China and the US have signed but not ratified the treaty, though it is believed these countries would fall in line if the U.S. did ratify it. Verification of this treaty, while not pertinent to this project, is still an active area of research. Verification is and will be done through a mix of seismic, hydroacoustic and infrasound monitoring stations around the world in addition to radionuclide stations to detect possible fallout (Comprehensive Test Ban Treaty Organization, 2012).

There are many other important problems in the nuclear-security mission that must be addressed alongside the disarmament effort. Once the number of warheads becomes the limiting measure of nuclear capability, greater restrictions will need to be placed on the generation and movement of plutonium worldwide. The IAEA has worked for decades to inspect nuclear civilian facilities in non-nuclear weapon states, and prevented the movement of special nuclear material to weapons facilities (Gibson, 1997). However, they do not do any surveillance of nuclear-weapon states, and as the number of weapons decrease further, the U.S. and Russia will need to submit to inspections of their energy facilities as well.

Another part of the nuclear-security mission is spent-fuel assay, currently done through some simple neutron and gamma gross count rates. A small percentage of spent fuel from reactors is fissile plutonium and uranium, but overall 90% of the world's plutonium reserves are contained in fuel dumps. High-accuracy, non-destructive measurement of the fission isotopes in spent-fuel rods is a task that current research is focusing on.

1.2 A Brief Introduction to Image Science

Some notational definitions are necessary for the reader. Throughout this work, vectors and matrices are bolded. As an example, the three-dimensional cartesian coordinate vector $\mathbf{r}$ is used to represent the $\{x, y, z\}$ coordinate space. Averages are represented using both angle brackets $<,>$ and in shorthand with a line over the variable.

To perform tasks such as discriminating an object as one of two types, data needs to be measured on that object. Expressed mathematically, a system h maps an object f to some data g. This measurement process can be represented in different forms depending on how the object and imager are described. This section begins with a description of those components in Section 1.2.1. After that, Section 1.2.2 introduces the reader to the concept of task-based assessment.

1.2.1 Object, System, and Image Description

For now, only the linear form of the imaging equation is discussed. As an example, a pinhole camera (pg 629 of Barrett's "Foundations of Image Science" (Barrett and Myers, 2003)) maps an object with an emitted spatial distribution $f(\mathbf{r})$ through an imaging system $h(\mathbf{r}, \mathbf{r}')$ to an image $g(\mathbf{r}')$. Shown in equation form,

$$g(\mathbf{r}') = \int h(\mathbf{r}, \mathbf{r}') f(\mathbf{r}) \mathrm{d}\mathbf{r}. \tag{1.1}$$

Here, h is example of a continuous-to-continuous (C-C) system. More generally, taking the object f as a function of variables $\mathbf{X}$, and the image data g a function of variables $\mathbf{X}'$, a C-C system can be represented as,

$$g(\mathbf{X}') = \int h(\mathbf{X}, \mathbf{X}') f(\mathbf{X}) \mathrm{d}\mathbf{X}. \tag{1.2}$$

Most current imaging systems, whether they are used to image electromagnetic radiation or neutrons, use a digital output. A fast-neutron coded-aperture detector is used throughout this thesis. While more detail on this system is discussed in Section 3.1, the imager ultimately outputs a detected energy for each photomultiplier tube in the imaging system. Binning this data by pixel index and energy, the system can be represented in a discrete-to-continuous (D-C) format,

$$g_m = \int h_m(\mathbf{X}) f(\mathbf{X}) \mathrm{d}\mathbf{X}. \tag{1.3}$$

Here, the data vector $\mathbf{g}$ would consist of all of the g_ms, where m goes from 1 to M total bins. $h_m(\mathbf{X})$ is the sensitivity of the m^{th} detector bin to a given object with parameters $\mathbf{X}$.

The imaging equation is still missing one critical component—the noise. Noise is randomness in the output due to any stochastic processes in the object or imaging system. The noise scales with the number of detected counts and therefore the strength of the object intensity. Accounting for noise is critical when it comes to task performance and the noise is represented in the imaging system as,

$$g_m = \int h_m(\mathbf{X}) f(\mathbf{X}) \mathrm{d}\mathbf{X} + n_m. \tag{1.4}$$

While reconstruction does not play a role in this project for reasons to be explained later in this chapter, many imaging related tasks use known information on g and h to reconstruct the object f_{rec}. This reconstruction is then used to make decisions. There is an enormous amount of literature devoted to this subject, but some basic reconstruction algorithms are maximum-likelihood estimation maximization (MLEM) (Shepp and Vardi, 1982) and filtered back projection (FBP) (Zeng, 2012).

In the following subsections, the source distribution f, imaging system h, data $\mathbf{g}$, and noise $\mathbf{n}$ are discussed in detail. A thorough discussion of imaging systems can be found in Chapter 7 (Barrett and Myers, 2003). A statistical description of the objects and image data can be found in Chapter 8 (Barrett and Myers, 2003).

1.2.1.1 Object Description

The object f is a continuous function representing the dependence of the emission rate over many variables. Particles of a certain type p_{name} are emitted from Cartesian location $\mathbf{r}$ with momentum $\mathbf{p}$ and energy E. As such, $\mathbf{X}$ in (1.3) could be represented as $\{p_{name}, \mathbf{r}, \mathbf{p}, E\}$. This f can be represented as a sum of a background term, b, and source term, s, both dependent on the same parameters as f.

When performing treaty-verification tasks, there are nuisance parameters present in the object, which are labeled γ in this work. Nuisance parameters are sources of variability that affect the data acquired, but are not of interest in performing the task. These unknowns in the system degrade task performance. Understanding the role nuisance parameters play in the forward model and properly accounting for them in the observer models helps to compensate for the performance losses they introduce. One treaty-verification task involves the imaging of an unknown object

inside a drum or container. This object could have unknown orientation inside the drum, and the placement of the drum itself is often imprecise. Another nuisance parameter could be variation in the age of the material among multiple TAIs of the same type. This causes variation in the detected gamma and neutron intensities and energy distributions. γ_j is defined as the set of of nuisance parameters for object j, e.g.,

$$\gamma_j = \{\text{object orientation, object location, source age, etc.}\}. \tag{1.5}$$

1.2.1.2 Imaging System Description

The discussed imaging equations in this chapter have all been linear. Ignoring noise for the moment, they have the property,

$$\begin{aligned} g_{m,1} &= \int h_m(\mathbf{X}) f_1(\mathbf{X}) \mathrm{d}\mathbf{X} \\ g_{m,2} &= \int h_m(\mathbf{X}) f_2(\mathbf{X}) \mathrm{d}\mathbf{X} \\ (g_{m,1} + g_{m,2}) &= \int h_m(\mathbf{X})(f_1(\mathbf{X}) + f_2(\mathbf{X})) \mathrm{d}\mathbf{X} \end{aligned} \tag{1.6}$$

Unfortunately, this linearity does not hold for gamma ray or neutron imaging. Gamma rays are attenuated by any material between their emission location and the detector, and are scattered by surrounding geometries. Neutron imaging deviates even further from the simple linear mapping; neutrons emitted through spontaneous fission can cause induce fission events in the object geometry, leading to a chain reaction. Such a system is highly nonlinear, and would more accurately be represented by the imaging equation,

$$g_m = \int h_m(\mathbf{X}; f) f(\mathbf{X}) \mathrm{d}\mathbf{X}. \tag{1.7}$$

In this system, the imaging operator is also a function of the object. This thesis will not delve deeper into nonlinear imaging systems, but this highly non-linear behavior led in part to the choice to use a Monte Carlo transport code to simulate detector data. A more thorough description of nonlinear systems can be found in section 7.5 of "Foundations" (Barrett and Myers, 2003).

1.2.1.3 Noise Description and Probability Theory

Up to this point, focus has been on the deterministic aspects of the imaging equation. In a real-life experiment, there is always some randomness associated with $\mathbf{g}$. A random variable x takes on values governed by its probability distribution.

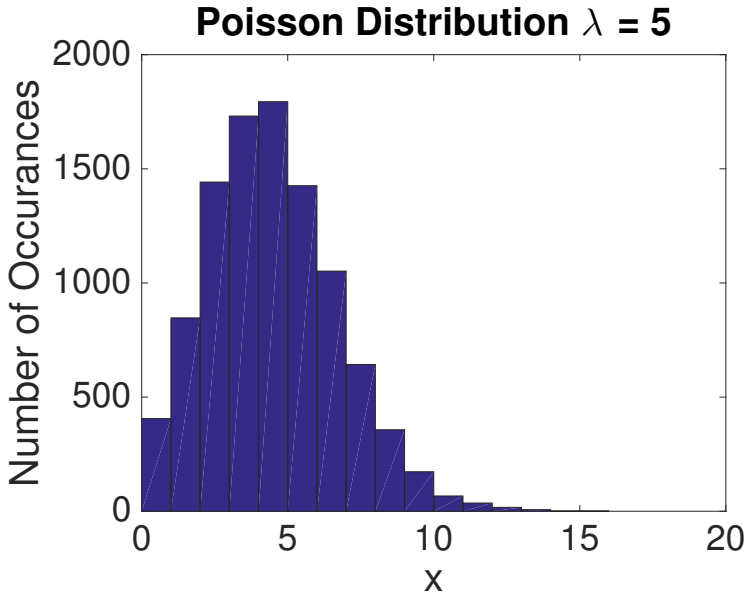


Figure 1.4: A Poisson distribution with mean 5. The x axis is the value that the random variable x takes on. The y axis shows the number of times x is sampled out of 10,000 total trials.

Discrete random variables can only take on discrete values and are described by a probability $Pr(\cdot)$, while continuous random variables can take on infinitely many and uncountable values and are represented by a probability density function (pdf) $pr(\cdot)$. This section discusses two prominent sources of noise in imaging.

Poisson noise (Good, 1986), also known as shot noise, is always present in an imaging system, and is due to the discrete nature of the emitted particles. When counting the number of detected particles, only non-negative integer values are possible. One parameter (called λ) is used to describe a Poisson distribution. λ is equal to both the mean and variance of the distribution. The discrete probability distribution for a Poisson random variable x is shown below,

$$Pr(x) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{1.8}$$

An example Poisson distribution with $\lambda = 5$ is shown in Figure 1.4. The Poisson distribution is positively skewed (the mean value is greater than the median value).

The Gaussian distribution (Gauss, 1809), defined below, also plays an important role in image science. Gaussian distributions are defined by their mean $\overline{x}$ and variance (spread) σ_x^2, though they are often described by their standard deviation σ_x as well. A Gaussian pdf is,

$$pr(x) = \frac{1}{2\pi\sigma_x^2} e^{-\frac{(x-\overline{x})^2}{2\sigma_x^2}}. \tag{1.9}$$

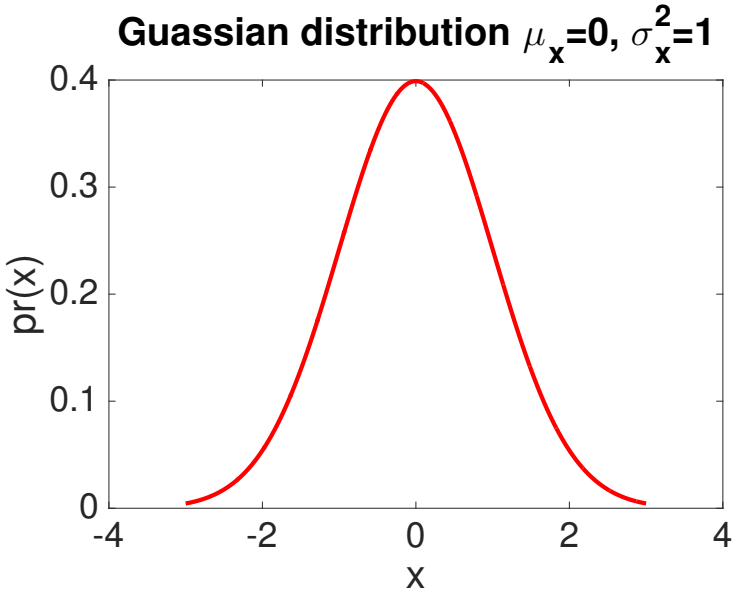


Figure 1.5: A Gaussian distribution with mean 0 and variance (and standard deviation) 1. As a rule of thumb, 68% of the values fall within one standard deviation, 95% within two and 99.7% within three.

The probability of selecting a value of x within a certain range can be found by integrating this probability density over that range of x values.

The pdf for a Gaussian distribution with mean 0 and variance 1 is plotted in Figure 1.5. Gaussian distributions are also known as normal distributions and are occasionally represented in this work as $\mathcal{N}(\overline{x}, \sigma_x^2)$. Gaussian noise is present in imaging due to energy smearing (p. 756 of (Barrett and Myers, 2003)). For example, when a gamma ray of energy E_0 enters the detector and is absorbed, the output of a given PMT in the imaging system is a voltage V. V is actually Poisson due to photon-counting statistics, but for enough detected photons it can be modeled as a Gaussian random variable with mean and variance dependent upon the absorbed energy and detector quality.

Gaussian distributions also arise from the Central Limit Theorem (Araujo and Giné, 1980), which roughly states that the sum of independent, identically distributed random variables x_n with mean $\overline{x_n}$ and variance $\sigma_{x_n}^2$ yields an approximately normal distribution with mean $\sum_{n=1}^{N} \overline{x_n}$ and variance $\sum_{n=1}^{N} \sigma_{x_n}^2$. The observer models presented in this paper can generally be presented as a sum of independent and identically distributed random variables, and hence, with a few exceptions, the resulting test statistic distributions are Gaussian.

Finally, the multivariate Gaussian distribution plays a role in the channelized Hotelling observer. The pdf on a set of random variables $\mathbf{x}$, whether they are

correlated or uncorrelated, is,

$$pr(\mathbf{x}) = \frac{1}{(2\pi)^N |\mathbf{K_x}|} e^{\frac{-1}{2}((\mathbf{x}-\overline{\mathbf{x}})^\dagger)\mathbf{K_x}^{-1}(\mathbf{x}-\overline{\mathbf{x}})^\dagger)}. \tag{1.10}$$

$\mathbf{K_x}$ is the covariance matrix, which contains the variance on each element of $\mathbf{x}$ along the diagonal and the covariance between each pair x_i and x_j where $i \neq j$ in the off-diagonal elements. $|\mathbf{K_x}|$ represents the determinant of that covariance matrix.

1.2.1.4 Image Description

Photon-counting detectors often bin the data into a vector $\mathbf{g}$ that contains the number of detected events attributed to a certain pixel interaction and in a certain energy range for a given acquisition time. Photon-processing detectors process data in LM format. Throughout chapter 4, the data is represented by a total number of detected events, N, and the detectable data for each event, $\{A_n\}$.

The statistics of the image data are governed by those of the object and imaging system. Poisson noise is present due to counting statistics. If an object is imaged for a certain acquisition time, only a certain number of particles N_d are detected, and the distribution on the number of detected particles is Poisson with mean $\overline{N_d}$. The mean number of detected particles $\overline{N_d}$ is related to the mean number of emitted particles $\overline{N_e}$ by a detection efficiency η_{e-d}. For a point source and a detector without any surrounding geometries to scatter particles, there are generally two components to this detector efficiency; a factor representing the geometric likelihood that an emitted particle hits the detector η_{geom} and a detector efficiency representing the probability that a particle entering the detector interacts in the detector η_{det}.

Object variability due to nuisance parameters is another source of randomness in the image data. When nuisance parameters are present, the data $\mathbf{g}$ is doubly stochastic, as there are two sources of variability. When determining the mean of $\mathbf{g}$, which is denoted $\overline{\overline{\mathbf{g}}}$, two averages must be taken:

$$\overline{\overline{\mathbf{g}}} = \left\langle \langle \mathbf{g} \rangle_{g|\gamma} \right\rangle_\gamma \tag{1.11}$$

The first average is over the Poisson noise in the imaging system given knowledge of the nuisance parameter. The second average is over the nuisance parameters.

1.2.2 Task-Based Assessment

Image quality is a term that can take on many different meanings depending on the way it is used. Some define image quality by the resolution of a reconstructed object;

the more capable the imaging system is of resolving small features in the object, the better the system for that task. Measures such as the signal-to-noise ratio (SNR, defined in (1.12)), contrast to noise, mean squared error between a reconstructed object and the original object, and other statistics are also useful measures of image quality. In the below definition of SNR, the variables x_1, x_2 are random variables with means $\overline{x_1}, \overline{x_2}$ and variances $\sigma^2_{x_1}, \sigma^2_{x_2}$,

$$SNR = \frac{\overline{x_2} - \overline{x_1}}{\sqrt{\sigma^2_{x_2} + \sigma^2_{x_1}}} \tag{1.12}$$

A multitude of research groups in the national-security mission are currently working on novel detectors to achieve better resolution, or acquire more complete data. The University of Michigan, for example, is working on a Compton neutron imager to perform long range localization (Poitrasson-Rivière et al., 2015). Oak Ridge National Laboratory has developed a detector (Archer et al., 2010) that uses time-correlated measurements to count coincident detections due to fission. These systems all provide useful contributions to the nuclear-security mission and potential verification of TAIs.

This thesis, however, takes a task-based approach to imaging, emphasizing detectors and decision making models that best perform certain objectives. Task-based imaging requires a well-defined task to be performed on objects, an observer to perform that task, and a figure of merit to judge the observer performance. There are many different tasks applicable to arms-control-treaty verification that will be explored in Section 1.3. The most prominent are null-hypothesis tasks, binary-classification tasks, and counting tasks.

An observer is the person or mathematical model that performs the defined task. A great summary on observer models, including those relevant to this work, can be found in (Barrett et al., 1993). In a task such as tumor detection in medical imaging diagnostics, the human observer uses the reconstructed object to make decisions. X-ray CT, SPECT, and PET imaging often use a human observer (radiologist) to make a decision in a given task. This thesis, however, will focus on the application of mathematical observer models to treaty-verification tasks. These mathematical models act on the data, ultimately returning a scalar test statistic which is thresholded or compared to some range of values to make a decision

There also must be a figure of merit to judge the ability of the observer model to perform a given task. The figure of merit used will be task-dependent; a deeper explanation for analyzing model performance for tasks related to treaty verification

is explained later in Section 1.3.

1.3 Relevant Treaty-Verification Tasks

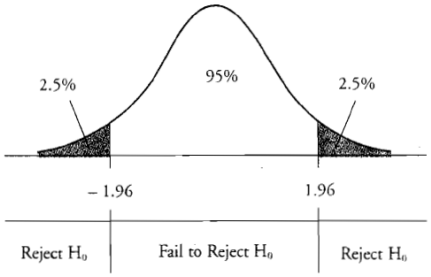
There are many potential tasks that are necessary to perform in arms-control-treaty verification and for other purposes in the nuclear-security mission. Null-hypothesis (Section 1.3.1), binary-discrimination (Section 1.3.2), counting (Section 1.3.3), estimation (Section 1.3.4) and other important tasks (Section 1.3.5) will be discussed in this section, and examples will be given for each.

1.3.1 Null Hypothesis Tasks

To perform a hypothesis test, one begins by defining the null hypothesis and alternative hypothesis. The null hypothesis, for example, could be that the measured TAI is of a single type, or has certain physical characteristics. The alternative hypothesis would be that the null hypothesis is not true. A statistical test is used to declare whether a tested object's measured data is consistent with the null hypothesis statement within statistical chance. If so, the the null hypothesis is not rejected. If not, the null hypothesis is rejected.

Calibration data on a trusted TAI would be measured, and the mathematical observer model would be trained on this data. This observer returns a test statistic when an independent measurement is taken from the same source. The test statistic itself is a random variable due to Poisson noise and randomness in the source term. The spread on the test-statistic distribution will depend on the variability of the detector data from one measurement to the next, and over many measurements of the trusted source a distribution on test statistic values $pr(t|H_0)$ is built for the null-hypothesis object. When measuring an unknown source, the test statistic returned by the model can compared to $pr(t|H_0)$ and rejected if the value is unlikely.

An intuitive explanation of a two-tailed null hypothesis test is shown in Figure 1.6. A type I error occurs when the value of the test statistic (resulting from performing the model on H_0) falls outside the accepted range of values. This occurs with probability α. Because the range of values for t to be accepted can not be infinite, there will always be a finite probability of rejecting H_0 even if it is true. A correct acceptance of H_0 then occurs with probability $(1-\alpha)$ when testing the the object used to train the statistical test. A type II error would represent a successful spoof of the trusted TAI and occurs with probability β. $(1-\beta)$ is referred to as the



	State of Null (Truth)	
Decision:	**Null True**	**Null False**
Fail to reject null	**Correct Decision**	**Type II error** (β)
Reject null	**Type I error** (α)	**Correct Decision (Power)**

Figure 1.6: The left image (Subudhi, 2013) shows an example test-statistic distribution $pr(t|H_0)$, which in this case is standard normal. 95% of the values fall within two standard deviations, or for this distribution, a value within $\pm$ 2. A test statistic with an absolute value greater than 2 is rejected. On the right is a decision table. A type II error occurs when the measured item is inconsistent with H_0, but model fails to reject it. A type I error occurs when the null hypothesis is true but the model rejects the item.

"power" of the test. A greater "power" implies that the test is stronger in rejecting H_a.

In this dissertation, null-hypothesis performance plots consist of the percent of the time H_0 is rejected as a function of acquisition time. As more data is read in, it becomes increasingly likely that the model will correctly reject any source that is not from H_0. This is exemplified in Figure 1.7. When a source belonging to a class other than H_0 is tested, the observer ideally rejects it. As the acquisition time increases, the test statistic should deviate further from $pr(t|H_0)$, and be rejected more often. When the H_0 source is tested, it is rejected with a probability determined by the power of the test. In this case, values greater than 2 standard deviations from the mean were rejected, which corresponds to 5% of of the samples.

It is important to recognize that a null-hypothesis test can never be used to confirm a hypothesis, only reject it. There is always a finite probability that the tested object was not of the null-hypothesis type. There are two potential causes for this. First, the detector may not be able to distinguish the null hypothesis object from the tested object due to poor resolution or lack of some other discriminatory capability. Second, the object itself could just be a very well done spoof that fools the observer model. These pitfalls require a more practical definition of "successful" result. One potential answer is that a spoof must be prohibitively expensive to produce so that it is not in the host country's best interest to do so.

Examples of null-hypothesis tests include the Chi-Squared Goodness of Fit Test (Greenwood and Nikulin, 1996) and the Mahalonobis distance (De Maess-

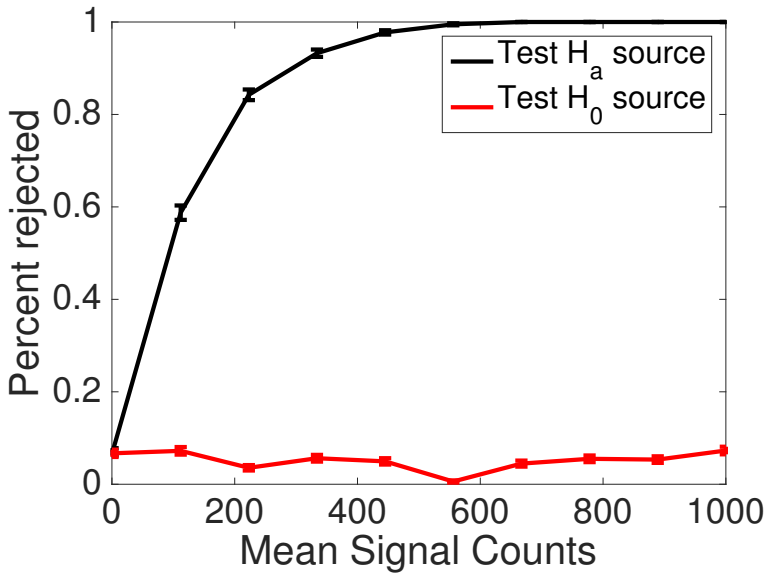


Figure 1.7: Example performance plot for a null hypothesis test. The probability of the tested object being rejected is plotted as a function of time. In this task, there are only two outcomes (reject/not reject). The error bars are therefore governed by binomial statistics. One standard deviation is defined as $sqrt(\frac{pr*(1-pr)}{N})$, where N is the number of samples taken to test the model.

chalck et al., 2000). These are discussed in detail in Section 6.1. Next, specific hypothesis-testing tasks are introduced.

1.3.1.1 Is the Imaged Object a Warhead?

In this task, the host country places a warhead inside some container (such as the one shown in Figure 1.8) in front of a measurement system. The monitor wants to confirm the presence of a warhead inside the container. This is a question that would likely be central to any warhead-counting procedure, but is inherently difficult to answer. Different individuals and groups have different answers as to what constitutes a warhead. The IAEA, for example, declares a "significant quantity" of certain isotopes and elements necessary to create a bomb, such as 8kg of plutonium or 25kg of highly enriched uranium (defined as >20% U235) (IAEA in Austria, June). A mass of plutonium on a similar order of magnitude to these numbers could be treated as a warhead, but that disregards how difficult the construction of a functional warhead is, and the necessary geometry. Alternatively, a warhead could be defined as SNM arranged to produce yield? If this were the chosen definition, the geometric construction would need to be taken into account.

Figure 1.8: An example storage container for a warhead.

1.3.1.2 Was the Object Changed in Transport?

In this task, an object is measured prior to transport, then is loaded into a vehicle that is outside a monitor's purview (Hauck et al., 2012). It is measured again when it reaches its arrival destination, and the monitor must verify that the imaged object is the same as the object that left the initial facility. This is a well-defined task in comparison to warhead verification. A template-matching technique would be the desired approach to this problem; the monitor would compare the first and second measurements to make a decision. However, there are inherent hurdles to overcome in this specific task not present in a warhead verification task. The background distribution, for example, would be different due to the change in location. In addition, it is possible that the object moved in transport, causing extra variability in the source term.

1.3.2 Classification Tasks

Binary-classification and N-type classification tasks also have a role to play in treaty verification. In a binary-classification task, there are two hypotheses. For example, in nuclear-threat detection, one hypothesis, denoted H_1, is that only a background is present and the second hypothesis, H_2, is that a source and background are present. Calibration data is acquired for the two hypotheses and used to train the mathematical model. Then the model is tested on independent signal present and signal absent distributions, resulting in two test-statistic distributions, $pr(t|H_1)$ and $pr(t|H_2)$, as shown in Figure 1.9.

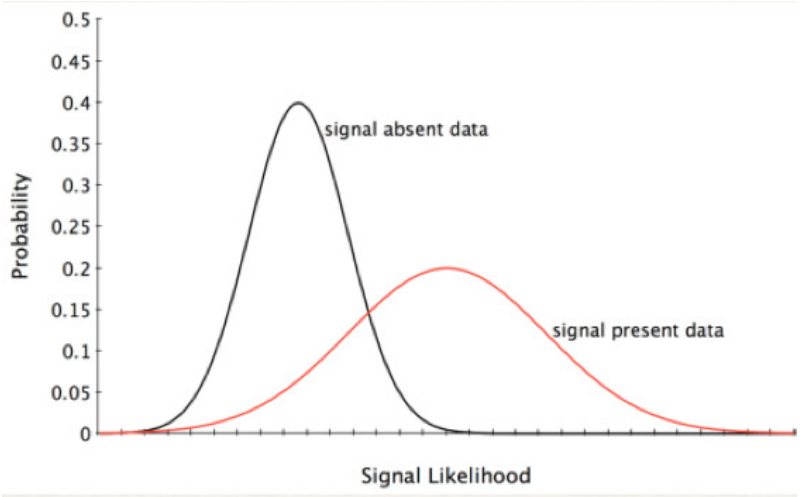


Figure 1.9: Example test statistic distributions when performing the observer model on signal-absent and signal-present scenes. The signal-absent distribution is $pr(t|H_1)$ and signal-present distribution is $pr(t|H_2)$.

	H_1 is true	H_2 is true
Choose H_1	Correct	Type II error
Choose H_2	Type I error	Correct

Table 1.1: Decision table for binary-classification tasks. There are two hypotheses, H_1 and H_2, that the observer must decide between. Incorrectly diagnosing source 1 as source 2 is a type I error and incorrectly diagnosing source 2 as source 1 is a type II error.

A threshold is set based on these distributions. Any test statistic greater than the threshold causes a signal-present declaration; a lower test statistic yields a signal-absent declaration. An example decision table is shown in Table 1.1. The penalties for a type 1 error and type 2 error are task-dependent. In a threat-detection task, the costs of incorrect decisions are very different. The type 1 error raises an alarm, wasting valuable time and taxpayer money. The type 2 error allows for the possibility of a nuclear attack. Each of these would need to be assigned their own costs. If the task is instead the binary classification of a tested item as one of two types, the penalty for either of the two errors would be the same—a misclassification of the object. Choice of the test-statistic threshold changes the probabilities of each of these errors. For example, if the threshold is set at the far left of the distribution in Figure 1.9, the model always decides H_2. If the threshold is set at the far right, it always decides H_1.

One figure of merit to gauge the performance of an observer model in binary-classification tasks is the Receiver-Operating-Characteristic (ROC) curve (Hanley and McNeil, 1982). This plots the true-positive fraction (probability of correctly

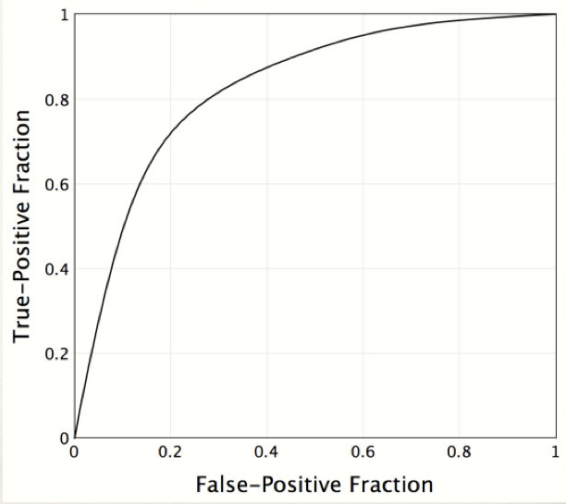


Figure 1.10: Example ROC curve for overlapping test-statistic distributions, as in Figure 1.9. If the two test-statistic distributions completely overlap, the ROC curve would be a diagonal line from (0,0) to (1,1). If the two test statistic distributions are completely separated, the ROC curve goes from (0,0) to (0,1) then over to (1,1).

choosing the H_2 outcome) as a function of the false-positive fraction (type II error probability) as the threshold is varied. An example ROC curve is shown in Figure 1.10. Using this curve and the cost functions associated with type 1 and type 2 errors, one can choose an optimal threshold. Further discussion on this topic can be found in section 13.2 of (Barrett and Myers, 2003).

This work is not immediately concerned with how to define the cost functions for incorrect decisions or where exactly to set the threshold for each task. Therefore, the figure of merit used in this thesis is the area under the ROC curve, which can be seen as a measure of the separation of the two test-statistic distributions, closely related to the SNR (see p. 819 of (Barrett and Myers, 2003)). When the test-statistic distributions overlap, the AUC takes on a value of 0.5. When they are completely separated, it has a value of 1. In this dissertation, binary-discrimination task performance isjudged by plots of the AUC as a function of acquisition time (see Figure 1.11)—as more counts are received, the observer model becomes more certain in declaring an unknown source as type H_1 or H_2.

One advantage to using the AUC as a metric is that the two-alternative forced-choice test (2AFC) can be used to determine the AUC value (Fechner et al., 1966). With this method, the observer is presented with a series of pairs of testing datasets. One data set is a measurement of source 1, and the other is a measurement of source 2. For each dataset, the observer calculates a test statistic that is intended to have

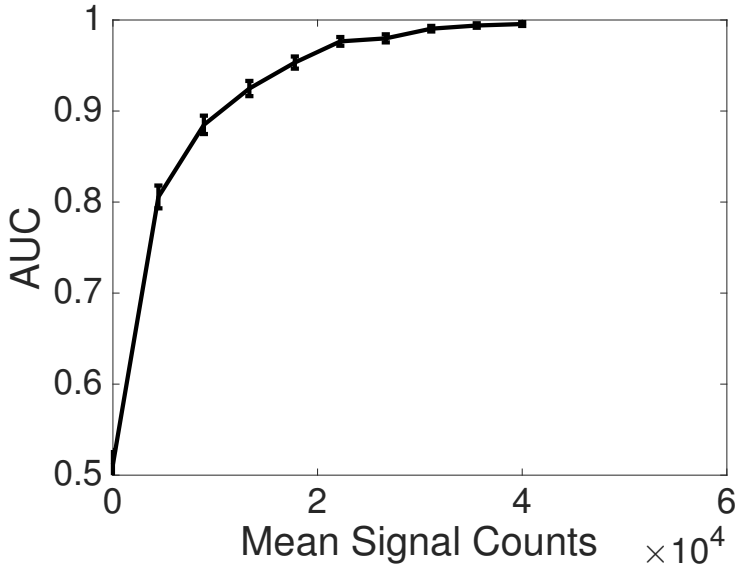


Figure 1.11: Example performance plot for a binary classification task.

a higher value for source 2 than for source 1. The AUC is equivalent to the fraction of the time that the source 2 test statistic is greater.

N-type classification can be accomplished in a similar matter. Rather than use a single threshold, the test-statistic space can be segmented into N (or more) decision sections. There would be a probability of the test statistic falling into each decision range, $pr(\Delta t_n, |H_j)$. A decision table analogous to Table 1.1 would have $N \times N$ outcomes, and there would be a corresponding cost for each specific incorrect outcome. Ideally, multiple test statistics would be used, each corresponding to a different aspect of the measurement data, and this $M < N$ dimensional space could be used to classify the items.

While the development of models that can perform binary or N source classification tasks is important, the ideal model would be able to answer the question "is this tested source of type 1,2,...N or a spoof?" A combination of the classification and null-hypothesis tests would be needed to answer the question.

1.3.2.1 Explosive Dismantlement

One type of binary-classification task useful for treaty verification is explosive dismantlement. The reader can consider a warhead as being composed of two components—a primary object that is used to start the nuclear reaction and a secondary that is compressed by the energy released by the primary, drastically increasing the yield. The dismantlement step involves the removal of the primary (see Fig-

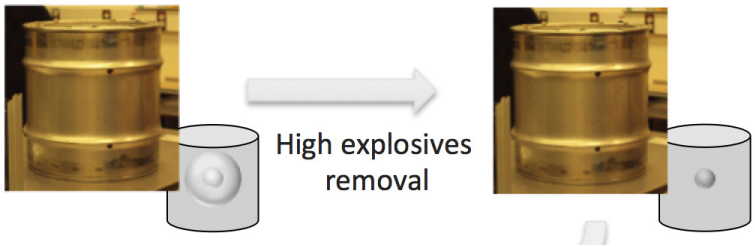


Figure 1.12: Illustration of the removal of the explosive from a pit.

ure 1.12). More information on this task can be found at `http://nnsa.energy.gov/ourmission/managingthestockpile/dismantlementanddisposition`. A monitor would be performing a binary classification task, determining whether the tested TAI does or does not still have the high explosive attached. This could be achieved by using a template-matching approach, with the bare pit serving as one hypothesis, while the pit surrounded by the explosive serves as the second hypothesis.

1.3.2.2 Categorize Warhead Type

The U.S. has a limited number of nuclear warheads and each type gets assigned to a certain missile (Norris and Kristensen, 2010). A model that can perform N-source classification tasks would be ideal.

1.3.3 Counting Tasks

While the work accomplished in this thesis has focused on the prior two tasks, the development of models that can count the number of warheads is desired. Ideally, an imaging detector could be placed at a storage site and image every object in its field of view, counting the number of warheads. This is a very difficult task. The largest obstacle to overcome is the lack of linearity in the imaging system. Gamma-rays emitted from further objects are attenuated by closer objects. Neutrons emitted from farther objects induces fission in closer objects.

1.3.4 Estimation Tasks

While not a focus of this thesis, estimation tasks are common in medical imaging and have a role to play in the nuclear-security mission. Rather than detecting the presence of a material, these tasks estimate some feature of the object, whether that be mass, size or location. Performance can be judged by an EROC curve, where the y axis is the number of times a true positive result occurs with correct estimation

of the source parameter.

1.3.5 Other Necessary Tasks for the Nuclear Security Mission

There are many other tasks that prove useful for the nuclear security mission. These range from detection and localization tasks to portal monitoring.

1.3.5.1 Threat Detection and Localization

The first project I worked on was the detection and localization of nuclear threats in a cluttered radiation environment (The APS Panel of Public Affairs, 2013), such as a city (see Figure 1.13). It is a critical mission to ensure civilian safety. Specifically, this project emphasized the detection of dirty bombs (United States Nuclear Regulatory Comission, 2012), which are radiological dispersal devices that terrorists could construct. This is a hard problem—the signal is generally weak and the background distribution changes with time and location. To perform detection and localization, an imaging detector was placed inside a vehicle and driven down a street. Eleven images were taken, once every three meters. This setup differs significantly from medical imaging devices, as the projections of the object scene onto the detector come from a limited number of angles, making reconstruction of the object scene difficult.

For this task, the object was discretized, and split into 1m^3 voxels, with each voxel having a certain intensity (emission energy was ignored in this study). The background was modeled as a lumpy-Gaussian distribution. To generate this model, the number of Gaussian peaks was randomly selected from a predefined distribution, with the center of each peak being randomly selected from the object scene. The object was modeled as a point source, randomly located in the field of view. Localization, as discussed above, is an estimation task, and model performance was judged to be correct when it correctly predicted the object location within a tolerance of 1-3m.

MLEM reconstruction was performed and if the peak reconstructed intensity was above some threshold, the object was declared present at that location. The estimated location was compared to the known value. Alternatively, the scanning linear observer (SLO) (Whitaker et al., 2008) assumes a multivariate normal distribution on the data. Because the SLO uses knowledge of the correlations in the detector data between different source locations, it is able to more effectively detect

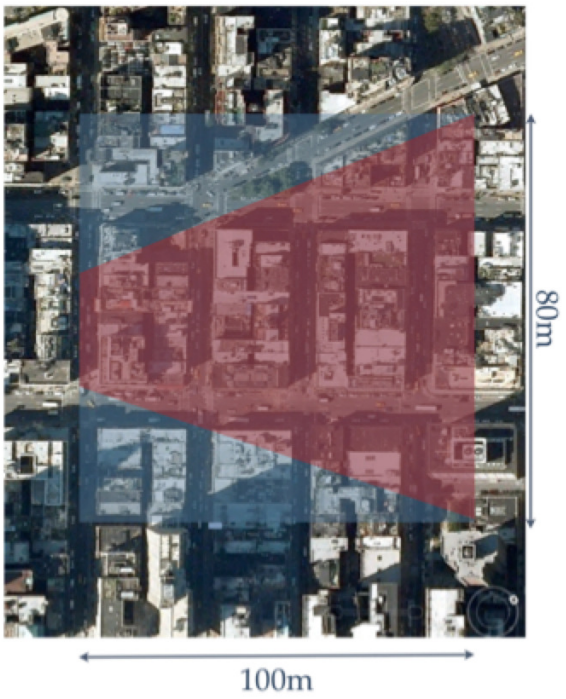


Figure 1.13: A detector is driven through an environment, with different regions of the object scene in its field of view at different times. In simulation, it reconstructed a 80m x 100m scene.

and locate the source. Performance of these methods is shown in Figure 1.14.

1.3.5.2 Cargo Screening and Portal Monitoring

A good summary for this task can be found at `https://missions.llnl.gov/counterterrorism/cargo-containers`. Each year, roughly 6 million cargo containers enter the U.S. and a small fraction are physically inspected. Officials are concerned that a party might hide uranium, plutonium or other special nuclear material (SNM) in these containers, but physically searching each container would be impossible due to time constraints. In this task, cargo containers shipped from overseas are imaged for potential explosives. This is difficult in part due to the significant shielding in these containers and the many different object types that could be stored inside the container. In addition, there is background supression—it is hard to acquire background calibration data because the containers themselves serve to suppress the background over the field of view. Developed detectors need to quickly image the contents and verify the absence of nuclear materials.

Both the threat-detection and cargo-screening missions have added practical constraints. The cost of a false alarm is expensive; it requires professionals to investigate the alarms themselves. Given that the likelihood of SNM detection is generally rare,

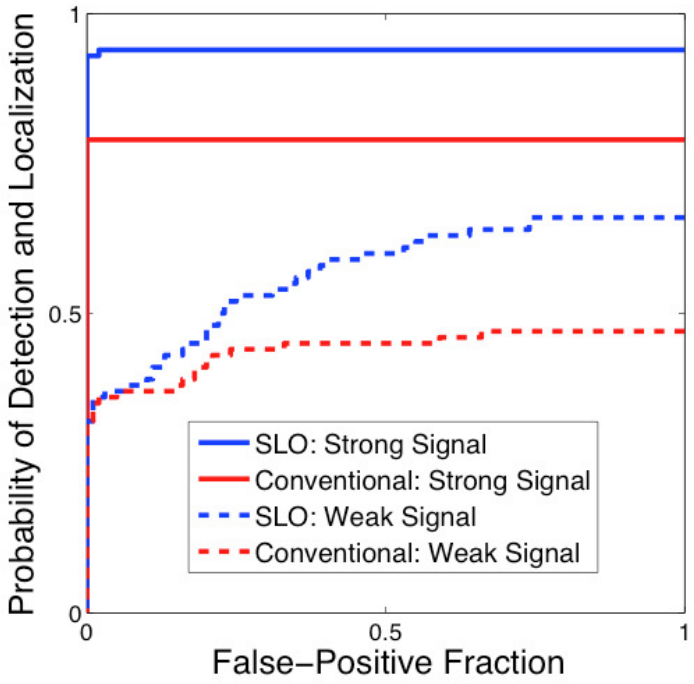


Figure 1.14: Plot of threat detection and localization performance. The scanning linear model, which utilizes the statistics of the detector data set, represents an improvement over MLEM reconstruction.

a high false-alarm rate wastes operator time and has a negative impact on commerce. In addition, nuclear threats are not the only concern for cargo screening as drug and human trafficking are more common, so a nuclear detector is one of many modalities that assess the container.

1.3.5.3 Spent-Fuel Assay

The IAEA is responsible for verifying that nuclear material dedicated to, and generated from, civilian facilities is not repurposed for weapons programs. To confirm this, methods must be developed that can accurately estimate the amount of plutonium contained in fuel cells. Some examples of systems developed for spent nuclear fuel assay, as well as deeper discussion of the motivation, can be found in (Willman et al., 2006; Quiter et al., 2010).

1.4 Medical Imaging Applications

The mathematical models used in this work have previously been applied to various tasks in the field of medical imaging. I will discuss two of these in this section.

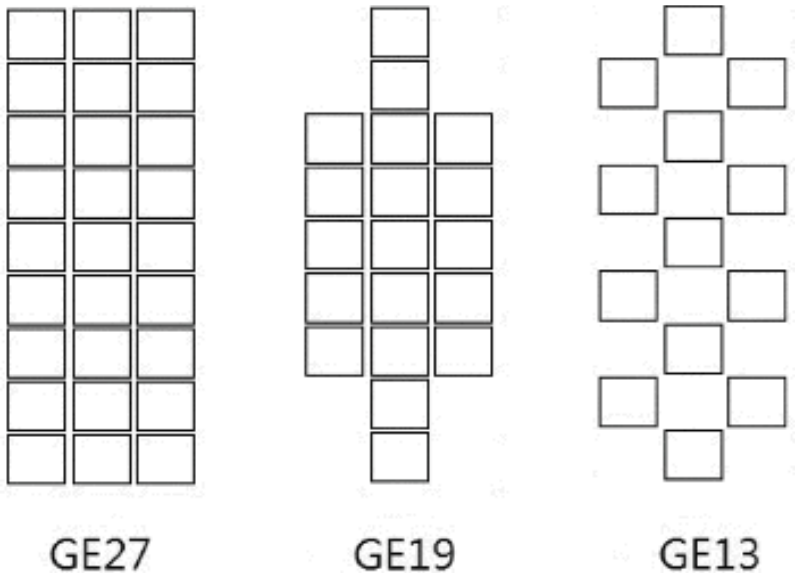


Figure 1.15: The different SPECT detector designs considered by Lee.

1.4.1 Detector Optimization

Dr. Chih-Jie Lee worked on a project for General Electric (GE) to do a cost-benefit analysis of different detectors for the task of detecting and localizing myocardial infarction (Lee et al., 2013). The pixels that GE uses in their SPECT imagers are expensive to manufacture, and GE wanted to gauge performance in detection and localization tasks using various other pixel arrangements (see Figure 1.15). Patients were simulated with the NURBS-based cardiac torso software package. The background consisted of several different organs in his study. Standard geometries and ranges in radiotracer uptake were simulated for all of these organs. Attenuation by the organs was included as well. Defects were varied in size and location in the heart. Radiotracer uptake changes due to these defects, and this was the signal that the models looked for. The scanning linear observer was applied to perform the detection and estimation tasks on these defects. Lee found that GE could substantially reduce the number of pixels by up to 30% while still maintaining optimal performance.

1.4.2 Modeling Human Performance

Park and Clarkson (Park et al., 2005) studied the ability of the channelized Hotelling observer to detect signals in signal-known-exactly and signal-known-statistically tasks. A description of the channelized Hotelling observer will not be given here as it is discussed later in this dissertation. Performance was compared to human ob-

servers at those same tasks. The motivation for this work, and similar work (Kupinski and Clarkson, 2005), is that a radiologist's time is expensive and mathematical observers that show superior performance could be used to develop improved detector systems. A lumpy background model was used throughout the study, combining a series of gaussian curves over the image plane. The signal was treated as a fixed (signal-known-exactly) or variable (signal-known-statistically) peak in the projection data. Three different imaging systems were considered, trading off resolution for sensitivity. They found that the channelized Hotelling observer outperformed the human observer when the signal was known, but performed poorly when the signal location was unknown.

1.5 Current Approaches to Warhead Verification and the Necessity for an Information Barrier

While the need to verify individual warheads and closely monitor and track plutonium production and storage may be a few years away, the difficulties inherent to warhead verification have been recognized for decades. In this section, two standard approaches to treaty verification are discussed (Section 1.5.1). The need for IBs in treaty-verification tasks is further expanded on (Section 1.5.2). One example is discussed of a system that utilizes an IB to categorize TAIs without giving the monitor knowledge of sensitive characteristics of the objects. Finally, some competing approaches to treaty verification are outlined (Section 1.5.3).

1.5.1 Template Matching vs Attribute Estimation

There are two common approaches that can be taken to verification tasks. One is an attribute-verification approach. This could extract features such as the mass, isotopic composition or size of the warhead. The system would output information that would give the monitor confidence that what it is measuring is a warhead and not a spoof.

The second approach to these tasks is a template-matching approach, and this dissertation focuses on those methods. This approach assumes that calibration data has been taken on previously verified TAIs. The observer models are built from these trusted TAIs' data, then used to classify independent objects. The medical imaging community has been using methods to perform similar tasks for decades, such as the Bayesian ideal observer and channelized Hotelling observer.

1.5.2 Need for Information Barriers

Basic neutron count-rate detectors have been used in the past to verify the absence of SNM on the missiles. Combined with an estimate of the mass and size of the warhead, this can give information on the composition and amount of plutonium. So far, this information has been sufficient to perform verification in the limit of a large number of missiles. As future treaties reduce the stockpile further and warheads become the limiting factor, detecting the difference between a chunk of plutonium or other neutron emitting substance and a warhead is critical. A gamma spectrum for example can give information such as the uranium and plutonium isotopic composition. Geometric properties can be determined through a particle imager, whether for neutrons or gamma rays. Generally, it would be ideal to allow for as many detection modalities as possible to perform the verification task.

However, the act of taking these measurements, and imaging in particular, would release sensitive information to the monitor; through reconstruction methods such as FBP or MLEM, the monitor could gain access to the shape or other geometric properties of the warhead. The monitor wishes to verify that the imaged item is a TAI, while the host wants to avoid disclosing any sensitive aspects of the objects (Fuller, 2010). Because of this, there has been a focus on developing systems and methods that utilize an IB to prevent transmission of sensitive information to the monitor.

One such system that utilizes an IB is the Controlled Intrusiveness Verification Technology (CIVET) system (Zuhoski et al., 1999), developed in part by Zuhoski, Indusi and Vanier at Brookhaven National Laboratory. Essentially, CIVET is an intelligent system, with software jointly developed by the host and monitor, that can analyze data and make declarations about the objects being measured without revealing any sensitive information to the monitor. The implementation of the system outlined in the referenced paper only uses a high-resolution gamma spectrometer to test sources, though other instruments could also be included behind the IB. The objectives for the development of this system are:

1. The system must be unable to transmit data.
2. The system must be unable to covertly store data.
3. The system must assure proper program (software) execution.
4. The system must verify proper sensor operation.

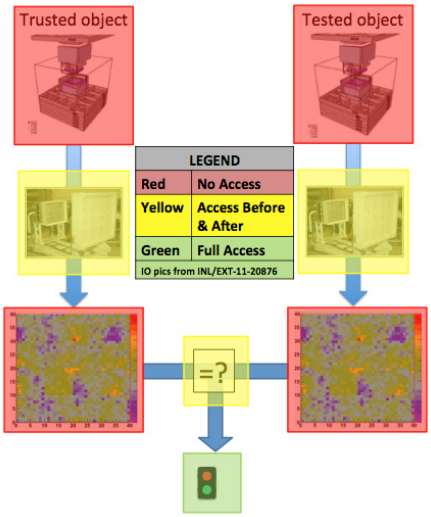


Figure 1.16: An intuitive picture summarizing a template-matching approach to treaty verification that utilizes an IB. The red blocks signify need for an IB, yellow the need at certain times, and green is open to the monitor.

5. The system must securely protect collected data.
6. The system must be composed of exportable technology.
7. Both hardware and software must be inspectable.

Due to specialized components and a lack of documentation, commercial hardware cannot be used—the host and monitor must therefore jointly develop their own hardware. The benefit is that there is not a limit on the number of sensors to be used, as they all would be behind this IB. Additional examples of IBs are the Trusted Radiation Identification System (TRIS) and Trusted Radiation Attribute Demonstration System (TRADS) systems (Seager et al., 2001; Mitchell and Tolk, 2000; Geelhood et al., 2000) developed by Sandia National Laboratories. These systems are fundamentally similar to the CIVET system, placing a measurement behind an IB. Figure 1.16 gives an intuitive summary of the CIVET system and any others that utilize an overarching IB while taking a template-matching approach.

IBs are very costly. They would require significant joint effort on the part of the U.S. and Russia to develop these methods and authenticate the system. In comparison, a mathematical model that stores only non-sensitive information that is still sufficient for confirmation (to be discussed in more detail in the following section) would allow the host and monitor to share all information in the model.

The ideal measurement system would be unable to distinguish objects that differ along predefined sensitive parameters regardless of the model used to perform the

task. The output of such a measurement device could be shared with the monitor.

1.5.3 Competing Work in the Field of Information Barrier-less Imaging

This section explores some approaches that other research groups have taken to performing treaty verification tasks with a significantly reduced IB.

1.5.3.1 Zero Knowledge Protocol

A “Zero Knowledge Protocol" (ZKP) has been developed by Glaser, Barak and Goldston at Princeton (Glaser et al., 2014). A template that corresponds to the negative of a neutron measurement is preloaded into the measurement device. An example could be a neutron image of an object, where the data vector $\boldsymbol{g_1}$ consists of the counts detected in each pixel. The host and monitor would agree on a number of desired counts in each bin, g_{max}. The detector output would then be preloaded with $g_{max} - \boldsymbol{g_1}$ counts in each bin. This initial value would not be available to the monitor. A tested item is then imaged, and the template value for the detected bin would be incremented by one for each detected event. At the end of the acquisition time, every bin would have g_{max} counts (ignoring Poisson noise). If the number of counts in each bin is not g_{max}, then the template did not match the tested item.

The group at Princeton proposes to use multiple preloads along with one previously verified and one unknown item. The monitor would choose which preload to use with which item. The monitor does not know the preloads but does know it is designed to yield g_{max} counts in each bin. If the preload that is the negative acts on a trusted item or an unknown item of the same type, it will yield the correct result. The other preload will not. This procedure would occur many times. If the host was to cheat and design the second preload to match the second object, multiple measurements mixing the different preloads and objects would pick out the cases when the preloads and measurements are mismatched, proving to the monitor that the host was cheating. Otherwise, if the unknown item was of the same type as the verified item, an "accepted" result would occur when the preload for the trusted item was used, and a negative result would occur when the preload for a different item was used. Therefore, the monitor could declare that the two objects are the same and the host did not cheat if they have a positive result 50% of the time. More information, including an intuitive explanation of the protocol using marbles in a bucket, is presented in (Glaser et al., 2014).

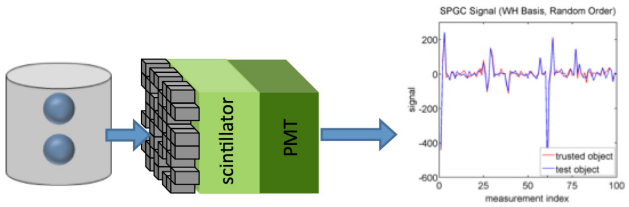


Figure 1.17: The SPGC uses a single pixel and randomly changing mask pattern to encode a measurement in time.

In this protocol, an IB would be required to hide the preloads from the monitor. In addition, it's not entirely clear how this model could account for nuisance parameters; its possible that many distinct preloads could be used, one corresponding to each specific nuisance parameter realization. The expected fraction of positive results would then be $\frac{1}{2*N_{NP}}$ where N_{NP} is the number of nuisance parameter realizations used.

1.5.3.2 Single-Pixel Gamma Camera

A group at Pacific Northwest National Lab has worked on a single-pixel gamma camera (SPGC), shown in Figure 1.17 to perform treaty verification tasks Gilbert et al. (2016). This detector would consist of a single large-volume pixel and have a mask pattern that changes randomly with time. In order to decode the output of the detector pixel and reconstruct an image of the object, the mask pattern's initial seed would need to be known. Only the host would have access to the mask sequence, but by starting the mask at the same initial seed, a measurement of a trusted item could be compared to a tested item.

1.6 Task-Based Approach to Limiting Dispersal of Sensitive Information in Treaty Verification

This work focuses on methods that can be used to overcome the IB requirement, and if not that, significantly diminish the amount of measurements that need to be taken behind an IB. It serves as a big picture summary for the models developed in this thesis.

1.6.1 Use of Projection Data

Projection data is the data that results from measuring an object with a detector. The methods developed in this dissertation only use projection data to verify the

inspection objects. This adds a minor benefit to treaty verification, as the monitor would not see aggregated measurement data at the time that the test is performed. However, any information given to the monitor at the time should ultimately be considered theirs to own, so the projection data would still reveal sensitive information on the objects. The monitor could use the projection data and knowledge of the imaging system to reconstruct sensitive details on the object.

More importantly, with a well defined task, the statistical nature of the projection data can be utilized to improve task performance. This is because reconstruction methods, such as FBP or MLEM, ignore that statistical information. Past studies, including the threat-detection results discussed in Section 1.3, have demonstrated the advantages that can be provided by an observer that utilizes the statistical nature of the detector data.

1.6.2 List-Mode Processing

In addition to using projection data, methods were developed that process LM data in an attempt to remove the necessity of an IB when performing measurements of the unknown item. The use, and subsequent disposal, of LM events—which can be defined as the interaction of a particle in the detector—as they are acquired by the system means that an "image" is never actually formed either in terms of a projection image or a reconstruction of the object. As data is never aggregated, there is no sensitive information available to the monitor when testing sources. The execution of an example observer model that uses LM data is shown in Figure 1.18. In this example, a probability model is developed from calibration data acquired from imaging a scene with the signal present and absent. The model used in this study is the ideal observer, which is described in chapter 4. When testing an object, the test statistic is updated with each detected event. The greater the number of detected events, the more likely that the test statistic is correctly above or below the threshold and that the chosen decision is the correct one. The processing of testing data on the right side of Figure 1.18 is nonsensitive, as shown in Figure 1.19.

The requirement to process data in LM is very restrictive. It limits the mathematical models to linear models and models that can be represented by a product of terms containing the LM data. This is emphasized further by considering advanced classification techniques that have been developed in recent years. Some examples of these methods are tree classifiers (Safavian and Landgrebe, 1991), support vector

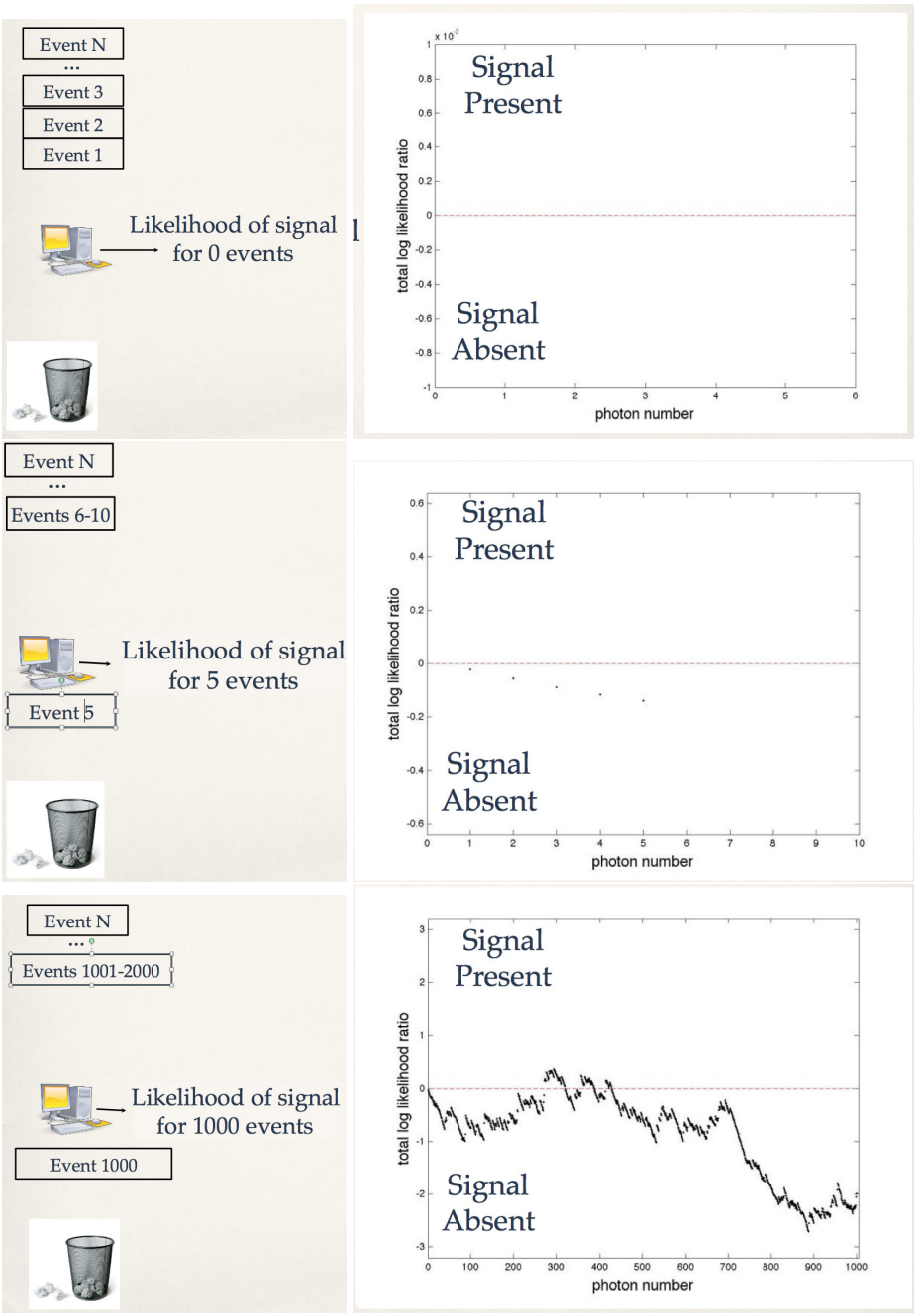


Figure 1.18: The left column depicts the LM processing of particles. Each detected particle is read in, its data updates a test-statistic, and then that data is purged from memory. On the right is the the ideal observer's test statistic (the log of the likelihood ratio), updated as each event is read in. In this example the threshold is set to zero to declare signal present or signal absent.

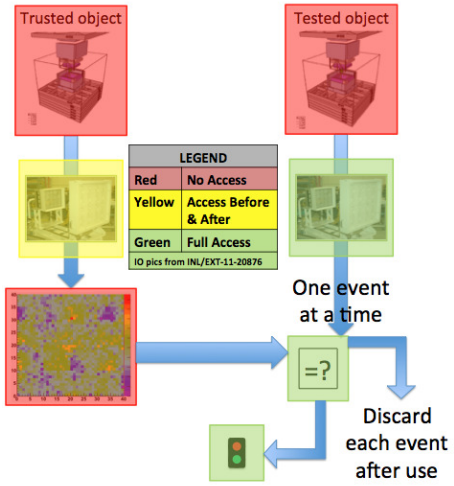


Figure 1.19: An intuitive picture exemplifying an observer model that processes LM data to make a decision. In this graphic, the model utilizes sensitive calibration data. When testing a source, only the test statistic, which is nonsensitive, is updated.

machines (SVM) (Schölkopf and Smola, 1998), and artificial neural networks (Dreiseitl and Ohno-Machado, 2002). It appears difficult to adapt some of these routines to process LM data. Random forest classifiers (Breiman, 2001; Liaw and Wiener, 2002) operate by constructing many random decision trees (where decisions could consist of a comparison of one of the data variables with a number) from bootstrapped samples of calibration data and then aggregating the tree results to make a decision. This classifier requires the aggregation of testing data to make decisions and such information would necessitate an IB. SVMs take the M dimensional data space and segment it using a function to make decisions. SVMs can take many forms, but for this purpose it is useful to split them into two categories—linear and non-linear. Non-linear SVMs would require knowledge of the complete data set. Only a linear SVM would satisfy the LM requirement. Similarly, artificial neural networks cannot process LM data because they use a nonlinear sigmoidal function after each node.

This dissertation does not discuss in significant detail the methods that could be used to enforce LM processing. One could imagine an electronic board attached to the PMT output that performs some mathematical operations from a predefined template and outputs a single number for the test statistic. The host would need to verify that the system does not aggregate spectral and spatial information.

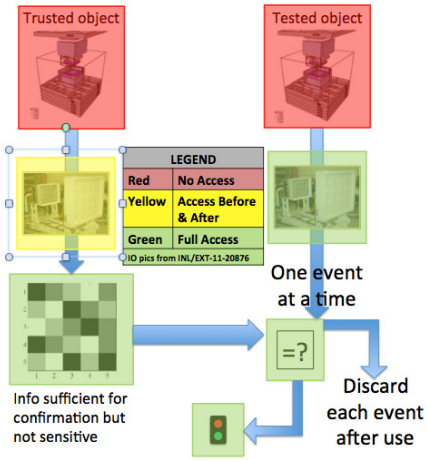


Figure 1.20: An intuitive picture demonstrating a nonsensitive observer model. As the acquired information necessary to build the model would still be sensitive, the only IB would be on this acquisition of calibration data. Otherwise, the monitor would have access to all of the information in the mathematical model.

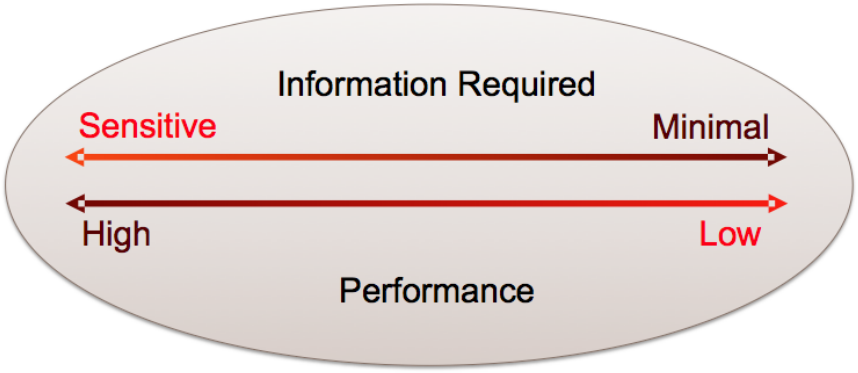


Figure 1.21: Model performance should scale roughly with the amount of sensitive information required.

1.6.3 Development of Observer Models that Store Nonsensitive Information

The ultimate goal for this work is the development of observer models that can perform various tasks without revealing sensitive information on the TAIs. This would allow the host to share the observer with the monitor without revealing sensitive aspects of its own TAIs. Such a model is depicted in Figure 1.20. This work discusses a range of observer models that store varying amounts of information. Generally, as more information is included, performance is expected to improve, as shown in Figure 1.21. The only information barrier required with such a model would be when acquiring calibration data on the trusted items.

While this thesis focuses on a template-matching approach, a similar framework could still be used for an attribute-estimation approach. Nonsensitive characteristics

related to sensitive details on the objects could be measured. This could be done by simply returning a red light/green light for certain features, such as if the mass of the plutonium in an object is greater than 500g. Significantly more nuclear material than this would be required for a high-yield weapon, so it may not be deemed a sensitive measurement.

CHAPTER 2

Radiation Detection for Arms-Control-Treaty Verification

Radiation is the transmission of energy in the form of waves and particles through a medium. For the purpose of arms-control-treaty verification, it is desirable to detect electromagnetic radiation (in the form of gamma rays) and particle radiation (neutrons) that were emitted from the TAI. Three specific types of radiation are discussed:

- Photons are quanta of electromagnetic radiation. In particular, gamma rays—photons with energies greater than 100keV—are considered in this work.
- Neutrons are neutrally charged subatomic particles that can travel long distances through dense materials and interact with other nuclei.
- Alpha particles, denoted by α, consist of two protons and neutrons bound together and are identical to the Helium nucleus (He4). As they have no electrons, they are positively charged.

In this chapter, photon, neutron, and alpha radiation are discussed and their relevant physical processes summarized (Section 2.1). Then, in Section 2.2, the important detectable features of TAIs are briefly discussed along with the usefulness of gamma ray and neutron measurements. The cause of background detections for gamma-ray and neutron measurements are explained in Section 2.3. Finally, various measurement systems designed for detecting gamma rays and neutrons are discussed in Section 2.4. This section borrows liberally from Reilly, Enselin, Smith and Kreiner's work, "Passive Non-destructive Assay of Nuclear Material", and is occasionally referred to as PANDA throughout this chapter. See (Reilly et al., 1991) for more details.

2.1 Physics of Fundamental Particles

This section discusses the physics processes relevant to treaty-verification tasks. Gamma rays (Section 2.1.1), neutrons (Section 2.1.2) and alpha particles (Section 2.1.3) are discussed in detail.

2.1.1 Gamma Rays

Gamma rays are emitted by radioactive decay processes. They are quanta of electromagnetic radiation, travel at the speed of light, and their energy is proportional to their wave frequency. They are not effectively attenuated by most material, making them a prime source of radiation to be measured.

The cumulative attenuation that does occur can be represented by Beer's law,

$$I(L) = I_0 e^{(-\mu_l L)}. \tag{2.1}$$

In (2.1), the intensity, I is the transmitted energy per unit second per area, I_0 is the intensity of the radiation entering a medium, μ_l is the attenuation coefficient (units 1/length) and L is the distance traveled through the medium. μ_l is dependent on the material type, density, and the energy of the gamma rays. The inverse of μ_l is often referred to as the path length or attenuation length of the gamma ray. Throughout this chapter, the mass attenuation coefficient $\mu = \mu_l/\rho$ is also used, where ρ is the material density. The mass attenuation coefficient is independent of density and directly relates the interaction probability to element number.

A plot of the attenuation coefficient of photoelectric absorption, Compton scattering, and pair production for a NaI crystal is shown in Figure 2.1. These physics processes are discussed in more detail in the following subsubsections.

2.1.1.1 Emission Processes

Certain isotopes, specifically those of uranium and plutonium, are unstable and naturally decay through alpha and beta radiation to daughter isotopes (see Figure 2.2). After decay, the resulting isotopes are often left in an excited state and decay through gamma emission to a more stable state. The emissions occur only on certain energy lines. The emission distribution of an object changes with isotope age and material composition.

`http://www.radiochemistry.org/periodictable/gamma_spectra/` is a good source to find the emission spectra for various materials. An example for U235 is shown in Figure 2.3. Of particular importance in this spectra is the 186 keV line, the most active line with a high enough energy to avoid being completely self-shielded in uranium objects. Additionally, U238 has a moderately intense peak at 1001 keV and plutonium 239 has important peaks in the 300-400 keV, 639-648 keV, and 756-769 keV ranges.

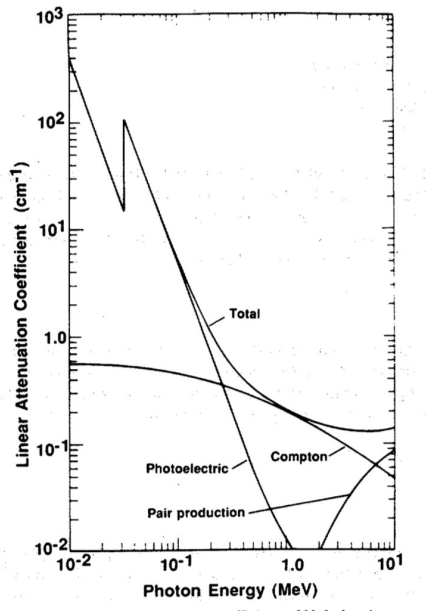


Figure 2.1: Plot of the attenuation coefficient for photoelectric absorption, Compton scattering and pair production in NaI scintillator (Reilly et al., 1991). Photoelectric absorption is the dominant process up to roughly 300 keV but falls off drastically with energy. Compton scattering is the dominant process from about 400 keV to 3 MeV. Pair production dominates at higher energies.

2.1.1.2 Physics in Transport

There are two processes that cause attenuation of gamma rays within the energy range of interest to this project; photoelectric absorption and Compton scattering.

Low-energy gamma rays primarily undergo photoelectric absorption (discovered by Heinrich Hertz in 1887 using ultraviolet light (Hertz, 1887)). The gamma ray interacts with a bound atomic electron, as shown in Figure 2.4, transferring all energy to the electron. Some energy goes to overcoming the binding energy, and the rest increases the kinetic energy of the ejected electron,

$$h\nu = E_e + E_b, \tag{2.2}$$

where ν is the frequency of the photon, $h\nu$ is the energy of the photon, E_e is the kinetic energy of the ejected electron and E_b is the binding energy of that electron to the nucleus. The probability of this interaction occurring increases with Z^4 (Z being the proton number) and falls off with $(h\nu)^3$ (Reilly et al., 1991).

Compton scattering (Compton, 1923) (shown in Figure 2.5) is the interaction of a gamma ray with a free or weakly bound electron, transferring a fraction of its energy to an electron initially at rest. The gamma ray loses energy and changes trajectory by an angle ϕ. That direction is related to the initial and final energy by

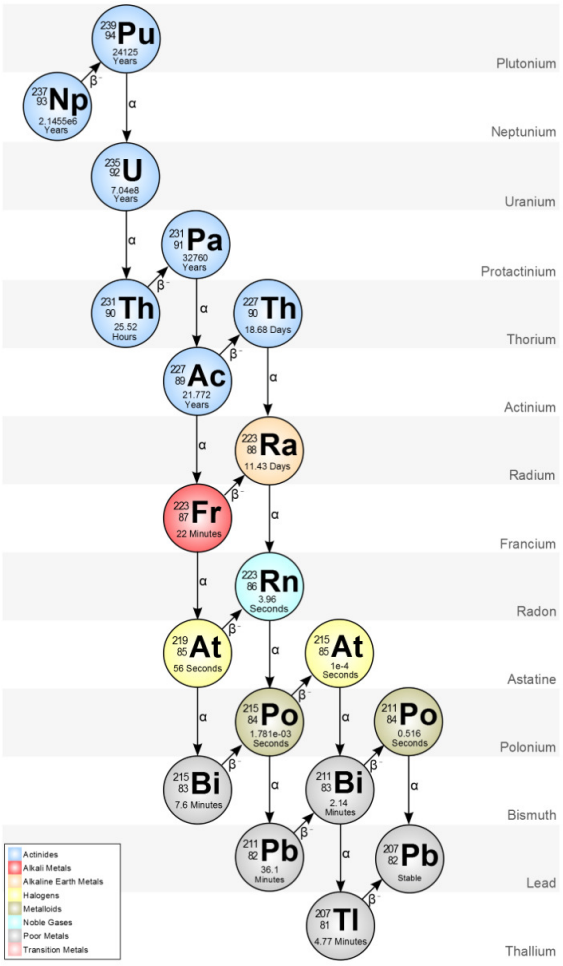


Figure 2.2: A diagram of the Pu-239 decay chain. Lawrence Berkeley's Nuclear Forensic Search Project has more decay chain examples at http://metadata.berkeley.edu/nuclear-forensics/Decay%20Chains.html.

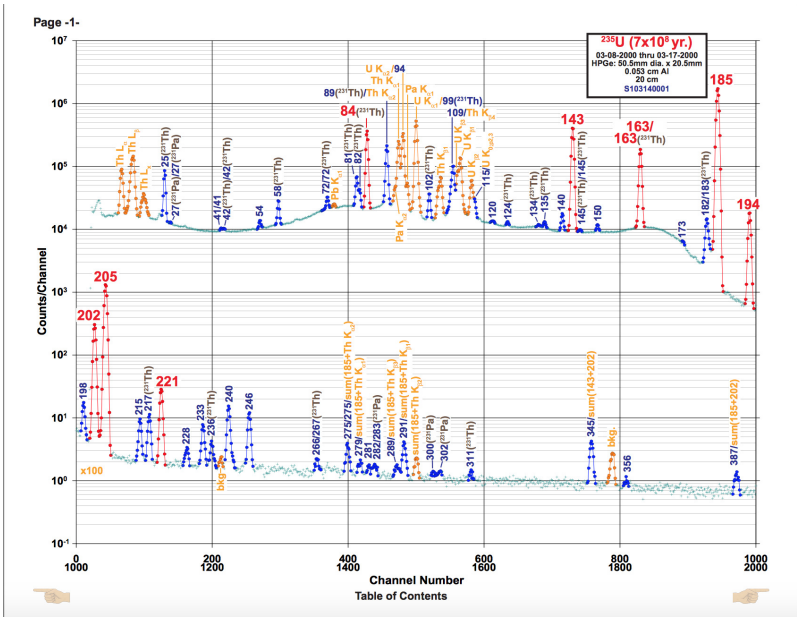


Figure 2.3: An example low-energy spectral measurement of U235 with a high-purity germanium (HPGe) detector.

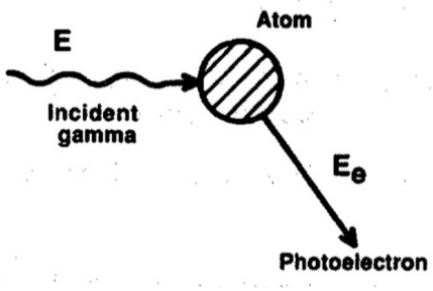


Figure 2.4: Diagram of photoelectric absorption—the incoming gamma ray deposits all of its energy into the electron, ejecting it from its atomic binding (Reilly et al., 1991).

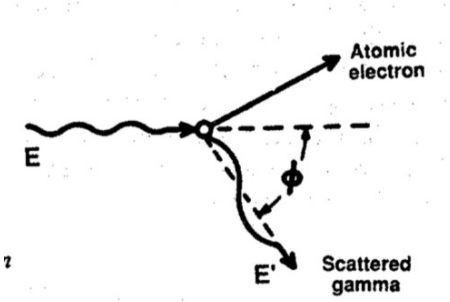


Figure 2.5: Diagram of Compton scattering. An incoming high-energy gamma ray scatters off a free at-rest electron, resulting in an energetic electron and scattered gamma. (Reilly et al., 1991).

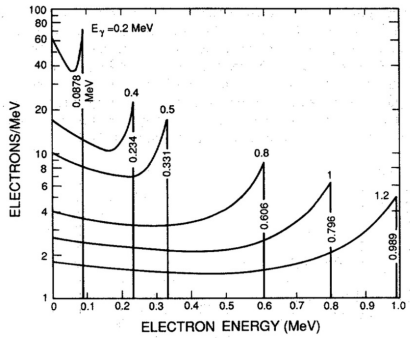


Figure 2.6: Example detected spectra when only Compton scattering interactions occur (Reilly et al., 1991). The maximum energy corresponds to a 180° deviation for the incident gamma ray. The number above each of the curves corresponds to the energy of the gamma ray before Compton scattering occurs.

the Klein-Nishina distribution (Klein and Nishina, 1929). Using the conservation of energy and momentum, one can derive the final energy E' of the gamma ray as a function of its incident energy E, the electron's rest energy $m_e c^2$, and ϕ,

$$\frac{1}{E'} - \frac{1}{E} = \frac{h}{m_e c^2}(1 - cos(\phi)). \tag{2.3}$$

When $\phi = \pi$, the resulting gamma is scattered backwards and the maximum amount of energy is transferred to the electron. A plot of the resulting energy distribution on the scattered gamma rays is shown in Figure 2.6.

Two other physics processes that involve gamma rays are pair production and photofission. Pair production occurs when a gamma ray passes close to the nucleus, creating an electron-positron pair. After the pair anhilates, two 511 keV gammas are created. Photofission occurs at very high energies, when an incoming gamma causes the ejection of nuclear particles. These two processes happen at high energies, mostly not of interest to the detection of TAIs. Combining the various processes, a total mass-attenuation coefficient is found (shown in Figure 2.7). A particular point

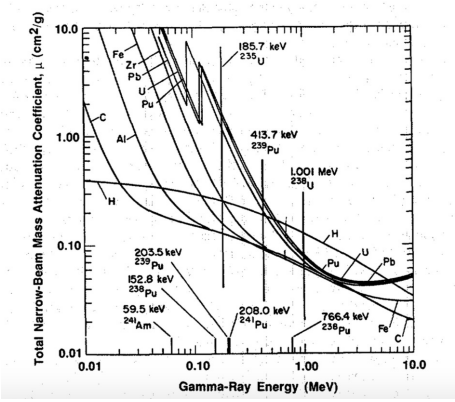


Figure 2.7: The total mass attenuation coefficient for many relevant elements (Reilly et al., 1991).

of interest is the 186 keV line, emitted from U235. This emission line is heavily attenuated by uranium, plutonium and lead.

2.1.2 Neutrons

Neutrons are particles that do not have a charge and are therefore able to travel long distances through most matter. The energy of a neutron is its kinetic energy, proportional to its velocity squared. Because neutrons vary drastically in energy and their interaction rates change accordingly, neutrons in different energy ranges are often given different names. Only fast (1-20 MeV) and thermal (room temperature, or around .025 eV) neutrons are discussed in this thesis. In this subsection, the physics of neutron creation are discussed as well as the important physics processes involved in transport. More details on the physics of neutron interactions in matter can be found in chapters 11 and 12 of (Reilly et al., 1991).

2.1.2.1 Emission Processes

Neutron emission occurs mostly through fission, either spontaneous or induced. Spontaneous fission happens when an isotope splits into two or more fragments, overcoming the nuclear force that binds the nucleus together and releasing a significant sum of energy. There are often multiple neutrons that result from a single fission event; the number of neutrons produced is denoted the multiplicity and follows an isotope-dependent probability distribution. Figure 2.8 shows the neutron emission rates for various isotopes. Generally, even isotopes are significantly more likely to fission than odd isotopes by three or more orders of magnitude. Another important fact to note is that plutonium isotopes, specifically Pu240 and Pu242,

Table 11-1. Spontaneous fission neutron yields

Isotope A	Number of Protons Z	Number of Neutrons N	Total Half-Life[a]	Spontaneous Fission Half-Life[b] (yr)	Spontaneous Fission Yield[b] (n/s-g)	Spontaneous Fission Multiplicity[b,c] ν	Induced Thermal Fission Multiplicity[c] ν
^{232}Th	90	142	1.41×10^{10} yr	$>1 \times 10^{21}$	$>6 \times 10^{-8}$	2.14	1.9
^{232}U	92	140	71.7 yr	8×10^{13}	1.3	1.71	3.13
^{233}U	92	141	1.59×10^{5} yr	1.2×10^{17}	8.6×10^{-4}	1.76	2.4
^{234}U	92	142	2.45×10^{5} yr	2.1×10^{16}	5.02×10^{-3}	1.81	2.4
^{235}U	92	143	7.04×10^{8} yr	3.5×10^{17}	2.99×10^{-4}	1.86	2.41
^{236}U	92	144	2.34×10^{7} yr	1.95×10^{16}	5.49×10^{-3}	1.91	2.2
^{238}U	92	146	4.47×10^{9} yr	8.20×10^{15}	1.36×10^{-2}	2.01	2.3
^{237}Np	93	144	2.14×10^{6} yr	1.0×10^{18}	1.14×10^{-4}	2.05	2.70
^{238}Pu	94	144	87.74 yr	4.77×10^{10}	2.59×10^{3}	2.21	2.9
^{239}Pu	94	145	2.41×10^{4} yr	5.48×10^{15}	2.18×10^{-2}	2.16	2.88
^{240}Pu	94	146	6.56×10^{3} yr	1.16×10^{11}	1.02×10^{3}	2.16	2.8
^{241}Pu	94	147	14.35 yr	(2.5×10^{15})	(5×10^{-2})	2.25	2.8
^{242}Pu	94	148	3.76×10^{5} yr	6.84×10^{10}	1.72×10^{3}	2.15	2.81
^{241}Am	95	146	433.6 yr	1.05×10^{14}	1.18	3.22	3.09
^{242}Cm	96	146	163 days	6.56×10^{6}	2.10×10^{7}	2.54	3.44
^{244}Cm	96	148	18.1 yr	1.35×10^{7}	1.08×10^{7}	2.72	3.46
^{249}Bk	97	152	320 days	1.90×10^{9}	1.0×10^{5}	3.40	3.7
^{252}Cf	98	154	2.646 yr	85.5	2.34×10^{12}	3.757	4.06

Figure 2.8: This chart shows the neutron emission rates per unit mass and the fission multiplicity. Of particular interest in neutron measurements is Pu240 and Pu242, each of which have a far higher emission rate than the remaining uranium and plutonium isotopes (Reilly et al., 1991).

have much higher emission rates than uranium isotopes. This is why when simulating neutron data in Section 3.7, neutron measurements are ignored for the two objects that only differ in their uranium composition. In addition to neutrons, fission events generally release 7 to 10 prompt high-energy gamma rays (Reilly et al., 1991). Delayed neutrons are also emitted far more rarely; these result from beta decay of the fission products.

While even isotopes are more likely to undergo spontaneous fission, odd isotopes of Pu and U are more likely to fission when interacting with a neutron. Figure 2.9 shows the fission cross-section of the various elements. Low-energy neutrons are drastically more likely to induce fission in the odd isotopes than high-energy neutrons.

The energy spectra of the emitted neutrons is defined by Watt's equation (Watt, 1952),

$$pr(E) = e^{(\frac{-E}{A})} sinh(\sqrt{BE}). \tag{2.4}$$

This form holds for many isotopes for both spontaneous and stimulated fission. Pu240, for example, has A=0.795 and B=4.69. Most isotopes have similar parameter values for the Watt spectra, and ultimately this makes the neutron-energy spectra somewhat useless for assay of TAIs.

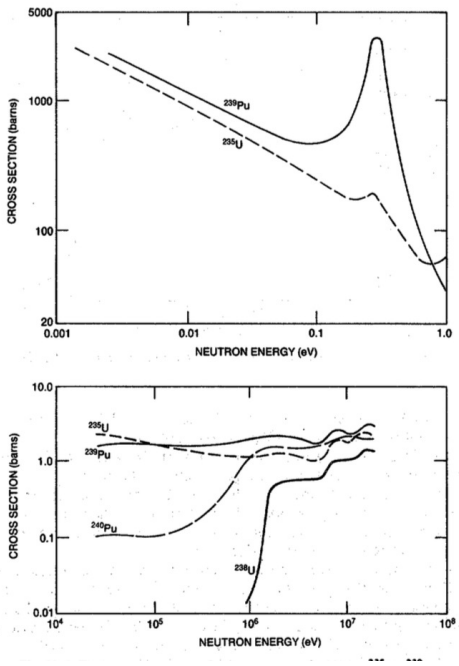


Figure 2.9: Fission cross sections for various isotopes (Reilly et al., 1991). Both Pu239 and U235 have significantly higher cross section values at higher energies.

2.1.2.2 Physics in Transport

In this section, the physics of the transport of neutrons through non-fissile material is discussed. In particular, elastic scattering and neutron capture are explained.

Elastic scattering happens when an incoming neutron collides with a nucleus, transferring some of its energy to that nucleus. Conservation of energy and momentum shows that greater energy transfer happens when the neutron collides with low Z nuclei. This is why hydrocarbons are often used as both a moderator (to dampen neutron energy) and in detector materials. Neutron capture occurs when a low-energy neutron is absorbed into a nucleus. The resulting isotope is generally in a higher energy state and releases a high-energy gamma ray (Reilly et al., 1991). Figure 2.10 gives scattering and capture information on various materials. Of particular interest is CH_2 in this table, which is comparable to the mask material for the detector used in this project. One attenuation length corresponds to roughly 0.37 cm at room-temperature and 2.22 cm around 1 MeV.

2.1.3 Alpha Particles

An alpha particle is essentially a He nucleus and is produced by alpha decay processes along the decay chain for the radioactive isotopes (Figure 2.2). It quickly loses energy when traveling through any medium due to electronic excitation and ionization.

Material	Atomic or Molecular Weight	Density (g/cm^3)	Cross Sections[b] E = 0.0253 eV σ_t (b)	σ_a (b)	Σ_t (cm^{-1})	Σ_a (cm^{-1})	E = 1 MeV σ_t (b)	σ_a (b)	Σ_t (cm^{-1})	Σ_a (cm^{-1})
Al	27	2.7	1.61	0.232	0.097	0.014	2.37	0.000	0.143	0.000
B	10	2.3	3845	3843	533	532	2.68	0.189	0.371	0.0262
B	11	2.3	5.28	0.005	0.665	0.0006	2.13	0.000	0.268	0.000
Be	9	9.0	6.35	0.010	3.82	0.0060	3.25	0.003	1.96	0.0018
C	12	1.9	4.95	0.003	0.472	0.0003	2.58	0.000	0.246	0.000
Nat Ca	40.08	1.55	3.46	0.433	0.081	0.101	1.14	0.004	0.027	0.0001
Cd	112	8.7	2470	2462	115.5	115.2	6.50	0.058	0.304	0.0027
Nat Cl	35.45	Gas	50.2	33.4	Gas	Gas	2.30	0.0005	Gas	Gas
Nat Cu	63.55	8.94	12.5	3.80	1.06	0.322	3.40	0.011	0.288	0.0009
F	19	Gas	3.72	0.010	Gas	Gas	3.15	0.000	Gas	Gas
Fe	56	7.9	14.07	2.56	1.19	0.217	5.19	0.003	0.441	0.0003
Nat Gd	157.25	7.95	49 153	48 981	1496	1491	7.33	0.223	0.223	0.0068
H	1	Gas	30.62	0.33	Gas	Gas	4.26	0.000	Gas	Gas
H	2	Gas	4.25	0.000	Gas	Gas	2.87	0.000	Gas	Gas
He	3	Gas	5337	5336	Gas	Gas	2.87	0.879	Gas	Gas
He	4	Gas	0.86	0.000	Gas	Gas	7.08	0.000	Gas	Gas
Li	6	0.534	938	937	50.3	50.2	1.28	0.230	0.069	0.0123
Li	7	0.534	1.16	0.036	0.053	0.0017	1.57	0.000	0.072	0.0000
Nat Mg	24.31	1.74	3.47	0.063	0.150	0.0027	2.66	0.001	0.115	0.0000
Mn	55	7.2	14.5	13.2	1.14	1.04	3.17	0.003	0.250	0.0002
N	14	Gas	12.22	1.9	Gas	Gas	2.39	0.021	Gas	Gas
Na	23	0.971	3.92	0.529	0.100	0.0134	3.17	0.000	0.081	0.0000
Ni	59	8.9	23.08	4.58	2.10	0.416	3.66	0.0008	0.322	0.0001
O	16	Gas	3.87	0.000	Gas	Gas	8.22	0.000	Gas	Gas
Pb	204	11.34	11.40	0.18	0.381	0.0060	4.39	0.0033	0.147	0.0001

Material	Atomic or Molecular Weight	Density (g/cm^3)	Cross Sections[b] E = 0.0253 eV σ_t (b)	σ_a (b)	Σ_t (cm^{-1})	Σ_a (cm^{-1})	E = 1 MeV σ_t (b)	σ_a (b)	Σ_t (cm^{-1})	Σ_a (cm^{-1})
Pu	238.05	19.6	599.3	562.0	29.72	27.87	6.66	0.190	0.330	0.0094
Pu	239.05	19.6	1021	270	50.4	13.3	7.01	0.026	0.346	0.0013
Pu	240.05	19.6	294	293	14.5	14.4	7.15	0.108	0.352	0.0053
Pu	241.06	19.6	1390	362	68.1	17.7	7.98	0.117	0.391	0.0057
Pu	242.06	19.6	26.7	18.9	1.30	0.922	7.31	0.098	0.357	0.0048
Nat Si	28.09	2.42	2.24	0.161	0.116	0.0084	4.43	0.001	0.230	0.0001
Th	232	11.3	20.4	7.50	0.598	0.220	7.00	0.135	0.205	0.0040
U	233.04	19.1	587	45.8	29.0	2.26	6.78	0.069	0.335	0.0034
U	234.04	19.1	116	103	5.70	5.07	8.02	0.363	0.394	0.0178
U	235.04	19.1	703	96.9	34.3	4.74	6.84	0.117	0.335	0.0057
U	236.05	19.1	13.3	5.16	0.648	0.251	7.73	0.363	0.377	0.0177
U	237.05	19.1	487.5	476.4	23.6	23.1	6.72	0.135	0.326	0.0066
U	238.05	19.1	11.63	2.71	0.562	0.131	7.10	0.123	0.343	0.0059
Nat U	238.03	19.1	16.49	3.39	0.797	0.1637	7.01	0.120	0.343	0.0058
Nat W	183.85	19.3	23.08	18.05	1.459	1.141	6.95	0.057	0.439	0.0036
CH_2	14	0.94			2.68	0.027			0.449	0.0000
H_2O	18	1.0			2.18	0.022			0.560	0.0000
D_2O	20	1.1			0.410	0.000			0.420	0.0000
Average Fission Products of:										
^{235}U	117		4496	4486			7.43	0.00036		
^{239}Pu	119		2087	2086			7.48	0.00093		

Figure 2.10: Table of cross sections and interaction lengths for various low Z and high Z isotopes. Σ_t gives the total attenuation coefficient and Σ_a the absorption coefficient of neutron capture. (Reilly et al., 1991)

Ultimately, this limits the utility of alpha particles in detection and classification of TAIs. Alpha particles have roughly a 4 cm path length in air and much less in denser materials.

2.2 Detectable Features of TAIs

This section summarizes the value that gamma-ray (Section 2.2.1) and neutron (Section 2.2.1) measurements can provide in performing warhead verification tasks.

2.2.1 Gamma Ray Measurements

Only certain isotopes of high-Z materials have a high probability of fissioning when interacting with low-energy neutrons, and these are referred to as fissile. An object that produces yield must contain fissile plutonium (Pu239) and/or uranium (U235). These isotopes also have a strong gamma-emission rate. If a monitor was able to take a high resolution measurement of the gamma spectra, they could back out the isotopic ratios of the uranium and plutonium in the TAI, determining the composition of U235 vs U238 and Pu239 vs. other plutonium isotopes. Similarly, if the statistics are high enough and spatial resolution good enough, gamma-ray imaging would offer the monitor confirmation that the imaged item could produce yield and is not just a hunk of SNM. Most of the intense emission lines for these fissile isotopes are low-energy photons, and these lines tend to be heavily shielded (Figure 2.7).

2.2.2 Neutron Measurements

Unfortunately, in part due to a low signal because of shielding, gamma-ray measurements are generally not sufficient for warhead verification. Even though spontaneous-fission events are far less likely to occur than gamma emission from radioactive decay, neutrons are easier to detect because the majority escape the TAI. The neutron background is also significantly less intense than the gamma background.

Neutron measurements provide multiple advantages. As fission produces multiple neutrons per event, the recording of multiple neutrons within a certain small time window is evidence of SNM being measured and not a potential background source. In addition, because neutrons are neutrally charged and are generally unlikely to interact in the TAIs, which mostly consist of high Z materials, they offer higher fidelity information on the geometric construction of the sources.

2.3 Background

Any measurement of gamma rays or neutrons also contains detected background particles. Natural radioactivity in the surrounding environment, cosmic-ray interactions and possible nearby objects that aren't of interest to the task produce a signal on the detector. Cosmic rays are high-energy particles that originate outside the solar system and produce secondary particles upon interacting in the earth's atmosphere. Other naturally occurring elements such as uranium, thorium and potassium also contribute to the gamma-ray spectra (Reilly et al., 1991).

In treaty-verification tasks, where measurements are taken with a stationary detector, the background should be constant with time. In a threat-detection task, as discussed in the introduction chapter, the background consists of many individual sources located throughout the environment. In addition, this background changes with time. Performing detection tasks with a locally-varying, time-dependent background is significantly more difficult.

Background neutrons contribute to the data as well, though less often and in fact were ignored in the simulation studies in this thesis. They are also produced through cosmic radiation and are naturally occurring in the ocean (Yamashita et al., 1966).

2.4 Detection

Radiation detection is playing an ever increasing role in our society. Possible uses range from early detection of certain cancers or diseases, to the detection of threats at home or abroad, to verification of nuclear warheads. In order to detect particles, there must be some way of converting the energy of the incoming particle into a digital signal. This section explores various methodologies to achieve this. It is ultimately desirable to have a detector that can measure the intensity of the source, energy spectra and be able to localize where the particle is coming from within the field of view.

Section 2.4.1 and Section 2.4.2 discuss methodologies for detecting gamma rays and neutrons, respectively. Section 2.4.3 provides a relatively brief summary on imaging high-energy particles with focus on a coded-aperture system. Section 2.4.4 gives more information on the detector response for scintillator detectors.

2.4.1 Gamma Ray Detection

Gamma rays strongly interact with matter due to their electromagnetic nature. Scintillation detectors and solid-state detectors are two detectors useful for performing measurements of gamma rays.

2.4.1.1 Scintillation Detector

A scintillator is a volume of material that can absorb energy from certain radiation and convert it to light waves (electromagnetic radiation with an energy on the order of 1eV). Gamma rays interact in the scintillator via Compton scattering and photoelectric absorption, producing ionized atoms and energetic electrons. These particles travel through the material, giving off photons as they interact at different locations throughout the scintillator. A photomultiplier tube (PMT) (Hamamatsu Photonics K.K., 2007) is often used to record the energy from the light waves; it converts the initial photon energy to electrons, then amplifies them (with low noise) to get an output signal proportional to the detected energy. It should also be pointed out that the ideal scintillator needs to emit photons at a different range of frequencies than it absorbs them; if the absorption and emission distributions overlapped, any produced signal would be lost. To force this separation between absorption and emission lines, impurities are often added to change the electronic band-gap structure. An example of this is thallium-doped sodium iodide. An example schematic of a scintillator detector can be found in Figure 2.11. Generally, scintillators do not have superb energy resolving capabilities. For example, NaI(Tl) achieves a resolution of about 14 keV at a deposited energy of 122keV (Reilly et al., 1991).

2.4.1.2 Solid State Detector

A solid-state detector (Figure 2.12) directly collects the charge produced by the gamma ray interactions inside the material. An example of a solid state detector is high-purity germanium (Twomey, 2003), which has very good energy resolution—about 0.5 keV for 122keV of deposited energy. Solid-state detectors also tend to be far more expensive.

2.4.2 Neutron Detection

As discussed in Section 2.1.2, neutrons only interact with other nuclei. There are therefore only two ways to perform neutron detection.

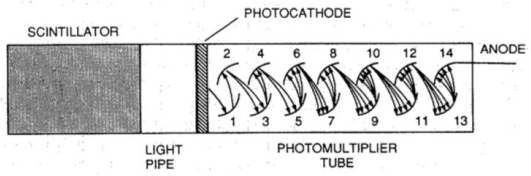


Figure 2.11: Schematic of a scintillation detector. The scintillator produces a light signal that passes through an intermediary light guide until it reaches the PMT (Reilly et al., 1991). In order to prevent loss of light leaving the scintillator, mirrors or reflective substances are usually placed around the remaining sides. In order to pass the light from the scintillator to the PMT, a light guide is often used to allow the light rays exiting the scintillator to spread out, proportionally hitting more than one PMT (only one is shown here). To prevent loss due to significant changes in index of refraction between the various components, the light guide must be chosen carefully.

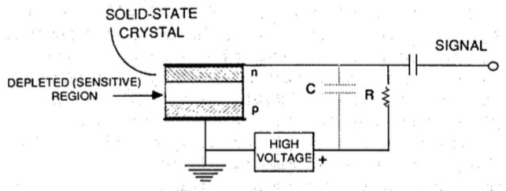


Figure 2.12: Schematic of solid-state detector (Reilly et al., 1991)

One option is to use a material that induces elastic scattering, and detect the energy of the resulting ionized particles. This method generally works well with fast neutrons, which have a higher probability of scatter than capture or fission. Both solid state and scintillator detectors can be used for this purpose. The downside to using a scintillator is that it generally has high gamma sensitivity as well. In addition, while the neutron-hydrogen energy conversion efficiency can be up to 100%, only a maximum of 28% can be transferred to a carbon nucleus due to the difference in masses. The advantage gained through a scintillator detector is that a dense material can be used, dramatically increasing detection efficiency.

The second detection strategy is to detect the fission products, gamma rays, alpha particles and protons that result from nuclear reactions. As shown in Figure 2.9, these reactions are far more common for lower energy neutrons. A common detector used to count neutrons is a He3 gas detector (Batchelor et al., 1955). He3 interacts with slow-moving neutrons, yielding hydrogen, tritium, and excess energy in the daughter products that are detected.

Coincidence detectors (Archer et al., 2010; Enqvist et al., 2008) are gaining prominence as well. The detection of multiple neutrons within a sufficiently small

time window is a sign that SNM is nearby. These detectors are placed in close proximity to the objects being measured, and are often themselves cylindrical in nature to surround the object.

2.4.3 Imaging

High-energy particle imaging works quite differently from optical imaging. The imaging of light (Barrett and Myers, 2003; Greivenkamp, 2004) can be done using curved glass (lenses), which serve to focus the incoming light onto an image plane. Meanwhile, it is impossible to bend gamma rays and neutrons through the use of some intermediary material, so imaging needs to be done mathematically through the methods discussed in this section. In order to localize the detected-particle interaction, the detector plane is discretized, yielding many output signals across a range of pixel locations.

To properly categorize the object's spatial emission distribution, many projections of the object onto the imager at certain angles (called slices) must be taken. This is called tomography, and tomographic reconstructions are possible with either of the following image methodologies (Hsieh, 2009; Cree and Bones, 1994).

2.4.3.1 Coded Aperture

A coded-aperture imaging systems prevents the passage of gamma rays and neutrons at certain locations and momentum directions. Lead (for gamma rays) and polyethylene (for neutrons) are two mask materials that effectively attenuate incoming radiation from specific directions. When imaging, there is a tradeoff between sensitivity and spatial resolution. For example, in a wide-area search application where the detected signal is minimal, it behooves the monitor to use large detector pixels to acquire high statistics in order to effectively perform detection tasks. This comes at the cost of reduced spatial sensitivity. My work, discussed in Section 1.3.5, used the NaI detector developed by Ziock et al (Ziock et al., 2006).

A simple pinhole mask would offer high resolution but very little sensitivity to the imaged object. To overcome this, coded aperture masks (Fenimore and Cannon, 1978) are commonly used to image high-energy radiation. Roughly 50% of the mask is filled, and the other 50% is holes, allowing for much higher sensitivity than a simple pinhole imager. These masks are uniformly redundant arrays, designed so that any shift in the mask from its initial state results in roughly half of the initial

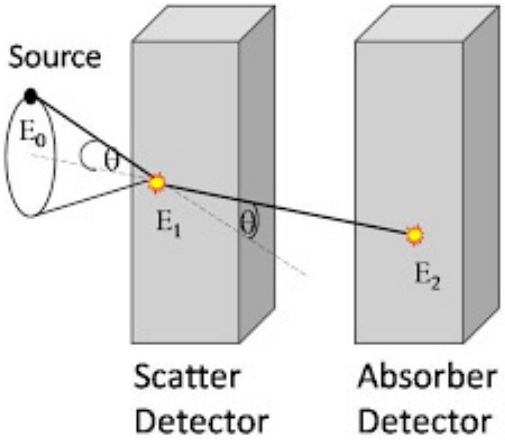


Figure 2.13: Schematic of the Compoton camera. Given the absorbed energy values in the scattering and absorption plane, the angle ϕ in (2.3) can be determined. The emission location must have ocurred somewhere on a cone in the object plane.

holes being covered. Expressed another way, the autocorrelation function peaks at zero, then sees a significant drop for any shift in mask location. A mask allows for imaging in all three dimensions, as a change in the location of the object leads to a change in the mask shadow.

2.4.3.2 Compton Imaging

Localization can also be done via Compton scattering if gamma rays are being detected. In the case of neutrons, two plane cameras can be used by taking advantage of elastic scattering (Poitrasson-Rivière et al., 2015). In this approach to detect high energy particles, two image planes are used, as in Figure 2.13. The first plane serves to scatter the incoming particles, and the second absorbs the lower-energy remnants. Using knowledge of the trajectory between the two planes and the deposited energies, it is possible to back out a cone of likely interaction locations (Cree and Bones, 1994).

2.4.4 Detector Response for Scintillation Detector

The detector used throughout this thesis (see Section 3.1) is designed to detect fast neutrons and utilizes a liquid scintillator. Three components of a scintillator detector's response to absorbed radiation are discussed: the particle recoil conversion efficiency, energy resolution, and pulse shape discrimination. More information on this detector can be found in the following chapter.

2.4.4.1 Detected Recoil Energy

After a scattering event, an energetic electron (Compton scattering) or proton (elastic scattering off H) travels through the medium. The light output for the two particle types is energy and material dependent, and the proton energy conversion efficiency is much less than that of electrons. As a quick example, a 60keV electron produces roughly the same light output as a 500keV proton (Reilly et al., 1991). For this reason, the detected neutron energies are expressed in keVee, with the last two es standing for "electron equivalent". Using the above example, a 500keV neutron would be detected at 60 keVee.

2.4.4.2 Energy Resolution

A detector's energy resolution is defined by its ability to resolve two energy peaks that are close together. This is inherently limited by the statistics of the produced electrons during scintillation. A NaI scintillator, for example, has a certain scintillation efficiency and produces roughly 1,000 electrons for a detected 300 keV photon (Reilly et al., 1991). The number of observed electrons is actually a random variable due to various inefficiencies with the detection process, and is governed by Poisson statistics. For a NaI scintillator, the energy resolution of a 300 keV photon is roughly 22.6 keV.

2.4.4.3 Pulse Shape Discrimination

Because organic scintillators have a high sensitivity to gamma rays, there needs to be a methodology to distinguish detected gamma rays from neutrons. This is done based on the electronic output signature from the PMTs. Neutron and gamma-ray detection results in different decay times of the pulses that arrive when a particle is detected. PSD is often done based on a measure of their delayed fluorescence (Adams and White, 1978), but classification is often imperfect and results in misclassification of gamma rays as neutrons, an important factor to take into account because the neutron detection rate is often significantly less than the gamma ray detection rate. This is discussed further in chapter 7.

CHAPTER 3

Data Simulation

Ideally, the data used to test the observer models developed in this work would be acquired from real life experiments. However, the cost of acquiring experimental data was prohibitively expensive for this project. The number of approved, unclassified inspection objects is also limited, and any existing data proved to be difficult to obtain. Instead, all data was simulated using the Geometry And Tracking (GEANT4) toolkit (Allison et al., 2006; Agostinelli et al., 2003), used to simulate the passage of particles through matter. GEANT4 is open source software, written in c++ and developed by the high-energy physics community. It is often used in particle-physics and nuclear-science applications. It contains approximately one million lines of code and over two thousand classes. In this chapter, all GEANT4 class names are italicized for clarity.

This chapter begins with a description of the detector in Section 3.1. Section 3.2 defines the different tasks that the developed models are performed on. A brief introduction to GEANT4 is given in Section 3.3 in order to familiarize the reader with the software. A summary of the variance reduction (VR) techniques used to speed up the simulations is described in Section 3.4 along with the results from a case study. Section 3.5 discusses the various classes chosen for particle emission, physics, transport, VR, and detection for the simulations used in this thesis.

3.1 Detector Description

The detector used throughout this thesis was the fast-neutron coded-aperture detector, developed by Sandia National Laboratories (SNL) and Oak Ridge National Laboratories Hausladen et al. (2012). This detector was chosen for a few reasons. First, two of the members of our collaboration contributed to its design. Second, as it is occasionally housed at the lab, the opportunity existed to acquire experimental data on some of these objects and see how the models developed in this thesis perform in practice.

The detector was designed to image high-energy neutrons (>1MeV) for arms-control-treaty verification tasks. It uses a high-density polyethylene coded aperture

mask to image neutrons. The mask attenuate neutrons via elastic scatter (PANDA cites a path length of 2.22 cm for CH_2) (Reilly et al., 1991). It is a rank 19 modified uniformly redundant array. The detector is split into 4x4 segregated blocks. Each block has 10x10 pixels filled with a liquid scintillator, with each pixel being 1 cm x 1 cm x 5 cm. The pixels are optically separated. Each detector block has its own light guide and 2x2 array of PMTs. The detector response software determined the pixel ID by using the relationship between the four PMT outputs, though spatial resolution is somewhat poor in part due to the low number of PMTs used. More information on the detector response calibration measurements can be found in Section 7.1.1.

Though the detector for this project is designed to detect neutrons, it also serves as a low-resolution gamma detector. The mask also provides some ability to image gammas due to a small but not insignificant scattering rate in the mask. Hydrogen, for example, has a mass attenuation coefficient of 0.2 cm^2/g at 400keV. For polyethylene, that would correspond (roughly) to a 5 cm attenuation length. A picture of the detector is shown in Figure 3.1. The front face of the detector is covered by a quarter inch lead plate; this plate effectively serves to attenuate low energy gammas (see Figure 2.7). Neutrons at 1MeV have about an 8 cm, or 3.2 inch interaction length in lead, so this lead plate is transparent to them.

The detector used in the GEANT4 simulations is somewhat different from the current iteration of the detector. The simulated detector uses a liquid scintillator of material EJ-309 (Eijen Technology, 2010) while the current iteration uses a plastic scintillator. The current detector has improved PSD performance. The scintillator is mostly composed of carbon and hydrogen atoms. Incoming neutrons elastically scatter as discussed in chapter 2, ejecting protons from the nucleus.

The detector system was simulated with a source-to-mask distance of 70.5 cm, mask-to-detector distance of 60 cm, mask-element size of 1.21 cm and mask thickness of 6.95 cm. In the simulations, the imaging axis was the $\hat{z}$ axis.

3.2 Treaty Verification Tasks

This section describes the various tasks used to gauge the observer models that are discussed in chapter 4 and chapter 5. All measurements in these tasks are passive.

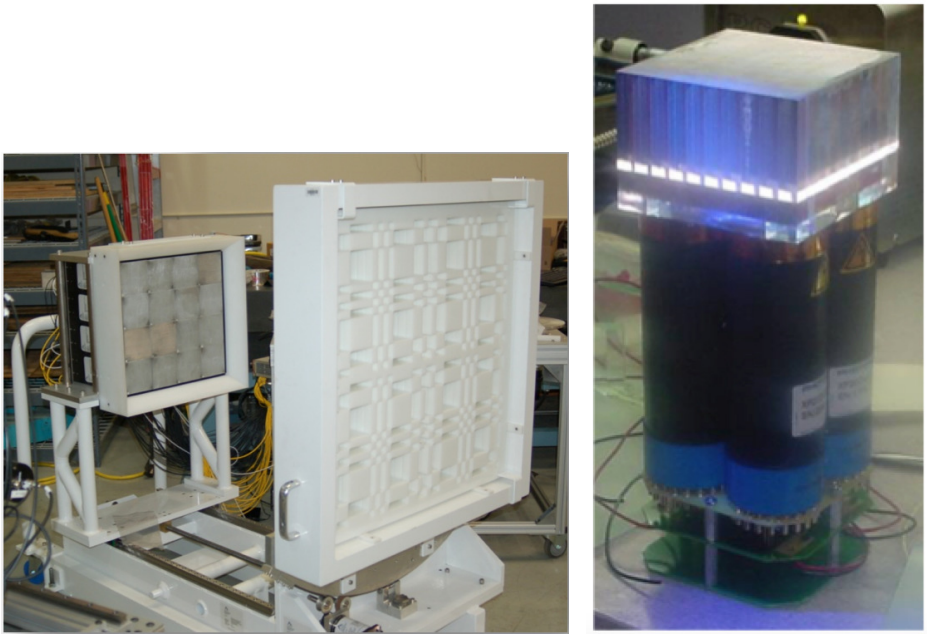


Figure 3.1: On the left is a picture taken of the fast-neutron coded aperture detector. The polyethylene mask is in front of the detector on the right. The quarter-inch lead plate that is normally in front of the detector is not shown in this picture so that the detector blocks are visible. On the right is a close up of the composition of an individual detector block.

3.2.1 Idaho Inspection Objects

Binary-classification and null-hypothesis tasks were performed using inspection objects developed by Idaho National Laboratory (INL) (Neibert et al., 2010); this thesis uses inspection objects labeled 8 and 9 (see Figure 3.2), which are referred to here as IO8 and IO9. These geometries are built by stacking rectangular plates of similar size. Both objects have a geometrically identical hollow plutonium core. The Pu material consisted of 94% Pu239 and 4.1% Pu240 by mass with other elements accounting for the remainder. IO8 surrounds the Pu core with depleted uranium (99.8% U238) and IO9 shields the Pu core with highly-enriched uranium (93% U235). As discussed in the physics section, this causes a significant difference in emitted gamma spectra. U235's 186 keV line, while very intense, is hard to detect as it is self shielded by the geometry. The 1001 keV line of U238, while less intense, is only lightly attenuated when traveling through uranium. There is also a slight difference in the gamma images of these sources due to the different shielding geometries. The neutron information for this task was ignored due to the overall similarity of the geometries.

The objects were imaged with their vertical axis in construction aligned with the

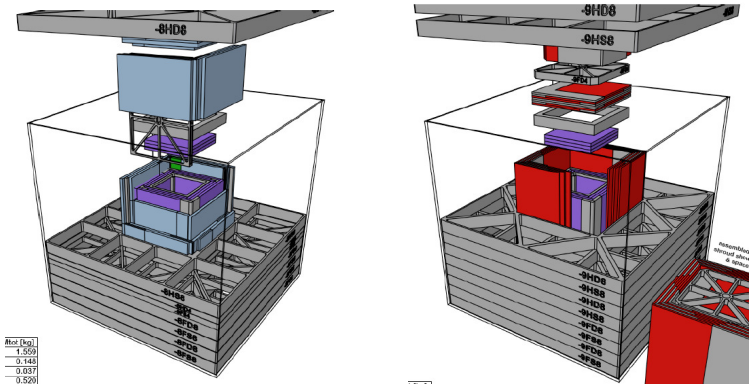


Figure 3.2: IO8 and IO9 developed by INL (Neibert et al., 2010). IO8 is plutonium shielded by depleted uranium (DU) while IO9 is plutonium shielded by highly-enriched uranium (HEU). Both assemblies are supported by an aluminum framework inside an 8” × 8” × 8” aluminum box that is 1" thick.

imaging axis.

3.2.2 BeRP Ball Location Study

In the second task the Beryllium Reflected Plutonium (BeRP) ball (Mattingly, 2009) was modeled. It is a 3.79 cm radius solid plutonium sphere. In this work, only the bare plutonium sphere was simulated. See Figure 3.3 for details. It is composed of 93.7% Pu239, 5.9% Pu240, 0.3% Pu241, and the remaining tenth of a percent consisted mostly of Pu238, Pu242, and Am241. The source was imaged at two locations in the x-y plane—once at (0 cm, 0 cm) and once at (2 cm, 2 cm). The models were trained on this data and asked to discriminate future images as being at one of the two locations.

3.2.3 2D Circle vs. Square Source

In this study, plutonium ring and square sources were simulated with lengths ranging from 10 cm to 30 cm. The sources were 1mm thick in the $\hat{z}$ dimension (along the imaging axis) and 1mm thick in the transverse dimension. The same plutonium composition as the BeRP ball was used. The observer models were tasked with discriminating an unknown object as one of the two geometries.

3.3 Introduction to GEANT4

Monte Carlo physics simulation was accomplished using the GEANT4 toolkit; particles are processed one by one, from emission until the particle either loses all of

Figure 3.3: The BeRP ball was developed by Los Alamos National Laboratory in 1980. The acronym stands for beryllium-reflected plutonium, but that is an anachronism. Today, the Pu sphere is surrounded by polyethylene shells which serve to scatter neutrons and increase the k_{eff} of the source. It is often used as a source to verify Monte Carlo transport code as well as an inspection object for treaty verification.

its energy or exits the world geometry. Every GEANT4 emission process is labeled as an event, and the particles resulting from this process labeled primary particles (of which there may be more than one). Every GEANT4 simulation requires the user to define a run manager (*G4RunManager*) that needs a primary generator action (*G4VUserPrimaryGeneratorAction*), a set of particles to track and physics processes to simulate (*G4VModularPhysicsList*), and a detector and physical geometry description (*G4VUserDetectorConstruction*). There are also a multitude of optional user classes that allow the user to interact with the simulation. This section introduces the reader to some of the important features of GEANT4.

3.3.1 Physics

GEANT4 has a large number of physical processes available for the user. The user chooses which particles to track when defining the physics list. Using the particle's process manager, the user can add various physical processes. For example, the user can register a high-precision elastic-scattering process to the neutron particle. There are also commonly used physics lists that can be referenced rather than defining each particle and process separately (such as *G4EmLivermorePhysics*, which simulates high-fidelity EM physics processes down to low energies).

GEANT4 includes methods for production cuts for secondary particles. The user sets a minimum distance that the secondary particle needs to travel. If GEANT4 calculates that it will not reach that minimum distance, the manager instead deposits

that energy at the current position. This can be used to speed up simulation time rather than simulating all particles down to very low energies.

3.3.2 Tracks and Steps

A step is defined as the distance from one physics process or geometry boundary crossing to the next, and the *G4Step* class contains information on how the particle's data changed from one step point to the next. The *G4StepPoint* class includes the particle's information at that location of the geometric boundary crossing or physics interaction. The user can define a *G4UserSteppingAction* class that retrieves information from the particle's track at each step, or terminate the particle if it reaches a certain set of conditions.

The *G4Track* class contains all of the current information about the particle at the end of its current step. It tracks the position, momentum, energy, time, current volume and material, and the next volume and material.

3.3.3 Geometry

Geometries in GEANT4 are defined by three classes—*G4VSolid*, *G4LogicalVolume*, and *G4VPhysicalVolume*. The *G4VSolid* class is the base class used to define physical geometries. GEANT4 has most desired geometries already available—rectangular solid, sphere, tube, cone, tetrahedra, and generic polygon classes— all of which are inherited from *G4VSolid*. There are also classes to unify or subtract solids. Each *G4VSolid* is assigned to a *G4LogicalVolume* class along with a material type. The *G4VPhysicalVolume* class is a placed instance of a logical volume inside the world volume.

3.3.4 Detector

The user-defined *G4VUserDetectorConstruction* class contains the world geometry and all object and detector geometries. The run manager calls the *Construct()* function in this class before initializing a run, and the various physical geometries are constructed and placed. The detector construction class requires at least one *G4VSensitiveDetector* object to be assigned to a geometry; otherwise, the run manager aborts and returns an error. The sensitive detector class takes a physical volume as an input. For each event, GEANT4 records a collection of the particles interacting in each detector. Each detected particle causes the creation of a user-defined

G4VHit object in GEANT4. The user can define what information gets recorded with each detector interaction, such as the location of interaction, change in energy, physics process, etc. The user can then access this collection of *G4VHit* objects for each event. The user can then define methods to process this information and output the data.

3.3.5 Primary Generator Action

GEANT4 has built-in classes and libraries to handle radioactive-decay processes for gammas and spontaneous-fission processes for neutrons. The user needs to define the location, momentum direction, particle type and energy of every emitted particle. It can do this through a basic *G4Gun* class that GEANT4 provides or through its own custom built methods. Regardless, each *G4VPrimaryGeneratorAction* class must include a *GeneratePrimaries()* function that calls *GeneratePrimaryVertex()* to create an emission event.

3.4 Variance Reduction in GEANT4

This section presents an overview of methods that can be used to speed up GEANT4 simulations. The goal of VR techniques is to reduce the variation in each data bin for the same amount of simulation time. VR techniques increase the likelihood that each emitted particle is detected, and assign a weight W to each emitted particle to prevent biasing of the detector data. An unbiased simulation would have all weights equal to one. A deeper summary of VR measures can be found in the MCNP primer (Shultis and Faw, 2011). MCNP (Briesmeister et al., 1986) is an alternative particle transport code.

This section begins with a discussion on the statistics used to gauge VR efforts (Section 3.4.1). Then, primary particle biasing (Section 3.4.2), importance sampling (Section 3.4.3) and weight windowing (Section 3.4.4) are explained. Finally results from my own efforts to utilize VR in GEANT4 are discussed in Section 3.4.5.

3.4.1 Statistical Measures to Gauge the Effect of Variance Reduction

The MCNP manual outlines ten statistical checks that can be used to decide whether the statistics are high enough, and VR methods effective enough to trust the simulated data. There are five metrics analyzed—the mean, relative error, variance of the variance, VR figure of merit, and a history score pdf (Tatsumi, 2012; Arce

et al., 2007). The statistic examining the history score pdf was not included in this work. GEANT4 does not have built in methods to track these statistics for a detector; as part of this work, these statistics have been coded into a post-processing routine that happens after the detector-response code is implemented. Ideally, the conditions themselves would be the stopping criteria for the simulation.

In the following subsections, the detector data is denote a vector $\mathbf{x}$ that contains all of the binned data. The user can decide whether each x_m should be the number of counts that are detected by a given pixel, or the number of counts in any pixel-energy bin, or some other definition.

3.4.1.1 Mean Data

The mean is the average weight in each detector bin,

$$\overline{x_m} = \frac{\sum_{n=1}^{N} W_{m,n}}{N} \tag{3.1}$$

where N is the total number of events processed in GEANT. The first condition is,

1. There is a nonmonotonic behavior in the estimated mean for each detector bin as a function of the number of events N over the last half of the problem.

3.4.1.2 Relative Error

The relative error is widely considered the most important statistical check, and was the check emphasized throughout this work. The relative error is defined as the standard deviation on the estimate of the mean divided by the mean (it can also be thought of as the inverse of the SNR),

$$R_m = \frac{S_{\overline{x_m}}}{\overline{x_m}} \tag{3.2}$$

There are three tests related to the relative error,

2. An acceptably low magnitude on the relative error (0.05 is a standard value).

3. Monotonically decreasing R as a function of the number of histories N for the last half of the problem

4. A $1/sqrt(N)$ decrease in R as a function of N over the last half of the problem.

For a simulation without VR, a relative error below 0.05 would correspond to 400 detected counts.

3.4.1.3 Figure of Merit

The figure of merit is defined as,

$$FOM = \frac{1}{R^2 T} \tag{3.3}$$

where R is the relative error, defined in (3.2) and T is the simulation time. More intuitively, it could be thought of as the SNR^2 (which increases with N) divided by N. The conditions on the figure of merit are,

5. A statistically constant value of the FOM as a function of N for the last half of the problem.

6. A nonmonotonic behavior in the FOM as a function of N over the last half of the problem.

3.4.1.4 Variance of the Variance

The variance of the variance (VoV) can be thought of as the accuracy of the estimation of the relative error R and is defined as,

$$VOV = \frac{S^2(S_{\bar{x}}^2)}{S_{\bar{x}}^2} \tag{3.4}$$

The VoV uses the third and fourth moments of the weight distribution for each bin. The statistical conditions for the VoV are:

7. Magnitude should be less than 0.1

8. Monotonically decreasing VoV as a function of N for the last half of the problem

9. A $1/N$ decrease in the VoV as a function of N for the last half of the problem

3.4.1.5 Discussion on Sufficient Data for Task Performance

These VR statistical measures give the user confidence in the data measured in simulation for a given object. For confidence in a binary-classification task-performance metric, these statistics could be used to gauge the difference between two data sets. This is a much stricter condition. The same statistics could be used, but rather than finding statistics of the data $\mathbf{x}$, the difference between the two data sets, $\Delta\mathbf{x}$, would be the chosen random variable. Any averages would be over the data distributions for each object.

3.4.2 Primary Particle Biasing

Primary-particle biasing is a biasing of the emission distribution for the particles. Properly done, it emphasizes the emission of particles that are more likely to hit the detector. Generally, this is a distribution on the location $\mathbf{r}$, momentum $\mathbf{p}$ and energy E for a given source. The emission probability can be expressed as $pr_{unbiased}(\mathbf{r}, \mathbf{p}, E)$. The user creates a new sampling distribution $pr_{biased}(\mathbf{r}, \mathbf{p}, E)$. Because the user is sampling more often from more interesting sections of the phase space, more particles are detected given the same number of emissions. To offset this, each particle is assigned a weight,

$$W = \frac{pr_{unbiased}(\mathbf{r}, \mathbf{p}, E)}{pr_{biased}(\mathbf{r}, \mathbf{p}, E)} \tag{3.5}$$

As an example, this thesis considered a linear bias on the energy distribution of emitted gamma rays. This is because lower-energy gammas are unlikely to escape the object and furthermore, unlikely to be detected. The equations describing the energy biasing are below,

$$\begin{aligned} Pr_{unbiased}(E_m) &= \frac{I(E_m)}{\sum_{m=1}^{M} I(E_m)} \\ Pr_{biased}(E_m) &= \frac{E_m I(E_m)}{\sum_{m=1}^{M} E_m I(E_m)} \\ W_m &= \frac{\sum_{m=1}^{M} E_m I(E_m)}{E_m \sum_{m=1}^{M} I(E_m)}. \end{aligned} \tag{3.6}$$

In the above equations, M is the total number of emission lines, E_m is the m^{th} emission energy and $I(E_m)$ is the intensity of that line. For complex geometries, location and momentum direction biasing is riskier. Due to scattering inside the object, particles emitted in opposite directions can still hit the detector. Hence, any biasing on these probabilities needs to be moderate.

Primary particle biasing is often used for a small increase in VR. Excessive primary biasing leads to unlikely, but not insignificant, parts of the emission phase space being sampled rarely and the corresponding weight being too high, leading to a simulation that increases the time needed to achieve confidence in the data set.

3.4.3 Geometric Importance Sampling

Importance sampling increases the likelihood that a particle reaches the detector regardless of where it is emitted. In GEANT4, the user can overlay parallel geometries

on top of the mass geometry, which is where the physics and detection processes occur. In the parallel geometry, the user can define a mesh of geometries with different importance values, where a higher importance value corresponds to a higher probability of detection at that location. When a particle moves from geometry m, with importance I_m to a new geometry n with importance I_n, a ratio $r = \frac{I_n}{I_m}$ is defined. Generally, it is easiest to assign importance values that are a power of 2. The behavior of the system then depends on the value r takes on,

$$\begin{aligned} &\text{If } r > 1, \text{ split into } r \text{ tracks, reduce track weight by } 1/r. \\ &\text{If } r = 1, \text{ continue tracking.} \\ &\text{If } r < 1, \text{ kill track with probability } (1 - r). \end{aligned} \tag{3.7}$$

The user can set up different parallel geometries for different energy ranges and particle types. It is often recommended to choose distances between importance cells roughly equal to the particle's interaction length at that energy. This procedure results in on average one particle escaping the object for each emission.

3.4.4 Weight Windowing

Weight windowing is a method used to keep the weights within a user defined range. Weight windowing can be applied to every physical cell and energy range independently. The procedure is outlined in Figure 3.4. Weights below the lower weight bound are killed off with probability probability $\frac{W}{W_L}$, while weights above the upper weight bound are split into a number of particles related to the ratio between the initial weight and upper weight limit. The resulting weight (if the initial weight was outside the range) is set to the survival weight. Weight windowing is often used with supplementary VR methods, such as cross-section biasing on the scattering distribution.

3.4.5 Brief Comments on Implementation in GEANT

GEANT4 has built in methods to perform importance sampling and weight windowing. A secondary parallel geometry can be laid on top of the physical geometry, with different geometries being given different importance values or weight windows. Using importance sampling, with the necessary classes defined, a given particle moves into a high importance cell and is split. The initial particle's weight changes and a

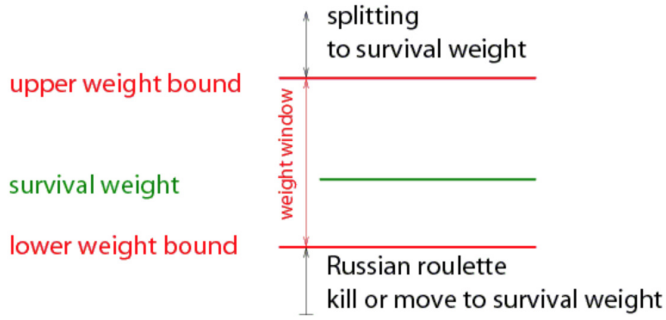


Figure 3.4: The user defines a survival factor C_s and upper limit factor C_u for the whole problem. The user supplies each space-energy cell with a lower weight bound W_L.

second particle is produced. GEANT4 also has scoring classes available to the user to calculate the sum of the weights and other statistics.

Implementation is difficult due to the detector response. For an unbiased simulation, the detector response for an event sums the detected energies and bins the total energy into a mean pixel ID. For a simulation with importance sampling or weight windowing, all particles created from a splitting along an importance boundary must be treated as a unique event. I developed these methods myself. In the *G4UserSteppingAction* class, I tracked all new particles created due to splitting along a parallel-geometry boundary. These new particles were assigned their own "scoring track" IDs. Secondary particles resulting from the split of the first particle were given the same scoring ID. In the detector response stage, the particle type, mean energy and pixel ID were found for the data for each scoring track.

Simulation studies were performed on a simple HEU sphere, with a parallel mesh set so that each layer had the thickness of the attenuation length for the 186 keV line. In addition, primary-particle biases on momentum and location were considered as well as an energy-location cut, where only particles emitted within a certain number of interaction lengths of the edge of the sphere were simulated. With the cutoff and momentum and energy biases, R^2T decreased by a factor of 25. Importance sampling was considered in addition to this, with shell thicknesses equal to the scattering length at 186 keV; this resulted in a less efficient simulation (using the FOM metric) than one that used a minimum energy cutoff and primary particle biases. In twice the time, the relative error dropped by around 20% (the expected drop would be greater than 40% for improved VR). A table of the performance of the various attempted VR methods can be found in Table 3.1. I believe that either my user code was inefficient or the added time due to transporting the particle through

VR Method	FOM
none	1
PP	25
IS+WW	12.5
IS+WW+PP	15

Table 3.1: Table exploring speed improvement for VR techniques. PP corresponds to primary particle biasing, IS importance sampling and WW weight windowing.

the various cells and performing the VR algorithms was a drag on the simulations that prevented an improvement in VR.

While a significant amount of time was spent speeding up the GEANT4 simulations, alternative methods were also considered. These are discussed in the following section.

3.5 Simulation Features for Each Task

The user-defined functions used in the GEANT4 simulations are discussed in this section. A picture of the simulation can be found in Figure 3.5. In this picture, IO8 is stored inside an aluminum box and is imaged by the fast-neutron detector. A close-up of IO8 is shown in Figure 3.6. There are no geometries other than the source and detector in this simulation. This decision was made to keep the simulation times reasonable; with four walls, a floor and a ceiling surrounding the geometry, every emitted particle that escapes the source would need to be processed through some material, whether that is the detector or the room. The room would scatter both gammas and neutrons back to the detector, which could be thought of as a second background term related to the room geometry. The inclusion of the room geometry slowed the simulations down considerably and subsequently was ignored. Discussion on how inclusion of other physical geometries in the environment would impact model performance is discussed in chapter 7.

Orientation was chosen as a nuisance parameter in some task-performance studies. In these studies, a random rotation method developed by Arvo (Arvo, 1992) was used. Three random numbers between zero and one are chosen. The object is first rotated a random amount around the $\hat{z}$ axis using the first number; then the $\hat{z}$ axis is rotated to a random location in ϕ, θ space using the last two. To generate stratified samples, evenly spaced values of the three random numbers were used. Three initial rotations around $\hat{z}$ were chosen. Then $\hat{z}$ was rotated into twenty different points (five in ϕ, four in θ) on the unit sphere for sixty total orientations. In

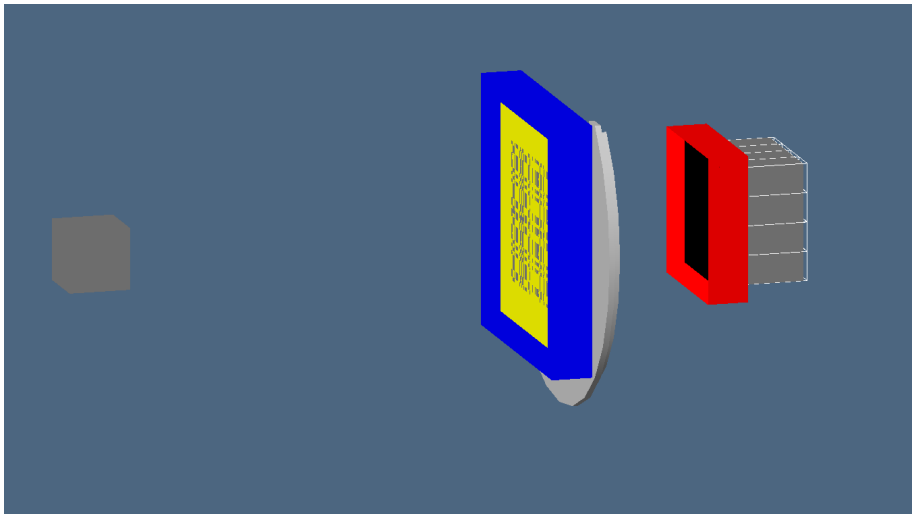


Figure 3.5: Geant4 model of system. An inspection object is stored inside an aluminum cube on the left (gray). The polyethylene mask is shown in yellow and the gray geometries in the mask are holes. On the right is the detector.

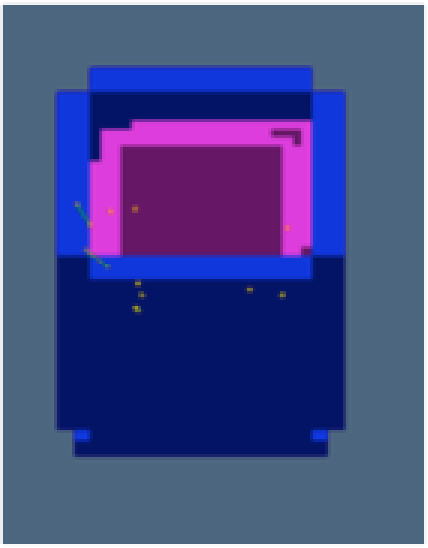


Figure 3.6: Geant4 model of IO8, removing the aluminum case and the top DU plate. The magenta geometry is the plutonium and the blue geometry is depleted uranium.

the studies in the experiments section, a stated rotation number $x_1x_2x_3$ corresponds to Arvo random numbers $x_1/3$, $x_2/5$, and $x_3/4$. As an example, Arvo rotation 111 corresponds to Arvo random numbers 1/3,1/5, and 1/4.

3.5.1 Particle Emission

The object geometries were coded into the transport application. Any materials with a significant gamma-ray emission rate, such as plutonium and uranium, were treated as source geometries. Particles were emitted from these geometries by randomly selecting a location from inside a large box surrounding the geometry and verifying that the particle was inside that geometry. Particles were emitted isotropically and with a biased energy distribution that is described in more detail in the VR subsection. Separate simulations were set up for fission emissions and radioactive-decay processes.

3.5.1.1 Radioactive Decay Processes

Radioactive-decay processes were modeled using a Sandia library, “SandiaDecay", created by SNL employee and project team member Will Johnson, that is based on Evaluated Nuclear Structure Data File (National Nucear Data Center, 2016) data. The library allows for custom mixing and aging of isotopes and includes over 3,000 nuclides. Each source geometry was read into the *G4VPrimaryGeneratorAction* class, and the emission rate was calculated for each geometry in the object based on its size and material. A source information class was set up to contain the emission spectra data for each material, output from “SandiaDecay". The *G4VUserPrimaryGeneratorAction* class randomly selects a geometry from an intensity distribution, finds the material, and samples the energy.

3.5.1.2 Spontaneous Fission

All items were imaged passively in this dissertation; alternatively, the host and monitor could agree to actively image the objects, sending in a beam of neutrons that induce fission in the object geometry. Spontaneous fission was simulated using a Lawrence Livermore National Lab fission library (Wright, 2015), *G4FissLib*, that is included in the GEANT4 source code. The user initializes predefined *SponFissIsotope* classes and assigns them an isotope number and an intensity. Multiple *SponFissIsotope* classes can be added to a *MultipleSource* class that contains different

isotopes. This library does not have the necessary data to calculate the spontaneous fission intensity for a given mass of a substance, so I included a function that found the spontaneous rate for each isotope from PANDA (Reilly et al., 1991).

The LLNLFission library randomly samples an isotope number from the *MultipleSourceClass*. The number of produced gammas and neutrons is found from that fission reaction's multiplicity distribution, and the energy for each neutron from the Watt fission spectra. The library does allow for correlated emissions. A flag can be set to either sample each neutron's energy independently (as was done in this simulation) or set the total emitted energy based on experimental data.

3.5.2 Physics

The *G4EmLivermorePhysics* class was used for all electromagnetic processes. The class includes models based on Livermore datasets for gammas and electrons. These physics processes were always simulated regardless of whether the simulation was set up for photon or neutron emission.

Neutron simulations used predefined data sets for elastic (*G4NeutronHPElastic*) and inelastic (*G4NeutronHPInelastic*) collisions, fission (*G4NeutronHPFission*) and capture (*G4NeutronHPCapture*). These are high precision models designed to simulate neutron transport down to low energies. In addition, the various cross sections for high Z isotopes ($Z > 92$) were requested and received from CERN, though no claims were made on the accuracy of this data.

The Lawrence Livermore Simulation site (Wright, 2015) does offer updated physics libraries with more thoroughly tested data. That neutron data was not used in this work.

3.5.3 Transport, Detection and Detector Response

A user stepping-action class was defined that terminated any gammas when they dipped below 100keV as these were unlikely to be detected. In addition, any neutrons attenuated below 50keV by the mask were terminated. At the end of each event, a user event action obtains the hit collection for the given event for each sensitive detector. The simulations did not model the the transport of the light resulting from the energetic electron or proton interactions in the scintillator. Instead, we were given a lookup table, derived from experimental data, that returns the light output for a given detected proton or electron. After that, an energy smearing

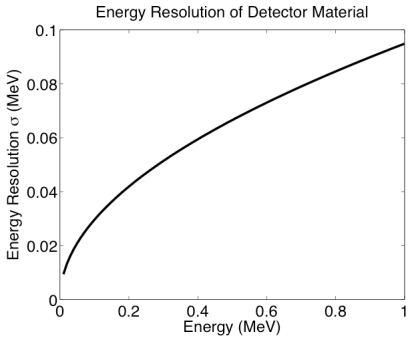


Figure 3.7: Plot of the standard deviation of the Gaussian energy smear as a function of deposited energy.

was applied to the light output. The distribution of the energy smear is Gaussian with a standard deviation given in Figure 3.7. For each event, the average position, weighted by deposited energy, of each detected particle was found and binned into a pixel. PSD was assumed to be perfect, even though it is often not, as described in the Physics chapter. The pixel ID, detected particle, energy type, and time since particle emission were all recorded.

3.5.4 Variance Reduction Techniques

The VR techniques used were simulation dependent. The sources where only neutron images were considered, such as the BeRP ball location study and 2D geometric source tasks, were unbiased. This is because the neutrons do not interact significantly in the source geometries, and therefore most escape and the effect of biasing is limited. A neutron-emission cut was considered, but low energy neutrons do induce fission in the objects which needs to be accounted for. A primary momentum-direction bias would likely have been helpful, especially for the thin 2D sources where interaction inside the object geometry is limited. Regardless, only about 200 hours were needed to simulate the desired number of counts for a given configuration of the 2D geometric sources.

However, the gamma simulations of the INL inspection objects were biased. This is necessary due to heavy self-shielding. When Pu241 beta decays, Am241 is produced. Am241 has a very intense peak (corresponding to 27 emissions per second per gram) at 60keV and another intense peak at 26.4 keV (1.8 emissions per second per gram). These gammas are extremely unlikely to escape the object or pass through the lead plate on the detector. Between these peaks and other low energy peaks (the highest intensity emission line above 100keV was a 105keV line

VR Method	CPU hours	R^2T
none	6.4e6	1
100keV cutoff	17200	372
100keV cutoff + Energy bias	4800	1333

Table 3.2: Table exploring speed improvement for VR techniques for the inspection objects. The 100keV energy cut provided the greatest decrease in number of CPU hours required and the energy bias provided another factor of 4.

at 0.0048/s), only roughly 1 in 1 billion emitted particles were detected for IO9 in the unbiased simulations, which took roughly 10 hours to simulate. To achieve a relative error of 0.05 for each of the 1600 pixels, this would require 6.4 million CPU hours.

To speed up the simulations, no gamma emissions below 100keV were considered. This is conservative—a 200keV gamma must travel through roughly seven path lengths in the lead plate in front of the detector. However, the 186 keV peak of U235 is active and some gammas pass through the lead plate. A linear energy bias was also included, defined by,

$$\begin{aligned} Pr_{biased}(E_m) &= \frac{C_m I(E_m)}{\sum_{m=1}^{M} C_m I(E_m)} \\ C_m &= E_m \text{ when } E_m < \text{ 1 MeV} \\ &= \text{ 1MeV when } E_m > 1\text{MeV} \end{aligned} \tag{3.8}$$

The effect of these choices is shown in Table 3.2. This provided a dramatic improvement in simulation time

3.5.5 Parallel Processing

GEANT4.10 was released in December, 2014. This significant update offers the capability of multi-threaded processing, though it does not have a GPU implementation available yet. A helpful code-migration reference can be found at `https://twiki.cern.ch/twiki/bin/view/Geant4/QuickMigrationGuideForGeant4V10`. The multithreaded GEANT4 parallelizes processing, tracking each event on a separate CPU. A new user action class *G4VUserActionInitialization* has been defined, and the primary-generator-action, stepping, event, and run action classes all get wrapped up into this new class that is thread local.

The user must be careful to keep the code thread safe using multi-threaded GEANT. In particular, thread locking through mutexes is necessary for any input

Threads	CPU hours	Real Time	Speedup
1	5e7	3543	1
4	5e7	934	3.8
8	5e7	514	6.9
12	5e7	390	9.1
16	5e7	345	10.3

Table 3.3: Parallelizability of GEANT4 code using multithreaded build. This performance analysis was done on the Sandia glory cluster, but an analysis on the modern red-sky cluster achieved a maximum speedup of about 13.

being read in from a file or output saved to a file. Parallelization results for the simulations can be found in Table 3.3.

3.5.6 Splitting up Simulations

A critical aspect of this work is accounting for the role that nuisance parameters play in treaty-verification tasks. To do so, it was necessary to simulate data sets for many different object orientations and locations. While simulating a set of detector data took roughly 4,800 hours for IO8, only 16 hours were necessary to transport the particles from the surface of the object to the detector. For this reason, the simulations were split into two components; one simulation transports the emitted particles to a sphere surrounding the object (see Figure 3.8), and the second reads in the LM data from the output of the first simulation and transports those particles to the detector (see Figure 3.9). As in the other VR efforts, this primarily provided an improvement for the IO8 and IO9 gamma simulations. The simulation time changed from 4,800 hours (in the case of the inspection objects) for each simulated image to about 4,800 hours for the initial source flux calculation and then only 16 hours for each realization of the source (before the parallel processing speedup). It provided a factor of 8 improvement when processing neutrons for IO8 and IO9, though the improvement was minimal for the neutron sources considered in this thesis.

The downside to this method is that each measurement required a large amount of storage. The LM detector data only requires on the order of 1 GB of storage for sufficient data; after splitting the simulations up, the storage requirement on the flux exiting the object was on the order of 100GB.

To store the LM data output from the first simulation, the ROOT framework (Brun and Rademakers, 1997) was used. All of the data for each detected particle was stored in an object. ROOT provides methods to visualize large amounts

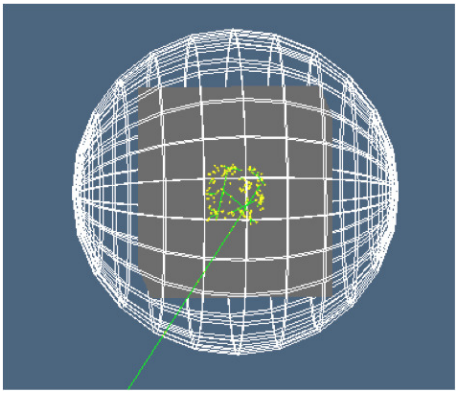


Figure 3.8: First GEANT4 simulation. A spherical detector surrounds the object, shown in wireframe visualization here.

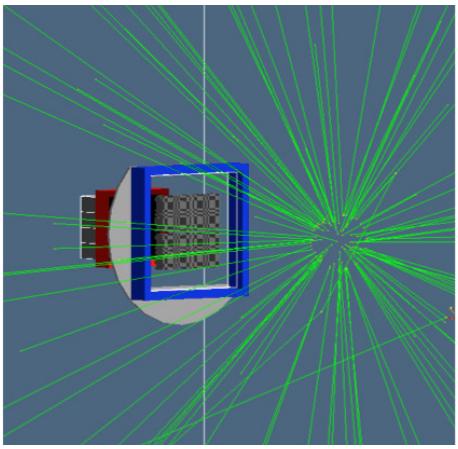


Figure 3.9: Second GEANT4 simulation, reading in the LM data file output from the first simulation and transporting those particles to the detector.

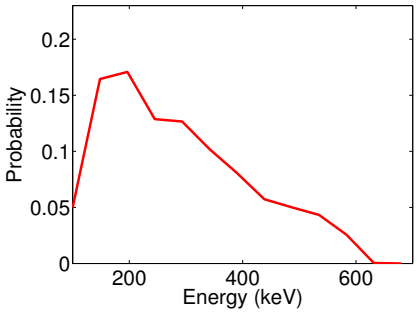


Figure 3.10: Background spectrum created with 2.60% K40, 3.49 ppm of uranium, and 11.09 ppm of thorium.

of data which can be used to verify the simulations are working as expected.

3.5.7 High Performance Computing

The steps taken in this section show a drastic improvement in simulation speed; energy biasing, parallelization and splitting up the simulation into two make the simulation time more manageable. However, 4800 hours (roughly 400 after parallelization) to simulate high enough statistics on the flux exiting the inspection objects is still an unmanageable number. To make these simulations manageable, the transport applications were run on Sandia's high performance computing clusters. The Red-Sky cluster, for example, has roughly 2800 nodes and 22,000 cores.

3.6 Background

The gamma-radiation background spectrum was generated using the Gamma Detector and Response Software (GADRAS)(Mitchell, 1988). Because GADRAS does not include liquid scintillators in its list of detector materials, NaI was used. Spectral templates for this geometry were created for 1.01% K40 (from the earth's mantle), 10 ppm of thorium, and 5 ppm of uranium (both from soil). Using these templates, background spectra can be created for different outdoor locations. An example gamma-ray background that was used throughout this work is shown in Figure 3.10. This background spectrum was applied equally to all pixels. This is a significant assumption that often is not true in real life. In actuality, background particles coming from the direction of the source would be "imaged" by the detector, leading to a shadow pattern due to the mask. In addition, any imaged TAI suppresses the background behind it. No neutron background was used.

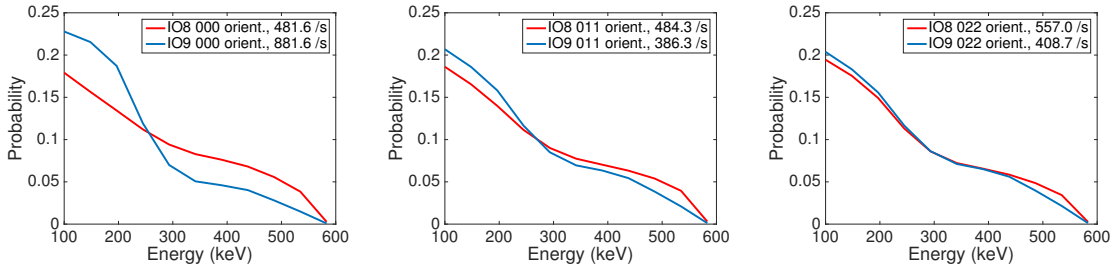


Figure 3.11: A comparison of the gamma spectra and count rates for IO8 and IO9 under three different orientations. When IO9 is imaged with the cube face perpendicular to the imaging axis (as in the 000 orientation), the low energy photons travel through the minimum shielding, and the count rate is highest and spectra shifts most towards low energies. The IO8 spectra and count rate are fairly consistent regardless of orientation chosen (count rate varies by 15%). In all three cases, photons detected from IO8 are more likely to be of higher energy than IO9.

3.7 Simulation Data

The simulated data used in each of the various tasks is discussed in this section.

3.7.1 Idaho Inspection Objects

IO8 and IO9 were measured in simulation under many orientations. The detected count rate and gamma-energy spectra for IO8, with depleted uranium, is less sensitive to changes in orientation than IO9 (Figure 3.11), which sees significant shifts in both spectra and count rate. With the VR techniques, roughly 2 million gamma rays were detected. The detected weights summed to around 2,000 for each of the two sources. The particles were binned into 64 energy bins, ranging from 100keV to 3MeV.

3.7.2 BeRP Ball Location Study

Neutrons emitted from the BeRP ball were imaged in simulation at two locations in the x-y plane. Roughly 5.5 million detected neutrons were recorded (all with weight 1) for each of the two simulations. Events were binned into an energy-weighted pixel ID. Count maps can be found in Figure 3.11. Unless stated otherwise, the count-rates were always set equal for these tasks to gauge the observer's ability to discriminate on spatial data rather than count rate.

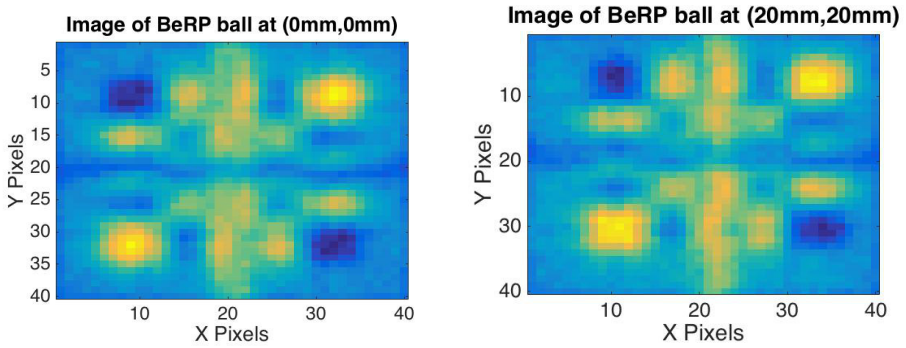


Figure 3.12: Images of the BeRP ball at (0 cm,0 cm,0 cm) on the left and (2 cm,2 cm,0 cm) on the right. Each pixel is 1 cm in length, so the difference in image location corresponds to a 2 pixel shift in both the vertical and horizontal directions.

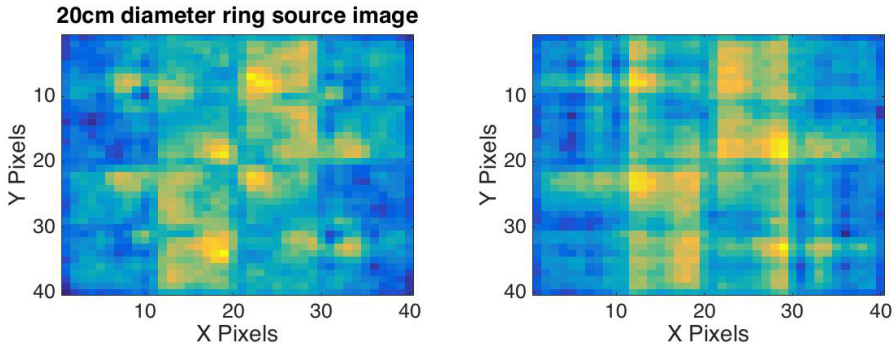


Figure 3.13: Images of the 20 cm ring source (on the left) and square source (on the right).

3.7.3 2D Circle vs Square Source

Plutonium ring and hollow square sources were simulated, each with a length of 20 cm. The neutron count maps can be found in Figure 3.13. Smaller and larger sources with these geometries were also simulated for the developed models that penalize storage of sensitive information. 12.5 million counts (with weight 1) were simulated for each of these sources. The detected count rate scales roughly with diameter, though in all of the performance studies, the count rates of any imaged sources were set equal to focus on the geometrical nature of the sources instead of the mass.

CHAPTER 4

Bayesian Ideal Observer

The Bayesian ideal observer is the optimal observer for binary-classification tasks, as it has complete probabilistic knowledge of the measured data. Importantly for this project, it can be represented in a form that processes LM data. Section 4.1 discusses multiple forms of the ideal observer, beginning with the signal-known-exactly (SKE) model and then expanding the theory to incorporate nuisance parameters. Section 4.2 demonstrates how the LM ideal observer models perform in the various tasks discussed in Section 3.2. This chapter expands on a JOSA-A publication (MacGahan et al., 2016d) and IEEE conference proceedings (MacGahan et al., 2014). Much of the theory presented in those publications is repeated in this chapter, but with additional emphasis on implementation of the models and ways a cheating host and monitor could trick the other party.

4.1 Theory and Model Implementation

In advance of the theory discussion, it is useful to formalize the notation for LM data. This notation is taken from work developed by Barrett, Parra, and Caucci (Barrett et al., 1997; Caucci and Barrett, 2012). Additional discussion on list-mode theory and its applications can be found in work by Clarkson (Clarkson, 2012) and Jha et al. (Jha et al., 2013). There is a fundamental difference in the motivations for their work and this work. While they utilize LM data to prevent the loss of information that comes with binning data, the desire here is to overcome the need for an IB by discriminating sources with LM data. Because this work is not utilizing the full information of the LM data by binning it, it is true that performance is not optimal compared to the methods developed in the above cited papers. However, the focus of this project is on deriving observer models that do not aggregate data. This is because as the limit of the individual bin size (in the case of energy) goes to zero, the discrete distribution will approximate a continuous one.

In this chapter, the data is represented by the total number of counts N and the set of LM data $\{A_n\}$. Each A_n contains all of the detectable information for the n^{th}

event. For a neutron-coded aperture imager, this data is,

$$A_n = \{\text{particle type, pixel number, energy deposited}\}. \tag{4.1}$$

For a HPGe detector (Twomey, 2003), which is a high-resolution gamma detector, the data would consist of only the gamma-ray energy deposited, as the detector does a poor job detecting neutrons and does not have any imaging capabilities. Notice that (4.1) contains both discrete (particle type and pixel number) and continuous (energy) random variables. In all of the experiments in this section, the energy is binned, meaning all distributions are discrete. Despite this, all probabilities on the LM data are treated as continuous.

The ideal observer (Barrett and Myers, 2003) is defined as

$$\Lambda(\{A_n\}, N) = \frac{pr(\{A_n\}, N|H_2)}{pr(\{A_n\}, N|H_1)}. \tag{4.2}$$

In (4.2), the arguments are a mixture of discrete and continuous random variables and the $pr(\cdot)$ notation was used in these cases. The ideal observer thresholds the likelihood ratio to make decisions and declare the data from class 1 or class 2. Note that the likelihood includes the LM data as well as the number of detected events N, which is not LM data, as it requires accumulating information (the event count) over many events. Though not explicitly stated in the above equation, the likelihoods and ideal observer are dependent on acquisition time.

In the following subsections, forms of the ideal observer are developed for an SKE discrimination task (Section 4.1.1), a task where nuisance parameters are present (Section 4.1.2), and an alternative, occasionally more useful form to incorporate nuisance parameters (Section 4.1.3). Finally, Section 4.1.4 discusses methods to account for the inherent variability in performance due to imperfect calibration data.

4.1.1 Signal-Known-Exactly Ideal Observer

The SKE ideal observer assumes that the nuisance parameters are known and thus the likelihood ratio is given by,

$$\Lambda_{SKE}(\{A_n\}, N|\gamma_1, \gamma_2) = \frac{pr(\{A_n\}, N|\gamma_2, H_2)}{pr(\{A_n\}, N|\gamma_1, H_1)}. \tag{4.3}$$

The SKE likelihoods in (4.3) can be represented in LM format. Under a known set of nuisance parameters, the likelihood that the data $(\{A_n\}, N)$ comes from hypothesis

H_j is

$$pr(\{A_n\}, N|\gamma_j, H_j) = pr(\{A_n\}|N, \gamma_j, H_j)Pr(N|\gamma_j, H_j). \tag{4.4}$$

The first term is a pdf on observing some set of LM data given full knowledge of the nuisance parameters relevant to hypothesis H_j. The second term is a Poisson probability on the number of counts observed. As each event is independent, (4.4) can be written as

$$pr(\{A_n\}, N|\gamma_j, H_j) = Pr(N|\gamma_j, H_j) \prod_{n=1}^{N} pr(A_n|\gamma_j, H_j), \tag{4.5}$$

where the last term $pr(A_n|\gamma_j, H_j)$ is the probability of observing the LM event data A_n given that object j is being imaged and with known nuisance parameters γ_j.

The Poisson probabilities depend on the mean count rate for events originating from the object being imaged (the source) $\overline{N}_j^{(s)}$ and the mean count rate for events originating outside the object (background events) $\overline{N}^{(b)}$, both of which depend on the set of nuisance parameters γ_j. The overall mean count rate for hypothesis H_j is defined as $\overline{N}_j = \overline{N}_j^{(s)} + \overline{N}^{(b)}$. In (4.4), $Pr(N|\gamma_j, H_j)$ is a Poisson probability with mean $\overline{N}_j$, i.e., $Pr(N|\overline{N}_j)$.

To make sense of the second term, $pr(A_n|\gamma_j, H_j)$, a variable h_n is defined. h_n describes the origin of the n^{th} detected particle. The probability that the detected event is a background event ($h_n = h^{(b)}$) can be differentiated from a source event ($h_n = h^{(s)}$). Including these conditional probabilities in the LM term,

$$\begin{aligned} &pr(A_n|\gamma_j, H_j) = \\ &pr(A_n|\gamma_j, h_n = h^{(b)})Pr(h_n = h^{(b)}|\gamma_j, H_j) + \\ &pr(A_n|\gamma_j, H_j, h_n = h^{(s)})Pr(h_n = h^{(s)}|\gamma_j, H_j), \end{aligned} \tag{4.6}$$

where $Pr(h_n = h^{(b)}|\gamma_j, H_j)$ is the probability that the detected event came from the background, and $Pr(h_n = h^{(s)}|\gamma_j, H_j) = 1 - Pr(h_n = h^{(b)}|\gamma_j, H_j)$ the probability that the detected event originated from the source. These probabilities are equal to the ratio between the mean number of background or signal counts and the total mean number of counts. The dependence of the LM data for a background event on H_j was dropped because the background distribution is the same for either object being imaged.

The likelihoods found in the numerator and denominator of (4.3) have now been expanded to include a Poisson term on the total detected counts, terms which account for the probability of a background or a source event (another nuisance

parameter related to the count rate), and the probability density of observing LM data for a source event and a background event. These last two probabilities include the distribution of where an event occurs in the detector (i.e., the imaging aspect of the system) as well as the distribution associated with the energy of the event (i.e., the spectral aspect of the system). These distributions must be determined either through calibration or through Monte Carlo simulations. Replacing the likelihoods in (4.3) with the discussed expressions reveals that,

$$\Lambda_{SKE}(\{A_n,\},N|\gamma_1,\gamma_2) = \frac{Pr(N|\overline{N}_2)}{Pr(N|\overline{N}_1)}\prod_{n=1}^{N}\frac{pr(A_n|\gamma_2,H_2)}{pr(A_n|\gamma_1,H_1)}. \tag{4.7}$$

Both the numerator and denominator inside the product utilize the background and source decomposition shown in (4.6).

4.1.1.1 Implementation

Implementation of this SKE ideal observer would occur in three stages, consistent with the binary-classification task discussion in Section 1.3. First, in the calibration stage, a pair of high-statistics LM-data sets are measured (or simulated) for two different trusted TAIs. These data sets are used to find $\overline{N_j}$ and are binned by energy and pixel number and then normalized to find a probability density on observing the LM data $pr(A_n|\gamma_j,H_j)$.

Second, the monitor would need to measure the trusted TAIs many times (enough to properly account for the Poisson noise on the test statistic) and perform the ideal observer on that data to generate a distribution on Λ for each of the two hypotheses, $pr(\Lambda|H_1)$ and $pr(\Lambda|H_2)$. For each measurement, the test statistic Λ would be initialized to one. For each detected event, the test statistic is multiplied by the ratio of observing that event's data A_n given the two hypotheses and the known nuisance parameters. At the end of the acquisition time, Λ is multiplied by the ratio of the probabilities for observing N counts under the two hypothesis, resulting in a final value for Λ. This distribution would be used, once the appropriate cost functions are decided upon for incorrect decisions, to decide on a threshold value t_{thresh}. t_{thresh} could be designed to maximize the true-positive rate for a given acceptable level of incorrect classifications.

Finally, in the testing stage, an unknown object would be placed in front of the imager for the monitor to perform a verification measurement. The model

would be performed on the data as described in the above paragraph. In this case, however, the LM data A_n would update the model and then be deleted. Finally, Λ is thresholded to make a decision.

The choice to compute the test statistic rather than the individual likelihoods was made for computational ease. The LM probabilities used in this model are often very small (the smaller the bin size, the lower the probability values are). Calculation of an individual likelihood expression such as $pr(A_n, N|\gamma_1, H_1)$ is difficult as it is the product of small numbers and goes to zero computationally after a handful of list-mode events. This can be overcome through the use of log likelihoods or by tracking only the test statistic.

In practice, the LM ideal observer could be implemented with an electronic board that updates the test statistic with a certain value for each detected particle read in. The system would only need to have enough memory to store $\overline{N_j}$ and $pr(A_n|\gamma_j, H_j)$, determined from the calibration data, and t. In addition, it would need to perform any mathematical calculations necessary as events are read in.

Another important component to the implementation of this model is addressing the variability in detector response. This is an important task that has not been considered in detail due to lack of experimental data. It is addressed further in Section 7.1.1.

4.1.1.2 Storage and Need for an Information Barrier

The storage requirement for the Bayesian ideal observer is significant. The SKE ideal observer requires storage of both spatial and spectral measurements. In a treaty-verification application, this information could be used to determine sensitive isotopic and spatial information about the object. Storage of this calibration data would need to be behind an IB, as in Figure 1.19. The monitor would therefore not be able to access this information, reducing its confidence that a useful measurement is being performed.

4.1.1.3 A Cheating Host

The host country could try to fool the monitor by using a spoof that the model would declare is H_1 or H_2. The monitor would only be given access to the test-statistic distributions for the trusted TAIs, $pr(t|H_1)$ and $pr(t|H_2)$, and the test statistic for the unknown item. If an unverified object returns a test statistic that is statistically unlikely for either of the two trusted TAIs, the tested source could be defined as

neither of the two. A probability density on the log of the test statistic distribution, $\log(t|H_j)$ can be found easily; $\log(t|H_j)$ is the sum of many independent random variables in the form of $\log(A_n|H_j)$. In the SKE case, it is a normal distribution due to the central limit theorem. However, it is fairly easy to spoof given the massive dimensionality reduction involved, going from a total of M detector bins to a single test statistic. An example is discussed in Section 4.2.2.

Furthermore, the fact that the monitor can not access this information leads to an incentive for the host to bake in spoofs into H_1 and H_2 in order to trick the monitor. The monitor would not be able to determine if the host did so.

4.1.1.4 A Cheating Monitor

It could be of interest to the monitor to perform the model on unclassified items with known data distributions. Given enough measurements, especially if done with single emission lines to back out the value the model assigns to that energy, this could begin to reveal specifics of the model.

4.1.2 Ideal Observer Incorporating Nuisance Parameters

When the nuisance parameters γ_1 and γ_2 are not known exactly, the likelihood expressions in the ideal observer must be integrated over the probability densities of those nuisance parameters for optimal performance. (4.2) then becomes,

$$\Lambda(\{A_n\}, N) = \frac{\int pr(\{A_n\}, N|\gamma_2, H_2) pr(\gamma_2) \mathrm{d}\gamma_2}{\int pr(\{A_n\}, N|\gamma_1, H_1) pr(\gamma_1) \mathrm{d}\gamma_1}. \tag{4.8}$$

The probability density on the LM data $\{A_n\}$ is conditioned on knowledge of the nuisance parameters and then averaged over the distribution of the nuisance parameters $pr(\gamma_j)$. This prior distribution would be very difficult to determine empirically. For example, if the location was unknown, the host could make a best guess on the distribution of the location nuisance parameter, but any deviation of the actual tested locations from this distribution would degrade task performance.

The incorporation of nuisance parameters brings a practical concern. The dependence of the true detected spatial and energy distributions on certain nuisance parameters is complex. In order to use the ideal observer in practice, the host country would need to measure their TAIs under many different conditions to acquire the probability densities on $pr(A_n, N|\gamma_j, H_j)$ in order to properly train the ideal observer. While the nuisance parameter priors discussed in this section are continuous,

in reality they would likely need to be treated as discrete as the host would not be able to measure $pr(A_n, N|\gamma_j, H_j)$ for all γ_j, or analytically determine the likelihood. In addition, the host would need to investigate the objects themselves to properly identify the prior distributions. Performance degrades for any deviation between the prior distributions chosen by the host and the actual distribution of those nuisance parameters when the monitor is testing the objects.

This model could be evaluated through Monte Carlo sampling of the numerator and denominator in (4.8). This is explained in more detail in the following subsubsection.

4.1.2.1 Observer Evaluation

The observer model must maintain the ability to process LM data to avoid aggregating data when testing a source. With the SKE ideal observer, this is relatively trivial. But it becomes a more difficult problem when nuisance parameters are incorporated. Using Monte Carlo integration methods, integration over γ_j can be represented as a sum over the number of Monte Carlo samples for that source S_j,

$$\Lambda(\{A_n\}, N) = \frac{\sum_{s_2=1}^{S_2} pr(\{A_n\}, N|\gamma_{(2,s_2)}, H_2)}{\sum_{s_1=1}^{S_1}, pr(\{A_n\}, N|\gamma_{(1,s_1)}, H_1)}, \tag{4.9}$$

where each γ_{j,s_j} have been sampled from $pr(\gamma_j)$. Note that this integral requires two separate summations, which is why the sample numbers for the nuisance parameters in source 2, s_2 have been differentiated from the sample numbers from source 1, s_1. Each γ_{j,s_j} must be determined before processing of the tested source' data begins. Then, each $pr(A_n, N|\gamma_{j,s_j}, H_j)$ would be updated with the LM data.

Monte Carlo sampling of the integrals in the numerator and denominator offers the most straightforward solution, but proves mathematically difficult in instances where the numerator and denominator are both very close to zero computationally, as in this application. This can be circumvented by using the log likelihood. To proceed without the storage of LM data, all samples of γ_j must be taken before the data is processed, and each log likelihood expression $\log(pr(A_n, N|\gamma_{j,s_j}, H_j))$ must be updated as the LM data is read in. At the end, a common factor is subtracted from both numerator and denominator of (4.9), and the terms are re-exponentiated and added. This procedure makes the problem computationally feasible. In testing, the method requires storage of the individual log likelihood values $\log(pr(\{A_n\}, N|\gamma_{(j,s_j)}, H_j))$ for each of the two sources and chosen samples of γ_{j,s_j}.

Monte Carlo integration proved sufficient to evaluate the ideal observer for the experiments chosen in this dissertation, as only one nuisance parameter was treated at a time. As the dimensionality of the nuisance parameters γ_1 and γ_2 increases, evaluation of this integral through standard Monte Carlo methods becomes difficult due to slow convergence (Asmussen and Glynn, 2007). This can be improved through Quasi-Monte Carlo methods, but a faster technique to calculate this integral is Markov-Chain Monte Carlo (MCMC) (Gilks, 2005). MCMC integration continuously resamples the nuisance parameters based on a proposal density $pr(\gamma_{j,new}|\gamma_{j,old})$, performing the following operation,

1. Sample $\gamma_{j,new}$ from $pr(\gamma_{j,new}|\gamma_{j,old})$
2. If $pr(A_n, N|\gamma_{j,new}, H_j) > pr(A_n, N|\gamma_{j,old}, H_j)$, accept $\gamma_{j,new}$
   Otherwise, accept $\gamma_{j,new}$ with probability $\dfrac{pr(A_n, N|\gamma_{j,new}, H_j)}{pr(A_n, N|\gamma_{j,old}, H_j)}$ (4.10)
3. Recalculate $pr(A_n, N|\gamma_j, H_j)$ with current γ_j
4. Repeat

The correlations between consecutive evaluations of $pr(A_n, N|\gamma_j, H_j)$ bring practical concerns. γ_j is initialized to a guess. Most MCMC guides recommend a burn-in time to allow γ_j to find a more likely value of $pr(A_n, N|\gamma_j, H_j)$, preventing a large number of unlikely probabilities from stopping quick convergence. In addition, it is often recommended to take one of every 100 or so evaluations to prevent sample to sample correlations.

4.1.2.2 Implementation

Implementation of the ideal observer when nuisance parameters are present is a painstaking task on the host country's part. First, the host would need to estimate the priors $pr(\gamma_j)$. To repeat the earlier example, location variability could be present, whether due to inconsistent packaging in its container or inconsistent placing of the container in front of the detector. The prior could then be treated as a Gaussian distribution in the coordinate plane, centered at (0,0,0) with a certain estimated variance in all directions. The host country would need to measure the object in many different locations to find $pr(\{A_n\}, N|\gamma_j, H_j)$. To use this method while keeping the LM requirement, $pr(\{A_n\}, N|\gamma_j, H_j)$ would need to be calculated over a predefined grid on the nuisance parameter values. Likewise, if MCMC was used, the proposal density $pr(\gamma_{j,new}|\gamma_{j,old})$ would also need to have discrete values of γ_j.

As in the SKE case, the host country would also need to test the trained ideal observer on trusted objects to generate the test statistic distributions. For best performance, these trusted objects (and the tested items) must also exhibit the same variability as the sources used to calibrate the observer model.

4.1.2.3 Storage and Need for an Information Barrier

The fact that many measurements must be taken to train the ideal observer enforces the need for an information barrier. For example, if the orientation of the TAI is unknown, the observer needs to train on many measurements of the TAI at different orientations. Such data could be used in a MLEM routine or FBP to reconstruct more sensitive information than a single measurement could provide.

4.1.2.4 A Cheating Host and Monitor

Like the SKE model, the ideal observer that incorporates nuisance parameters also offers some ability to distinguish spoofs. However, the distribution on $pr(t|H_1)$ and $pr(t|H_2)$ are broader than the SKE distributions, making it harder to reject an item other than H_1 and H_2.

The monitor does gain an advantage in its ability to access the test-statistic distribution. While the log of the ideal observer in the SKE case is normal, the log ideal observer when testing objects with nuisance parameters is a weighted sum of normal random variables and can actually be highly non-normal. The nature of the distribution will give the monitor some information about how the SKE likelihoods vary with nuisance parameter values. This is discussed further in Section 4.2.3.

4.1.3 Ideal Observer Using Posterior Probability Density

Through some manipulation of (4.8), it is possible to derive a more manageable form of the ideal observer. The following derivation is similar to work done by Kupinski et al. in a previous paper (Kupinski et al., 2003). Before beginning, a notational change is made for convenience. The symbol γ_0 is now the set of nuisance parameters shared by source 1 and 2, such as variability in background activity, orientation, and location. The symbols γ_1 and γ_2 will now be used to describe the nuisance parameters unique to sources 1 and 2, such as the material composition of each

source if not exactly known. (4.8) now becomes,

$$\Lambda(\{A_n\}, N) = \frac{\int\int pr(\{A_n\}, N|\gamma_0, \gamma_2, H_2)pr(\gamma_0)pr(\gamma_2)\mathrm{d}\gamma_0\mathrm{d}\gamma_2}{\int\int pr(\{A_n\}, N|\gamma_0, \gamma_1, H_1)pr(\gamma_0)pr(\gamma_1)\mathrm{d}\gamma_0\mathrm{d}\gamma_1}. \tag{4.11}$$

Beginning with (4.11), simplify the denominator back to $pr(\{A_n\}, N|H_1)$ and marginalize the numerator over the remaining nuisance parameters γ_1,

$$\begin{aligned}\Lambda(\{A_n\}, N) =& \frac{1}{pr(\{A_n\}, N|H_1)} \\ &\times \int\int\int pr(\{A_n\}, N|\gamma_0, \gamma_2, H_2)... \\ &pr(\gamma_0)pr(\gamma_1)pr(\gamma_2)\mathrm{d}\gamma_0\mathrm{d}\gamma_1\mathrm{d}\gamma_2.\end{aligned} \tag{4.12}$$

Next, multiply both the numerator and the denominator inside the integral by the SKE likelihood for source 1, $pr(A_n, N|\gamma_0, \gamma_1, H_1)$, and acknowledge that $pr(A_n, N|\gamma_0, \gamma_2, H_2)/pr(A_n, N|\gamma_0, \gamma_1, H_1)$ for a specific γ_1 and γ_2 is the SKE observer as in (4.7),

$$\begin{aligned}\Lambda(\{A_n\}, N) =& \frac{1}{pr(\{A_n\}, N|H_1)} \\ &\times \int\int\int \Lambda_{SKE}(\{A_n\}, N|\gamma_0, \gamma_1, \gamma_2) \\ &pr(\{A_n\}, N|\gamma_0, \gamma_1, H_1)pr(\gamma_0)pr(\gamma_1) \\ &pr(\gamma_2)\mathrm{d}\gamma_0\mathrm{d}\gamma_1\mathrm{d}\gamma_2.\end{aligned} \tag{4.13}$$

Next simplify further using Bayes' rule, creating a posterior probability density,

$$pr(\gamma_0, \gamma_1|\{A_n\}, N, H_1) = \frac{pr(\{A_n\}, N|\gamma_0, \gamma_1, H_1)pr(\gamma_0)pr(\gamma_1)}{pr(\{A_n\}, N|H_1)}. \tag{4.14}$$

Substituting (4.14) into (4.13), we arrive at the final result,

$$\begin{aligned}\Lambda(\{A_n\}, N) = \int& \Lambda_{SKE}(\{A_n\}, N|\gamma_0, \gamma_1, \gamma_2)pr(\gamma_2) \\ &pr(\gamma_0, \gamma_1|\{A_n\}, N, H_1)\mathrm{d}\gamma_0\mathrm{d}\gamma_1\mathrm{d}\gamma_2.\end{aligned} \tag{4.15}$$

This form of the ideal observer presents various advantages and disadvantages over (4.8) that are explored in the following subsubsection.

4.1.3.1 Observer Evaluation

This integral can be evaluated by sampling the nuisance parameters γ_0 and γ_1 from the posterior density $pr(\gamma_0, \gamma_1|\{A_n\}, N, H_1)$ and the γ_2 nuisance parameters from

their respective probability densities and performing Monte Carlo integration. This provides an advantage over (4.8) because the SKE ideal observer is generally more computationally feasible than the individual likelihoods and does not require the use of log likelihoods. Another advantage to this method is that the posterior pdf provides a distribution on γ_0, γ_1 values more consistent with the data $(\{A_n\}, N)$ than a simple Monte Carlo sample. The integral should therefore converge faster.

4.1.3.2 Implementation, Storage, and Ability to Discriminate Spoofs

The posterior pdf $pr(\gamma_0, \gamma_1 | \{A_n\}, N, H_1)$ is dependent on the data set $(\{A_n\}, N)$. However, to keep the LM storage requirement, the full $\{A_n\}$ can not be known in advance. Computing the posterior pdf requires evaluating the likelihood $pr(\{A_n\}, N | \gamma_0, \gamma_1, H_1)$ over a large number of points on the γ_1, γ_0 nuisance parameter grid. The full $(\{A_n\}, N)$ are not known before this nuisance parameter grid is set up, requiring a large range of values for the γ_0, γ_1 grid. In the end, it is not clear that this method would provide a significant speed increase as a similar number of grid points may be required to evaluate (4.15) as samples needed to effectively evaluate (4.8).

This model presents a new method to calculate the ideal observer that is beneficial in certain situations. The storage requirements and ability to discriminate spoofs does not change compared to the Monte Carlo integration in (4.8). Regardless of the approach chosen, the ideal observer would need to store many sets of calibration data under different object configurations in order to incorporate nuisance parameters, increasing storage requirements and complexity of operations behind an IB.

4.1.4 Method to Account for Imperfect Calibration Data

In both simulation and a real-life application, the accuracy of the calibration data is limited by the statistics; this limitation degrades task performance due to the mismatch between the calibration data and the "true" distribution that the tested source's detector data would take on given unlimited statistics. To address this problem, the variation in the calibration data can be treated as a nuisance parameter. A few variables are defined to better understand this problem:

- The mean values of the true data distribution when imaging source j are denoted as $\mathbf{g}_{j,t}$. This is non-random (assuming a constant detector response

for now) and unknown.

- The calibration data corresponds to a random sample from a Poisson (or normal if the statistics are high enough) distribution with mean $\mathbf{g}_{j,t}$. The calibration data for source j is denoted as $\mathbf{g}_{j,c}$, containing values $g_{j,c,m}$. These are non-random and known.

- The vector $\mathbf{g}_{j,cs}$ is a random variable used to represent the range in possible values the true mean could take on given the calibration data. Individual realizations of $\mathbf{g}_{j,cs}$ correspond to possible values of the true mean.

As the only known data is the single set of calibration data, we must assume $\mathbf{g}_{j,cs}$ is normal with mean and variance equal to $\mathbf{g}_{j,c}$. This proposed distribution is an approximation and imperfect. If the true mean for a bin was smaller than the sampled calibration data, it would have a lower variance. If the true mean was higher, it would have a higher variance. Therefore, a proper sampling distribution would emphasize higher $\mathbf{g}_{j,cs}$ values over lower ones. Ignoring this, the SKE ideal observer, with truth data known, would be represented as,

$$\Lambda(\{A_n\}, N) = \frac{pr(\{A_n\}, N|\mathbf{g}_{2,t})}{pr(\{A_n\}, N|\mathbf{g}_{1,t})}. \tag{4.16}$$

However, as already stated, the truth data is not known. Therefore, it can be represented as an integral over the random variable $\mathbf{g}_{j,cs}$.

$$\Lambda(\{A_n\}, N) = \frac{\int pr(\{A_n\}, N|\mathbf{g}_{2,cs})pr(\mathbf{g}_{2,cs})d\mathbf{g}_{2,cs}}{\int pr(\{A_n\}, N|\mathbf{g}_{1,cs})pr(\mathbf{g}_{1,cs})d\mathbf{g}_{1,cs}}. \tag{4.17}$$

In the following subsubsection, I present a method to evaluate the likelihood integrals. I also present a method to gauge the variability in performance due to unknown count rate.

4.1.4.1 Evaluating Likelihood Integrals

The likelihood of observing a data set given source j was imaged is,

$$pr(\{A_n\}, N|H_j) = \int pr(\{A_n\}, N|\mathbf{g}_{j,cs})pr(\mathbf{g}_{j,cs})d\mathbf{g}_{j,cs}. \tag{4.18}$$

Breaking this down into the LM and Poisson components,

$$pr(\{A_n\}, N|H_j) = \int Pr(N|\mathbf{g}_{j,cs}) \prod_{n=1}^{N} pr(A_n|\mathbf{g}_{j,cs})pr(\mathbf{g}_{j,cs})d\mathbf{g}_{j,cs}. \tag{4.19}$$

The mean count rate for a certain $\mathbf{g}_{j,cs}$ is then equal to,

$$\overline{N}_{cs} = \sum_{m=1}^{M} g_{j,cs,m}. \tag{4.20}$$

$\overline{N_{cs}}$ is $\mathcal{N}(\sum_{m=1}^{M} g_{j,c,m}, \sum_{m=1}^{M} g_{j,c,m})$. It should be noted that $\overline{N_{cs}}$ has a smaller relative error than any individual bin. If the n^{th} detected particle falls into bin m_n, the probability on the LM component can be expressed as,

$$pr(A_n|\mathbf{g}_{j,cs}) = \frac{g_{j,cs,m_n}}{\sum_{m=1}^{M} g_{j,cs,m}}. \tag{4.21}$$

Plugging (4.20) and (4.21) into (4.19), and assuming the Poisson term is relatively constant to the LM term, yields the final form,

$$\begin{aligned} pr(\{A_n\}, N|H_j) =& Pr(N|\sum_{m=1}^{M} g_{j,cs,m}) \int \frac{\prod_{n=1}^{N} g_{j,cs,m_n}}{(\sum_{m=1}^{M} g_{j,cs,m})^N} \times \\ & pr(g_{j,cs,1}) pr(g_{j,cs,2}) ... pr(g_{j,cs,M}) dg_{j,cs,1} dg_{j,cs,2} ... dg_{j,cs,M}. \end{aligned} \tag{4.22}$$

This is the expectation value of a ratio of correlated random variables. This expression was not evaluated analytically, though a useful source that utilizes a Taylor series expansion can be found at `http://www.stat.cmu.edu/~hseltman/files/ratio.pdf`. This was evaluated in simulation with Monte-Carlo techniques, randomly sampling each $\mathbf{g}_{j,cs}$ from $\mathbf{g}_{j,c}$. The results are not included with this thesis, as they show no difference in returned ideal observer value from the SKE model. I believe that because $pr(\mathbf{g}_{2,cs})$ is normal, it is equally likely that a sampled value is higher or lower than the calibration mean. The result is that when integrating over the randomness in each bin, the returned average is the same as if the SKE model was used. After seeing this result, emphasis was put on accounting for the variability in the ideal observer due to imperfect calibration data.

4.1.4.2 Using Known Variability in Calibration Data to Gauge Performance Variability

Data $\mathbf{g}_{1,c'}$ and $\mathbf{g}_{1,c'}$ was sampled from $pr(\mathbf{g}_{1,cs})$ and $pr(\mathbf{g}_{2,cs})$. The ideal observer was then calibrated using these new data sets and performed on independent testing data. This procedure generates a number of AUC values, which allows for uncertainty quantification of the AUC value.

This method can be compared to the SKE method (ignoring calibration data variability) to gauge whether the statistics of the calibration data are sufficient. A

significant drop in model performance would imply that the statistics are not high enough and the AUC value changes when using a different set for the calibration data. In the experiment section where these methods are implemented in practice, the output plots are a distribution of resulting AUC values.

An alternative to this method would be to double the acquisition time for the objects, and create what is called an antler plot. An antler plot would show the AUC for a certain acquisition time as a function of the number of counts N in the calibration data set. As N increased, the AUC should move towards the true value for that acquisition time.

4.2 Experiments and Results

In this section, the ideal observer models discussed in the theory section are applied to the various tasks described in Section 3.2. The experimental methodology is discussed in Section 4.2.1. Then, the SKE ideal observer (Section 4.2.2) and the ideal observer incorporating nuisance parameters (Section 4.2.3, Section 4.2.4) are applied to specific tasks. Finally, the methods to account for uncertainty in the calibration data are discussed in Section 4.2.5.

4.2.1 Methodology

The methodology used to evaluate the performance of an observer model in binary-classification tasks can be found in Section 1.3.2. Calibration and testing data sets were simulated for the two objects in each task. These two sets have high statistics and both are essentially independent samples of the "true" data on the detector when a given object is imaged. The calibration data set is used to train the observer model. LM data is sampled from the testing data set to test the model. Therefore, the LM data actually takes on a statistical nature different from both the calibration data and the "true" data, though the impact of that is ignored in these experiments. The reason LM data was sampled from the binned simulation data rather than taken directly from a data file (GEANT4 can output detections event-by-event) is that it is unclear how to properly sample from the list when the LM events have non-unity weights, as in the IO8 vs. IO9 classification task.

4.2.2 SKE Ideal Observer

In this section, experiments are designed and data simulated to test the SKE ideal observer (4.3). In each case, the model is trained on a specific pair of data sets from measured objects under known conditions. The model is then judged by its ability to discriminate a second set of measurements. All of the discussed discrimination tasks are considered in this section. In addition, examples of the test statistic distributions and discussions of discriminating spoofs can be found in each subsection.

4.2.2.1 IO8 vs. IO9 Gamma Data Discrimination

In this task, the ideal observer was trained on IO8 and IO9 data measured under a certain orientation. When discussing the orientation of the objects being imaged, I refer to the Arvo rotation number described in Section 3.5. This model was then used to classify independently measured IO8 and IO9 data sets from that same orientation. This was done for Arvo orientation 000 and 111. Arvo orientation 000 puts the face of the cube parallel to the mask plane. IO9's 186 keV line undergoes the minimal attenuation in this case. This leads to drastically different count rates between IO8 and IO9 and a more disparate energy spectra (see Figure 3.11). The 111 orientation is a standard off-imaging axis orientation and the count rates and spectra for the two objects imaged at this orientation are generally more similar.

Figure 4.1 shows the performance of different components of the ideal observer in classifying the two items. The "All terms" lines correspond to the full SKE ideal observer. The "Poisson terms" lines correspond to an ideal observer based on only the Poisson probabilities on overall detected counts,

$$\Lambda(N) = \frac{Pr(N|H_2)}{Pr(N|H_1)}. \tag{4.23}$$

The "LM terms" lines correspond to an observer based on just the LM probabilities,

$$\Lambda(\{A_n\}) = \prod_{n=1}^{N} \frac{pr(A_n|H_2)}{pr(A_n|H_1)}. \tag{4.24}$$

Note that the likelihoods in eq. (4.24) are not proper probability densities as it they are not normalized. The plots in Figure 4.1 show how the Poisson and LM components of the ideal observer effect overall performance; in the 000 orientation study, the Poisson probabilities play a larger role in ideal observer performance than the 111 orientation study.

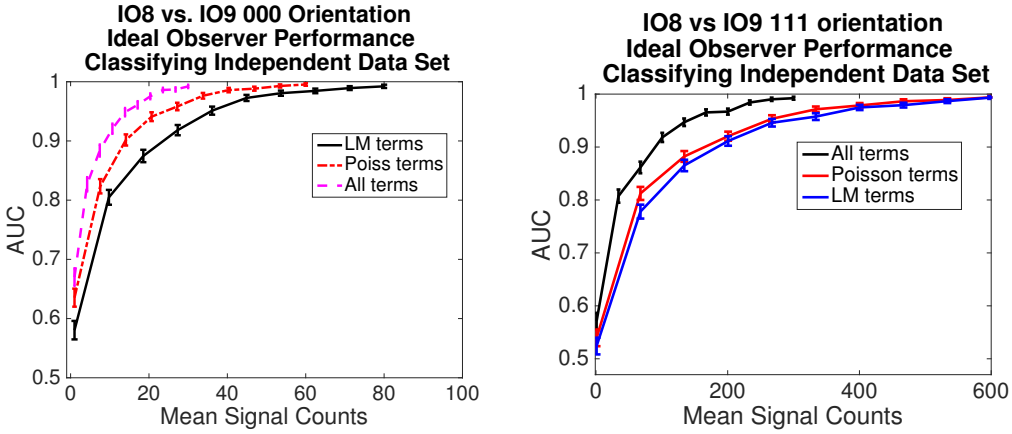


Figure 4.1: The left plot shows the performance of the total (4.3), LM (4.24) and Poisson (4.23) components of the ideal observer trained and tested on independent measurements of the 000 orientation. The right plot shows the performance when trained on and testing the 111 orientation. For these plots, the data sets were summed over pixel ID, producing high statistics data sets on the counts observed in each energy bin.

The error bars on the plots in Figure 4.1 are due to the limited number of testing data samples taken on IO8 and IO9 to calculate the AUC. The error is due to binomial statistics—the outcome is chosen to be H_1 or H_2, with a certain probability. The standard deviation on the AUC value is then $sqrt(\frac{AUC*(1-AUC)}{N})$. The impact of testing and training data that doesn't take on the "true" distribution on the AUC is ignored for now.

A second study was performed to check whether the LM observer performance for the 000 orientation in Figure 4.1 was due to spectral or spatial differences. In these studies, the count rates for the two measured objects were set equal to avoid decisions being made based on the count rate component. The data was binned different ways, once by just energy value, a second time by pixel ID, and a third using both the spatial and spectral information. The performance of all three is shown in Figure 4.2. As explained in Section 3.7.1, the difference in gamma images is minimal, and in this performance study it is apparent that while there is some signal to be found in the differences between the gamma images, it is low compared to the gamma spectra.

4.2.2.2 BeRP Ball Location Discrimination and Geometry Classification

Next, the ideal observer was performed on the two tasks based on neutron data. Ideal observer performance in both tasks is shown in Figure 4.3, binned in different

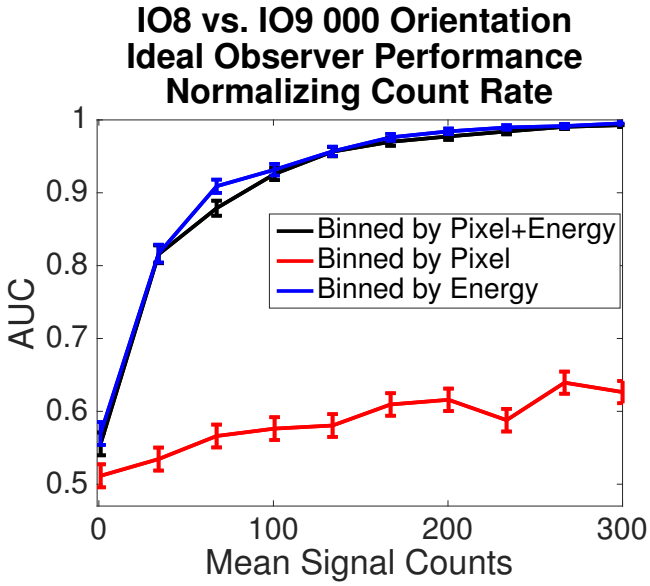


Figure 4.2: In this study, measurements of the Arvo 000 orientation were binned in different ways and the ideal observer was trained and tested on this data. The count rates for the data sets were set equal to properly explore the performance in classifying the spatial vs. spectral data. The "Pixel+Energy" line bins the data into 1600 pixels and roughly 50keV energy bins, the "Energy" line bins the data into just the energy bins, and "Pixel" bins the data into just the Pixels.

ways. The count rates for the data sets were set equal to properly gauge the model's ability to discriminate the two images. In both tasks, the detected neutron energy data is of little use to the ideal observer, with the decision being made almost entirely on the image data. The model trained and test on "Binned by Pixel" data generally outperforms the "Binned by Pixel+Energy" data. This is because the splitting up each pixel count bin into smaller energy bins (without increasing the size of the calibration data set) just adds noise to the model. This is explored further in Section 4.2.5

4.2.2.3 Spoof Rejection Example 1

In this subsubsection, the test statistic distributions are examined and spoofs are tested. The 20 cm ring and square source discrimination task was used in this study. The count rates were equalized to gauge the effect of the spatial information. For an acquisition time corresponding to 2,000 signal counts being observed, the resulting test-statistic distributions when testing independent data sets as well as the BeRP ball data sets is shown in Figure 4.4. The test-statistic distributions on the ring and square sources could be used to identify spoofs. For example, if any tested items return a test statistic that falls outside the middle 90% of likely values for the

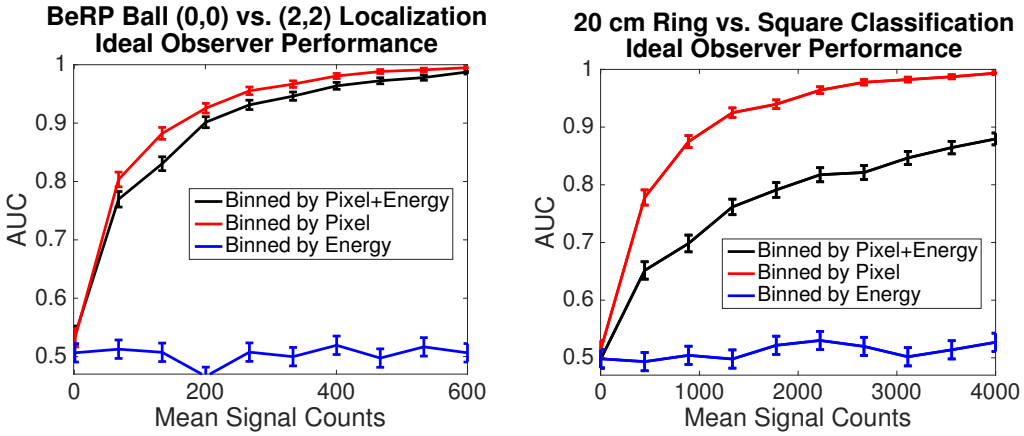


Figure 4.3: For the left plot, the ideal observer (with different binning choices) acts on BeRP ball data at (0 cm, 0 cm) and (2 cm, 2 cm). In the right plot, the ideal observer acts on the square and ring source data.

two distributions, that item could be rejected. Accepting values within the middle 90% for the two distributions, only 12.4% of the time the BeRP ball at (0 cm, 0 cm) would be rejected, and only 13.8% of the time the BeRP ball at (2 cm, 2 cm) would be rejected. If the middle 95% was accepted, the entire range of test statistics covered by the x axis in Figure 4.4 would fail to be rejected.

The monitor's ability to reject alternative hypotheses can be improved by using higher statistics. Upping the mean number of acquired counts to 20,000 significantly improved the probability of correctly rejecting the BeRP ball. Using the middle 90%, the rejection rates increase to 95.1% (BeRP ball at (0 cm, 0 cm)) and 94.5% (BeRP ball at (2 cm, 2 cm)). If the middle 95% was accepted, the rejection rates are 91.5% (BeRP ball at (0 cm, 0 cm)) and 91.9% (BeRP ball at (2 cm, 2 cm)). Using these higher statistics, the ideal observer trained for the geometric-source discrimination task could effectively distinguish and reject the BeRP ball.

4.2.2.4 Spoof Rejection Example 2

A second study was performed with the ideal observer trained to discriminate the 16 cm ring source from the 20 cm ring source. This study is complementary to the prior study. The count rates were set equal for all of the imaged sources. For an acquisition time corresponding to 20,000 counts observed and 1,000 samples of testing data taken, and rejecting values outside the middle 95% of the test-statistic distribution, the following results were found.

- The 24 cm ring source was rejected 31.1% of the time.

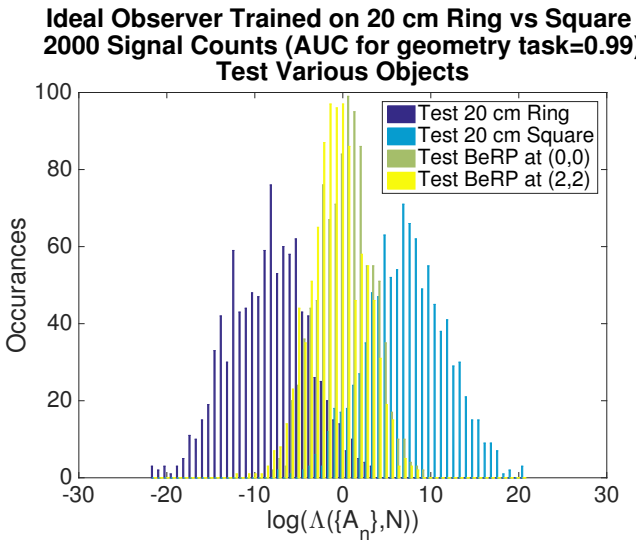


Figure 4.4: Test-statistic distributions for each object with an average of 2,000 signal counts being recorded. The AUC for the ring and square source distributions is over 0.99. The BeRP ball test statistic distributions fall between the ring and square source geometry distributions, but the test-statistic values are not rejected.

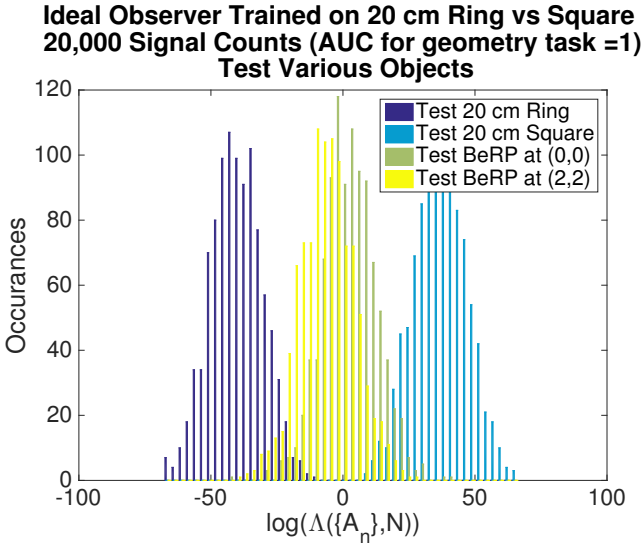


Figure 4.5: Test statistic distributions for the sources with an average of 20,000 signal counts being recorded for each distribution. The AUC for the ring and square source distributions is 1. The BeRP ball distributions fall between the ring and square source geometry distributions. For this acquisition time, both distributions correctly reject the BeRP ball.

- The 16 cm square source was rejected 100% of the time.
- The 20 cm square source was rejected 6.5% of the time.
- The 24 cm square source was rejected 84.8% of the time.
- The BeRP ball located at (0 cm, 0 cm) was rejected 96.5% of the time.
- The BeRP ball located at (2 cm, 2 cm) was rejected 91.5% of the time.

The ideal observer is therefore effectively able to reject most of the other simulated sources at 20,000 counts. The variation in rejection efficiency for these spoofs is not trivial to understand; the 16 cm ring and 20 cm ring are closest geometrically to the 16 cm square and 20 cm square, yet one of these sources is rejected easily and the other is not. Theoretically, these two should be the hardest objects to reject out of all 6, but this is not the case.

Intuitively, the mask is designed such that a delta function in object space produces an orthogonal mask pattern to that of a shifted delta function. In other words, an inner product between the image data for any two different locations in object space should be roughly constant. Likewise, the inner product between a difference of mask patterns for two different locations and the mask pattern for a third location should return a constant (roughly zero). It is not so clear how this information could be used to better understand the rejection rates for this model. The ideal observer is a ratio between mask patterns, not a simple difference, and the objects themselves are extended.

4.2.2.5 Procedure to Generate Spoofs

While higher statistics can help discriminate two distributions whose means vary significantly, they do not help to reject a tested item with a mean test-statistic value equal to that of the TAI. However, it is not difficult to create images that effectively spoof one of the test statistic distributions. In this section, the count rate was normalized, leaving only the LM terms to impact the ideal observer. The log of the SKE likelihood value is then,

$$\log(pr(\{A_n\}, N|H_j)) = \sum_{n=1}^{N} \log(pr(A_n|H_j)) \tag{4.25}$$

$pr(A_n|H_j)$ is given by the calibration data, and for the m^{th} bin, that probability density is equal to $\frac{g_{m,c}}{\sum_{m=1}^{M} g_{m,c}}$. When testing an unknown item, the number of detected

counts in each bin can be represented as g_m. The likelihood can then be represented as,

$$\log(pr(\{A_n\}, N|H_j)) = \sum_{m=1}^{M} g_m \log(\frac{g_{m,c}}{\sum_{m=1}^{M} g_{m,c}}) \tag{4.26}$$

The average likelihood value over multiple measurements of the unknown object is then equal to,

$$\langle pr(\{A_n\}, N|H_j)\rangle_{\mathbf{g}} = \sum_{m=1}^{M} \overline{g_m} \log(\frac{g_{m,c}}{\sum_{m=1}^{M} g_{m,c}}) \tag{4.27}$$

For example, consider a four pixel image where the calibration data is acquired and the likelihood of a particle being detected in each pixel is [0.1 0.25 0.25 0.4]. Imaging the *same* object that the model was trained on for 100 detections would result in roughly a likelihood mean of $(\log(0.1) * 10 + \log(0.25) * 25 + \log(0.25) * 25 + \log(0.4) * 40) = -56$. The tested item distribution could be manipulated from here; for the two bins with near equal probability, the tested item probability could be raised and lowered by the same amount. For example, the normalized data distribution on the tested item could potentially be [0.1 0.3 0.2 0.4]. This results in an average likelihood value of $(\log(0.1)*10+\log(0.25)*30+\log(0.25)*20+\log(0.4)*40 = -56)$. I am ignoring the change in variance here, and unlike the mean, the variance does change by manipulating the distribution on the tested item's data in this way.

For this study, the ideal observer was trained on the ring and square source data. The ring source was the item that was chosen to be spoofed in this study. The ideal observer was generated using an acquisition time corresponding to 20,000 signal counts being detected, which was high enough to properly reject the BeRP ball data in the prior subsubsection. I created a program that found the two probability values in $pr(A_n|H_j)$ closest together, added/subtracted the same number from those probabilities, and carried out this procedure a number of times. This number was chosen to be a random number between 0 and 10% of the minimum value of the two probabilities. The result of this procedure is spoofs 2, 3, and 4 shown in Figure 4.6. These were treated as data sets on the tested items. Note that with spoof 4, this was taken to the extreme, and the method starts to break down when the probability differences between the chosen bins becomes too large. In addition, spoof 1 was generated by subtracting a certain probability from the left side of the count map and adding it to the right. Given the symmetry of the count map, I expected that this would also lead to a successful spoof. Evidence that these spoofs were successful is shown in Table 4.1. Both the mean and variance of the spoof test

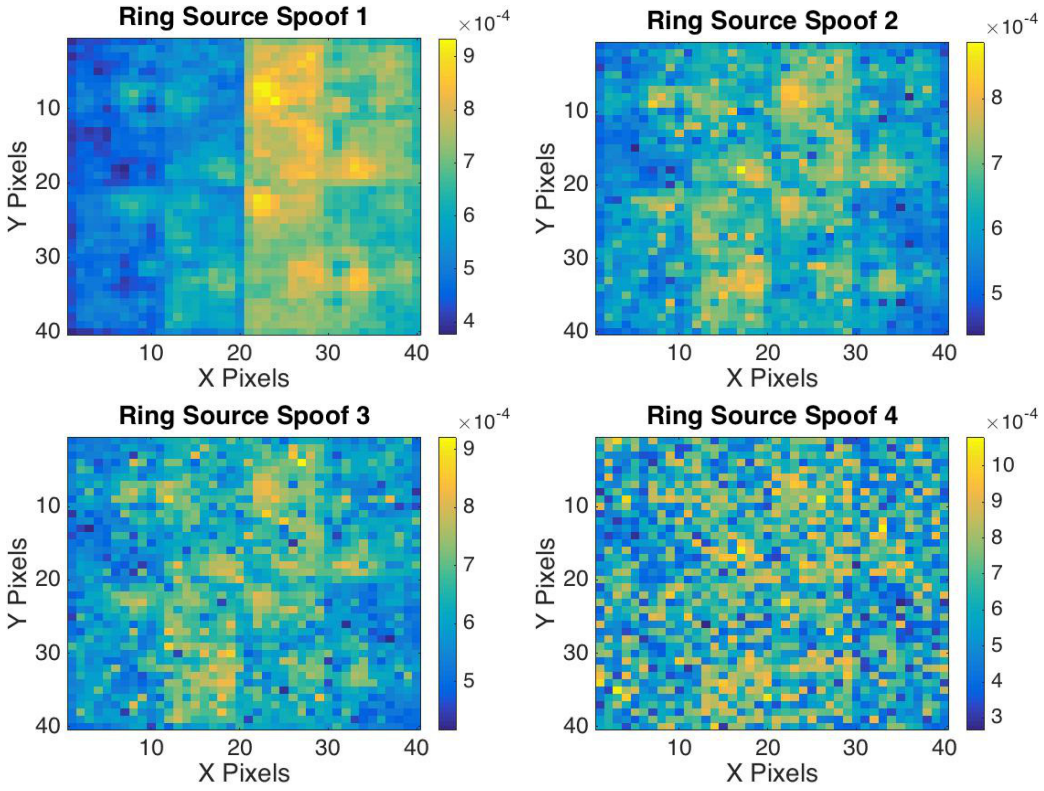


Figure 4.6: Four successful spoofs of the ring source for the ideal observer trained for the circle/ring discrimination task. The final three were run through the procedure outlined in this section. Spoof 2 used 100 iterations, spoof 3 used 400 iterations and spoof 4 used 4000 iterations.

statistic distributions align closely with the mean and variance of the ring's test statistic distribution.

This methodology ignored the fact that only certain images are possible with coded-aperture imaging. While this simplistic method leads to significantly noisy, non-realistic images, other methods could be used to create more physically realistic spoof images. A possible routine to do this is outlined below,

1. Change $pr(A_n|H_j)$ by adding and subtracting values in bins with near equal probability.

2. Reconstruct the object corresponding to this image.

3. In object space, the reconstructed objected is pushed towards a more physically realistic geometry.

4. Simulate data on this new model.

Tested item	$\overline{\Lambda}$	σ^2_Λ
20 cm Ring Source	-40.4	94
Spoof 1	-37.81	98.8
Spoof 2	-39.5	96
Spoof 3	-41	106
Spoof 4	-46	93.1

Table 4.1: The mean and variances for the test statistic distributions when imaging the four sources in Figure 4.6 in addition to the ring source that they are designed to spoof.

4.2.3 Monte Carlo Evaluation of Ideal Observer Incorporating Nuisance Parameters

This section provides a concrete example of the ideal observer that integrates over nuisance parameters as in (4.8).

4.2.3.1 IO8 vs. IO9 with Orientation Variability

Here it is assumed that TAIs of an unknown orientation are put inside opaque containers and are imaged by the detector, with every orientation of the source being equally likely. The goal of this experiment is to classify TAIs regardless of their orientation. A total of 60 evenly-spaced orientations of the objects were imaged. The Bayesian prior for the nuisance parameter was built assuming each of these orientations was equally likely. The assumption in this section is that the tested sources have the same pdf on the orientation nuisance parameter as the training sources. These studies were done with the strong background. Two separate studies are discussed.

The first study (see Figure 4.7) highlights the benefits of including the nuisance parameters in the observer model. An SKE ideal observer was found for the simulated sources with Arvo rotation 000. It was used to discriminate IO8 and IO9 testing data using Arvo rotation 000. Performance in this task is very strong. This model was then used to classify IO8 and IO9 data under rotation 111, and the observer performs worse than the guessing observer. These two orientations offer the most extreme disparity between the two sources, and predictably performance is quite poor in the case where the observer and testing data are mismatched. The observer model that averaged over all 60 orientations as in (4.8) was then used to discriminate simulated IO8 and IO9 data sets under each orientation. Performance improves upon using the observer that accounts for both orientations when classi-

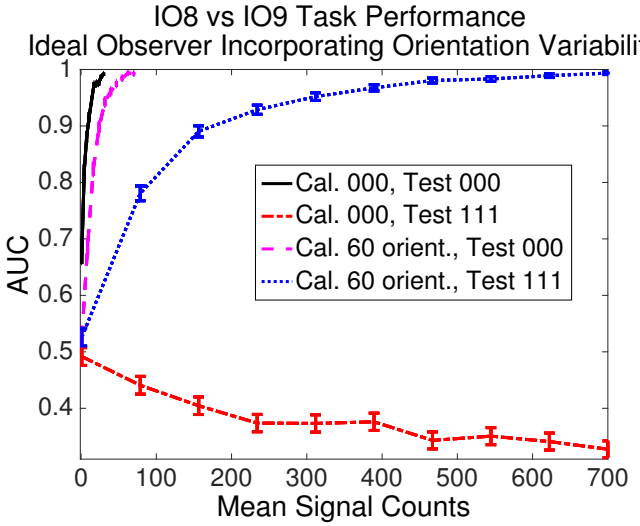


Figure 4.7: When the testing data is taken from a different orientation than the SKE ideal observer was derived, performance declines. Averaging over the orientation nuisance parameter in the observer model improves performance.

fying data from the 111 rotation.

In the second study, rather than test an individual orientation, the testing data was randomly sampled from one of the 60 orientations. An observer developed from the 000 orientation performs worse than the guessing observer in classifying all of the training data orientations, while the 111 rotation, whose data is more representative of most of the 60 rotations, performs fairly well. As expected, performance is best when the observer model that averages over the orientation nuisance parameter acts on the randomly sampled testing data, as in Figure 4.8. This study emphasizes that while strong (but non-optimal) performance can be retained without properly accounting for nuisance parameters, it is subject to the chosen nuisance parameter value or prior density on which the observer is built.

4.2.3.2 Test-Statistic Distributions

The incorporation of nuisance parameters allows for the possibility of non-normal distributions on the log of the ideal observer. As an example, for the IO8/IO9 discrimination task with orientation variability, the test statistic distributions are shown in Figure 4.9. IO8's is mostly normal, which is because the detector data does not vary much when IO8 is rotated. IO9's test statistic distribution, though, is decidedly non-normal. This holds up whether incorporating the count-rate differences or not. This non-normal behavior makes testing for spoofs, as shown in Section 4.2.2, even harder—the IO9 distribution is broad enough that it would be difficult to reject

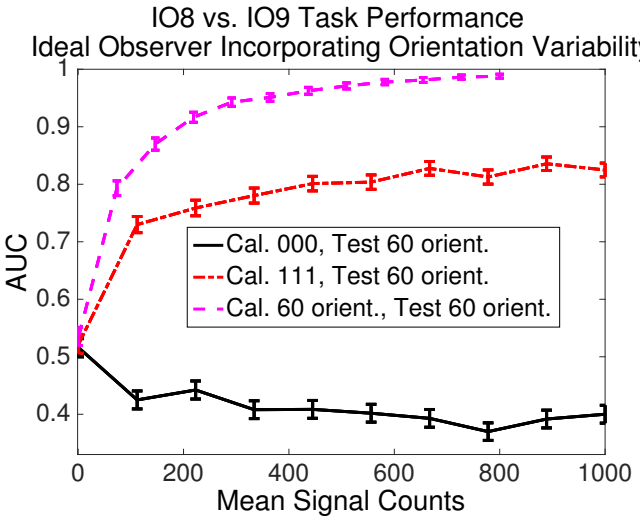


Figure 4.8: Performance varies significantly when using just a single orientation for IO8 and IO9 calibration data in the observer model to discriminate these sources when testing a random orientation. The observer integrating over all orientations performs the best.

a wide range of spoofs.

The test-statistic distribution itself does give the monitor hints as to the object's construction. If the monitor knows that orientation was a nuisance parameter, then they can discern from the test-statistic distribution that H_2 is an object with spatially dependent shielding and/or composed of SNM that prioritizes low-energy emissions.

4.2.4 Ideal Observer using Posterior Probability Density

This section presents a practical implementation of the observer model derived in the posterior pdf theory section.

4.2.4.1 IO8 vs. IO9 with Count-Rate Variability

This study represents a real-life scenario where a set of sources were created with the same geometry (and thus all should be classified as the same source) but with different emission rates. The observer assumes that there is a spread on activity rates, leading to a probability density over $\overline{N_1}$ and $\overline{N_2}$. A variable background strength is also accounted for in this study, and its corresponding detection rate $\overline{N_b}$ is an example of a shared nuisance parameter γ_0. The practical implementation

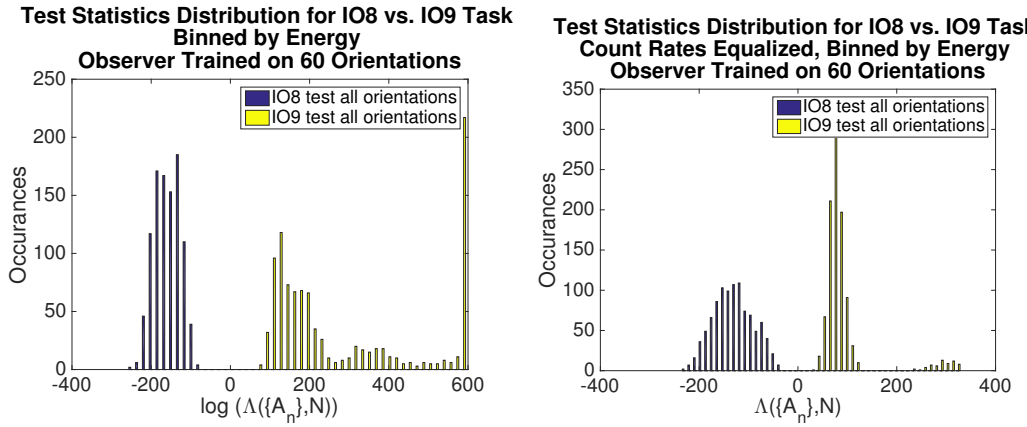


Figure 4.9: Test statistic distributions for IO8 and IO9 set at an acquisition time to acquire 20,000 signal counts. The left plot incorporates count-rate differences among the different orientations. In the right plot, all of the count rates for different orientations of the two sources were set equal so the ideal observer does not make decisions on that information. In both cases, the data was binned by energy. While the IO9 test-statistic distribution on the right does not appear continuous, a more thorough sampling of the orientation nuisance parameter space would yield a continuous distribution.

of (4.15) in this instance is,

$$\Lambda(\{A_n\}, N) = \int\int\int \Lambda_{SKE}(\{A_n\}, N|\overline{N_b}, \overline{N_1}, \overline{N_2}) \\ pr(\overline{N_b}, \overline{N_1}|\{A_n\}, N, H_1) pr(\overline{N_2}) \\ \mathrm{d}\overline{N_b}\mathrm{d}\overline{N_1}\mathrm{d}\overline{N_2}. \tag{4.28}$$

IO8 and IO9 calibration data from Arvo rotation 000 was read in and the observed count rate was assumed to be the mean of a normal pdf with a standard deviation equal to 40% of the mean. Sample data was found through randomly sampling the mean number of counts on each pixel according to this posterior density. Figure 4.10 shows that when only using the initial single set of calibration data, the observer model does a poor job classifying IO8 and IO9 objects with varying count rate, with an AUC value that flattens out around 0.9 from 100 to 500 signal counts. Using the observer model that incorporates count-rate variability in the above equation, improved performance is achieved.

4.2.5 Accounting for Imperfect Calibration Data

This subsection will present practical examples of the effect of using imperfect calibration data, and how the host could quantify the effect of this uncertainty on the performance.

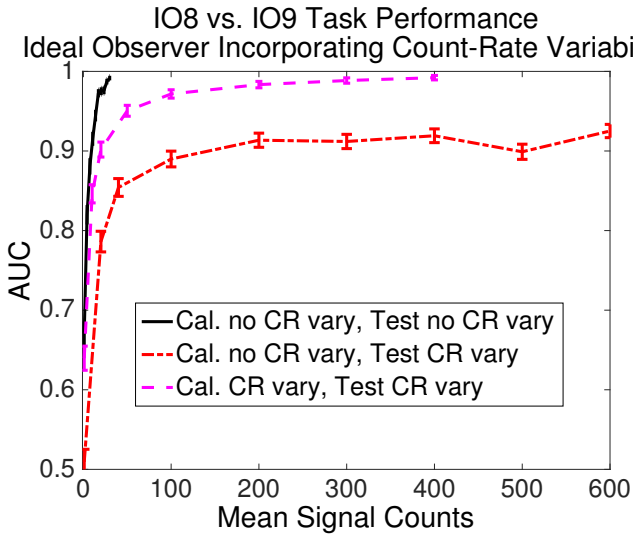


Figure 4.10: Performance of the SKE observer with known count rate (CR) declines when a range of source activity rates among the testing objects is assumed. Performance improves upon using the ideal observer that averages over the CR nuisance parameter (4.28) with the correct probability density on the CR.

4.2.5.1 Effect of Using Independent Testing Data

The models are trained on data simulated via Monte Carlo techniques. This data set is equivalent to a sample from the "true" probability distribution on the detector data. This section compares performance when the sampling distribution for the tested data comes from the training data or from an independently simulated set. When classifying data sets sampled from the training data distribution, performance is unrealistically good. Testing data sampled from the independently simulated distribution leads to worse, more physically realistic performance.

In reality, all of the measurements would be independent samples from the "true" detector distribution. If the training data set had high enough statistics, one would expect it would well approximate the "true" distribution and the performance would be roughly the same whether testing the sets sampled from the calibration or independently simulated.

The performance of the various ideal observer terms is shown for two separate tasks in Figure 4.11. In classifying IO8 and IO9, the performance of each of the terms stays relatively constant. This is because the data is binned into a relatively small number of bins (10 energy bins correspond to a range of roughly 500keV), allowing for higher statistics to accumulate in each bin. In comparison, the performance of the LM terms for the geometry classification task drops precipitously when testing on the calibration data compared to testing data. This is because the data is binned

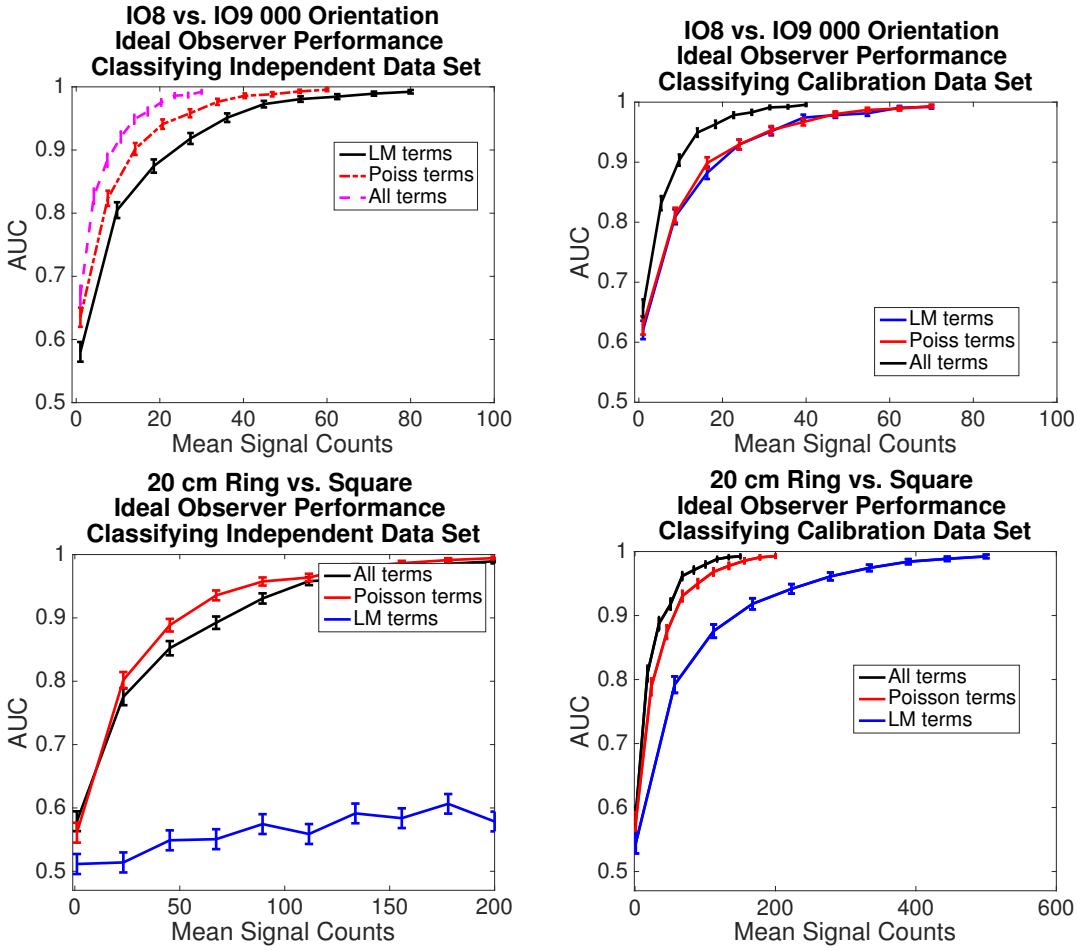


Figure 4.11: The top row of plots shows the effect of classifying the calibration data as opposed to independently simulated data for the IO8 and IO9 classification task. Performance of the LM component in this study drops when classifying an independent data set. The bottom row shows the effect when classifying the 20 cm ring and square sources. These plots did not assume the count rates were equal. These plots show a dramatic decrease in LM performance.

into 1,600 bins, requiring a much higher acquisition time to make decisions when the statistics of the testing and training data sets do not match. Meanwhile, in both studies, the Poisson component stays fairly consistent as it is essentially a single bin of data.

I do not believe that there is a method that can be used to produce an accurate AUC value despite the lack of sufficient statistics on the training data sets. As discussed in Section 4.1.4, an attempt was made to integrate over possible "true" detector-data distribution values, but this arrives at similar performance results to the SKE case.

An initial look at figures where the Poisson component outperforms the complete ideal observer (Figure 4.11), or the imaging data outperforms the spatio-spectral

data (Figure 4.3) brings up an interesting point in regards to the ideal observer. While this could be the result of limited statistics on the calibration data, it is also possible that the Poisson and LM component, when combined, yield worse performance. The Poisson and LM information are indiscriminately combined when evaluating the ideal observer. Taking the log of the ideal observer,

$$\begin{aligned}\log(\Lambda(\{A_n\}, N)) = &(\log(Pr(N|H_2)) - \log(Pr(N|H_1))) + \\ &\left(\sum_{n=1}^{N} \log(pr(A_n|H_2)) - \log(pr(A_n|H_1))\right)\end{aligned} \tag{4.29}$$

Each of these four terms can be treated as a normal distribution (in the SKE case). The distribution on $\log(\Lambda)$ is then a sum/difference of distributions on the Poisson and LM components. The difference of the log of the Poisson terms is roughly normal for a high acquisition time, as well the as the LM terms on the right. The logged ideal observer can then be represented as the sum of two normal distributions. It is possible for the SNR^2 that results from the sum of distributions for H_1 and H_2 to be less than the SNR^2 for one of those distributions. This is demonstrated using the Poisson, LM and complete ideal observer test statistic distributions. Throughout this discussion, I use SNR^2 essentially as a surrogate for the AUC, though in reality this is only true when the variances are equal. Still, the SNR^2 does give some information on the separation of the test-statistic distributions. First, the probability densities on the Poisson and LM test statistics are below,

$$\begin{aligned} pr(t_P|H_1) &= \mathcal{N}(\mu_{1,P}, \sigma^2_{1,P}) \\ pr(t_P|H_2) &= \mathcal{N}(\mu_{2,P}, \sigma^2_{2,P}) \\ pr(t_{LM}|H_1) &= \mathcal{N}(\mu_{1,LM}, \sigma^2_{1,LM}) \\ pr(t_{LM}|H_2) &= \mathcal{N}(\mu_{2,LM}, \sigma^2_{2,LM}). \end{aligned} \tag{4.30}$$

From these equations, the SNR^2 between the H_1 and H_2 test statistic distributions for the Poisson and LM components can be determined,

$$\begin{aligned} \Delta\mu_P &= \mu_{2.P} - \mu_{1,P} \\ \Delta\mu_{LM} &= \mu_{2,LM} - \mu_{1,LM} \\ \sigma^2_P &= \sigma^2_{2,P} + \sigma^2_{1,P} \\ \sigma^2_{LM} &= \sigma^2_{2,LM} + \sigma^2_{1,LM} \\ SNR^2_P &= \frac{(\Delta\mu_P)^2}{\sigma^2_P} \\ SNR^2_{LM} &= \frac{(\Delta\mu_{LM})^2}{\sigma^2_{LM}} \end{aligned} \tag{4.31}$$

The SNR^2 for the combined LM and Poisson distributions is,

$$SNR^2_{P+LM} = \frac{(\Delta\mu_P + \Delta\mu_{LM})^2}{\sigma^2_P + \sigma^2_{LM}} \tag{4.32}$$

Through some algebra, condition for when $SNR^2_{P+LM} < SNR^2_P$ can be determined,

$$\begin{aligned} \frac{(\Delta\mu_P + \Delta\mu_{LM})^2}{\sigma^2_P + \sigma^2_{LM}} &< \frac{(\Delta\mu_P)^2}{\sigma^2_P} \\ \frac{\Delta\mu^2_{LM}}{\Delta\mu^2_P} + 2\frac{\Delta\mu_{LM}}{\Delta\mu_P} &< \frac{\sigma^2_{LM}}{\sigma^2_P} \end{aligned} \tag{4.33}$$

In the worst case, $\Delta\mu_{LM} = 0$ and σ^2_{LM} is high, resulting in a higher SNR^2 using just the Poisson terms than using the Poisson and LM terms.

It should be mentioned that there is a relationship between the mean and variance of a normal log-likelihood distribution, as discussed in Barrett's work (Barrett et al., 1998). He derives that the mean for a normal log-likelihood distribution H_1 distribution should be equal to the negative of half of the variance for that distribution. The mean under H_2 should then be the negative of the mean underH_1, and the variance under H_2 should be equal to that under H_1. Using this, the variables in (4.31) can be redefined in terms of the hypothesis two parameters:

$$\begin{aligned} \Delta\mu_P &= 2\mu_{2,P} \\ \Delta\mu_{LM} &= 2\mu_{2,LM} \\ \sigma^2_P &= 2\sigma^2_{2,P} \\ \sigma^2_{LM} &= 2\sigma^2_{2,LM} \end{aligned} \tag{4.34}$$

This leads to a final condition for when the SNR^2 decreases when Poisson and LM distributions are combined in terms of the statistics under H_2:

$$\begin{aligned} \frac{(\Delta\mu_P + \Delta\mu_{LM})^2}{\sigma^2_P + \sigma^2_{LM}} &< \frac{(\Delta\mu_P)^2}{\sigma^2_P} \\ \frac{\mu^2_{2,LM}}{\mu^2_{2,P}} + 2\frac{\mu^2_{2,LM}}{\mu^2_{2,P}} &< \frac{\sigma^2_{2,LM}}{\sigma^2_{2,P}} \end{aligned} \tag{4.35}$$

4.2.5.2 Method to Account for Lack of Perfect Calibration Data

Here, I test the methodology to bound the variation in performance due to the fact that the calibration measurement is a sample from the unknown "true" detector data. This work is the implementation of the methods discussed in the last subsubsection of Section 4.1.4. The 20 cm circle/square geometry discrimination task was used to test this. The reason for this choice is that the performance plots in Figure 4.11 and Figure 4.3 show the complete ideal observer performing worse than the

Poisson component and full spatio-spectral binned information performing worse than just the spatially binned data.

$\mathbf{g}_{1,c}$ and $\mathbf{g}_{2,c}$ were set to the calibration data sets for H_1 and H_2. Calibration was read in and $pr(\mathbf{g}_{1,cs})$ and $pr(\mathbf{g}_{2,cs})$ were treated as normal random variables with mean and variance equal to $\mathbf{g}_{1,c}$ and $\mathbf{g}_{2,c}$. Data $\mathbf{g}_{1,c'}$ and $\mathbf{g}_{1,c'}$ was sampled from these $\mathbf{g}_{1,c}$ and $\mathbf{g}_{2,c}$, the ideal observer was trained on this data and then evaluated on the independent testing data set. This procedure was repeated for 400 samples from $\mathbf{g}_{1,c}$ and $\mathbf{g}_{2,c}$. The results are shown in Figure 4.12. For each of these plots, the acquisition time was set in advance.

The first study looks at the different components of the ideal observer. A slight drop in performance was observed compared to the SKE ideal observer for the LM and complete ideal observer models. Because the statistics are high enough, performance using just the Poisson component stayed approximately the same.

In the second study, the ideal observer was trained and tested on data binned in different ways. When binned by energy, the observer acts as a guessing observer for both the SKE ideal observer and the observer acting on resampled calibration data sets. When binned by pixel, performance is consistently strong in both cases. But a significant change is seen when binning into pixel and energy values—the AUC drops from 0.80 in the SKE case to 0.69 in the the resampled calibration data case. This implies that the performance seen in the SKE case is largely due to statistical chance and that a second simulation of the calibration data could yield significantly different results. These results give us one piece of evidence regarding whether the resulting AUC value returned by the ideal observer is reliable.

4.3 Conclusion and General Comments

The Bayesian ideal observer is able to process LM data and offers optimal performance, both in the SKE case and in the presence of nuisance parameters. The nuisance parameter distributions would need to be accurately estimated by the host, and any deviation from these distributions would result in nonoptimal classification of the objects. The ideal observer also has demonstrated the ability to reject other sources, specifically alternative neutron sources, based on their image data. Studies of IO8 and IO9 with orientation variability show that the ideal observer's ability to discriminate spoofs degrades as more nuisance parameters are simulated and the test-statistic distributions become broader.

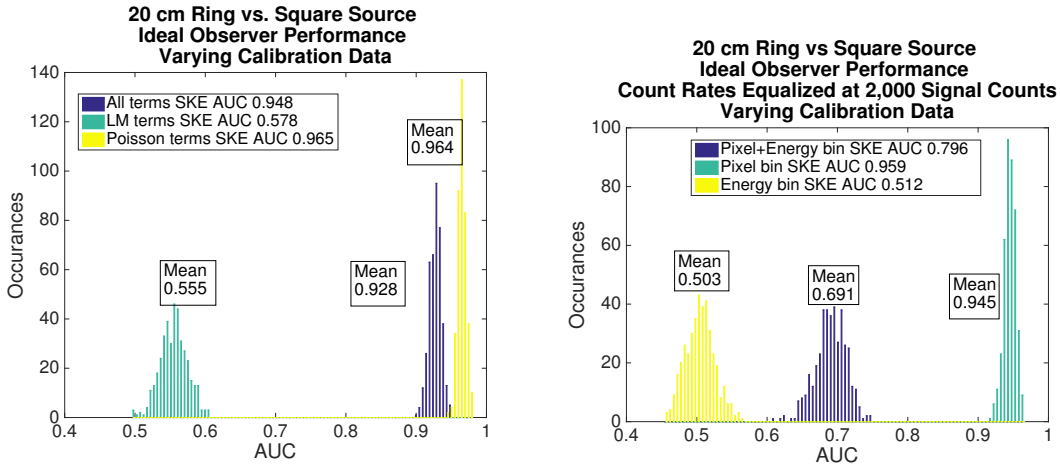


Figure 4.12: For each plot, the ideal observer was evaluated when resampling the calibration data from $pr(\mathbf{g}_{1,cs})$ and $pr(\mathbf{g}_{2,cs})$. In the left plot, the Poisson, LM and complete ideal observer models were evaluated for the circle/square classification task. The right plot shows the performance of the data when binned by energy, pixel ID or both. The legends show the SKE ideal observer's AUC value, based only on the simulated calibration data in addition to the mean of the shown distributions.

Ultimately, I am skeptical of how useful the ideal observer would be in a treaty-verification setting compared to the standard IB approach discussed in Section 1.5.2. In the CIVET system, the monitor and host jointly agree on intelligence for the system, and the monitor can verify that the system can properly categorize inspection objects based on these measurements. This system aggregates data, but the monitor never sees it. For the ideal observer, calibration measurements would be taken on trusted TAIs, and this sensitive calibration data and model would need to be stored behind an IB. The host would train the ideal observer with this data and without monitor oversight, because the monitor could not access this sensitive data. Then the model is tested on unverified TAIs, and the monitor must trust that the host did not bake a spoof into the training of the ideal observer. Furthermore, the testing of nonsensitive inspection objects does not help the monitor gain any knowledge due to massive dimensionality reduction from the measurement data to t.

CHAPTER 5

Development of Hotelling and Channelized Hotelling Observers that Prevent Discrimination on Sensitive Information

While the Bayesian ideal observer is a useful tool, and offers optimal performance, the fact that the model would need to be stored behind an information barrier would make a monitor hesitant to agree to using it in a treaty-verification setting. In this chapter, observer models that act as linear discriminants are discussed. These models process LM data linearly. The purpose of this chapter is threefold.

- The Hotelling observer (HO) (Hotelling, 1931) is a linear observer with equivalent performance to the ideal observer when the statistics of the data are Gaussian. The Hotelling weights are applied to binned testing data to yield a test statistic which is then thresholded to make a decision. This chapter demonstrates the advantages that this model provides over the ideal observer, including less stored information and an inability to back out the image data.

- There are advantages to be gained through utilization of the channelized Hotelling observer (CHO) (Barrett et al., 1993). With this method, a series of templates are applied to the image data, resulting in a set of channelized values; the optimally weighted sum of these values gives the test statistic. This method essentially gives the monitor access to multiple test statistics, which could be used for better discrimination of spoofs.

- Additional penalty terms can be incorporated into the CHO's optimization routine to either maximize the information available to the monitor while maintaining optimal performance or to create non-optimal channels that the monitor could access.

The theory behind HO and CHO models and the results for the IO8 vs. IO9 discrimination task were presented at the 2015 IEEE NSS/MIC Symposium (MacGahan et al., 2015). The work presented here will be submitted to Nuclear Instruments and Methods in Physics Research A in the upcoming months (MacGahan et al., 2016b). This chapter contains three sections. Section 5.1 introduces the theory for

the models and Section 5.2 shows how these models perform in practice. Section 5.3 presents a summary of the developed methods.

5.1 Theory

A linear model uses the binned form of the data $\mathbf{g}$, which can be represented in terms of the LM data $\{A_n\}$ and number of counts N by,

$$g_m = \sum_{n=1}^{N} f_m(A_n), \tag{5.1}$$

where f_m is the binning function for the m^{th} bin. A linear observer then applies a set of weights $\mathbf{W}$ to $\mathbf{g}$ to return a scalar test statistic t,

$$t = \mathbf{W}^{\dagger}\mathbf{g} = \sum_{n=1}^{N}\sum_{m=1}^{M} W_m f_m(A_n). \tag{5.2}$$

The weights for each bin W_m (also called the template throughout this chapter) determine how well the test statistic distributions are separated when performing the model. Both the HO and CHO are examples of linear template models.

Potential spoofs are more seriously considered in this chapter than in the previous one. To do this, it is important to know the statistics on the test-statistic distribution. The statistics on $\mathbf{g}$, the number of counts in each bin, are Poisson. The test statistic t is then a sum of many weighted (due to $\mathbf{W_g}$) Poisson random variables. If $\mathbf{g}$ is appreciably large, as is often the case when imaging an object or performing a high-resolution gamma measurement, t will be normal due to the central limit theorem with mean and variance given by,

$$\begin{aligned} \bar{t} &= \sum_{m=1}^{M} W_m \overline{g_m} \\ \sigma_t^2 &= \sum_{m=1}^{M} W_m^2 \overline{g_m}. \end{aligned} \tag{5.3}$$

5.1.1 Hotelling Observer

The HO is the linear observer that maximizes the SNR^2 (1.12) between the test-statistic distributions of the two objects being measured (Barrett and Myers, 2003). It is also equivalent to the Bayesian ideal observer when the statistics of the data are multivariate normal if the covariance matrices on the data are equal for the two

hypothesis. The Hotelling weights, which are denoted as $\mathbf{W_H}$, are defined by,

$$\begin{aligned} \mathbf{W_H} &= \mathbf{K_g^{-1}}\Delta\overline{\mathbf{g}} \\ \mathbf{K_g} &= \frac{\mathbf{K_1} + \mathbf{K_2}}{2} \\ \Delta\overline{\mathbf{g}} &= \boldsymbol{g_2} - \boldsymbol{g_1}. \end{aligned} \tag{5.4}$$

$\mathbf{g}_j$ is the mean data set for source j and $\mathbf{K_j}$ corresponds to the covariance matrix for source j. It is useful to think about the HO in terms of the operator $\mathbf{K_g^{-1/2}}$, where $\mathbf{K_g^{-1/2}K_g^{-1/2}} = \mathbf{K_g^{-1}}$. Using this operator, the HO can be presented as,

$$\begin{aligned} t =& \Delta\overline{\mathbf{g}}^\dagger \mathbf{K_g^{-1}}\mathbf{g} \\ t =& \Delta\overline{\mathbf{g}}^\dagger \mathbf{K_g^{-1/2}K_g^{-1/2}}\mathbf{g} \\ t =& (\mathbf{K_g^{-1/2}}\Delta\mathbf{g})^\dagger(\mathbf{K_g^{-1/2}}\mathbf{g}). \end{aligned} \tag{5.5}$$

There is a difference between $\mathbf{K_g^{-1/2}}$ in this equation and the whitening operator for a given data set $\mathbf{K}^{-1/2}$. When the whitening operator is applied to a data set as in $\mathbf{g}^{(1)} = \mathbf{K}^{-1/2}\mathbf{g}$, it returns a data set with unity variance. The HO is outlined in (5.5) is comparable but as it is the average of the covariance matrices for the two hypotheses, it will not return unity variance. Still, there is a decorrelating effect, and this method does produce a test statistic that is normal in nature.

5.1.1.1 Calculation of Hotelling Weights

The averages in (5.4) are over the unknowns in the imaging system. Always present is Poisson noise, and in an SKE study, where there is no other variability in the detector data, the covariance matrices are diagonal with values equal to the mean. In this case, taking the inverse of the covariance matrix is trivial; the inverse is also diagonal with values equal to the inverse of the covariance matrix values.

Upon incorporation of nuisance parameters, the averages become doubly stochastic or worse and the problem becomes more difficult. The terms in the Hotelling template are,

$$\begin{aligned} \mathbf{K_j} &= \left\langle \left\langle (\mathbf{g}_j - \overline{\overline{\mathbf{g}_j}})^\dagger (\mathbf{g}_j - \overline{\overline{\mathbf{g}_j}}) \right\rangle_{g_j|\gamma_j} \right\rangle_{\gamma_j} \\ \overline{\overline{\mathbf{g}_j}} &= \left\langle \left\langle \mathbf{g}_j \right\rangle_{\mathbf{g}_j|\gamma_j} \right\rangle_{\gamma_j}. \end{aligned} \tag{5.6}$$

The first average is over Poisson noise while the second is over the nuisance-parameter distributions. The resulting matrix is generally dense. Calculating the inverse of this matrix is usually impractical. Using a data set with 1,600 pixels and

64 energy bins, there are approximately 1e5 elements in $\mathbf{g}$ with 1e10 elements in $\mathbf{K_g}$. The number of samples will likely be far less than the size of the data vector, leading to a non-invertible matrix. Certain estimation techniques can be used to overcome this limitation (Bickel and Levina, 2008; Bai and Shi, 2011), but require often unrealistic assumptions. In simulation, this computational barrier can be overcome using the Matrix Inversion Lemma (Woodbury, 1950). To use this lemma, the covariance matrix must be able to be represented in the form,

$$\mathbf{K_g} = \mathbf{A} + \mathbf{BCD}, \tag{5.7}$$

where $\mathbf{A}$ is diagonal. To get the covariance matrix in this form, one can add and subtract $\overline{\mathbf{g}_j}$, the Poisson averaged data for a given nuisance-parameter realization, in the covariance matrix equation of (5.6). $\mathbf{K_j}$ can then be expressed as,

$$\begin{aligned}
\mathbf{K_j} &= \left\langle \left\langle (\mathbf{g}_j - \overline{\overline{\mathbf{g}_j}})^\dagger (\mathbf{g}_j - \overline{\overline{\mathbf{g}_j}}) \right\rangle_{g_j|\gamma_j} \right\rangle_{\gamma_j} \\
\mathbf{K_j} &= \left\langle \left\langle (\mathbf{g}_j - \overline{\mathbf{g}_j})^\dagger (\mathbf{g}_j - \overline{\mathbf{g}_j}) \right\rangle_{g_j|\gamma_j} \right\rangle_{\gamma_j} + \left\langle (\overline{\mathbf{g}_j} - \overline{\overline{\mathbf{g}_j}})^\dagger (\overline{\mathbf{g}_j} - \overline{\overline{\mathbf{g}_j}}) \right\rangle_{\gamma_j} \\
\mathbf{K_j} &= \left\langle \mathbf{K}_{\mathbf{g}_j|\gamma_j} \right\rangle_{\gamma_j} + \mathbf{K}_{\gamma_j} = Diag(\overline{\overline{\mathbf{g}_j}}) + \mathbf{K}_{\gamma_j}.
\end{aligned} \tag{5.8}$$

The left matrix is diagonal with values equal to the mean of the detector data, averaged over Poisson randomness and all nuisance parameter densities. The right term is dense and contains the variation of the Poisson-averaged data over the nuisance-parameter distribution. A deeper look at (5.8) brings up another interesting feature of the HO. The Poisson covariance matrix increases linearly with the number of detected counts N. The covariance matrix over nuisance parameters, however, has a component that increases with N^2. Therefore, the optimal Hotelling weights are dependent upon acquisition time.

Treating orientation as a nuisance parameter as an example, a vector $\boldsymbol{\theta}$ can be defined which corresponds to the orientation of the object being imaged, where [1 0 0 ....0] and [0 1 0 ....0] correspond to different object orientations. Then a system-response matrix $\mathbf{H}_\theta$ is defined and resulting detector data $\mathbf{g} = \mathbf{H}_\theta \boldsymbol{\theta}$. $\mathbf{H}_\theta$ in the above equation is slightly different from the $\mathbf{H}$ in the imaging equation ((1.3)), which generally reflects system sensitivity to a given emission location and energy. The $\mathbf{H}_\theta$ in (5.9) is the system response for an imaged object under a certain set of nuisance parameters. The covariance matrix for source j, $\mathbf{K}_{\gamma_j}$, can then be represented by the detector response function and the covariance matrix for the $\boldsymbol{\theta}$ vector,

$$\mathbf{K}_{\gamma_j} = \mathbf{H}_\theta \mathbf{K}_{\boldsymbol{\theta}_j} \mathbf{H}_\theta^\dagger. \tag{5.9}$$

In simulation, the known nuisance-parameter distributions are used to find $\mathbf{K}_{\boldsymbol{\theta}_j}$. $\mathbf{H}_\theta$ is determined by assuming the GEANT4 data is the "true" system response to each object. This technique enables the application of the Matrix Inversion Lemma, reducing the problem from a MxM inverse, where M can be thousands to millions, to a PxP inverse, where P is the number of orientations chosen to average over.

5.1.1.2 Implementation

In practice, the host would gather calibration data and determine the weights. As discussed in the previous subsections, if nuisance parameters are present then the host could choose to guess the prior density on the nuisance parameters and come up with a system response in order to use the Matrix Inversion Lemma to take the inverse of $\mathbf{K_g}$. The host could also choose to reduce the size of the dataset considerably and take enough measurements on their objects for the matrix to be invertible. If the weights were deemed sensitive, the monitor would only be able to verify the algorithm and have access to the test statistic. The weights would need to be stored behind an information barrier.

Testing would work much like the ideal observer. Independent measurements would be taken on objects of type H_1 and H_2 and test statistic distributions generated from these measurements. A threshold would be determined based on these distributions, either by maximizing the probability of correctly identifying the two sources for a certain type I or II error, or some other measure. The model would then act on an unknown source, updating the test statistic event by event for a certain acquisition time. This test statistic would then be thresholded to make a decision.

5.1.1.3 Storage

The differences in storage between the HO and Bayesian ideal observer are critical in regards to information security. While the ideal observer stores the detector data for each realization in the covered nuisance-parameter space, the HO simply contains a product of first and second order statistics over this space. As an example that is explored further in Section 5.2.1, if orientation is a nuisance parameter, the spatial and spectral information will be blurred out by averaging over the nuisance-parameter distributions. Because of this, even if the monitor was able to gain access to $\mathbf{K}_\mathbf{g}^{-1}$ and $\overline{\Delta\mathbf{g}}$, its ability to reconstruct the details of the object being imaged would be limited.

5.1.1.4 A Cheating Host

As discussed in the chapter 4, spoofs could be discriminated based on their test statistic value. As the monitor would not have access to the $\mathbf{W_H}$, they could only use the test statistic distributions for the two sources to identify spoofs. The distributions are normal with mean and variance stated in (5.3). Any measured test statistics would be compared to these distributions to see if they fall within that spread. The simple form of the test statistic distribution lends itself to a straightforward spoof attempt. One could manufacture a data set that has the same statistics as those for one of the TAIs as discussed in Section 4.2.2. Because the procedure for creating the spoof would be largely the same, successful spoof generation is not discussed in this chapter.

5.1.1.5 A Cheating Monitor

The HO in (5.4) is analogous to a secondary imaging system that only see the differences between the two objects. If the monitor somehow gained access to $\mathbf{W_H}$, and tried to reconstruct $\mathbf{g}_{rec} = \mathbf{W_H}^{-1}t$, all it could back out is projection data that looked like the Hotelling weights, not the highly sensitive detector data. All other information in $\mathbf{g}$ is in the matrix's null space.

The weights themselves are sensitive. In a spectral-discrimination task, the Hotelling weights could help the monitor determine the difference in composition of the two objects, such as the fraction of U235 vs. U238 in a uranium object. However, the individual data sets $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ are more sensitive as the monitor can directly relate these measurements back to the objects. There are conditions where the monitor could back out these data sets. If the monitor had knowledge of one of the two data sets, $\boldsymbol{g_2}$, they could solve for $\boldsymbol{g_1}$. This is especially easy in the SKE case, when the covariance matrix is diagonal. Simply rewriting the definition for the Hotelling weights in terms of $\boldsymbol{g_1}$,

$$\begin{aligned} \mathbf{W_H} &= 2 * \frac{\boldsymbol{g_2} - \boldsymbol{g_1}}{\boldsymbol{g_2} + \boldsymbol{g_1}} \\ \boldsymbol{g_1} &= \boldsymbol{g_2}\frac{2 - \mathbf{W_H}}{2 + \mathbf{W_H}}. \end{aligned} \tag{5.10}$$

The procedure is more difficult when nuisance parameters are present. If there are no nuisance parameters in source 2 (with a known data vector $\boldsymbol{g_2}$), but there are nuisance parameters in source 1 (unknown), the Hotelling weights are

$$\mathbf{W_H} = 2 * (Diag(\boldsymbol{g_2} + \boldsymbol{g_1}) + \mathbf{K}_{\gamma,1})^{-1} * (\boldsymbol{g_2} - \boldsymbol{g_1}). \tag{5.11}$$

Here, $\mathbf{K}_1$ has already been broken down into its diagonal and dense components (which are both unknown to the monitor). $\mathbf{K}_{\gamma,1}$ could be represented as a matrix product BCD (also unknown to the monitor) as in (5.9) and then the Matrix Inversion Lemma applied. Ultimately, there does not appear to be an easy way to find $\mathbf{K}_{\gamma,1}$ and $\boldsymbol{g_1}$ from the above equation, but the monitor could use prior knowledge on the object geometry and simulate perturbations of it to create an object resulting in the correct $\mathbf{W_H}$. If both objects used to form $\mathbf{W_H}$ are unknown, or if nuisance parameters are present in both objects, the procedure becomes even more difficult.

Finally, it could be of interest to the monitor to test a series of inspection objects to potentially bound the test statistic of the TAIs in order to increase confidence in the tested item. One could imagine the monitor imaging a series of objects ranging from the BeRP ball to a slab of plutonium to other potential sources. Due to the massive dimensionality reduction involved in taking $\mathbf{g}$ to t, bounding this value would be impossible. However there is some potential for concern on the host's part. Each measurement on a known object creates an additional constraint that the monitor could potentially use to back out $\mathbf{W_g}$. Due to the immense dimensionality reduction (for an image with this system, $\mathbf{g}$ has 1,600 values) it would take too many measurements for this procedure to put the Hotelling weights at risk.

5.1.2 Channelized Hotelling Observer

The CHO has become widespread in the field of medical imaging for a few reasons. First, it is a cheap alternative to a professional radiologist in image quality studies (Yao and Barrett, 1992; Wollenweber et al., 1999) and has proven to model human performance well in SKE tasks with a localized signal. In addition, it reduces the size of the data, requiring a far more practical number of data sets to train the model than the HO. In this thesis it is used for a much different purpose—information security. I believe this is the first time the CHO has ever been used for this purpose.

The CHO applies a channelizing matrix $\mathbf{T}$ to the binned data $\mathbf{g}$ (of size M), resulting in a much smaller dimensional vector $\mathbf{v}$ of length L. L can be as large or small as the host and monitor desire. Using calibration data for the two objects, an optimal set of channelized weights $\mathbf{W_v}$ are then found, and applied to channelized

testing data to make decisions,

$$\begin{aligned} \mathbf{v} &= \mathbf{T}\mathbf{g} \\ t &= \mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{v}, \end{aligned} \tag{5.12}$$

where the weights that best separate the resulting test statistic distributions are,

$$\mathbf{W}_{\mathbf{v}} = \mathbf{K}_{\mathbf{v}}^{-1}\overline{\Delta\mathbf{v}}. \tag{5.13}$$

$\mathbf{K}_{\mathbf{v}}^{-1}$ and $\overline{\Delta\mathbf{v}}$ are analogous to the terms in (5.4). The CHO can also be expressed as a template on $\mathbf{g}$ by defining,

$$\mathbf{W}_{\mathbf{g}}^{\dagger} = \mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{T}. \tag{5.14}$$

This form is used throughout this chapter.

The performance of this model depends on the $\mathbf{T}$ chosen. If $\mathbf{T}$ was set to a matrix of random numbers, performance (while better than the guessing observer) would be very poor. To achieve performance equivalent to the HO, it is necessary to find the matrix that best separates the test statistic distributions. This is done by maximizing the signal-to-noise ratio of the multivariate normal distributions on $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$,

$$SNR^2(\mathbf{T}) = \overline{\Delta\mathbf{v}(\mathbf{T})}^{\dagger}\mathbf{K}_{\mathbf{v}}^{-1}(\mathbf{T})\overline{\Delta\mathbf{v}(\mathbf{T})}. \tag{5.15}$$

The $\mathbf{T}$ that optimizes this function is found through a gradient descent optimization routine with backtracking (Boyd and Vandenberghe, 2004). This routine requires both an objective function (the SNR^2) and the derivative of that function. In order to take the derivative of the expression in (5.15), matrix calculus must be used (Bodewig, 2014). With an optimal $\mathbf{T}$,

$$\mathbf{W}_{\mathbf{H}} =\approx \mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{T}, \tag{5.16}$$

and performance equivalent to the HO is achieved. Because the optimization routine needs to be stopped to limit computer time, the resulting $\mathbf{W}_{\mathbf{H}}$ will only serve as a strong approximation for $\mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{T}$. The results section shows that a standard optimization of $\mathbf{T}$ results in sensitive channels for certain tasks. The addition of a penalty term to the objective function,

$$f_{obj}(\mathbf{T}) = SNR^2(\mathbf{T}) - f_{pen}(\mathbf{T}), \tag{5.17}$$

offers some possibilities in circumventing this barrier. In Section 5.1.3, Section 5.1.4, and Section 5.1.5, three different methods are presented that either reduce total information or prevent discrimination on sensitive information.

There is one last interesting note that is important to point out. The SNR^2 between the $\mathbf{v}$ distributions is maximized by this procedure. However, the test statistic is the inner product of $\mathbf{W_v}$ and $\mathbf{v}$. The difference between the mean test-statistic values can then be represented as,

$$\begin{aligned}\overline{\Delta t} &= \mathbf{W_g^\dagger}\Delta\overline{\mathbf{g}} \\ \overline{\Delta t} &= \mathbf{W_v^\dagger}\mathbf{T}\Delta\overline{\mathbf{g}} \\ \overline{\Delta t} &= \overline{\Delta\mathbf{v}}^\dagger\mathbf{K_v^{-1}}\mathbf{T}\Delta\overline{\mathbf{g}} \\ \overline{\Delta t} &= SNR^2.\end{aligned} \tag{5.18}$$

The SNR^2 between the $\mathbf{v}$ distributions is equivalent to the difference in mean test-statistic value. This is another effect of the decorrelating weights.

5.1.2.1 Implementation

Training of the CHO is a far more computationally practical task than the HO. Rather than needing M measurements to generate an invertible covariance matrix, the host can get by with (at a minimum) L measurements. $\mathbf{K_g}$ and $\overline{\mathbf{g}}$ can be found through the L samples, then the optimization occurs using $\mathbf{K_v} = \mathbf{T}^\dagger\mathbf{K_g}\mathbf{T}$ and $\overline{\mathbf{v}} = \mathbf{T}\overline{\mathbf{g}}$. This is a fundamental advantage for the CHO in a practical setting.

$\mathbf{T}$ would first be determined from calibration data for the two sources. Then, using the calibration data, the optimal weights $\mathbf{W_v}$ would be calculated for that $\mathbf{T}$. This model would be tested on the trusted items to generate a test-statistic distribution and the threshold set based on these distributions and the cost functions for incorrect outcomes, as discussed in the Implementation section of prior models. Future sources would then be classified with this model.

5.1.2.2 Storage

Compared to the HO, storage has been changed from the sensitive Hotelling weights $\mathbf{W_H}$ to two variables—a channelizing matrix $\mathbf{T}$ and a set of channelized weights $\mathbf{W_v}$. For an optimal $\mathbf{T}$, the monitor should therefore only access $\mathbf{T}$ or $\mathbf{W_v}$, or a subset of the two, but not the entirety of both. However, as shown in the results section, the optimization of $\mathbf{T}$ often results in sensitive channels, depending on the task. Even if $\mathbf{T}$ is nonsensitive, it is significantly easier in practice to hide the channelizing matrix than the weights. As $\mathbf{W_v}$ can be determined from aggregating $\boldsymbol{v_1}$s and $\boldsymbol{v_2}$s, it would be more convenient to hide $\mathbf{T}$. To hide $\mathbf{W_v}$, the host would need to hide the channel values $\mathbf{v}$ as well, which defeats the purpose of using the model in the

first place. Therefore, a standard implementation of the CHO in a treaty verification setting would have the host gather calibration data and determine the channelizing matrix. The monitor would only have access to $\mathbf{W_v}$, $\mathbf{v}$ and t.

5.1.2.3 A Cheating Host

The CHO does bring two additional benefits in discriminating spoofs:

- The L channel values themselves are nonsensitive. As the number of channels increases, the monitor has more information available to use to discriminate spoofs from the TAIs—the individual channel values can be aggregated as more sources are tested and the monitor could use that data to test for spoofs.

- There are an infinite number of channelizing matrices that maximize the SNR^2 in (5.15). This randomness goes a long way to increasing monitor faith that the host is not placing a spoof in front of the detector, as the host does know in advance what the channels are, and therefore cannot design a spoof to return the same value. The host country could store the calibration data, generating a new $\mathbf{T}$ for each new set of verification measurements.

The channelized values are also approximately normally distributed (for a large number of detector bins M), with each channel having mean and variance,

$$\begin{aligned} \overline{t_l} &= \sum_{m=1}^{M} T_{l,m}\overline{g_m} \\ \sigma_{t_l}^2 &= \sum_{m=1}^{M} T_{l,m}^2\overline{g_m}, \end{aligned} \tag{5.19}$$

where $T_{l,m}$ corresponds to the value in the l^{th} channel and m^{th} bin. An example of the CHO's ability to discriminate spoofs is discussed in the results section.

These extra channels provide advantages in detecting spoofs, but the channels themselves tend to be noisy versions of the Hotelling weights. If the monitor desired better detection of spoofs, additional terms would need to be added to the optimization of the channelizing matrix to maximize the separation of the test-statistic distributions between the TAIs and spoof objects. To accomplish this, an additional term could be added to eq. (5.17) in the form of,

$$f_{spoof} = \eta_{spoof} \sum_{n_{spoof}=1}^{N_{spoof}} (SNR_{1,n_{spoof}}^2 + SNR_{2,n_{spoof}}^2). \tag{5.20}$$

Such a penalty term would maximize the distance between the test-statistic distributions for the two TAIs and a set of N_{spoof} objects. η_{spoof} would determine how strongly the TAI-spoof optimization impacts the resulting $\mathbf{T}$ rather than the two TAIs in the discrimination task.

5.1.2.4 A Cheating Monitor

If the monitor was to cheat and gain access to $\mathbf{T}$, all would not be lost from the host's perspective. $\mathbf{T}$ is non-square and hence not invertible; instead, its Moore-Penrose pseudoinverse can be used (**?**), and it will be represented by $pinv(T)$. Similar to the HO discussion, $pinv(\mathbf{T})\mathbf{v}$ usually returns projection data that looks like the Hotelling weights. This happens because the optimization routine maximizes the SNR^2 and the individual channels take on the nature of the Hotelling weights. This is not true for a general $\mathbf{T}$. One could set each channel of $\mathbf{T}$ to a different basis function, and reconstruct far more information on each object's projection data than for the $\mathbf{T}$ that maximizes the SNR^2.

In its desire to create channels that can detect spoofs, the monitor could put the security of the host's objects at risk. A standard implementation of the CHO results in channels that are noisy versions of the Hotelling weights. If extra terms are added to the optimization routine to identify certain spoofs, these channels will have more varied information; such a channelizing matrix could do a better job reconstructing $\mathbf{g}$ from $\mathbf{v}$.

Hypothetically, host and monitor could agree on performing the model on a series of nonsensitive objects with known image data. If the monitor takes enough measurements, it would acquire $\mathbf{v}$ for each known object (with a known $\mathbf{g}$). Backing out information on the channelizing matrix is about as difficult as finding $\mathbf{W_H}$ for the HO, except now there are L templates rather than one.

5.1.3 Method to Generate Nonsensitive Channels

This approach attempts to maximize the amount of nonsensitive information available to the monitor while still maintaining optimal performance. Nonsensitive channels were created by choosing a penalty term that reduces the performance of each individual channel in distinguishing the objects in the binary-classification task,

$$f_{pen}(\mathbf{T}) = \eta \sum_{l=1}^{L} SNR^2_{l^{th}channel}(\mathbf{T}_{l^{th}channel}). \tag{5.21}$$

This penalty term is referred to as the "channel performance penalty" throughout this paper. Utilizing this penalty, it is possible to maintain optimal performance, as the optimization routine now focuses on the relationships between the channels rather than the channels themselves.

As the results section shows, this method is not a perfect answer for the information security problem. While the individual channels are nonsensitive, if the monitor was given the entire $\mathbf{T}$, a singular value decomposition (Golub and Reinsch, 1970) enables them to back out the Hotelling weights. However, it does allow the host to give the monitor a large number (but not all) of the channels. This would increase monitor confidence that the channelization process is working properly.

5.1.3.1 Implementation and Storage

Implementation and storage for this model would be functionally the same as the standard CHO. The only difference is that the monitor would be given a certain number of channels in the channelizing matrix. The monitor could also test benign inspection objects to verify that the channelization procedure is performing as expected.

5.1.3.2 A Cheating Host

The additional information available to the monitor is a net benefit here. The monitor would be given the channelized values for the host's items as well as the channels. The fact that the monitor has access to some of the channels would allow them to test their own spoofs and see how the spoof channelized-value distributions compare to the TAI distributions. This would allow them to find possible vulnerabilities in the channelizing matrix that they could explicitly test for when verifying tested items. However, the end result of the channel performance penalty tends to be very noisy channels. These channels would generally have difficulty distinguishing any measured items.

5.1.3.3 A Cheating Monitor

The host's objects could be put at risk due to the monitor's access to some of the channels. The monitor could measure their own objects, perform the channel templates on this data and arrive at channelized values v_l. By comparing the distribution on their measured item with the host's item, they could back out the nature

of the imaged item. Again, though, these channels tend to be noisy and would do a poor job distinguishing other sources.

Hypothetically, the host and monitor could agree on performing the model on a series of nonsensitive objects with known image data. This method, which would give the monitor a large number of channels, drastically reduces the number of degrees of freedom in the problem, though it would still take at least as many nonsensitive test item measurements as in the HO case.

5.1.4 Method to Gauge Storage-Information Tradeoff

Two routines are presented here that could be used to scale down the stored information used in the discrimination task, possibly allowing the host country to create a nonsensitive template. $\mathbf{T}$ is optimized through the same procedure for both methods,

$$\begin{aligned} f_{pen}(\mathbf{T}) =& \eta_1 (SNR^2_{\text{max channel}} - SNR^2_{\text{worst channel}}) \\ &+ \eta_2 \sum_{l=1}^{L} \sum_{l'=l+1}^{L} \left(T_{\text{channel } l} \cdot T_{\text{channel } (l+1)} \right)^2 . \end{aligned} \tag{5.22}$$

The first penalty term creates equally performing channels; the second enforces orthogonality. This penalty function is an effective way to spread information among the different channels. From here, the host could reduce the discriminatory power of the model, with the following method,

- The host could add noise to each bin of the resulting channels by,

$$\mathbf{T}_{newl,m} = \mathbf{T}_{l,m} + \mathcal{N}(0, C^2). \tag{5.23}$$

  As the standard deviation C of the noise increases, $\mathbf{W_v T}$ becomes increasingly noisy and less like $\mathbf{W_H}$. This does not require the optimization technique in (5.22).

- Individual channels in $\mathbf{T}$ could be zeroed out. This causes $\mathbf{W_v T}$ to become increasingly non-optimal and the overall task performance to decline. The initial optimization shown in (5.22) is important for this technique, as when information is spread out, dropping channels should lead to a mostly linear decline in performance.

The downside to this method is that performance in the classification task is not optimal. In that sense, it could be compared to a low-resolution measurement of an

object. Such measurements have been proposed for treaty-verification tasks in the past, but the fact that they are not optimal hurts task performance. An optimal way of penalizing out sensitive information is discussed in Section 5.1.5.

5.1.4.1 Implementation and Storage

Implementation is similar to the standard CHO. In this case, the host country would need to find what level of noise (or the number of channels if the channel reduction approach is chosen) is appropriate to identify the test items without handing the monitor revealing information on their TAIs. This methodology allows the host to share both the channelizing matrix and the weights with the monitor.

This model is one possible example of an acceptable observer model that classifies items based on LM data (see Figure 1.20). However, because it does not specifically penalize out certain sensitive information, it is also far from optimal performance-wise.

5.1.4.2 A Cheating Host

As always, the monitor would be able to access t, $\mathbf{v}$ and $\mathbf{W_v}$. This routine would also give the monitor complete access to the entire channelizing matrix. This presents numerous benefits. The monitor, with access to the model, has greater ability to determine possible successful spoofs and then test for them while performing the verification of the objects. The monitor can also verify that the system is working correctly by measuring a benign inspection object.

5.1.4.3 A Cheating Monitor

As described in Section 5.1.3, the monitor's access to $\mathbf{T}$ presents a vulnerability to the host's items. In addition, the monitor could use noise-reduction techniques to back out possible values of the Hotelling weights. For example, a median filter (Sun and Neuvo, 1994) could be applied or a smoothness constraint enacted. These methods are often imperfect. It would be easier to apply these methods to a $\mathbf{W_H}$ that slowly varies in space (when using an imaging detector) rather than a $\mathbf{W_H}$ that varies over a centimeter scale.

5.1.5 Method to Prevent Discrimination on Sensitive Parameters

The methods discussed in Section 5.1.3 and Section 5.1.4 give the monitor access to either individual channels or the entirety of the channelizing matrix. As discussed,

the monitor could hypothetically use knowledge of the channels to test their own items, attempting to create an object that results in the same channelized value and test-statistic distributions as the measured TAIs. The sharing of this information puts the host's sensitive items at risk. Ideally, a model could be developed that returns the exact same channelized values and test-statistic distribution for objects that differ along certain sensitive parameters. This section presents a method to achieve this goal.

If the host is able to explicitly declare what information is sensitive (such as mass or isotopic composition), the optimization of the channelization matrix can be used to prevent discrimination based on these sensitive parameters. If the host does not want the monitor to know the parameter p, which takes on a value p_0, within a tolerance Δp, the host can create the following objective function,

$$\begin{aligned} f_{obj}(\mathbf{T}) =& SNR^2_{1,2}(\mathbf{T}) \\ & - \eta SNR^2_{(1,p=p_0)-(1,p=p_0+\Delta p)}(\mathbf{T}). \end{aligned} \tag{5.24}$$

This penalty is referred to in this dissertation as the "sensitive information penalty". This objective function finds a $\mathbf{T}$ that maximizes the separation of the distributions on $\mathbf{v}$ between sources 1 and 2 while minimizing the separation between source 1 constructed with $p = p_0$ and source 1 constructed with $p = p_0 + \Delta p$. The sum of the optically weighted channels $\mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{T}$ no longer can distinguish imaged sources that differ along the penalized parameter. If both sources have more than one sensitive parameter, (5.24) can be generalized to,

$$\begin{aligned} f_{obj}(\mathbf{T}) =& SNR^2_{1,2}(\mathbf{T}) \\ & - \sum_{j=1}^{2}\sum_{k=1}^{K} \eta SNR^2_{j,(p_k=p_{k,0})-(j,p_k=p_{k,0}+\Delta p_k)}(\mathbf{T}). \end{aligned} \tag{5.25}$$

This method essentially pushes the differences in the measurements between pairs of objects into $\mathbf{T}$'s null space. Similar to (5.18), the effect of this penalty term is to create test-statistic distributions with overlapping means.

It should be noted that there are a limited number of degrees of freedom in the model. As the number of penalized pairs increases, the number of possible $\mathbf{W_g}$s resulting from the optimization routine decreases. Each penalized pair essentially puts an additional condition on $\mathbf{W_g}$, and too many would result in an optimization routine that can't possibly succeed. In the case of neutron detector used in this project, with 1600 pixels, it is unlikely that a prohibitive number of penalized

pairs would be required. Likewise, if the penalized space encompasses the difference between the optimized objects, the optimization routine will fail.

5.1.5.1 Implementation

Implementation of this model in real life presents a challenge. The host has access to the two items in the discrimination task. However, the sources that differ along the sensitive parameters from the true source would not be readily available and would be very expensive to construct. Ideally, simulation data would be accurate enough that this entire exercise could be done in simulation. Monte Carlo simulation can reliably produce the incident flux on the detector plane if all of the geometries are modeled accurately. A particular problem is the detector response. Detector response is often pixel-dependent and changes with time; hence, it is necessary for a calibration measurement to be taken to find the current detector response before a measurement has taken place. The simulated data could then be adjusted to account for the current detector response. More discussion on this point is presented in chapter 7.

In practice, multiple parameters would likely be deemed sensitive by the host, which may require an additional penalty term with a distinct η in (5.24). The inclusion of nuisance parameters adds another layer of difficulty to this problem, as we would need to know the effect of nuisance parameters on both objects in the penalty term.

Penalization is imperfect due to different statistics in the training and testing data sets. The test-statistic distributions given to the monitor could be found for the minimum acquisition time for maximum model performance for the discriminated pair of sources. Because it is unlikely the test-statistic distributions for the penalized objects completely overlap, the host would need to set some maximum allowable AUC value. There are a limited number of warheads of each type in the stockpile. The standard deviation on the estimate of the mean of the test-statistic value for all of the sources within one type can be denoted by σ_{TAI}. If the difference in mean test-statistic value for the true and penalized objects, $\Delta t_{TAI-penalizedTAI}$ is smaller than σ_{TAI}, the host would have confidence that the monitor could not glean useful information from $\mathbf{T}$. This would be an ideal result for the host, and this method could set the limit on the maximum allowed AUC.

5.1.5.2 Storage

The storage for this model is the ideal scenario, and a practical example of the model shown in Figure 1.20.

5.1.5.3 A Cheating Host

This methodology presents all of the benefits of the model that adds noise to the channels, except this model allows for optimal performance in the discrimination task. As there is more information in the channels than in Section 5.1.4, they should also do a better job discriminating certain spoofs on their own. As discussed in Section 5.1.2, additional terms could be incorporated into the optimization of the channelizing matrix to specifically identify certain spoofs.

5.1.5.4 A Cheating Monitor

The onus falls on the host to verify that their $\mathbf{T}$ is nonsensitive. One significant issue is that any changes from the calibration TAI measurement to the unknown item measurements result in imperfect penalization. The test-statistic distributions for the penalized sources may not overlap, allowing the monitor to attempt to back out the actual geometry. This emphasizes the need for high statistics to calibrate the model and for high-quality transport simulations and for realistic detector response.

Another concern is nuisance parameters. This procedure leads to minimum separation between the two test-statistic distributions for $pr(t|H_{1,p_0})$ and $pr(t|H_{1,p_0+\Delta p})$, but it remains possible that when testing certain individual realizations of the nuisance parameters, the TAI and penalized object will be more easily distinguishable by $\mathbf{T}$. The host may also desire to enforce a condition that the AUC for each realization is sufficiently low.

5.2 Experiments and Results

In the following studies, calibration data sets were simulated for the two sources in the discrimination task. Independent data sets were simulated for these same sources, and the 2AFC test was performed as described in Section 1.3.2. The chosen figure of merit—the AUC—was plotted as a function of acquisition time to gauge task performance. When the CHO was performed, the channelization matrix was always initialized to a random set of numbers before the optimization routine began.

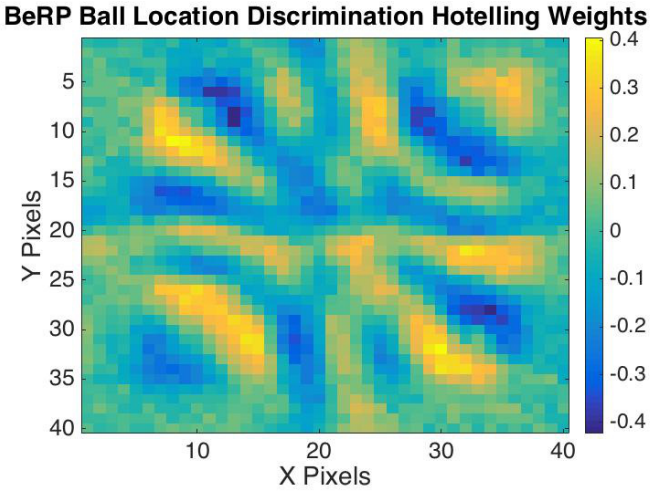


Figure 5.1: The Hotelling weights for the BeRP ball location-discrimination task. As expected, the optimized CHO performs equally well as the HO.

5.2.1 Hotelling Observer

The IO8 and IO9 discrimination task demonstrates how the presence of nuisance parameters affects the Hotelling weights and performance for a spectral-discrimination task. The BeRP ball location and circle/square studies show discrimination ability based on the neutron image shift. The Hotelling weights for each of these tasks will be shown because they represent the optimal linear discriminant which future results will be compared to.

5.2.1.1 BeRP Ball Location-Discrimination Hotelling Weights

In the BeRP ball location study, there were no nuisance parameters present. The Hotelling weights (Figure 5.1) can then be expressed as $\frac{g_2-g_1}{g_2+g_1}$ and correspond to a scaled version of the neutron image shift in Figure 3.12. Here, the BeRP ball at (2 cm, 2 cm) was source 2 in (5.4).

5.2.1.2 BeRP ball Location-Discrimination Inverse Problem

This section demonstrates what the monitor could gain by performing the inverse problem. If the monitor could somehow gain access to $\mathbf{W_H}$, they could perform $g_{rec} = \mathbf{W_H^{-1}}t$ to reconstruct a data set. Obviously, this reconstructed data set will look like the Hotelling weights.

An image of the BeRP ball at (0 cm, 0 cm) was taken with 100,000 detected counts. $\mathbf{W_H^{\dagger}g}$ was performed, yielding a test statistic of value -1,242. $pinv(\mathbf{W_H})t$ was then performed to acquire a reconstructed image. Taking the difference between

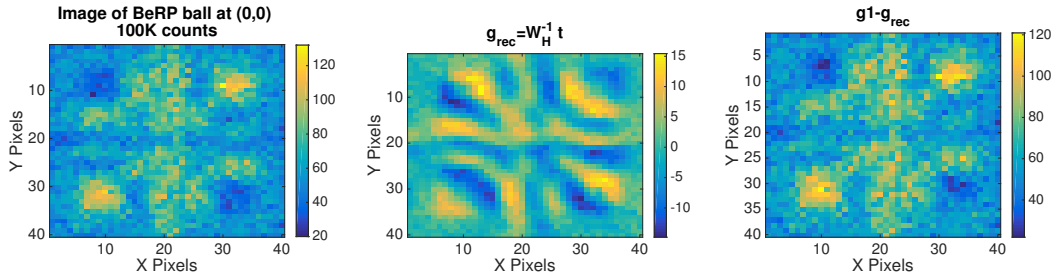


Figure 5.2: The left plot shows an example image taken of the BeRP ball at (0 cm, 0 cm), with 100,000 detected counts. The middle plot shows a reconstruction of that image using the test statistic t and Hotelling weights $\mathbf{W_H}$. The right plot shows the difference between the initial and reconstructed images. This delta image is in the null space of $\mathbf{W_H}$.

the true image and reconstructed image and passing that result through the system yields a test statistic of 4.5e-13 (approximately 0). This is because all of the data for that object other than the differences between $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ is filtered out when applying the Hotelling weights. The difference between the true projection data for that source and the delta image is almost imperceptible. Figure 5.2 contains the various plots for this study.

In a task such as this one, where there are no nuisance parameters present, the host could back out sensitive information on one object if it knows the data set for the second and the Hotelling weights. In this case, the BeRP ball at (0 cm, 0 cm) is treated as the unknown TAI and the BeRP ball at (2 cm, 2 cm) as the object that the monitor has knowledge about. The monitor can do this through the procedure shown in (5.10), resulting in the image shown in Figure 5.3. This is equivalent to the calibration image shown in Figure 3.12.

5.2.1.3 20 cm Ring Source vs. Square Source Hotelling Weights

The HO was tasked with discriminating a source as one of two geometric types—a ring or square. Like the BeRP ball study, they correspond to a scaled version of the image shift presented in Figure 3.13. The Hotelling weights are shown in Figure 5.4.

5.2.1.4 INL Inspection Object Classification

In this study, various orientations of the INL inspection objects were used to train the HO, which was then tested on independent data sampled from one of these orientations. As discussed in the theory section, the Hotelling weights are dependent on the acquisition time when nuisance parameters are present; Figure 5.5 shows the

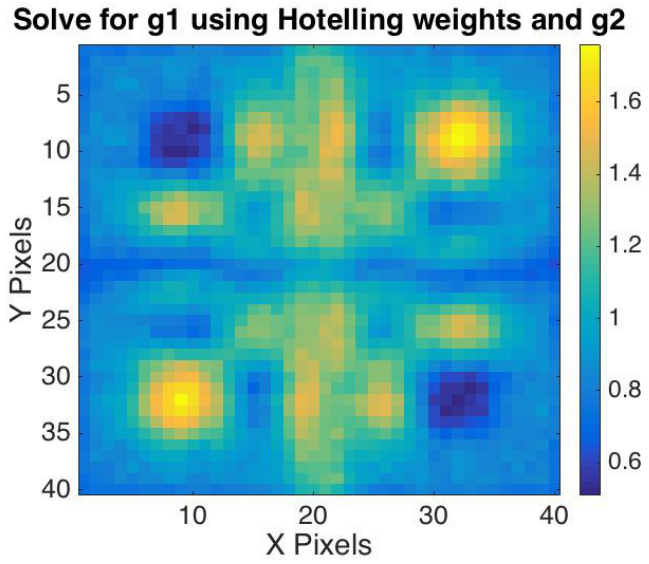


Figure 5.3: Using (5.10), and knowledge of $\boldsymbol{g_2}$ and $\mathbf{W_H}$, the monitor could back out the distribution on $\boldsymbol{g_1}$, as shown in this picture.

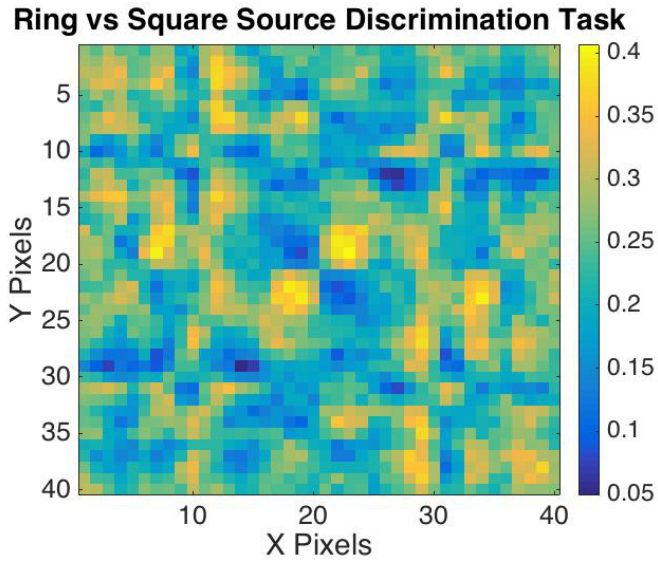


Figure 5.4: The Hotelling weights for the 20 cm ring and square source discrimination task.

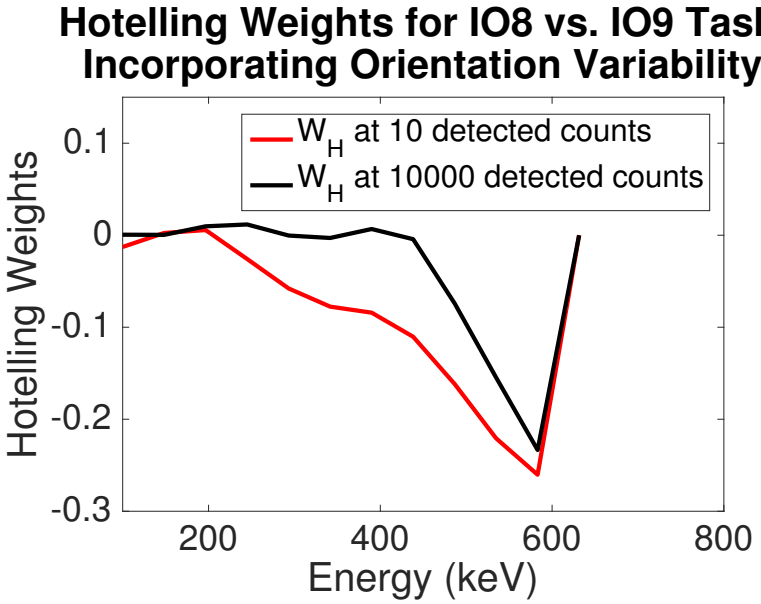


Figure 5.5: The Hotelling weights, $\mathbf{K}_{\mathbf{g}}^{-1}\Delta\overline{\mathbf{g}}$ are dependent on acquisition time due to the fact that the Poisson covariance term varies with N while the nuisance parameter matrix varies with N^2. The weights at two different acquisition times are shown in this figure.

change in Hotelling weights that incorporate orientation variability for an acquisition time corresponding to 10 counts as opposed to 10,000. The short-acquisition-time weights can be explained because IO8's measured spectra is consistently shifted towards higher energies when compared to IO9's. Therefore, a low-energy detection returns a low-positive weight while a high-energy detection returns a large-negative weight.

Figure 5.6 presents the performance of the HO, trained on different sets of data, in classifying data measured from a randomly chosen orientation of IO8 and IO9. This is the HO analog to the ideal observer performance shown in Figure 4.8. In comparing the two plots, the performance of the HO is substantially lower than the ideal observer when the model is trained and tested on data from all of the orientations. This is expected; the ideal observer is the optimal classifier and integrates over the likelihood distribution for each individual orientation, while the HO acts by decorrelating the calibration and testing data. The error bars for these performance curves are due to sampling of the testing data; 1000 total testing data samples were taken for each AUC point on each curve. The error is governed by binomial statistics as discussed in Section 4.2.2.

Finally, the test statistic distribution when the model data matches the testing data is always normal for the HO (shown in Figure 5.7), even when nuisance pa-

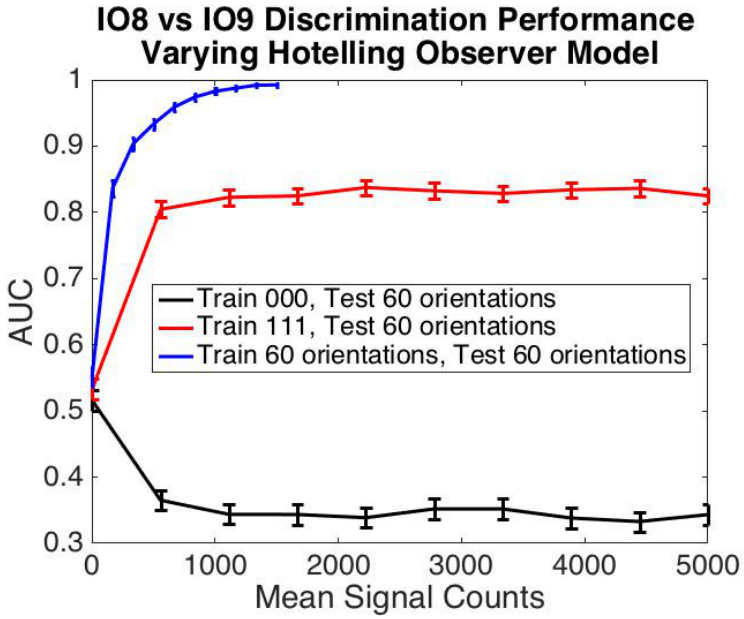


Figure 5.6: In all three studies, the developed model was tested by sampling data from a randomly chosen object's orientation. As in the ideal observer study, the observer trained on the 000 orientation performs poorly, 111 orientation performs better, and model trained on all 60 orientations performs the best.

rameters are present. This is by design due to the decorrelating process involved when applying $\mathbf{W_H}$ to $\mathbf{g}$. This could present an advantage over the ideal observer, which returns a non-normal test-statistic distribution that could reveal information on the object to the monitor. On the other hand, given that the monitor may be told the nuisance parameters, knowledge that the test-statistic distribution should be normal would give the monitor something to strive for if they tried to back out the host TAI's geometry.

5.2.2 Channelized Hotelling Observer

A four channel optimization using the optimization routine in (5.15) was performed for each task.

5.2.2.1 BeRP Ball Location Discrimination

An example channelization for the BeRP ball localization task is shown in Figure 5.8. Each channel clearly contains some component of the Hotelling weights in Figure 5.1, with the fourth channel looking especially similar. In this task, the channels themselves would constitute sensitive information and need to be put behind an IB. This is demonstrated by the performance study in Figure 5.9, showing equal performance between the best performing channel and CHO.

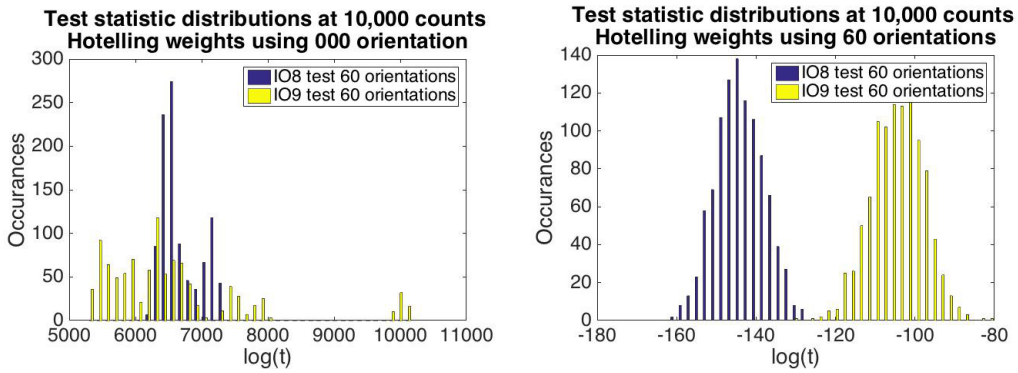


Figure 5.7: The left figure shows the test statistic distribution when a model trained on the 000 orientation of IO8 and IO9 is tasked with classifying data from one of 60 orientations. Notice the highly non-normal data distribution, similar to the result for the ideal observer (Figure 4.9). The right plot shows the test statistic distributions for the HO trained on all 60 orientations. Both resulting distributions are normal.

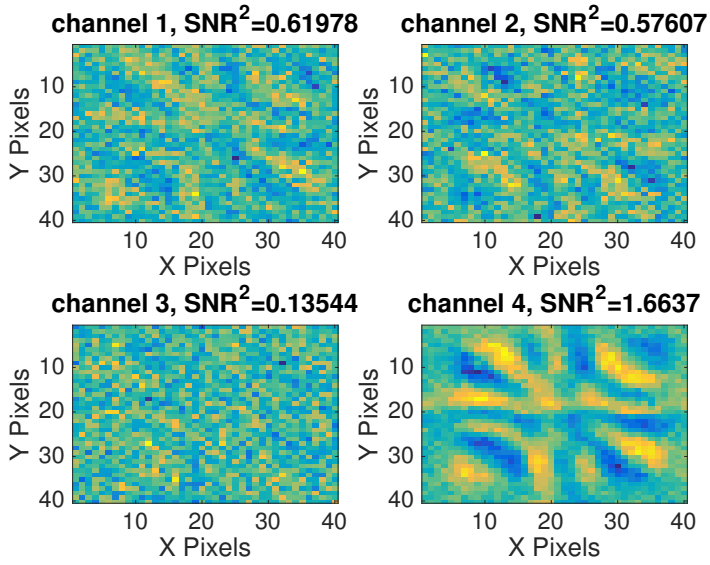


Figure 5.8: An example optimization of the channelizing matrix for the BeRP ball location study. Multiple channels show strong performance. Not shown here is the optimally weighted sum of channels, which results in an image very similar to the Hotelling weights for this task with an SNR^2 of 1.74.

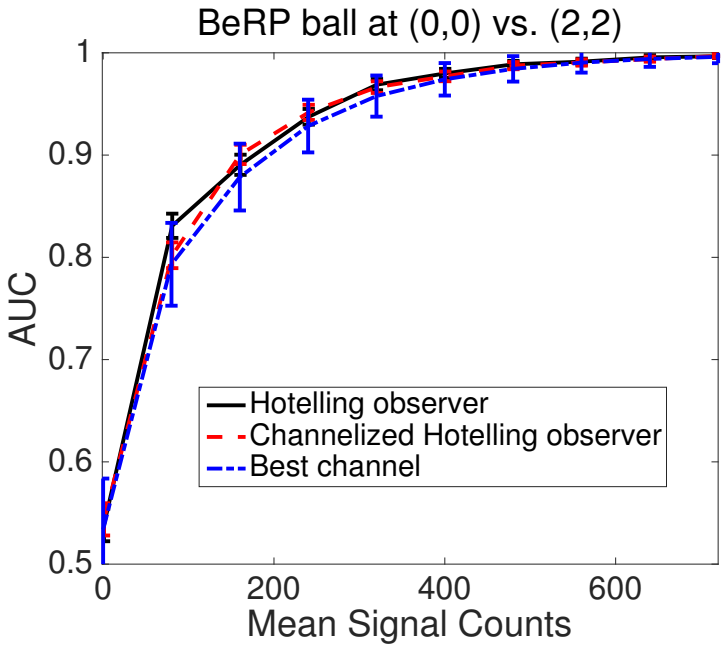


Figure 5.9: This plot shows how the HO, CHO (with an optimal $\mathbf{T}$) and the best performing channel performs at the task. This implies that at least one channel contains similar information to the sensitive Hotelling weights.

An extra comment needs to be made here about the variability in the performance curves using the CHO. A different $\mathbf{T}$ will result for a different $\mathbf{T}_{in}$ that is put into the optimization routine. $\mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{T}$ will be similar, but not equal, to the Hotelling weights. This model is optimal in performing the task on the *calibration* data that the model is trained for. When testing on the calibration data, each optimal $\mathbf{T}$ will perform the same. When testing on alternative data sets, even those independently sampled from the same objects the calibration data was simulated from, performance will vary with each optimization of $\mathbf{T}$. This effect is exaggerated when measuring the individual channel performance as in Figure 5.9; for each optimization of $\mathbf{T}$, vastly different channels will result.

The CHO performance curves shown throughout this chapter do not take this randomness into account—each AUC point corresponds to only one optimization of $\mathbf{T}$ tested on 1,000 samples from the testing data distribution. Therefore, the error bars for these AUC curves significantly underestimate the variation. Instead, the variability is partially accounted for by performing a single optimization for every AUC point.

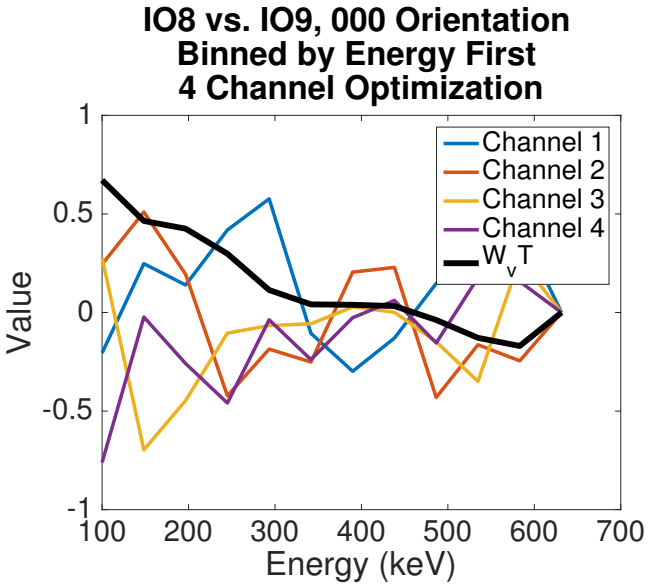


Figure 5.10: To generate this plot, the calibration data was binned by energy prior to channelization. None of the channels share much in common with the Hotelling weights, $\mathbf{W_v T}$.

5.2.2.2 IO8 vs. IO9 Discrimination

In the first study, data was simulated for IO8 and IO9 and the resulting data sets were binned by energy. The channelization matrix was found from this calibration data. It was then used to classify independently simulated IO8 and IO9 energy data. The channels resulting from the optimization (Figure 5.10) are quite different from the Hotelling weights for this task. The noisy channels seem to result from optimizing $\mathbf{T}$ for a low number of detector bins. Because these channels are noisy and do not necessarily have much in common with the Hotelling weights (as a non-object measure, they do not look similar visually), it is possible for the monitor to back out more information on the image data by performing $\mathbf{g}_{rec} = pinv(\mathbf{T})\mathbf{v}$ than in other examples in this section, where the inverse problem will result in the Hotelling weights. The result of the inverse problem is not shown here because the results are highly dependent on the particular channelization matrix optimization.

In the second study, the data was binned into spatio-spectral bins and the channelizing matrix was trained on this data (see Figure 5.11). When asked to classify independent data, the channels resulting from this procedure show equal performance to the HO and 4 channel CHO for the discrimination task. The resulting channels were also summed over pixel ID, and the left plot of Figure 5.11 shows that these rebinned channels are equivalent to the Hotelling weights for this task.

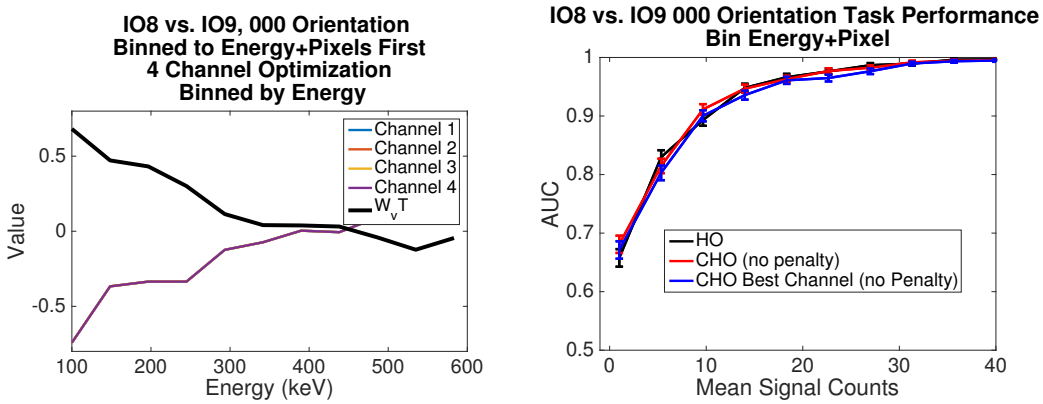


Figure 5.11: The channelization matrix was trained on full spatio-spectral binned data sets for IO8 and IO9. The left plot shows the result when the channels are binned by energy—the 4 channels are equivalent to (or the negative of) the Hotelling weights. The right plot shows that the best performing channel, and in fact all of the channels, have near equal performance to both the HO and CHO.

These first two studies show that while the channelization procedure creates noisy channels, recombining the binned data can yield sensitive data.

5.2.2.3 Spoofs

This study demonstrates how the individual channels can be used to discriminate spoofs. The model was trained to differentiate the 16 cm ring source from the 20 cm ring source, as in Section 4.2.2. A single four channel optimization of $\mathbf{T}$ was done, optimizing the separation between the training distributions on the channelized values. The CHO using this $\mathbf{T}$ was applied to independent data sets for the 16 cm ring and 20 cm ring sources, resulting in distributions on t, v_1, v_2, v_3, and v_4. The model was then tested against different data sets resulting from measurements of the other simulated neutron sources. Values that fell outside the 2.5 and 97.5 percentile of the distributions were rejected. Table 5.1 shows the percent rejected for each source using each value. It is rare that the channelized value does a better job rejecting the tested source than the whole test-statistic. This is because the channels that result from the optimization of $\mathbf{T}$ tend to be noisy versions of the Hotelling weights themselves. However, channel 2 for the 24 cm ring source is a counterexample, as it drastically outperforms the test-statistic in rejecting the object.

The monitor could also define specific spoofs to penalize against by adding additional terms to the optimization routine and optimizing the TAIs against these spoofs. However, as stated before, the difficult aspect in discriminating spoofs is

Spoof	$t\%$	$v_1\%$	$v_2\%$	$v_3\%$	$v_4\%$
16 cm Sq	100	100	94.2	100	100
20 cm Sq	6.1	6.1	7.6	5.6	6.0
24 cm Sq	73.5	72.4	69.0	74.7	76.4
24 cm R	20.4	19.8	28.6	21.3	23.0
(0,0) BeRP	96.7	96.3	90.6	96.6	96.3
(2,2) BeRP	94.8	94.9	81.8	94.1	94.5

Table 5.1: A single optimization of $\mathbf{T}$ was done, optimizing the distance between the test-statistic distribution for the 16 cm ring and 20 cm ring sources. Thresholds at the 2.5 and 97.5 percentiles were found for the total test statistic as well as the channelized values. The sources in the left column were measured 1,000 times in simulation, then the CHO was performed on them, and their channelized values and test statistic were compared to the corresponding thresholds.

that the monitor does not know in advance what object the host uses to spoof the TAI.

5.2.3 Method to Generate Nonsensitive Channels

This section presents multiple different implementations of the model that penalizes individual channel performance for a given task.

5.2.3.1 BeRP Ball Location Discrimination

The incorporation of the penalty term in (5.21) degrades the ability of each individual channel to discriminate the two sources. For this study, an acquisition time was set so that an average of 400 counts were read in from IO8. As Table 5.2 shows, when η is increased, the maximum channel performance decreases without reducing overall CHO performance. This behavior persists up to $\eta = 1$, after which overall performance drops. An example output of the channel optimization when η has a value of 1 is shown in Figure 5.12. The individual channels are nonsensitive—each performs very poorly in discriminating the two BeRP ball locations.

Though this routine can effectively generate nonsensitive channels, a singular value decomposition (Golub and Reinsch, 1970) of $\mathbf{T}$, as shown in Figure 5.13, reveals that the singular vector with the lowest singular value looks like the sensitive $\mathbf{W_H}$. This result emphasizes the fact that a $\mathbf{T}$ that optimizes the SNR^2 necessarily contains sensitive information on the objects. This implies that $\mathbf{T}$ is still sensitive and the host would not be able to share it with the monitor.

However, due to the importance this optimization routine places on the rela-

η	Mean SNR^2 for max channel	SNR^2 for all channels	Percent failed ($SNR^2 < 0.1$)
1e-4	8.13	8.63	0
1e-3	6.66	8.63	0
1e-2	1.17	8.63	0
1e-1	0.106	8.63	0
1	0.0042	8.63	0
1.1	0.0025	3.88	55

Table 5.2: 100 optimizations of channelizing matrix were performed for each row in this table. The acquisition time for this study was set so that an average of 400 mean signal counts were detected. This corresponds to an optimal SNR^2 of 8.63. As η increases, the best channel performance drops. When η rises above 1, the optimizations fail increasingly often.

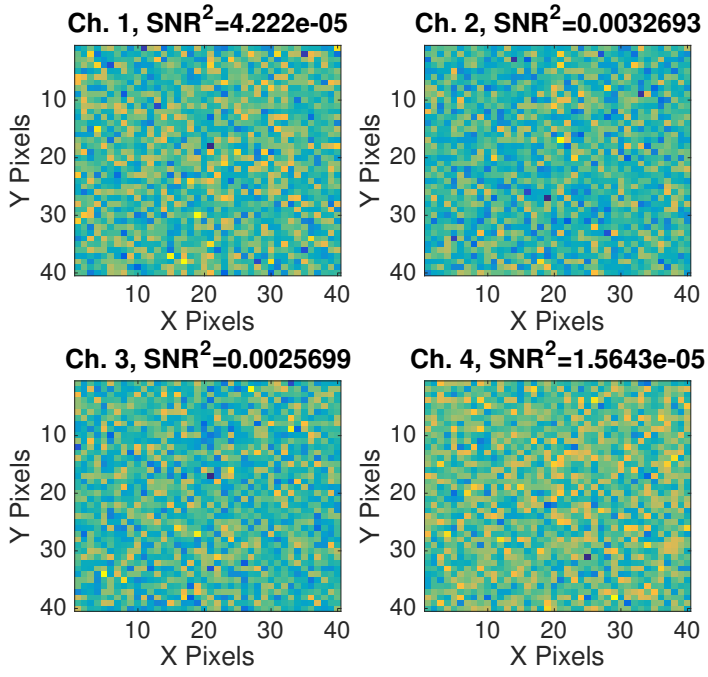


Figure 5.12: An example optimization of the channelizing matrix for the BeRP ball location study when the channel performance penalty $\eta = 1$. Each channel appears to be random, but when properly weighted, maximum performance is obtained and $\mathbf{W_v T}$ is equivalent to the Hotelling weights.

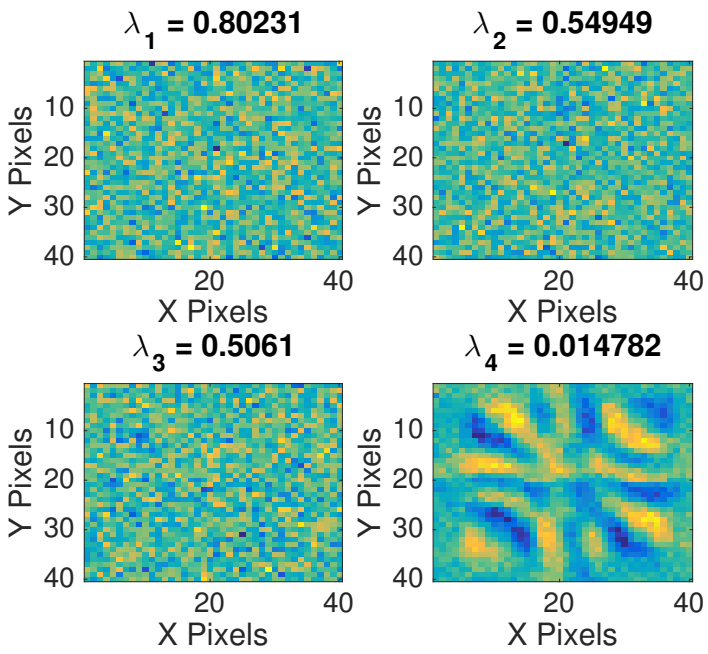


Figure 5.13: Singular value decomposition of the channelizing matrix shown in Figure 5.12. The singular vector with the lowest singular value contains the Hotelling weights.

tionship between channels, removing a single channel or small number of channels from the resulting $\mathbf{T}$ causes a debilitating effect on performance. This is evidenced in Table 5.3—regardless of how many channels there are, the performance of a large number $L_{mon} < L$ of the channels is very poor. When 90% of channels are known, performance is poor, but there are still occasional optimizations where the combined channels can effectively discriminate the sources. When only 75% of channels are given to the monitor, the resulting performance is very poor, and the $\mathbf{W_g}$ can no longer discriminate between the two sources. This creates an interesting application for treaty verification. The host can give the monitor L_{mon} channels of $\mathbf{T}$, the channel weights $\mathbf{W_v}$ and in testing the host can access all of the channelized data $\mathbf{v}$.

Overall, this method presents an advantage over the standard implementation of the CHO. Because the host can share L_{mon} channels, the monitor could hypothetically image a known nonsensitive test object in front of the detector, see the nonsensitive $\mathbf{g}$ and verify that the algorithm is working properly on the shared channels. This additional information at the monitor's disposal can also help the monitor to identify spoofs that could fool the known channels.

L	L_{mon}	SNR^2 for L_{mon} channels	% Performing Poorly (total $SNR^2 < 0.1$)
4	3	0.269	88
10	9	0.7105	50
10	7	0.0195	98
25	24	0.975	18
25	22	0.291	46
25	18	0.0314	98

Table 5.3: The acquisition time for this study was set so that an average of 400 mean signal counts were read in. This corresponds to an optimal SNR^2 of 8.63. 50 optimizations of $\mathbf{T}$ with L channels were performed for each row in this table. The second column L_{mon} is the number of channels shared with the monitor, chosen at random. The right two columns shows how the L_{mon} channels would perform at the task. Ideally, the L_{mon} channels would have no discriminatory ability, defined here as resulting in an SNR^2 under 0.1. The right column shows the percentage of the time that the L_{mon} channels led to an SNR^2 of 0,1 or under for this task.

5.2.3.2 Inspection Object Discrimination Task

The IO8 and IO9 data was binned into spatio-spectral bins and $\mathbf{T}$ was found. For an unpenalized optimization of $\mathbf{T}$, performance of the individual channels was very strong as shown in Figure 5.11. The channel-performance penalty was then included in the channelization-matrix optimization. η was chosen to be 0.2 to properly penalize the channels while retaining optimal performance. Figure 5.14 presents an example series of channels (summed over pixel ID) and a performance plot. This optimization results in 4 channels that look very similar, but none perform well.

5.2.4 Method to Gauge Storage-Information Tradeoff

This section presents multiple studies that explore the methods developed to reduce the discriminatory ability of the models.

5.2.4.1 BeRP Ball Location Discrimination

Here, an implementation of the optimization of $\mathbf{T}$ using the penalty terms in (5.22) is presented. The first penalty function effectively, though imperfectly, spreads information among the channels (see Table 5.4). Taking $\eta_1 = 0.2$, η_2 was optimized to minimize the sum over the inner products of the different channels (see Table 5.5). This method creates orthogonal channels that share information equally. After this optimization of $\mathbf{T}$, two methods of reducing the discriminatory ability of the CHO were explored, consistent with the theory section. The following methods don't make

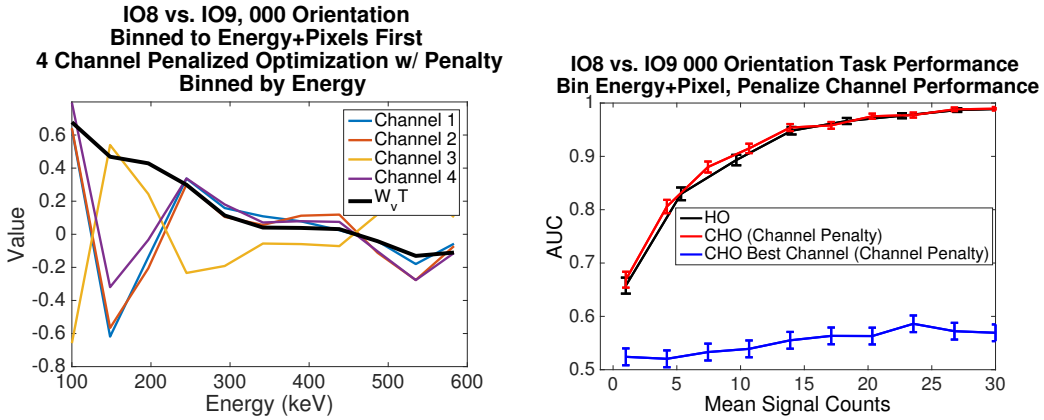


Figure 5.14: The left plot shows an example channelizing matrix, optimized for spatio-spectral data and then summed over pixel ID. The four resulting channels are all noticeably different from the Hotelling weights but are similar to each other. Each performs poorly, as shown in the right plot.

η	ΔSNR^2 for max vs. min channel	SNR^2_{tot}
1.00E-03	1.36	1.73
1.00E-02	1.12	1.73
1.00E-01	0.89	1.73
3.00E-01	0.75	1.70
5.00E-01	0.7	1.64
8.00E-01	0.72	1.56
1.00E+00	0.67	1.46
5.00E+00	0.0025	0.01

Table 5.4: 50 optimizations of the channelizing matrix with 10 channels were performed for each row in this table. The acquisition time for this study was set so an average of 80 mean signal counts were read in. This corresponds to an optimal SNR^2 of 1.73.

any distinction between what information in $\mathbf{W}_\mathbf{g}^\dagger = \mathbf{W}_\mathbf{v}^\dagger \mathbf{T}$ is sensitive or nonsensitive.

In the first method, individual channels are zeroed out in the channelizing matrix. For each resulting $\mathbf{T}$, the optimal weights $\mathbf{W_v}$ are found. As more channels are removed (see Figure 5.15), $\mathbf{W_v T}$ looks less like $\mathbf{W_H}$. $\mathbf{T}$ is therefore non-optimal. In the second method, zero mean Gaussian noise with standard deviation C is added to each bin of the channelizing matrix. The procedure follows (5.23), and the effect is shown in Figure 5.16. As C is ramped up, increasingly noisy weights $\mathbf{W_v T}$ result.

Hypothetically, one of these methods could offer the host a way of generating a nonsensitive channelizing matrix that could still differentiate the two objects. The second method has the advantage that the monitor keeps access to the different channels, which again is useful in distinguishing other objects from the two the

η_1	η_2	ΔSNR^2 for max vs. min channel	Sum of Inner Product Squared
0.2	1.00E-01	0.73	4.25
0.2	1.00E+00	0.48	2.15
0.2	1.00E+01	0.34	0.085
0.2	1.00E+02	0.33	0.0093
0.2	5.00E+02	0.32	0.0011

Table 5.5: 50 optimizations of the channelizing matrix with 10 channels were performed for each row in this table. The acquisition time for this study was set so an average of 80 mean signal counts were read in. This corresponds to an optimal SNR^2 of 1.73.

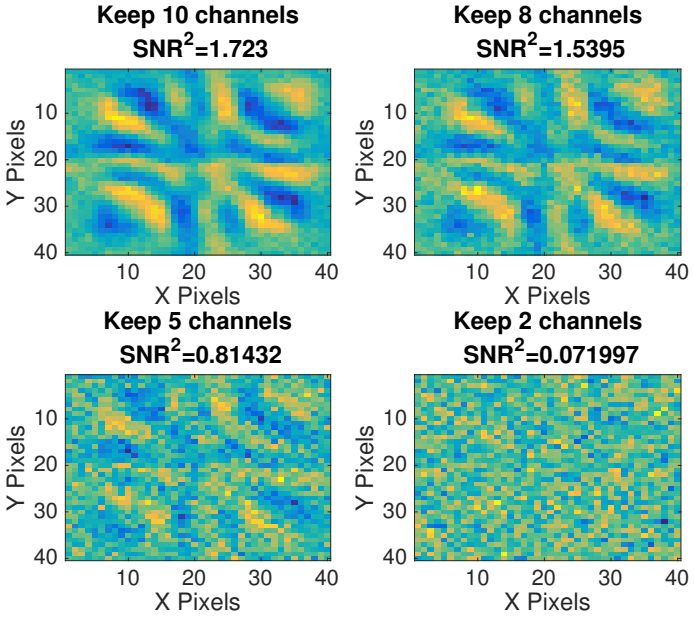


Figure 5.15: These plots show $\mathbf{W_v T}$ as channels are zeroed out in the channelizing matrix. The optimization was done for a mean of 80 signal counts. The SNR^2 totals correspond to AUCs of 0.810, 0.795, 0.724 and 0.562 respectively.

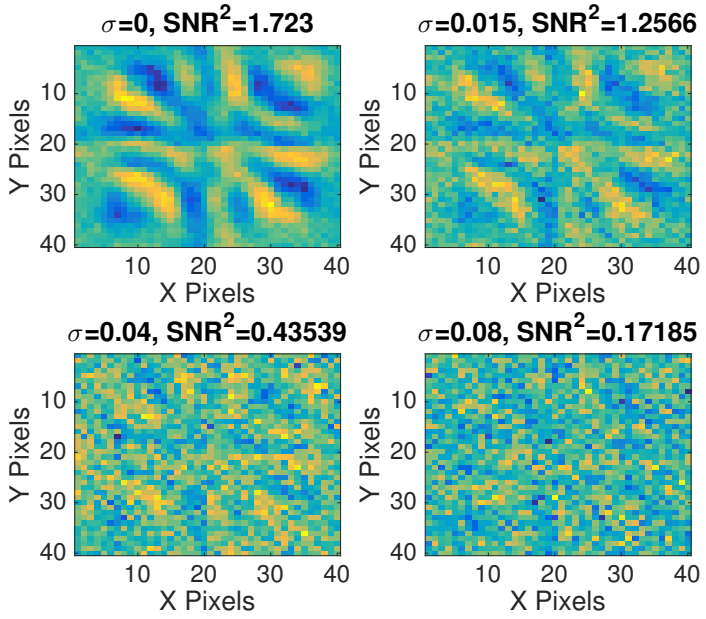


Figure 5.16: These plots show $\mathbf{W_v T}$ as zero mean Gaussian noise with various standard deviations is added to the channelizing matrix. The optimization was done for a mean of 80 signal counts. The SNR^2 totals correspond to AUCs of 0.828, 0.764, 0.665 and 0.625 respectively.

binary-classification task is designed for.

5.2.5 Method to Prevent Discrimination on Sensitive Parameters

This subsection discusses two implementations of the model that prevents discrimination along predefined sensitive parameters. In each section, a certain aspect of the object is explicitly defined as sensitive. The goal for each of these experiments is to create a $\mathbf{T}$ that returns the same test-statistic value for both the normal and penalized object.

5.2.5.1 BeRP Ball Location Discrimination Penalizing X Location

To carry out this task, a toy problem was created (see Figure 5.17) based on the BeRP ball location-discrimination task. The $\hat{x}$ parameter of the BeRP ball located at (0 cm, 0 cm) was treated as the sensitive information. The $\hat{x}$ location was declared sensitive from a value of 0 mm to 20 mm and differences in data due to the $\hat{y}$ location were treated as nonsensitive. The goal of this study was to create a model that is incapable of distinguishing a BeRP ball located at (0 mm, 0 mm) from (20 mm, 0 mm) yet still effectively classifies a BeRP ball at (0 mm, 0 mm) and (0 mm, 20 mm). The BeRP ball was simulated at (20 mm, 0 mm) to penalize $\hat{x}$ information

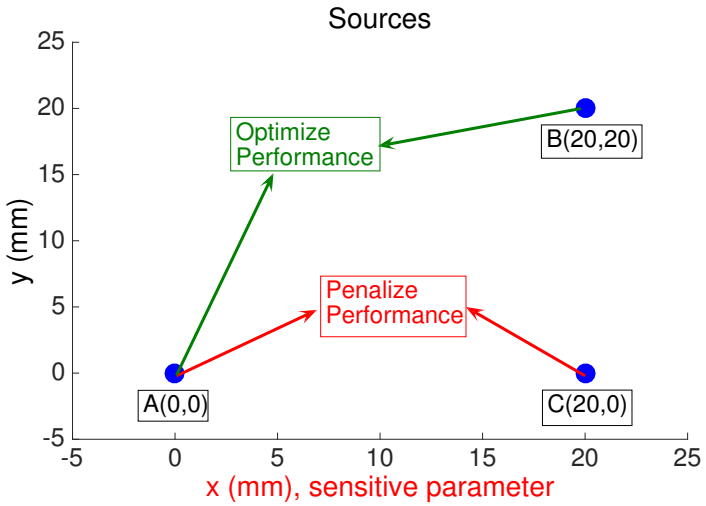


Figure 5.17: Diagram of BeRP ball locations (note that the BeRP ball size is larger than the difference between locations shown here). Performance in discriminating between BeRP balls located at (0 mm, 0 mm) and (20 mm, 20 mm) is optimized while (0 mm, 0 mm) and (20 mm, 0 mm) is penalized.

within this tolerance. The count rates for the sources were set equal to emphasize the role imaging plays in the model.

The necessity of incorporating a penalty term into the optimization of $\mathbf{T}$, as in (5.24), is demonstrated by Figure 5.18. This figure shows a performance study using the standard optimization of $\mathbf{T}$, optimizing the separation of the test-statistic distributions resulting from imaging the BeRP ball at the (0 mm, 0 mm) and (20 mm, 20 mm) locations. The performance curves show that with a high enough acquisition time, it is possible to differentiate a BeRP ball at (0 mm, 0 mm) and (20 mm, 0 mm) by their resulting test-statistic value. This is not ideal. In a real life study, the monitor could take home the test-statistic distribution from the testing site. In this case, the monitor could image a range of BeRP balls at various locations and find the resulting test-statistic distribution that corresponds to the BeRP ball at (0mm,0mm).

Penalization of the $\hat{x}$ location was accomplished through the following objective function,

$$\begin{aligned} f_{obj}(\mathbf{T}) =& SNR^2_{(0,0),(20,20)}(\mathbf{T}) \\ & - \eta SNR^2_{(0,0),(20,0)}(\mathbf{T}). \end{aligned} \tag{5.26}$$

Figure 5.19 shows the effect of the penalty coefficient η on the resulting channelizing matrix. Choosing a high value for the penalty coefficient, $\eta = 50$, the CHO is no longer able to distinguish the pair of sources that only differ in their $\hat{x}$ coordinate.

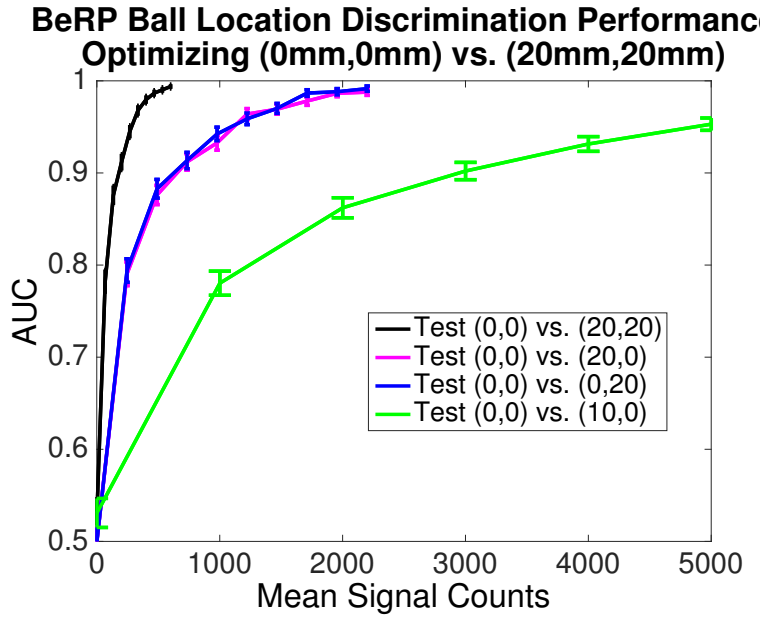


Figure 5.18: **T** was trained to optimize the separation between the test statistic distributions when performing the CHO on the data from the (0 mm, 0 mm) and (20 mm, 20 mm) BeRP ball images. This **T** was then used to discriminate different pairs of data sets.

A plot of $\mathbf{W_v T}$ is shown in Figure 5.20. The result corresponds to a simple $\hat{y}$ shift in the count maps.

The effectiveness of this method is demonstrated further in a performance study (see Figure 5.21). The **T** resulting from the optimization routine was tested on independently simulated data sets. When $\eta = 0$, the optimization routine only maximizes the separation of the two test-statistic distributions for the objects in the classification task, and the performance matches Figure 5.18. When $\eta = 50$, performance in discriminating the objects whose task performance was optimized is still very good while the performance between the (0mm,0mm) and (20mm,0mm) images is near the guessing observer. In comparison to the $\eta = 0$ performance study, it is easier to distinguish the BeRP balls at (0mm,20mm) from (0mm,0mm). This is because penalizing out the $\hat{x}$ differences results in a **T** that can only differentiate sources based on $\hat{y}$ information. Finally, this particular study carries the added benefit that a tested source inside the tolerance at (10mm,0mm) cannot be distinguished from (0mm,0mm). This is not always true, and the following subsubsection presents an example where multiple penalty terms would be required.

It is also important to note that the penalization of the test-statistic distributions was not perfect. While the SNR^2 shown in Figure 5.19 is essentially zero at $\eta = 50$, the AUC shown for the (0 mm, 0 mm) vs. (20 mm, 20 mm) task in Figure 5.21 is not

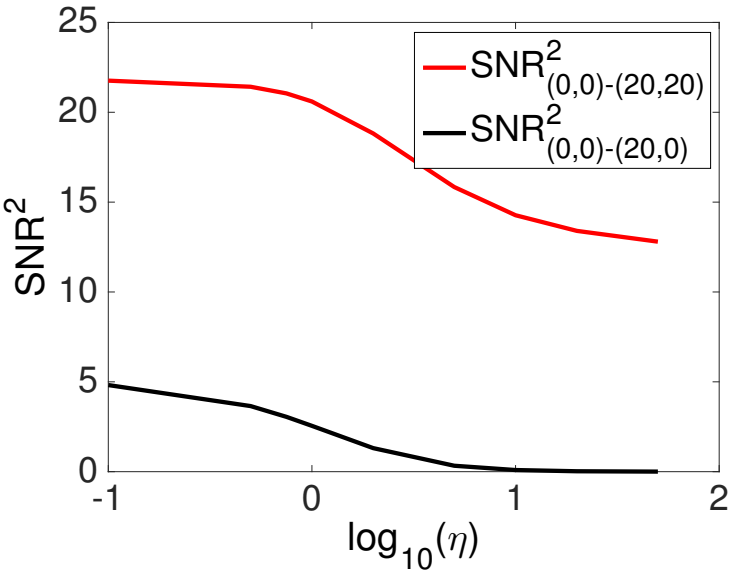


Figure 5.19: This study was done for an acquisition time corresponding to 1,000 detected signal counts. For each η, 20 optimizations of $\mathbf{T}$ were performed. The SNR^2 was plotted when performing this $\mathbf{T}$ on different pairs of calibration data sets. As η increases, the SNR^2 for the penalized pair of sources drops to zero while the SNR^2 for the optimized sources drops by a factor of about 1/3.

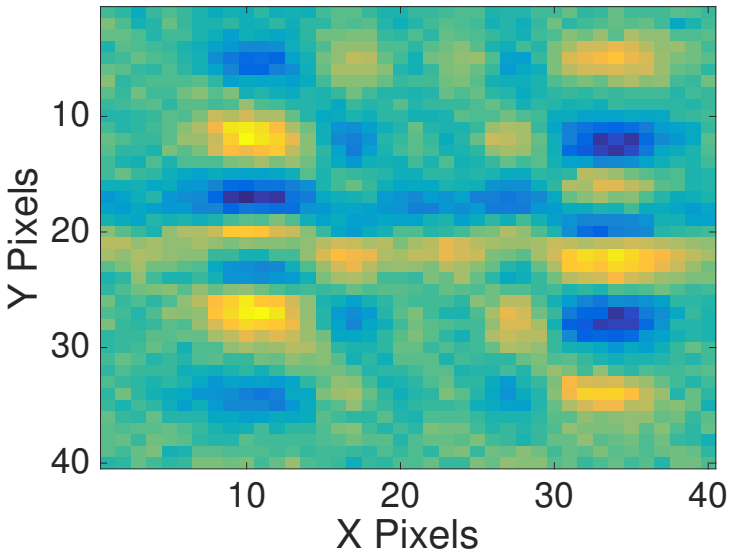


Figure 5.20: A plot of $\mathbf{W_v T}$ when the $\hat{X}$ information has been penalized with $\eta = 50$.

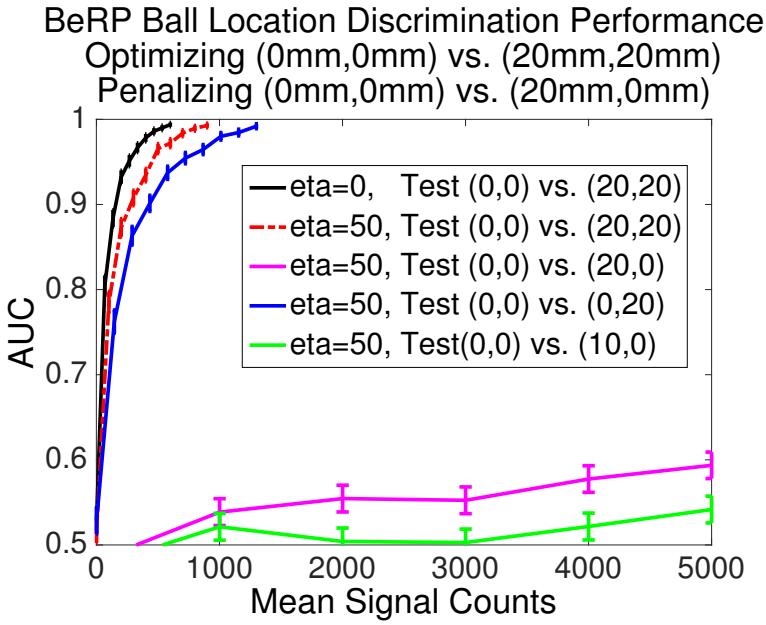


Figure 5.21: Performance of the CHO with channelization matrix optimized by (5.26). The black line shows the performance of a standard optimization without a penalty term. Including the penalty in the optimization of $\mathbf{T}$, good performance is maintained when classifying sources that different in their $\hat{y}$ component, while near guessing observer performance is seen when classifying sources that differ in their $\hat{x}$ component.

0.5 at 1,000 signal counts as expected. This brings up an important point. Training of $\mathbf{T}$ and optimization of η occurs using the same calibration data sets. This results in imperfect penalization when testing independent data sets, especially when the calibration data sets have limited statistics.

Finally, it is interesting to see what the monitor could back out of this procedure in terms of the image data. If the monitor were to gain access to the BeRP ball image at (20 mm, 20 mm), they could follow (5.10) to find the data on the BeRP ball at (0 mm, 0 mm). This resulting image (see Figure 5.22) looks like the BeRP ball imaged at (0mm, 0mm) smeared-out over x (see Figure 3.12 for a comparison).

The theory behind this work and the BeRP ball $\hat{x}$ location results will be presented at the 2016 Symposium for Radiation Measurements and Applications Conference MacGahan et al. (2016a).

5.2.5.2 Ring vs. Square Discrimination Penalizing Size

For this task, the lengths of the ring and square sources were treated as the sensitive parameter. Specifically, the problem was set up so the host wanted to prevent the monitor from gaining knowledge on the size of their sources up to a tolerance of $\pm$ 4

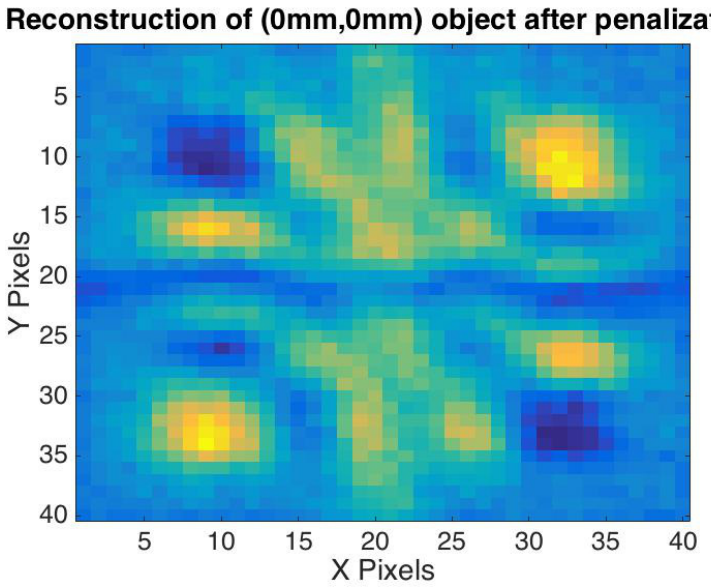


Figure 5.22: If the monitor were to gain access to $\boldsymbol{g_2}$ (the BeRP ball at (20 mm, 20 mm)), they could use (5.10) to back out the data on $\boldsymbol{g_1}$ (the BeRP ball at (0 mm, 0 mm)). This image is the result.

cm. A standard optimization of the channelizing matrix is able to distinguish ring and square sources of different sizes—this is emphasized in Figure 5.23. However, it is also able to easily distinguish sources of the same geometry with different sizes, implying their test statistic distributions are separable with a long enough acquisition time. To prevent $\mathbf{T}$ from being able to distinguish these sizes, the optimization routine in (5.27) is used.

$$\begin{aligned} f_{obj}(\mathbf{T}) =& SNR^2_{R_{20},S_{20}}(\mathbf{T}) - \eta\times \\ & (SNR^2_{R_{20},R_{16}}(\mathbf{T}) + SNR^2_{R_{20},R_{24}}(\mathbf{T}) \\ & + SNR^2_{S_{20},S_{16}}(\mathbf{T}) + SNR^2_{S_{20},S_{24}}(\mathbf{T})) \end{aligned} \tag{5.27}$$

This procedure penalizes the ability of the channelization matrix to differentiate objects of the same geometry but different sizes (results are in Figure 5.24). To best penalize the various pairs of objects, η was chosen to be 20. Again, note that the calibration data sets are penalized and not independently simulated sets. To explore the results of the penalization routine, two sets of plots are shown. At first, the model was trained with calibration data sets that included 5e8 detected counts. This proves to be insufficient—the model is essentially being trained on data that is too noisy for this particular task, leading to imperfect penalization when performing the model on independently simulated data. A second set of simulations was executed to increase the data by a factor of 5 to 25e8 detected counts. Figure 5.25 shows that

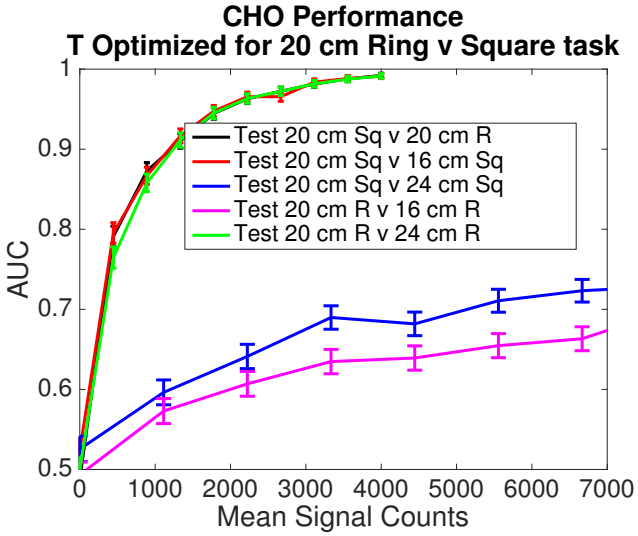


Figure 5.23: **T** was found using the unpenalized optimization routine for these performance curves. The CHO is able to distinguish the 20 cm ring vs. 20 cm square source in addition to pairs of objects of the same geometry but different size, particularly the 20 cm square vs. 16 cm square and 20 cm ring vs. 24 cm ring.

when the calibration data statistics are higher, the models more effectively penalized out the desired pairs of objects. Examining the right plots more carefully, at 5e8 detected counts, the model is still able to distinguish a 20 cm square source from the 24 cm square source. With higher statistics, performance in this pair decreases substantially, but the model can distinguish the 20 cm square source from an 18 and a 22 cm square source. If the monitor desired to penalize out this entire range, they would need to add more penalty terms to the optimization routine.

Alternatively, we can look at the model's ability to distinguish square and ring sources of varying sizes. The penalized model was used to classify different pairs

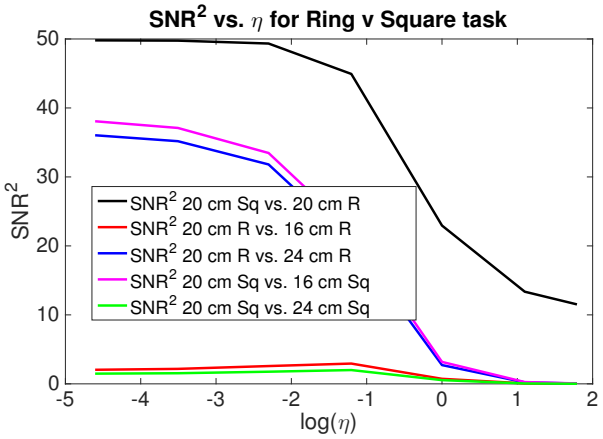


Figure 5.24: As η is increased, the SNR^2 for all penalized tasks decreases to zero. There is still signal left for the desired discrimination task.

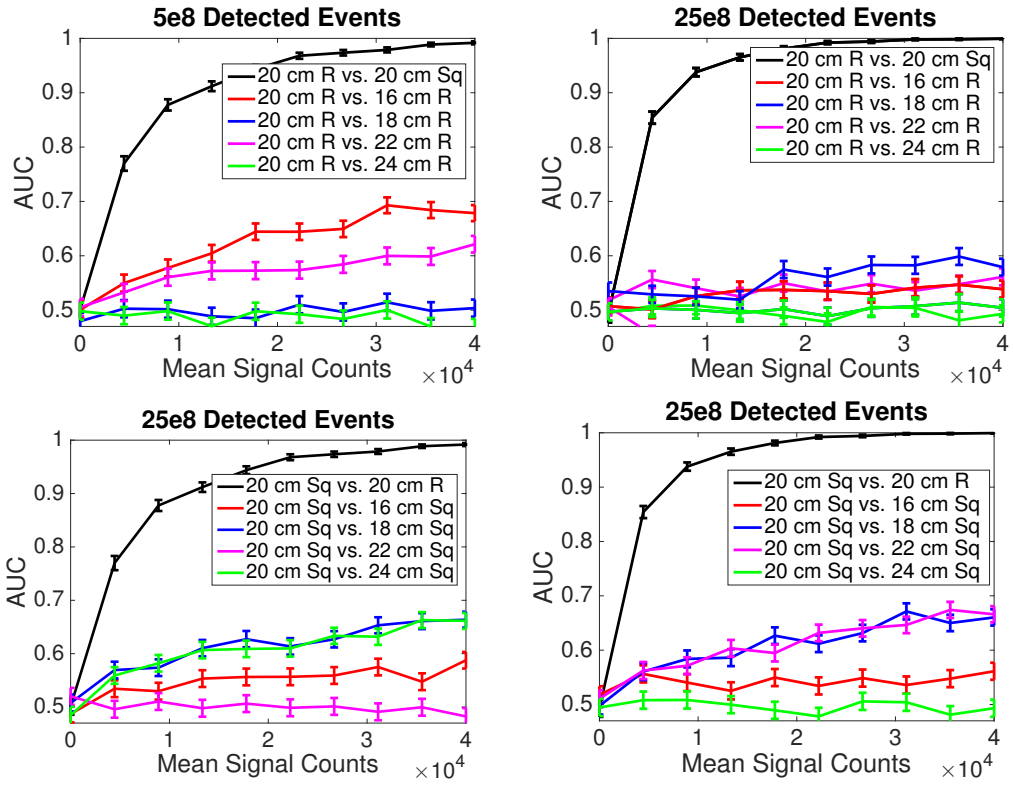


Figure 5.25: The plots show the penalized CHO model's (5.27) ability to distinguish ring sources (top) and square sources (bottom) of different sizes using independent data from the calibration data. When the amount of calibration data used to train the models is increased, the CHO's performance in penalizing the desired pairs decreases significantly.

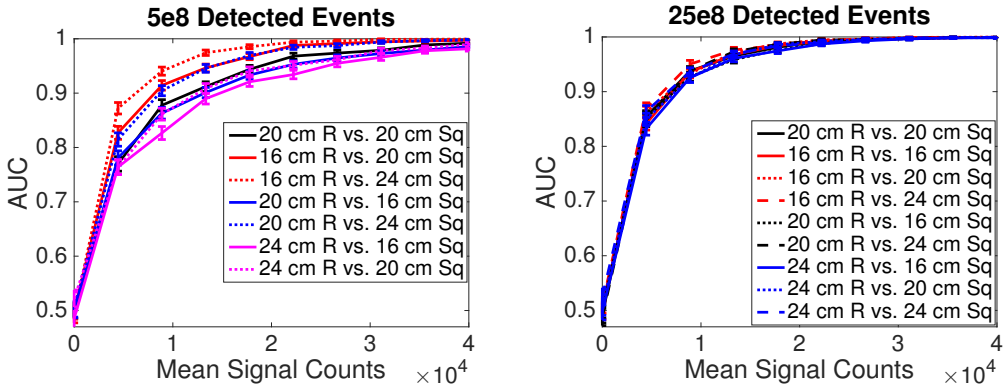


Figure 5.26: A plot of the performance of the penalized CHO model's ability to discriminate different geometries of sizes other than the 20 cm lengths it was optimized for. When the calibration data statistics are higher, penalization is improved and all tasks show closer to equal performance as a function of acquisition time.

with the results shown in Figure 5.26. For the penalized pairs to have an SNR^2 of zero, their test statistic distribution means need to overlap. The ideal result would put all of the means for the circle sources within a penalized size range at the same mean, and all of the square sources at a different mean. This should lead to near equal performance when classifying different geometries of non-equal sizes, and that behavior is demonstrated in Figure 5.26. The Hotelling weights are shown in Figure 5.5. The weights on $\mathbf{g}$ after penalization are shown in Figure 5.27.

The theory behind this work and the ring vs. square task with size penalization results will be presented at the 2016 INMM conference MacGahan et al. (2016c).

5.2.5.3 When Penalization Fails

The host would ideally be able to give $\mathbf{T}$, $\mathbf{W_v}$, the channelized value distributions for the two TAIs, and test-statistic distributions for the two TAIs to the monitor without the monitor being able to back out the sensitive data on $\mathbf{g}$. As shown in (5.18), the result of this optimization is test-statistic distributions with overlapping means. However, the variances of the distributions are not necessarily equivalent. This is important—the monitor could attempt to recreate the objects, trying to match the mean and variance of the test-statistic distributions to the true mean and variance. As (5.19) shows, a $\mathbf{W_g}$ yielding equal means does not necessarily imply equal variances.

A study was done where the CHO maximized the distance between the 20 cm ring and 20 cm square sources test-statistic distributions while penalizing the separation

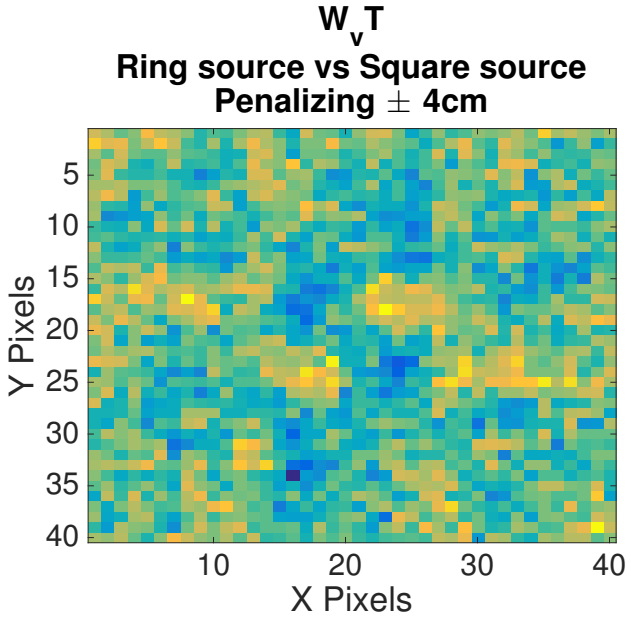


Figure 5.27: An example plot of $\mathbf{W}_{\mathbf{g}}^{\dagger} = \mathbf{W}_{\mathbf{v}}^{\dagger}\mathbf{T}$ for the geometric source discrimination task.

between the same geometries of different sizes, as in the prior subsubsection. The resulting $\mathbf{T}$ was tested on the calibration data to avoid extra uncertainty being introduced with different testing data. At first, the count rates were set equal for all of the sources. An acquisition time corresponding to 40,000 detections was chosen. The resulting test-statistic distribution for the 20 cm ring source was $\mathcal{N}(16.68, 27.32)$ and the distribution for the 16 cm ring source was $\mathcal{N}(16.39, 27.54)$. These two distributions essentially overlap as desired. However, when the count rates were *not* set equal prior to the optimization of $\mathbf{T}$, the distribution for the 20 cm ring source was $\mathcal{N}(314.7, 74.3)$ and the 16 cm ring source was $\mathcal{N}(314.76, 61.94)$. This is because when the count rates are different, the variation in magnitudes for each W_m tends to be larger, and the variance scales with W_m^2. This is a significant difference in variance (though the AUC in this case would still be very poor) that the monitor could use to back out information on the host's TAI. Methods to address this issue are presented in chapter 7.

5.3 Conclusion and General Comments

This chapter presents multiple linear observers that can be used to perform binary-discrimination tasks. Both the HO and CHO drastically reduce the storage compared to the ideal observer, a net benefit for treaty verification. Using a standard optimization of the channelizing matrix, the monitor would have the channelized

value and test-statistic distributions for the measured sources at its disposal, though it would be unable to access the channelizing matrix, which is often sensitive. The L channels could be optimized to distinguish certain likely spoofs.

Inclusion of penalty terms in the optimization routine of the channelizing matrix makes the CHO a more palatable option for treaty verification. The method to penalize individual channel performance leads to channels with no discriminatory power, yet optimal task performance is retained. This method would allow the host to share the channels with the monitor, increasing their confidence that the model is functioning as expected. A second method adds noise to the resulting channelization matrix, scaling down the discrimination ability of the model and reducing stored sensitive information. These two methods do not require the host to define precisely what components of the TAIs are sensitive.

The host can create a model that is unable to discriminate between the actual TAI and slightly altered items that differ along explicitly defined sensitive parameters. Such a model prevents the monitor from backing out the geometry of the constructed source. The two models cited in the above paragraph do not necessarily enforce poor discrimination along a sensitive parameter; this model does. This model could be shared between the host and monitor. This presents a significant benefit compared to the prior two methods. First, the channels are better able to discriminate spoofs as they have more structure than the channels resulting from the first two penalty routines. Second, performance in the discrimination task is optimal.

Putting all of this work together, an ideal model for using the CHO framework would:

1. Optimize the separation between the test-statistic distributions for the two items for the discrimination task.

2. Penalize the difference in test-statistic distributions for any stated sensitive parameters.

3. Design the model (or individual channels) to identify certain spoofs through an additional term in the optimization routine.

Such a model could optimally perform a task, be shared with the monitor, and be able to discriminate many different spoofs.

CHAPTER 6

Null Hypothesis Tests for Warhead Verification

While the methods developed so far in this thesis can perform binary-classification tasks such as verification of explosive dismantlement, the desire remains to develop null-hypothesis tests useful for treaty verification. The most important task required for future arms-control-treaties is the verification of the presence of a warhead. The monitor needs to answer the question "Is this object what the host says it is, or is it something else?" As with the binary-classification tasks, it would be ideal to develop a null-hypothesis model that (a) processes LM data to prevent aggregation of projection data for the tested item and (b) does not store sensitive information.

Unfortunately, it is difficult to create a model that processes LM data but is not spoofable. Section 6.1 explores some common null-hypothesis tests and explains why most are unable to process LM data. Section 6.2 introduces methods, using the LM likelihood from the ideal observer and the Hotelling observer as a framework, that while imperfect still satisfy the LM requirement. Section 6.3 presents some simulated applications of these methods and discuss the results.

6.1 Difficulty in Implementing Standard Null Hypothesis Methods with List-Mode Data

In this section there is a discussion of standard null hypothesis tests, and why they are unable to process LM data. The Chi-Squared test and Mahalanobis distance are discussed in detail, then a broader approach is taken.

6.1.1 Chi-Squared Test

The χ^2 distribution (Greenwood and Nikulin, 1996) is taught in most freshman statistics courses. It is the result of the sum of the square of independent, standard normal random variables (denoted in the equation below as Z_m). For M total data bins,

$$\chi^2 = \sum_{m=1}^{M} Z_m^2. \tag{6.1}$$

Depending on the number of bins M (which is also the number of degrees of freedom), the χ^2 value corresponds to a p value (the probability of observing that data given H_0) and a decision is made by thresholding the p value. When the measurement system is an imager, the data consists of the number of counts in each energy-pixel-particle bin, g_m. The distribution on each g_m is Poisson, but for a large number of counts approximates a normal distribution $\mathcal{N}(\mu = \overline{g_m}, \sigma^2 = \overline{g_m})$. Plugging these numbers into (6.1) results in the χ^2 Goodness of Fit Statistic,

$$\chi^2 = \sum_{m=1}^{M} \frac{(g_m - \overline{g_m})^2}{\overline{g_m}}. \tag{6.2}$$

Deviation of the tested g_m from its mean value increases the χ^2 value and reduces the probability that the tested data is from the H_0 distribution.

Unfortunately, the χ^2 test cannot be used to process list-mode data. The reason for this is the term g_m^2 that results from multiplying out the numerator in eq. (6.2). It is impossible to find the square of g_m using just list-mode data—the algorithm requires the aggregation of the projection data, which as previously discussed, is sensitive.

6.1.2 Mahalanobis Distance

The Mahalanobis distance (De Maesschalck et al., 2000) is essentially a more general description of the χ^2 distribution, which assumes independent data variables. The Mahalanobis distance takes the exponent of the multivariate normal distribution as a distance metric,

$$d = (\mathbf{g} - \overline{\mathbf{g}})^{\dagger} \mathbf{K}_{\mathbf{g}}^{-1} (\mathbf{g} - \overline{\mathbf{g}}). \tag{6.3}$$

When the data bins are independent, $\mathbf{K_g} = Diag(\overline{\mathbf{g}})$ and (6.2) and (6.3) are equivalent. While this distance metric is able to properly account for correlations between the variables, it ultimately has the same flaw as (6.2) in that it is unable to process LM data.

6.1.3 General Distance Metrics

Many other statistical tests have been analyzed. A useful list of standard similarity and distance metrics can be found in a paper by Sung-Hyuk Cha (Cha, 2007), which presents a series of metrics in terms of P, the probability density on the data

distribution for the null hypothesis, and Q, the density on the data distribution for the tested source. A quick scan of this paper reveals that only a single metric can be processed with LM data—the inner product,

$$S = \sum_{m=1}^{M} P_m Q_m \tag{6.4}$$

This is an unnormalized measure of the similarity of two distributions and can be thought of as the projection of $\mathbf{P}$ onto $\mathbf{Q}$. Similar to the ideal observer and Hotelling observer, this is inherently spoofable. Starting with $\mathbf{Q} = \mathbf{P}$, for two bins with equal probabilities of $\mathbf{P}$, one $\mathbf{Q}$ bin probability can be increased and the second decreased in such a manner to keep S constant.

Furthermore, a simple thought exercise presents a fundamental flaw with this model, or any hypothesis test that can be performed using LM data. If $\mathbf{P}$ is a flat distribution, then regardless of the shape of $\mathbf{Q}$, the same S is returned. In fact, this seems to be a fundamental issue with any null hypothesis test that processes list-mode data. However, there is no spatial or spectral measurement that would result in a flat distribution.

6.2 Theory

This section discusses the models that have been developed to perform null-hypothesis tasks.

6.2.1 Null Hypothesis Test Based on Likelihood Expression

The first step taken was the development of a model based on the probability density on (A_n, N). The likelihood from the ideal observer (4.8) was treated as the test statistic. In the equation below, it is explicitly noted that this probability is dependent on the acquisition time T,

$$\begin{aligned} t &= pr(\{A_n\}, N|H_0, T) \\ &= \int pr(\{A_n\}, N|H_0, \gamma) pr(\gamma) \mathrm{d}\gamma \\ &= \int Pr(N|H_0, \gamma, T) \prod_{n=1}^{N} pr(A_n|H_0, \gamma) pr(\gamma) \mathrm{d}\gamma. \end{aligned} \tag{6.5}$$

Once again, the reader should note that the total number of counts N is not LM data, though it is mostly nonsensitive. This model, while effective, is far from perfect. The

Poisson term has the nice feature that its value is greatest when N is closest to the calibration data mean $\overline{N}$. The probability drops off when measured objects have either higher or lower detection rates. The product of LM terms, meanwhile, is not so easily understood. First, it is possible to image a source that returns higher values of $pr(A_n|H_0,\gamma)$ on average than the null hypothesis source. An example of this is presented in the results section. Secondly, the count rate impacts the product of LM terms—as N increases, the number of likelihood products (all of which are less than one, due to the discrete nature of the probability densities in our work) increases, causing a decrease in overall probability.

For this reason, a second likelihood model was developed, this time conditioning the probability on the number of counts observed. This leads to a simpler model that is advantageous in certain circumstances,

$$\begin{aligned} t &= pr(\{A_n\}|H_0,N) \\ &= \int \prod_{n=1}^{N} pr(A_n|H_0,\gamma)pr(\gamma)\mathrm{d}\gamma \end{aligned} \tag{6.6}$$

This model performs better than (6.5) when the detected-count rates between H_0 and any tested sources are roughly equal. The reason for that is the test statistic distribution for (6.5) is broader than (6.6) as randomness in N increases the variance. This model also performs better in other circumstances. This is demonstrated further in the results section.

6.2.1.1 Implementation

Implementation of this model works similarly to the ideal observer. The null hypothesis object is imaged under a set of known nuisance parameters in the SKE case. The data is binned in some fashion, and the data is normalized to find the probability of detecting a particle in each bin. When nuisance parameters are present, the host must estimate $pr(\gamma)$ and find $pr(\{A_n\},N|\gamma,H_0)$ for many values of γ.

This model is then performed on independent data measured from the same trusted item, ideally measuring objects that take on the same nuisance parameter distribution as the calibration data was acquired for. This results in a test-statistic distribution. Thresholds are set based on this distribution, as described in Section 1.3.1, at the 97.5 percentile and 2.5 percentile.

Finally, unknown items are tested. Any tested items that return a test statistic greater than the 97.5 percentile or lower than the 2.5 percentile are rejected. Note

that a distribution comparison as discussed in Section 6.1 cannot be performed as the monitor could not guarantee that an entire set of imaged objects are of the same type, so objects would need to be tested and possibly rejected one by one.

6.2.1.2 Storage and Need for an Information Barrier

Similar to the ideal observer, model storage for the two likelihood-based null hypothesis tests would be very sensitive. This would require an IB. The only information visible to the monitor would be the test statistic, or when nuisance parameters are present, a series of likelihood values that are combined to result in a single test statistic used to make decisions.

6.2.1.3 A Cheating Host and Monitor

While designed spoofs are not covered in great detail in this chapter, one can refer to the discussion of spoofs in Section 4.2.2 for information on how a host could design a spoof to fool the likelihood model and return the same test statistic.

As stated in the ideal observer section, the SKE model would give a very limited amount of information to the monitor, just the normal test-statistic distributions. Upon incorporation of nuisance parameters, the model can become highly non-normal, and this distribution could be used by the monitor to back out some knowledge on the construction of the geometry or isotopics used.

6.2.2 Linear Models

In this section, the HO and CHO models are applied for a null-hypothesis task rather than for binary classification. The desire to use a linear model is based on the various advantages that optimization of the channelizing matrix $\mathbf{T}$ in the CHO can provide. The CHO's capability of penalizing sensitive information is crucial for null-hypothesis tasks as well. Penalty terms can be added to the routine, as in (5.17), to prevent discrimination of objects that differ along the sensitive parameter.

The difficulty in implementing this approach comes in finding a pair of sources to differentiate. As there is only one set of calibration data associated with TAI, the host would need to define an alternative hypothesis H_a to discriminate H_0 against. H_a could be a lack of sources in the field of view or a set of likely spoofs, or something else. Generally, the optimization routine for the channelizing matrix would look like,

$$SNR^2(\mathbf{T}) = SNR^2_{0-a}(\mathbf{T}) - \eta \sum_{k=1}^{K} SNR^2_{(0,p_k=p_{k,0})-(0,p_k=p_{k,0}+\Delta p_k)}(\mathbf{T}) \tag{6.7}$$

In the above equation, p_0 corresponds to the original value of the sensitive parameter and p_k corresponds to the parameter value for the k^{th} sensitive penalty term.The data sets for the penalized objects would all be simulated. This optimization of $\mathbf{T}$ would result in nonsensitive channelized values, however H_a is chosen.

The first consideration was to set H_a to a signal absent scene, which is the simplest definition of H_a. If a neutron measurement is considered, the background is minimal (it is ignored here). The flaw in this model is clear when analyzing the Hotelling weights (ignoring the penalty terms in (6.7) for now). In the absence of nuisance parameters, the first and second order statistics can then be represented as,

$$\begin{aligned} \Delta\overline{\mathbf{g}} &= \boldsymbol{g_0} \\ \mathbf{K_g} &= Diag(\overline{\boldsymbol{g_0}}/2) \end{aligned} \tag{6.8}$$

The Hotelling weights would then be $\mathbf{K_g}^{-1}\Delta\overline{\mathbf{g}} = \frac{\boldsymbol{g_0}}{\boldsymbol{g_0}/2} = 2$ for every single detector bin. The observer model would serve as a simple count-rate detector. This would be optimal for the task of distinguishing the detected image data from the absence of data, but would be spoofed easily. Any source with the same count rate as the null-hypothesis TAI would return the same test statistic.

A more optimal approach than taken in (6.8) would be to optimize the SNR^2 of the test-statistic distributions between H_0 and a summation of objects representing the space of likely spoofs. The spoofed sources could consist of slabs of SNM, non-weapons grade plutonium and uranium, or some other object that the host could create to fool the monitor. This space is extremely difficult to define; furthermore, the massive dimensionality reduction involved in the HO and CHO would likely mean that there are successful spoofs whose measurement data fall significantly far from any source that looks like a TAI. Hence, the space of all spoofs to penalize against would likely be too large.

As I have not developed a convincing linear model to use for hypothesis testing, there are no simulated experiments for this theory; more discussion on hypothesis tests can be found in the future work section.

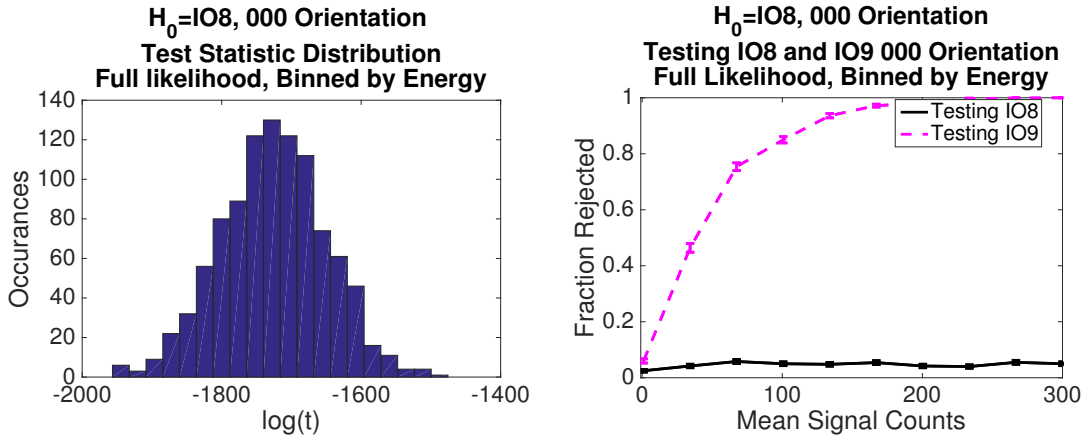


Figure 6.1: The left plot shows an example test-statistic distribution $pr(\log(t)|H_0)$ for IO8 using (6.5). Data was sampled from IO8 1,000 times and the test statistic found to generate this distribution. Cutoffs were set at roughly -1,575 and -1,825. Sources that produced test-statistic values outside this range were rejected. The second plot shows the percentage of the time a second tested item was rejected.

6.3 Experiments

In this section, the models discussed in Section 6.2.1 are employed to perform hypothesis-testing tasks on the INL inspection objects and various neutron sources.

6.3.1 Inspection Object Discrimination

In this subsection, the IO8 and IO9 inspection objects (Section 3.2.1) were considered. Throughout the following experiments, the data was binned by energy. To begin, IO8 was treated as the null hypothesis, with only the 000 orientation being considered. A test-statistic distribution was generated for independently simulated IO8 000 orientation data. A distribution over the log likelihood is shown in Figure 6.1. For large N, it is normal, as the log likelihood is the sum of many independent random variables. A cutoff for the 2.5 and 97.5 percentile were found from this distribution; any tested values outside this cutoff were rejected. Next, data was sampled from the independently simulated IO8 and IO9 simulation data, the model was performed on that data, and the resulting values were compared to the rejection cutoffs. Given enough counts, this model effectively rejects IO9 data while only rejecting IO8 5% of the time, as expected.

In the next set of experiments, various orientations of IO8 were treated as H_0 and various orientations of IO9 were tested, with the hope that IO9 would be rejected in all cases. In this study, both likelihood models discussed in Section 6.2.1 were

performed on the data. The reader should refer to Figure 3.11 for a refresher on the spectra for the two objects at various orientations. The results (see Figure 6.2) help to demonstrate the differences in the models. There are three components here that are worth discussing:

- The Poisson component in (6.5) pushes the IO9 distribution for the "full likelihood" model to lower test statistic values. This is generally a small contribution compared to the LM component.

- The measured IO9 gammas have higher probabilities of detection at lower energies than the IO8 gammas. For both models, this means that testing IO9 returns a higher LM likelihood $\log(pr(A_n|H_0))$ than IO8 on average.

- The magnitude of the full-likelihood test statistic depends most on the number of products (N) of $\log(pr(A_n|\gamma, H_0))$. When testing against a source with a higher detection rate, more particles are detected, leading to more LM products and an increasingly negative log likelihood. When testing against a source with a lower detection rate, less LM terms are recorded, resulting in a higher test statistic.

In the 111 discrimination task, the IO9 count rate is roughly 30% lower than I08's. In this case, the individual $\log(pr(A_n|H_0))$ values when testing IO9 are greater than for IO8, which affects both models, pushing the test-statistic distribution higher for IO9 than IO8. In addition, for the full likelihood model, less products go into the model than for IO9 than for IO8, further increasing the distance between the two distributions. This is why the full-likelihood model outperforms the LM product model for this orientation.

In the 000 discrimination task, when testing the IO9 source that causes an increase of 60% in detection rate over IO8 for that orientation, IO9's full likelihood is more negative due to more LM products entering the equation. However, the count rate difference is not great enough to pull the IO9 log-likelihood distribution far enough to the left to provide an improvement in performance over the LM only model.

The poor performance of the hypothesis test that integrates over nuisance parameters in Figure 6.2 can be related back to the test-statistic distributions. For testing data corresponding to a single realization of the nuisance parameters, the model produces a normal distribution (even the model that incorporates nuisance

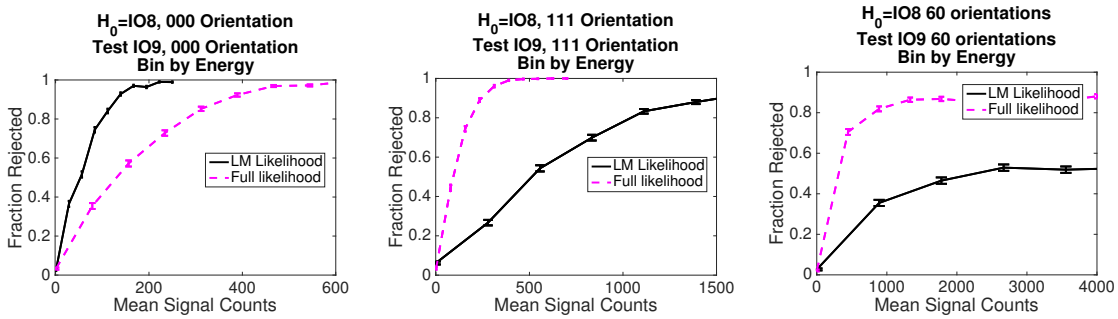


Figure 6.2: Performance for models built on certain orientations of IO8 in discriminating certain orientations of IO9. When H_0 is set to a single orientation of IO8, it is able to reject IO9 for the two studies shown. When H_0 is set to all IO8 orientations, it is only able to reject IO9 a fraction of the time even as the acquisition time goes to infinity.

parameters). The test statistic-distribution resulting from testing objects taking on many different realizations of the nuisance parameters results in a sum of normal distributions. The resulting distribution (Figure 6.3), as in the ideal observer section, can be non-normal. When IO8 is treated as H_0 and tested on independently simulated IO8 data using many nuisance parameter realizations, the result is a mostly normal test-statistic distribution. When testing IO9, however, the distribution is highly non-normal. While it appears that certain values of $\log(t)$ between the IO9 distribution peaks are negligible, this is due to the stratified orientation sampling method; if the orientation of IO9 was truly random, as in real life, the distributions would be continuous.

As the test statistic distributions show, regardless of how many counts are detected, the model would never be able to reject IO9 100% of the time. Likewise, if IO9 were chosen to be H_0, as in the right figure of Figure 6.3, the model would never be able to reject IO8's data because the 2.5% and 97.5% tiles would always encompass it.

6.3.2 Ring Source Hypothesis Testing

In this section, the 20 cm ring source is treated as H_0. Independent data measured from the 20 cm ring source was used to create the $pr(t|H_0)$ distribution for each of the two discussed models ((6.5) and (6.6)). The 2.5 and 97.5 percentiles were found for these test statistic distributions, and any tested items yielding test statistics outside these bounds were rejected. The various neutron sources discussed in Section 3.2.2 and Section 3.2.3 were tested against H_0. All of the data sets were binned by pixel

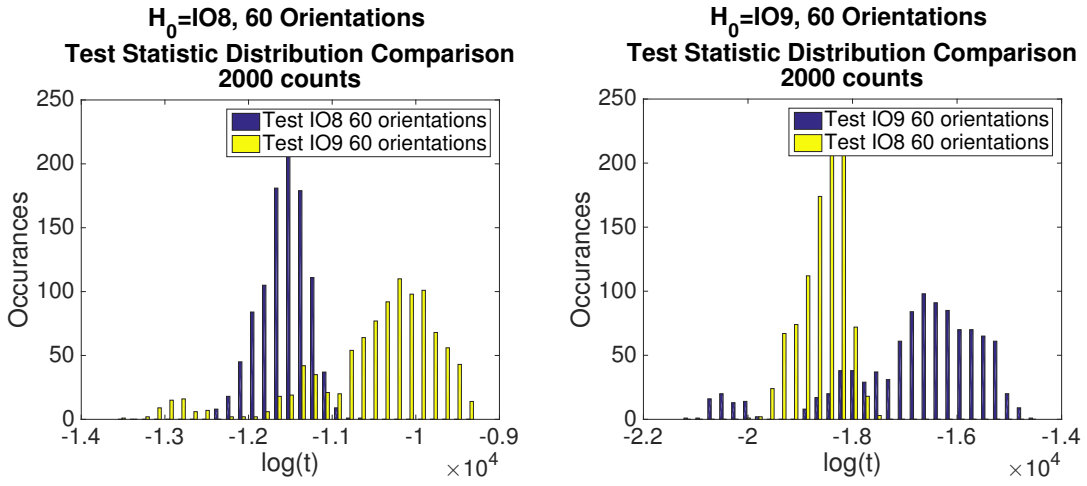


Figure 6.3: In the left plot, IO8 is treated as H_0. The model was built and tested on many orientations of IO8 and IO9. In the right plot, IO9 was treated as H_0 and IO8 and IO9 data were tested. Note that the distributions are approximately the same in the left and right plots. This demonstrates the fact that the distribution nature is caused more by the variability of the tested source's data than the model itself.

ID. Because emphasis was put on the neutron count maps in decision making for this task, all of the count rates were set equal. This approach benefits the LM ratio likelihood test (6.6) over the likelihood that incorporates the count rate (6.5). This is demonstrated in Figure 6.4. At 25,000 counts, the full likelihood model cannot distinguish a ring source from a square source based on the test statistic. Notice that the means for the distributions are the same regardless of model but the variance increases considerably for the model conditioned on acquisition time. This is because the number of detected counts is still a random variable when the acquisition time is held constant, leading to variability in the number of products in (6.5).

This is demonstrated further by the performance plots in Figure 6.5. The likelihood conditioned on acquisition time is not able to reject any of the other neutron sources. The likelihood conditioned on the number of observed counts, meanwhile, is able to reject all of them given a long enough acquisition time. This is not the first time we've seen this result; the ideal observer proved able to correctly reject certain spoofs (Figure 4.5) as well.

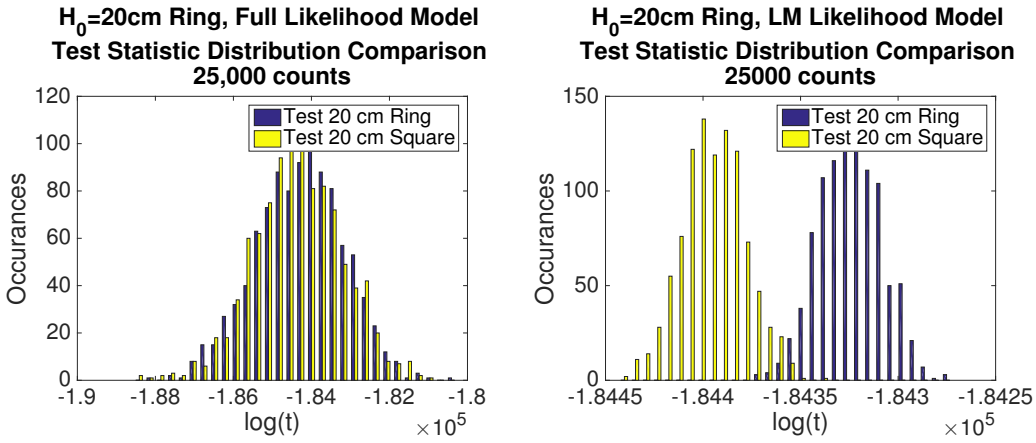


Figure 6.4: The left plot shows the test statistic distributions for the ring and square sources using the likelihood expression based on acquisition time (6.5) and the right plot shows the test statistic distributions for the likelihood using a predefined number of counts (6.6).

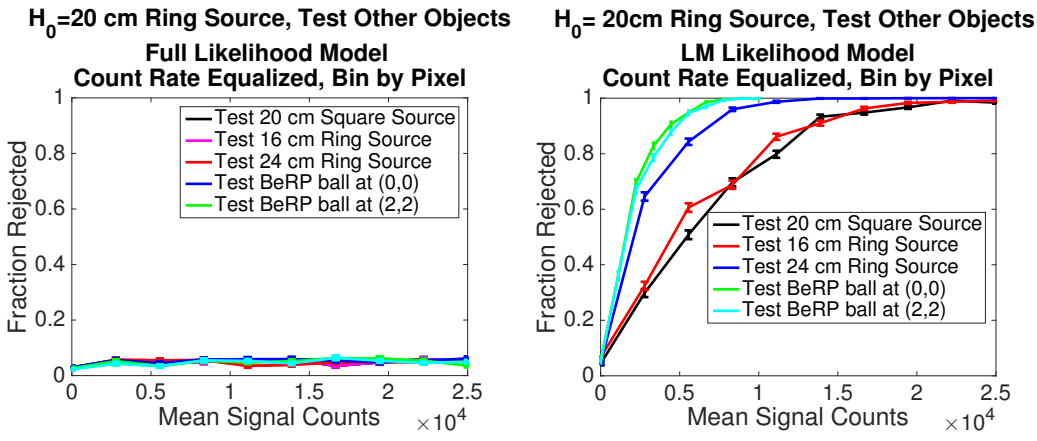


Figure 6.5: The left plot shows the performance of the likelihood expression based on acquisition time (6.5) when testing alternative neutron sources. The right plot shows the performance of the likelihood using a predefined number of counts (6.6).

6.4 Conclusion and General Comments

A null-hypothesis test was developed that uses the LM likelihood to reject possible spoofs. This test is not ideal, and can be tricked. This is especially true when nuisance parameters are present in the TAI and imaging system that lead to broad test-statistic distributions on the TAI, preventing the model from rejecting a large range of test-statistic values. Furthermore, the sensitive nature of the stored data would require an IB, preventing the monitor from accessing the model.

Linear models were also considered, and provide the advantage that the penalty term formalism discussed in chapter 5 can be applied. However, the definition of the alternative hypothesis is critical to this model's performance and it is unclear how to define H_a to best reject spoofs.

CHAPTER 7

Future work

This dissertation presents multiple methods that could be used to perform binary-discrimination tasks without revealing sensitive information to the monitoring party. However, there is more that can be done, whether that is expanding upon the models already developed or creating new methods for null-hypothesis tests. Section 7.1 and Section 7.2 consider the improvements that will need to be made to the simulation studies and observer models to perform real-life verification. Section 7.3 discusses new ideas for null-hypothesis tests that have yet to be be implemented. Section 7.4 expands on the CHO work to discuss improved penalty terms. Experimental data has been acquired on the ring and square sources, and Section 7.5 shows the first results for this data and summarize how to compare simulation and physically-acquired data. Finally, Section 7.6 demonstrates the advantages of a detector that is physically insensitive to certain parameters of the imaged items. One possible method to create such a detector is presented.

7.1 Simulation Studies

The simulations performed for this dissertation were fairly simplistic; if these models were going to be used to predict real-life performance, they would need to be improved. There are three components addressed here:

- Most importantly, variation in detector response is critical to model performance and was ignored in the simulation studies in this thesis. (Section 7.1.1)
- The simulation geometry is bare bones, including just the object and detector geometries. The impact of simulating a more physically realistic world is discussed here. (Section 7.1.2)
- Pulse-shape discrimination between gamma rays and neutrons was not considered in this work. Its inclusion would impact the performance of the models. (Section 7.1.3)

7.1.1 Variability in Detector Response

In the simulation studies, there was no image-to-image variability incorporated in the detector response other than energy smearing. In reality, there are changes associated with the detector response over short and long time scales. Significant temperature changes are a standard problem. If this detector is shipped from one location to another, the detector response could change (there are often numerous missile sites where testing would occur). During transport, the geometries of the detector could degrade (in the case of PMTs) or shift. In particular, shifts in the location of the light guides and PMTs could alter the light collection efficiency. This requires calibration measurements prior to measurements at a new site.

An example of the changing detector response over a large scale time period is shown in Figure 7.1 and Figure 7.2. These maps are generated by flooding a given detector block with neutrons. The location of interaction can be determined by the ratio of outputs for the four PMTs for each detector block. For a detector with all of the PMT's having equal gain and light-collection efficiency working as designed, there are 100 evenly spaced Gaussian peaks on the neutron flood map. The location of the calibration peaks shown in the two calibration measurements was found assuming the ratio of PMTs for a perfect detector response. When the measurement occurs, any detections in a given pixel block with the PMT ratio corresponding to that calibration measurement would be assigned to that pixel ID.

The two count maps differ significantly in some detector blocks. In these maps, blocks such as the one in row 1, column 2 and row 2, column 2 show a fairly ideal detector response, where the light coming from each pixel is easily distinguishable. The block in row 2, column 3 could be due to PMTs in the upper left corner and lower right corners having increased gain, or PMTs in the other corners having lower gain. The end effect is a stretching of the response along the upper left to lower right diagonal. Of particular concern when comparing two calibration measurements is the blocks in row 4, columns 1-3. It is not clear what is causing this behavior but it could potentially be that air gaps are being created in the light guide to scintillator and light guide to PMT connections, causing loss of light in certain regions.

In addition, a simple neutron-flood calibration measurement is essentially a first-order calibration. For example, if the location of the neutron source changes, the neutron travels through a different path in each pixel. For this particular detector, the pixels are fairly long and light bounces around and spread out before entering

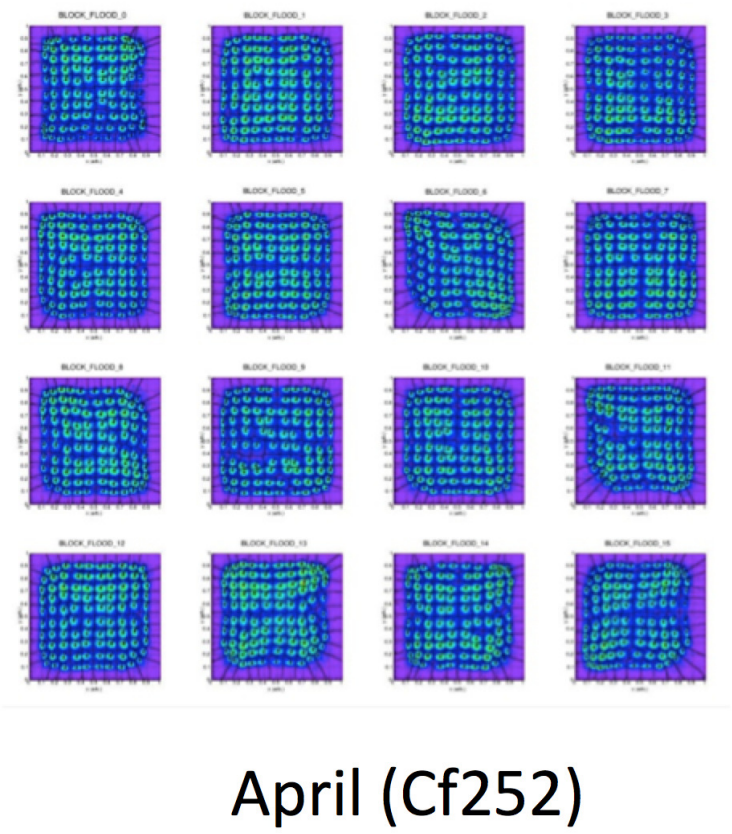


Figure 7.1: An example calibration measurement for the fast-neutron coded aperture detector taken in April 2015 using a Californium source.

the light guide, so location of the calibration source does not significantly effect the light collected.

7.1.2 Room Geometry

A concrete floor was considered in the early stages of the project. The detector was placed inside a 4 meter x 8 meter square room in the second stage of the simulation, which transports the flux exiting the inspection object to the detector. 6 inches of concrete was used as the floor. GEANT4 allows certain materials to be uploaded from the National Institute of Standards and Technology database. Concrete is composed of 52.9% oxygen, 33.7% silicon, 4.4% cadmium, 1% hydrogen, and other elements, all with Z<26. As expected, the addition of low Z materials to the world geometry causes a far greater impact on the neutron-simulation speed than the gamma-simulation speed.

The addition of the floor led to a roughly 11% decrease in the speed of gamma-ray simulations—a drop from 105,000 events per second to 93,000 events per second.

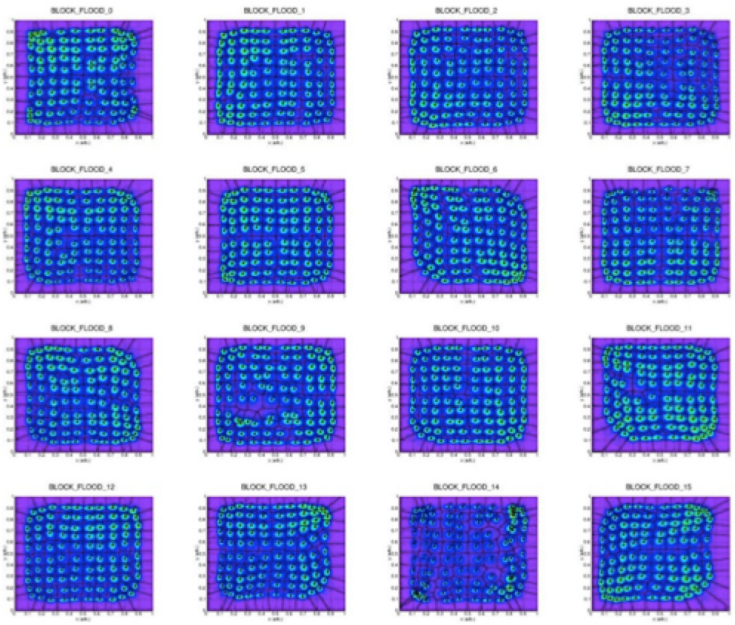


Figure 7.2: An example calibration measurement for the fast-neutron coded aperture detector taken in November 2015 using an AmBe source.

The effect on the neutron simulations was far more dramatic. Inclusion of the floor dropped the neutron simulation speed by over 90%, from 53,000 events per second to 4,700 events per second. This significant speed decrease combined with less than satisfactory results when attempting to utilize importance sampling with GEANT4 (see Section 3.4) is the reason the floor, and any other room geometries, were not simulated. Another important material for consideration in the room geometry is aluminum due to its high rate of thermal-neutron capture and subsequent gamma emission (Hardell et al., 1969).

7.1.3 Pulse-Shape Discrimination

In the simulation studies, misclassification of neutrons as gammas, and vice versa, was largely ignored. PSD is often done by observation of the electron output pulse from the PMT. Protons (produced by elastic scattering by neutrons off of hydrogen atoms) and electrons (produced by photoelectric absorption and Compton scattering of gamma rays) have different time-dependent light yields. An example method to accomplish PSD is presented in (Adams and White, 1978). In this procedure, the pulse is integrated over two time windows. T_1 (corresponding to the width of the

rising pulse, set to roughly 25 ns) has integrated charge Q_1 and T_2 (the total pulse width, set to roughly 400ns) has integrated charge Q_2. Normalization factors for the integrated charge, K_1 and K_2 are then found. A check is done to see if $K_1 Q_1$ is greater than $K_2 * (Q_1 + Q_2)$. If so, the particle is declared to be a gamma ray; if not, it is declared a neutron. Other PSD methods use a tail-to-total ratio.

Regardless of the approach chosen, PSD methods are prone to misclassification. Such misclassification leads to a smearing of the neutron and gamma detector data sets, which could degrade task performance for the models developed in this dissertation. Accounting for misclassification is particularly important when using reconstruction techniques. For example, in IO8 and IO9, where uranium (high gamma-emission rate, low neutron-emission rate) shields plutonium (high neutron-emission rate), misclassification of gammas from the uranium material as neutrons would result in neutrons that appear to be coming from the uranium material. This could lead the monitor to believe that the uranium geometry is actually another material.

7.2 Model Implementation

This section explains how the observer models would be adapted to incorporate the various physical processes explained in Section 7.1.

7.2.1 Variability in Detector Response

Accounting for such drastic variation in detector-response is impossible to do statistically using the nuisance-parameter formalism developed for the ideal observer and CHO. A calibration measurement must be taken prior to treaty-verification measurements. Sandia employees have developed code that takes in the detected neutron flux on each detector block (found in simulation) and uses the flood map calibration measurement to find the expected output ratio for the four PMTs. This can be used to create a realistic set of detector data from the GEANT4 simulated data.

If the models built on training data at a specific site are intended to be used to perform tasks on items at another site, there must be a methodology to adjust the data to the new detector response in order to perform the observer models. This could be done by taking the experimental data and using the flood map to back out an estimate on the number of interactions in each pixel. Then, when performing the

model at a second site, the absorbed flux for each pixel would be translated back to an electronic output by using the new detector-response code. Any errors in this process affect the task performance.

Accurately adjusting the data for the current detector response is doubly important for the CHO that penalizes storage of sensitive information. This model is trained on certain data sets, some of which are acquired in simulation and some through experiment. The simulated detected flux can be reliably generated through GEANT4 or MCNP. The detector-response code would be applied to the simulation output, ideally resulting in an accurate set of detector data for the penalized objects. The model's ability to discriminate the TAI from the penalized sources would then be penalized in the channelizing-matrix optimization routine.

If this observer and detector will be used in future verification measurements, penalization of the differences in these data sets must be consistently strong and adapt to the changing detector response. Imperfect penalization would result in the monitor being able to observe differences in the test-statistic distributions and use the TAI's distribution to reverse-engineer the geometry. There are multiple procedures that could be used to accomplish this; two are outlined below.

7.2.1.1 Assume the Penalty Direction Vector is Constant

When nuisance parameters are not present, the penalization routine degrades **T**'s ability to detect a change in measured data between the TAI and the penalized object. This data-difference vector between measurements would be in the null space of **T**. Using the calibration of the detector response, the simulated GEANT4 data can be turned into a simulated measurement. The difference between the experimentally measured TAI and the simulated penalized object would be found and the model built off of this. The second set of verification measurements would be done at some later time with a different detector response. A measurement of a trusted TAI could be done, then the simulated penalized source found by assuming the difference vector is constant despite the changing detector response. The model would be trained on this data. This would require a second measurement of the trusted TAI with the new detector response, which the host may not want to agree to.

This procedure is a bit simplistic. First of all, the presence of nuisance parameters makes the penalization routine more complicated than simply putting a single difference vector in **T**'s null space. Second, as Figure 7.1 and Figure 7.2 show, the

response varies with location on the detector. If there are certain pixels where less light is collected between calibration measurements, both the TAI and altered TAI would have less light collected, leading to a smaller difference in measurement data for the second TAI measurement than the first. Assuming a constant data-difference vector would yield inaccurate penalization.

7.2.1.2 Method to Adjust Experimental and Simulated Data to New Detector Response

The flood map calibration measurement would be performed before any measurements of the TAIs. The two TAIs in the discrimination task would then be measured and any penalized objects simulated. An estimate on the detected flux in each pixel would be found from the experimental data on the TAI and the flood map. This detected flux should be consistent across measurements (ignoring the room geometry effects for now). When the detector is taken to a new site, a new flood map measurement would be taken and the detector response recalibrated. The simulation and experimental detector flux from the first set of measurements would then be processed with this new detector response to best model the new measurement data. Then, the model would be trained on this new data and unknown items tested. This procedure is probably more accurate than the first and does not require a second measurement of a trusted TAI.

7.2.2 Room Geometry

To account for the room geometry's effect on the detector data, a prior measurement would be required. Such a measurement could be taken with the TAI in place. A chunk of attenuating material could be placed between the TAI and detector, as in Figure 7.3. All neutrons traveling on a straight line path to the detector would be attenuated while the other neutrons interacting in the room would not be shielded. Then, a measurement without the attenuator could be taken and the difference between these measurements would be treated as the calibration and testing data. This routine would account for both room effects and the locally varying background. The issue with this approach is that the attenuating material could suppress the background coming from behind the item. When the attenuation material is removed, background particles could pass through the source and interact in the detector.

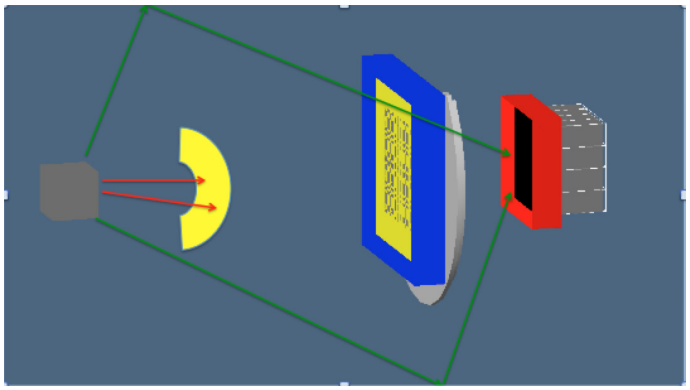


Figure 7.3: An example measurement that could be used to gauge the impact of the geometries of the room on the detector data. The inspected object is shielded by the polyethylene hemisphere.

7.2.3 Pulse-Shape Discrimination

Accounting for imperfect PSD in the observer models is actually unnecessary. This is a benefit using the template-matching approach with projection data. Both the training and testing data contain the misclassified gammas and neutrons. The detector software decides whether the detected radiation was from a gamma ray or a neutron and then that event's data is binned by particle type, pixel ID and energy. In fact, the host and monitor could instead choose to bin the data by PMT ratio and the two measures for the integrated charge of the pulse, rather than pixel ID, total energy and particle type.

7.3 Quadratic Approximation for Null Hypothesis Test

As discussed in Section 6.1, the large majority of distribution-distance metrics do not naturally process data in LM format. However, a quadratic approximation to some of these methods seems like a viable alternative. The work presented here was undertaken by Mohammed Khalil, a Sandia employee. Assuming normalized calibration data $\mathbf{P}$, and testing data $\mathbf{Q}$, the Mahalanobis distance can be represented as,

$$d = (\mathbf{Q} - \mathbf{P})^{\dagger} \mathbf{K_P}^{-1} (\mathbf{Q} - \mathbf{P}), \tag{7.1}$$

where $\mathbf{K_P}$ is the covariance matrix of the normalized data $\mathbf{P}$. This is a good model to start with. It returns a non-negative distance d, and is 0 when $\mathbf{P} = \mathbf{Q}$. An SVD can be done of the (symmetric) covariance matrix so that it can be represented by,

$$\mathbf{K_P} = \mathbf{X}^{\dagger} \Lambda \mathbf{X}. \tag{7.2}$$

Here, $\mathbf{X}$ is a matrix of eigenvectors and Λ is a diagonal matrix with the eigenvalues λ_k along the diagonal. There are K total eigenvectors; for an invertible covariance matrix, K is equal to the number of bins in the data, M. The inverse of the covariance matrix is,

$$\mathbf{K_P}^{-1} = \mathbf{X}^{\dagger}\Lambda^{-1}\mathbf{X}. \tag{7.3}$$

Using this representation, (7.1) can be represented as,

$$\begin{aligned} d &= \sum_{k=1}^{K} (\mathbf{Q}-\mathbf{P})^{\dagger}\mathbf{X}_k^{\dagger} * 1/\lambda_k * \mathbf{X}_k(\mathbf{Q}-\mathbf{P}) \\ d &= \sum_{k=1}^{K} 1/\lambda_k * ((\mathbf{Q}-\mathbf{P})^{\dagger}\mathbf{X}_k)^2 \\ d &= \sum_{k=1}^{K} 1/\lambda_k * (\mathbf{Q}^{\dagger}\mathbf{X}_k - \mathbf{P}^{\dagger}\mathbf{X}_k)^2. \end{aligned} \tag{7.4}$$

Taking this a step further, the inner product between $\mathbf{Q}$ and $\mathbf{X}_k$ can be expressed in terms of the counts N_m (out of N total) detected in each of the M bins,

$$\begin{aligned} \mathbf{Q}^{\dagger}\mathbf{X}_k &= \sum_{m=1}^{M} Q_m X_{k,m} \\ \mathbf{Q}^{\dagger}\mathbf{X}_k &= \sum_{m=1}^{M} \frac{N_m}{N} X_{k,m} \\ \mathbf{Q}^{\dagger}\mathbf{X}_k &= \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} X_{k,m_n}, \end{aligned} \tag{7.5}$$

where X_{k,m_n} is the value of the m_n^{th} bin (corresponding to the n^{th} detected particle) for the k^{th} eigenvector. This derivation proves that (7.4) can process data in LM format, with the $\mathbf{Q}^{\dagger}\mathbf{X}_k$ product being updated as events are read in. K total LM products would need to be stored for this method. Even if the monitor was to be given each eigenvector, the dimensionality reduction would make it hard to back out $\mathbf{Q}$. This procedure shares some similarities to the CHO—the CHO projects a set of channels onto the observed data, while this method projects eigenvectors onto the data.

The question then becomes one of spectral analysis. To best distinguish $\mathbf{Q}$ from $\mathbf{P}$, all $K = M$ eigenvectors would be needed to best approximate the true distance d. However, if all were used, the model would likely be deemed sensitive. Instead, if $K < M$ eigenvectors were used, the returned d in (7.4) would be an approximation to the d using the complete decomposition. Though some eigenvectors (those with

larger λ_k) are more important to the decomposition of $\mathbf{K_P}$ than others, that doesn't necessarily mean they would do a better job identifying spoofs. Spoofs can take on many geometries and ideally the eigenvectors would be able to reject many different types. However, the eigenvectors for a certain $\mathbf{K_P}$ are not something that can be optimized as they are fixed—the best eigenvectors to discriminate spoofs would need to be selected.

Instead, $\mathbf{K_P}$ can be approximated generally through,

$$\mathbf{K_P}^{-1} = \mathbf{\Phi}^{\dagger}\Lambda\mathbf{\Phi}. \tag{7.6}$$

where $\mathbf{\Phi}$ is a normalized vector not related to the eigenvector $\mathbf{X}_k$. Using this formalism, the Φ vectors could be optimized in some way to best reject a set of spoofs.

7.4 Channelized Hotelling Observer

There are a few tasks left to accomplish in the development of a CHO model that can effectively penalize the model's discriminatory ability on sensitive parameters of the object. This section covers two in particular—preventing discrimination based on the channel distributions for two objects, and preventing discrimination on the variances of the two distributions. The ideal result would be a $\mathbf{T}$ that returns the exact same test-statistic and channelized value distributions when the penalized pairs of objects are measured.

7.4.1 Preventing Discrimination of Channelized Value Distributions

The techniques discussed in Section 5.1.5 serve to equalize the mean test-statistic value for objects that differ along a sensitive parameter. This is a good start, but ideally, the monitor would also be unable to back out information from the channelized values, $\mathbf{v}$, either. Each channelized value can be presented as the inner product of a channel and the data vector,

$$v_l = \mathbf{T}_l\mathbf{g} \tag{7.7}$$

An example penalty term is below,

$$f_{pen}(\mathbf{T}) = \eta \sum_{j=1}^{2}\sum_{k=1}^{K}\sum_{l=1}^{L} SNR^2_{(j,p_k=p_{k,0})-(j,p_k=p_{k,0}+\Delta p_k)}(\mathbf{T}_{l^{th}channel}). \tag{7.8}$$

Note that this penalty is similar to (5.21), except now the discrimination ability of each individual channel is being penalized for the pair of objects that the host does

not want the model to discriminate, rather than the pair of objects that the model needs to discriminate. It should be noted that equal means on the channelized data for the channels would imply equal means on the test statistics. This penalty term would replace eq. (5.25).

7.4.2 Preventing Discrimination on Distribution Variance

Prior work has shown that penalizing the SNR^2 of the test-statistic distributions between the two objects serves to equate the mean of the test-statistic distributions. However, as explained at the end of Section 5.2.5, in studies where the count rates for the penalized pair of objects are not set equal, the variances of the two test-statistic distributions can be notably different (in the stated example, one had a variance of 62 and the other 74). The monitor could use the fact that they are different to reverse engineer the objects and back out the true value of that sensitive parameter.

This is a difficult problem to overcome. Penalizing SNR^2 is not the same as the difference in mean data, $\overline{\Delta t} = \mathbf{W}_{\mathbf{g}}^{\dagger}\Delta\overline{\mathbf{g}}$ for the penalized objects. This is because $\mathbf{W}_{\mathbf{v}}$ is found for the performance-optimized pair of sources for the task, not the performance-penalized pair. When using the penalized pair to determine $\mathbf{W}_{\mathbf{v}}$, the magnitude of the variance dropped drastically, but the ratio between variances (and hence, the ability to differentiate the two) stayed approximately the same.

A second attempt was made to penalize the variance based on equation (5.3). The difference in variance on the two test-statistic distributions can be expressed in matrix form as,

$$\Delta\sigma_t^2(\mathbf{T}) = \mathbf{W}_{\mathbf{g}}(\mathbf{T})^{\dagger} Diag(\Delta\overline{\mathbf{g}})\mathbf{W}_{\mathbf{g}}(\mathbf{T}). \tag{7.9}$$

In this equation, $\mathbf{W}_{\mathbf{g}}$ is a function of the channelizing matrix. However, this is not a distance metric, so the square was taken to create a penalty function,

$$f_{pen}(\mathbf{T}) = (\mathbf{W}_{\mathbf{g}}(\mathbf{T})^{\dagger} Diag(\Delta\overline{\mathbf{g}})\mathbf{W}_{\mathbf{g}}(\mathbf{T}))^2. \tag{7.10}$$

This form is easily differentiable with matrix calculus. Unfortunately, an optimization routine based on this objective function and gradient has not returned test-statistic distributions with more similar variances. Instead, it returns distributions with variances lower in magnitude. Alternative penalty terms that equalize the variances of the distributions are still being considered.

Another approach could be to utilize a distance metric between the distributions of the measured objects that differ along the sensitive parameter. Some possible

techniques can be found in (Cha, 2007). Despite their inability to process LM data, these metrics can still be used in the optimization routine.

7.5 Experimental Study on Ring vs Square Source Size Penalization

This dissertation presented a method to discriminate two objects by geometry type while avoiding discrimination based on the object size. Experimental measurements on extended Californium sources have been taken to compare to the simulation studies. Small Californium sources were placed on a moving surface and imaged by the detector. The surface was programmed to move in both circular and square motions of different sizes. The results after a matched filter reconstruction technique (Ye et al., 2006) are shown in Figure 7.4. Reconstructions of the objects show that the location was not held constant. This is not a concern for this study, but it does mean a comparison of classification performance (and penalization of the size parameter) won't be exact between simulation and experiment. Furthermore, this study emphasizes the importance of accounting for nuisance parameters in an actual verification measurement. These images could be used to classify independently imaged objects, which could vary in location just as these objects did. If so, the orientation nuisance parameter would need to be accounted for, requiring more calibration images.

Each measurement will be split up into training and testing data sets. The optimization routine will optimize $\mathbf{T}$ to perform the geometric discrimination task for the 20cm geometries while penalizing the ability to discriminate a single geometric source of different sizes. $\mathbf{T}$ will be tested on the independent data sets. We plan to publish these results in IEEE Transactions on Nuclear Science in late 2016.

7.5.1 Simulation Validation

These measurements also present an opportunity to perform validations of the GEANT4 simulations. The reconstructed objects could be run through GEANT4 and the resulting projection data paired with the calibrated detector response to create a simulated experimental measurement. This could be compared to the measured data to check model consistency.

7.6 Detector Insensitive to Certain Information

The downside to relying on electronics to apply the observer model to the data is that there is always the opportunity for one side to fool the other. A dishonest host

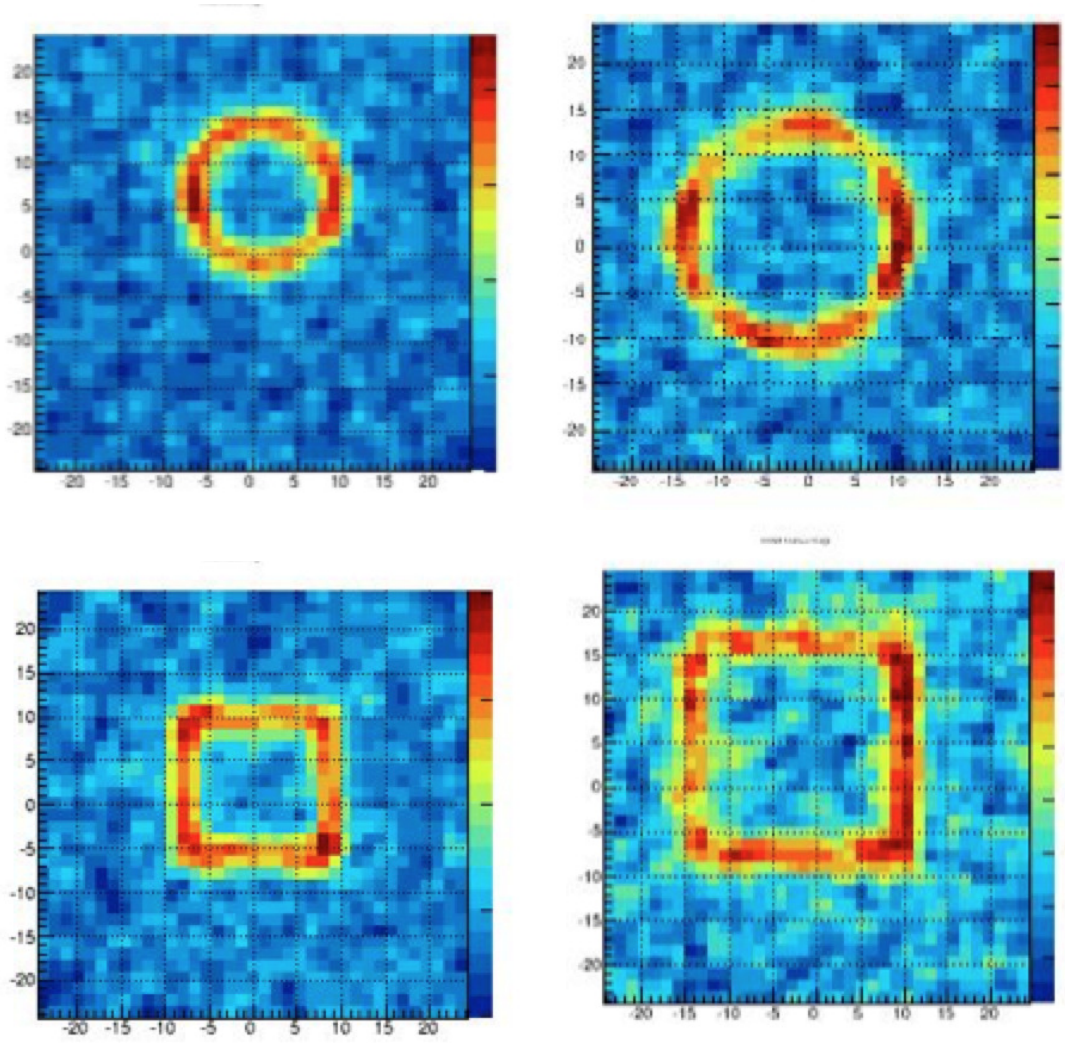


Figure 7.4: Reconstructions of the extended ring sources (16cm and 24cm) and square sources (16cm and 24cm) are shown here. The 16cm ring and 24cm square were imaged a little off center by accident.

could implement an electronic switch that changes how LM data is processed, or process data other than what is read in from the detector. This is an especially significant worry when implementing the ideal observer model, because the monitor does not have access to the model itself. Likewise, the monitor could try to fool the host by including a method to secretly aggregate the projection data.

The ideal end result for this project would be the construction of a detector specifically designed to be insensitive to certain aspects of the TAIs. This would allow the monitor to take measurements without requiring an IB. This is a difficult objective, but could be accomplished by taking advantage of the non-sensitive weights on $\mathbf{g}$ that the CHO can produce. The following subsection is one possible implementation.

7.6.1 Building Non-Sensitive Weights into an Attenuation Plate

The host would start by utilizing the standard imaging detector, including the standard mask designed to optimally differentiate signals at different locations in object space. For a binary-discrimination task such as explosive dismantlement, the host could acquire calibration data on the TAIs and simulate objects that differ from those TAIs along predefined sensitive parameters. The host could then follow the CHO procedure to create a set of non-sensitive weights that are unable to differentiate objects that differ along a sensitive parameter. The ideal set of weights $\mathbf{W_g}$ would have values between zero and one, allowing the monitor to build an attenuation plate that can absorb that amount in front of each pixel. This procedure is outlined below.

The test statistic returned by the performing the CHO on measured objects is $t = \mathbf{W_g^\dagger g}$, where $\mathbf{W_g}$ is the weights on $\mathbf{g}$ that are non-sensitive to the penalized objects. If the count rates for the two objects are equal, the sum of the values in $\mathbf{g}$ will also be equal for a given acquisition time. Adding a constant to $\mathbf{W_g}$ results in a vector $\mathbf{W'_g}$ that should also return equal means for the two tested items. This procedure could be used to create a positive $\mathbf{W'_g}$. An alternative if the count rates are not equal for the various objects would be to design an optimization routine for $\mathbf{T}$ that enforces positive weights on $\mathbf{g}$.

From there, the values in $\mathbf{W'_g}$ could be linearly scaled down until the maximum value is one, a procedure that also yields equal means for the various test statistic distributions. These new weights are denoted by $\mathbf{W''_g}$. These weights can be imple-

mented physically by putting an attenuating material right up against the detector plane. Each individual pixel could have a different attenuator thickness. Pixels with a low $\mathbf{W}_{\mathbf{g}}''$ would have a higher attenuator thickness so that the number of detected counts decreases.

Ultimately, this idea is probably easier to implement in theory than in practice. Neutrons are hard to stop and require at least a couple centimeters of an attenuating material such as polyethylene. This thickness could cause neutrons directed at one pixel to be attenuated by the moderator for an adjacent pixel. Furthermore, when location is a nuisance parameter, the object may be located away from the center of the field of view, exaggerating this effect. The ideal attenuator would be as thin as possible.

This procedure is advantageous over the electronic board channelizing procedure for a number of reasons. The sensitive TAI parameters would fall into the "null space" of the imager itself—the monitor would never have the possibility of accessing this information as long as the attenuating material is in front of the detector. The monitor could use this detector with any of the discussed observer models (or even models that do not require LM processing) as the detector data is always non-sensitive.

There are pitfalls to this procedure, however. Similar to the discussion on the CHO in Section 7.4, while it is possible to set the means of the test-statistic distributions equal, it is a more difficult task to set the variances equal.

REFERENCES

(1946). Atomic Energy Act. Online: http://www.legisworks.org/congress/79/publaw-585.pdf.

(1991). *Public Papers of the Presidents of the United States, George Bush, 1991*. Federal Register.

Adams, J. and G. White (1978). A Versatile Pulse Shape Discriminator for Charged Particle Separation and its Application to Fast Neutron Time-of-Flight Spectroscopy. *Nuclear Instruments and Methods*, **156**(3), pp. 459–476.

Agostinelli, S., J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boundreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell'Acqua, G. Depaolo, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J. Gomez Cadenas, I. Gonzalez, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumpliner, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov, H. Kurashige, E. Lamanna, T. Lampen, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. More de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O'Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. Di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J. Wellisch, T. Wenaus, D. Williams, D. Wright, T. Yamada, H. Yoshida, and D. Zschiesche (2003). GEANT4—A Simulation Toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **506**(3), pp. 250–303.

Allison, J., K. Amako, J. Apostolakis, H. Araujo, P. Arce Dubois, M. Asai, G. Barrand, R. Capra, S. Chauvie, R. Chytracek, G. Cirrone, G. Cooperman, G. Cosmo, G. Cuttone, G. Daquino, M. Donszelmann, M. Dressel, G. Folger, F. Foppiano, J. Generowicz, V. Grichine, S. Guatelli, G. P, A. Heikkinen, I. Hrivnacova, A. Howard, S. Incerti, V. Ivanchenko, T. Johnson, F. Jones, T. Koi, R. Kokoulin, M. Kossov, H. Kurashige, V. Lara, S. Larsson, F. Lei, O. Link, F. Longo, M. Maire, A. Mantero, B. Mascialino, I. McLaren, P. Mendez Lorenzo, K. Minamimoto, K. Murakami, P. Nieminen, L. Pandola, S. Parlati, L. Peralta, J. Perl, A. Pfeiffer, M. Pia, A. Ribon, P. Rodrigues, G. Russo, S. Sadilov, G. Santin, T. Sasaki, D. Smith, N. Starkov, S. Tanaka, E. Tcherniaev, B. Tome, A. Trindade, P. Truscott, L. Urban, M. Verderi, A. Walkden, J. Wellisch, D. Williams, D. Wright,

and H. Yoshida (2006). GEANT4 Developments and Applications. *Nuclear Science, IEEE Transactions on*, **53**(1), pp. 270–278.

Araujo, A. and E. Giné (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*, volume 431. John Wiley and Sons Inc, New York.

Arce, P., M. Embid, and J. Ignacio Lagares (2007). Point Detector Scoring in GEANT4. GEANT4 Event Biasing and Scoring Mini-Workshop.

Archer, D. E., C. L. Britton Jr, R. J. Carter, R. F. Lind, J. T. Mihalczo, J. A. Mullens, J. E. Radle, and M. C. Wright (2010). Fieldable Nuclear Material Identification System. Technical Report ORNL/TM-2012/22, Oak Ridge National Laboratory.

Arms Control Association (2012). New START at a Glance. Online: https://www.armscontrol.org/factsheets/NewSTART.

Arms Control Association (2014a). The Intermediate-Range Nuclear Forces (INF) Treaty at a Glance. Online: https://www.armscontrol.org/factsheets/INFtreaty.

Arms Control Association (2014b). U.S. Russia Nuclear Arms Control Agreements at a Glance. Online: https://www.armscontrol.org/factsheets/USRussiaNuclearAgreementsMarch2010.

Arvo, J. (1992). Fast Random Rotation Matrices. In *Graphics Gems III*, pp. 117–120. Academic Press.

Asmussen, S. and P. W. Glynn (2007). *Stochastic Simulation: Algorithms and Analysis*, volume 57. Springer Science & Business Media.

Bai, J. and S. Shi (2011). Estimating high dimensional covariance matrices and its applications.

Barrett, H. H., C. K. Abbey, and E. Clarkson (1998). Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions. *JOSA A*, **15**(6), pp. 1520–1535.

Barrett, H. H. and K. J. Myers (2003). *Foundations of Image Science*, volume 1. John Wiley and Sons Inc.

Barrett, H. H., T. White, and L. C. Parra (1997). List-Mode Likelihood. *JOSA A*, **14**(11), pp. 2914–2923.

Barrett, H. H., J. Yao, J. P. Rolland, and K. J. Myers (1993). Model Observers for Assessment of Image Quality. *Proceedings of the National Academy of Sciences*, **90**(21), pp. 9758–9765.

Batchelor, R., R. Aves, and T. Skyrme (1955). Helium-3 Filled Proportional Counter for Neutron Spectroscopy. *Review of Scientific Instruments*, **26**(11), pp. 1037–1047.

BBC News (2015). North Korea's Nuclear Tests. Online: http://www.bbc.com/news/world-asia-17823706.

Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pp. 199–227.

Bird, H. G. (2015). Australian National Statement. The Ninth Review of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons.

Bodewig, E. (2014). *Matrix calculus*. Elsevier.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge university press.

Breiman, L. (2001). Random Forests. *Machine learning*, **45**(1), pp. 5–32.

Briesmeister, J. F. et al. (1986). *MCNP–A General Monte Carlo Code for Neutron and Photon Transport*. Los Alamos National Laboratory.

Brun, R. and F. Rademakers (1997). ROOT—An Object Oriented Data Analysis Framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **389**(1), pp. 81–86.

Bureau of International Security and Nonproliferation (1972). Interim Agreement Between The United States of America and The Union of Soviet Socialist Republics on Certain Measures With Respect to the Limitation of Strategic Offensive Arms. Online at: http://www.state.gov/t/isn/4795.htm.

Caucci, L. and H. H. Barrett (2012). Objective Assessment of Image Quality. V. Photon-Counting Detectors and List-Mode Data. *JOSA A*, **29**(6), pp. 1003–1016.

Cha, S. H. (2007). Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, **1**(4).

Clarkson, E. (2012). Asymptotic Ideal Observers and Surrogate Figures of Merit for Signal Detection with List-Mode Data. *JOSA A*, **29**(10), pp. 2204–2216.

Comprehensive Test Ban Treaty Organization (2012). Overview of the Verification Regime. Online: https://www.ctbto.org/verification-regime/background/overview-of-the-verification-regime/.

Compton, A. H. (1923). A Quantum Theory of the Scattering of X-Rays by Light Elements. *Physical review*, **21**(5), p. 483.

Cree, M. J. and P. J. Bones (1994). Towards Direct Reconstruction from a Gamma Camera Based on Compton Scattering. *Medical Imaging, IEEE Transactions on*, **13**(2), pp. 398–407.

CTBTO Preparatory Commission (2012). General Overview of the Effects of Nuclear Testing. Online: https://www.ctbto.org/nuclear-testing/the-effects-of-nuclear-testing/general-overview-of-theeffects-of-nuclear-testing/.

CTBTO Preparatory Commission and others (1996). Comprehensive Nuclear Test-Ban Treaty (CTBT). *Preparatory Commission for the CTBT Organization, Provisional Technical Secretariat, Vienna, Italy*.

De Maesschalck, R., D. Jouan-Rimbaud, and D. L. Massart (2000). The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, **50**(1), pp. 1–18.

Dreiseitl, S. and L. Ohno-Machado (2002). Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review. *Journal of biomedical informatics*, **35**(5), pp. 352–359.

Eijen Technology (2010). EJ-309. Online: http://www.eljentechnology.com/index.php/products/liquid-scintillators/73-ej-309.

Enqvist, A., M. Flaska, and S. Pozzi (2008). Measurement and Simulation of Neutron/Gamma-Ray Cross-Correlation Functions from Spontaneous Fission. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **595**(2), pp. 426–430.

Fechner, G. T., E. G. Boring, D. H. Howes, and H. E. Adler (1966). *Elements of Psychophysics. Translated by Helmut E. Adler, With an Introd. by Edwin G. Boring.* Holt, Rinehart and Winston.

Federation of American Scientists (1999). Strategic Arms Reduction Treaty (START1). Online: http://fas.org/nuke/control/start1/.

Federation of American Scientists (2015). LGM-30 Minuteman III. Online: `http://fas.org/nuke/guide/usa/icbm/lgm-30_3.htm`.

Fenimore, E. E. and T. Cannon (1978). Coded Aperture Imaging with Uniformly Redundant Arrays. *Applied optics*, **17**(3), pp. 337–347.

Fuller, J. (2010). Verification on the Road to Zero: Issues for Nuclear Warhead Dismantlement. *Arms Control Today*, **40**(10), pp. 19–27.

Gale, T. (2008). Cuban Missile Crisis. *International Encyclopedia of Social Sciences*.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium Auctore Carolo Friderico Gauss*. sumtibus Frid. Perthes et IH Besser.

Geelhood, B., B. Bartos, R. Comerford, D. Lee, J. Mullens, and J. Wolford (2000). Information Barrier Working Group Evaluation of TRADS with Particular Attention to Its Authentication Merits. Technical Report PNNL-13259, Pacific Northwest National Laboratory.

Gibson, J. (1997). History of the International Atomic Energy Agency. The First Forty Years. *Journal of Radiological Protection*, **18**(1).

Gilbert, A. J., B. W. Miller, and T. A. White (2016). Information Barriers for Imaging: The Single-Pixel Gamma Camera. In *Institute of Nuclear Materials Management (INMM) 57th Annual Meeting*.

Gilks, W. R. (2005). *Markov Chain Monte Carlo*. Wiley Online Library.

Gilman, L. (2004). *Bay of Pigs*. The Gale Group Inc.

Glaser, A., B. Barak, and R. J. Goldston (2014). A Zero-Knowledge Protocol for Nuclear Warhead Verification. *Nature*, **510**(7506), pp. 497–502.

Golub, G. H. and C. Reinsch (1970). Singular Value Decomposition and Least Squares Solutions. *Numerische mathematik*, **14**(5), pp. 403–420.

Good, I. (1986). Some Statistical Applications of Poisson's Work. *Statistical science*, pp. 157–170.

Greenwood, P. E. and M. S. Nikulin (1996). *A Guide to Chi-Squared Testing*, volume 280. John Wiley and Sons Inc.

Greivenkamp, J. E. (2004). *Field Guide to Geometrical Optics*, volume 1. SPIE Press Bellingham, Washington.

Groves, L. R. (1983). *Now It Can Be Told: The Story of the Manhattan Project*. Da Capo Press.

Hamamatsu Photonics K.K., E. T. D. (2007). Photomultiplier Tubes: Basics and Applications. Technical report, Hamamatsu Electronics K.K.

Hanley, J. A. and B. J. McNeil (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**(1), pp. 29–36.

Hardell, R., S. Idetjärn, and H. Ahlgren (1969). Thermal-Neutron Capture Gamma Rays from the 27 Al (n, γ) 28 Al Reaction. *Nuclear Physics A*, **126**(2), pp. 392–400.

Hauck, D. K., D. W. MacArthur, M. C. Browne, and R. F. Parker (2012). The Role of Portal Monitors in Arms Control and Development Needs. In *53rd Annual INMM Meeting*. Orlando, Florida.

Hausladen, P., M. A. Blackston, E. Brubaker, D. Chichester, P. Marleau, and R. J. Newby (2012). Fast-neutron coded-aperture imaging of special nuclear material configurations. In *53rd Annual Meeting of the INMM, Orlando, FL, USA*.

Haynes, J. E. and A. Vassiliev (2009). *Spies: The Rise and Fall of the KGB in America*. Yale University Press.

Hertz, H. (1887). Ueber einen Einfluss des ultravioletten Lichtes auf die electrische Entladung. *Annalen der Physik*, **267**(8), pp. 983–1000.

Hotelling, H. (1931). The Generalization of Student's Ratio. *Ann. Math. Statist.*, **2**(3), pp. 360–378. doi:10.1214/aoms/1177732979.

Hsieh, J. (2009). *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. Wile.

IAEA in Austria (ed.) (June). *IAEA Safeguards Glossary*. 3. IAEA.

Jha, A. K., H. H. Barrett, E. Clarkson, L. Caucci, and M. A. Kupinski (2013). Analytic methods for list-mode reconstruction. *Intl Meet Fully Three-Dim Image Recon Rad Nucl Med, California*.

Kennedy, J. F. (1963). The Strategy for Peace. Commencement Address at American University.

Klein, O. and Y. Nishina (1929). Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac. *Zeitschrift für Physik*, **52**(11-12), pp. 853–868.

Kristensen, H. M. (2015). US Drops Below New START Warhead Limit for First Time. Online: https://fas.org/blogs/security/2015/10/newstart2015-2/.

Kupinski, M. A. and E. Clarkson (2005). Extending the channelized hotelling observer to account for signal uncertainty and estimation tasks. In *Medical Imaging*, pp. 183–190. International Society for Optics and Photonics.

Kupinski, M. A., J. W. Hoppin, E. Clarkson, and H. H. Barrett (2003). Ideal-Observer Computation in Medical Imaging with Use of Markov-Chain Monte Carlo Techniques. *JOSA A*, **20**(3), pp. 430–438.

Lee, C.-J., M. A. Kupinski, and L. Volokh (2013). Assessment of cardiac single-photon emission computed tomography performance using a scanning linear observer. *Medical physics*, **40**(1), p. 011906.

Liaw, A. and M. Wiener (2002). Classification and Regression by randomForest. *R news*, **2**(3), pp. 18–22.

Loeber, C. R. (2005). *Building the Bombs: a History of the Nuclear Weapons Complex*. Sandia National Laboratories.

MacGahan, C. J., M. A. Kupinski, E. M. Brubaker, and N. R. Hilton (2015). A Channelized Hotelling Observer for Treaty Verification Tasks. In *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, p. 1.

MacGahan, C. J., M. A. Kupinski, E. M. Brubaker, N. R. Hilton, and P. Marleau (2016a). Development of a Nonsensitive Template for Arms-Control-Treaty-Verification Tasks. In *Symposium on Radiation Measurements and Applications*.

MacGahan, C. J., M. A. Kupinski, E. M. Brubaker, N. R. Hilton, and P. Marleau (2016b). Manuscript in Progress: A Channelized Hotelling Observer for Treaty Verification Tasks. *Nuclear Instruments and Methods in Physics Research Section A*.

MacGahan, C. J., M. A. Kupinski, E. M. Brubaker, N. R. Hilton, and P. Marleau (2016c). Nuclear Imaging for Treaty Verification without an Information Barrier. In *Institute of Nuclear Materials Management (INMM) 57th Annual Meeting*.

MacGahan, C. J., M. A. Kupinski, N. R. Hilton, E. M. Brubaker, and W. C. Johnson (2016d). Development of an Ideal Observer that Incorporates Nuisance Parameters and Processes List Mode Data. *JOSA A*, **33**(4), pp. 689–697.

MacGahan, C. J., M. A. Kupinski, N. R. Hilton, W. C. Johnson, and E. M. Brubaker (2014). Development of a list-mode ideal observer to perform classification tasks when imaging nuclear inspection objects under signal-known-exactly conditions. In *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pp. 1–5. doi:10.1109/NSSMIC.2014.7431051.

Mattingly, J. (2009). Polyethylene-Reflected Plutonium Metal Sphere: Subcritical Neutron and Gamma Measurements. *SAND2009-5804 Revision*, **2**.

Mitchell, D. and K. Tolk (2000). Trusted Radiation Attribution Demonstration. In *Proceedings of the 41st Annual INMM Meeting*. New Orleans, LA.

Mitchell, D. J. (1988). Gamma Detector Response and Analysis Software (GADRAS). Technical Report SAND88-2519, Sandia National Laboratories.

National Resources Defense Council (2002). Table of U.S. Nuclear Warheads.

National Nucear Data Center (2016). Evaluated Nuclear Structure Data File Search and Retrieval. Online: http://www.nndc.bnl.gov/ensdf/.

Neibert, R., J. Zabriskie, C. Knight, and J. L. Jones (2010). Passive and Active Radiation Measurements Capability at the INL Zero Power Physics Reactor (ZPPR) Facility. Technical Report INL/EXT-11-20876, Idaho National Laboratory (INL).

Norris, R. S. and H. M. Kristensen (2010). US Nuclear Forces, 2010. *Bulletin of the Atomic Scientists*, **66**(3), pp. 57–71.

Office of the Historian (1963). The Limited Test Ban Treaty. Online: https://history.state.gov/milestones/1961-1968/limited-ban.

OTA Project Staff (1990). Verification Technologies: Measures for Monitoring Compliance with the START Treaty. Technical Report OT A-ISC-479, Congress of the United States: Office of Technology Assessment.

Park, S., E. Clarkson, M. A. Kupinski, and H. H. Barrett (2005). Efficiency of the human observer detecting random signals in random backgrounds. *JOSA A*, **22**(1), pp. 3–16.

Pérez-Stable, M. (1999). *The Cuban Revolution: Origins, Course, and Legacy*. Oxford University Press.

Poitrasson-Rivière, A., B. A. Maestas, M. C. Hamel, S. D. Clarke, M. Flaska, S. A. Pozzi, G. Pausch, C.-M. Herbach, A. Gueorguiev, M. F. Ohmes, et al. (2015). Monte Carlo Investigation of a High-Efficiency, Two-Plane Compton Camera for Long-Range Localization of Radioactive Materials. *Progress in Nuclear Energy*, **81**, pp. 127–133.

Quiter, B., B. Ludewight, V. Mozin, and S. Tobin (2010). Nondestructive Spent Fuel Assay Using Nuclear Resonance Fluorescence. Online at: `http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/41/097/41097511.pdf`.

Reilly, D., N. Ensslin, H. Smith Jr, and S. Kreiner (1991). Passive Nondestructive Assay of Nuclear Materials. Technical report, Nuclear Regulatory Commission, Washington, DC (United States). Office of Nuclear Regulatory Research; Los Alamos National Lab., NM (United States).

Safavian, S. R. and D. Landgrebe (1991). A Survey of Decision Tree Classifier Methodology. *IEEE transactions on systems, man, and cybernetics*, **21**(3), pp. 660–674.

Schölkopf, B. and A. Smola (1998). Support Vector Machines. *Encyclopedia of Biostatistics*.

Schwartz, S. I. (2011). *Atomic Audit: The Costs and Consequences of US Nuclear Weapons Since 1940*. Brookings Institution Press.

Seager, K., D. Mitchell, T. Laub, K. Tolk, R. Lucero, and K. Insch (2001). Trusted Radiation Identification System. In *Proceedings of the 42nd Annual INMM Meeting*. Indian Wells, CA.

Shepp, L. A. and Y. Vardi (1982). Maximum Likelihood Reconstruction for Emission Tomography. *Medical Imaging, IEEE Transactions on*, **1**(2), pp. 113–122.

Shultis, J. and R. Faw (2011). *An MCNP Primer*. Kansas State University.

Subudhi, S. (2013). Hypothesis Testing - CFA Level - I. Online: http://www.simplilearn.com/cfa-level-1-hypothesis-testing-article.

Sun, T. and Y. Neuvo (1994). Detail-preserving median based filters in image processing. *Pattern Recognition Letters*, **15**(4), pp. 341–347.

Tatsumi, K. (2012). A General Assistant Tool for the Checking Results from Monte Carlo Simulations. Powerpoint presentation.

The APS Panel of Public Affairs (2013). A Technical Review: The Domestic Nuclear Detection Office Transformational and Applied Research Directorate R and D Program.

Twomey, T. (2003). The Best Choice of High Purity Germanium (HPGe) Detector. Technical report, ORTEC.

United States and Union of Soviet Socialist Republics (1963). Treaty Banning Nuclear Weapon Tests in the Atmosphere, in Outer Space and Under Water. Online at: http://www.state.gov/t/isn/4797.htm.

United States Department of State (2009). United States Relations with Russia: The Cold War. Online: http://2001-2009.state.gov/r/pa/ho/pubs/fs/85895.htm.

United States Nuclear Regulatory Comission (2012). Fact Sheet on Dirty Bombs. Online: http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/fs-dirty-bombs.html.

United Nations Office of Disarmament Affairs (1968). Treaty on the Non-Proliferation of Nucear Weapons. Online: http://www.un.org/disarmament/WMD/Nuclear/NPTtext.shtml.

United States and Union of Soviet Socialist Republics (1974). Treaty Between The United States of America and The Union of Soviet Socialist Republics on the Limitation of Underground Nuclear Weapon Tests. Online: http://www.state.gov/t/isn/5204.htm.

United States and Union of Soviet Socialist Republics (1979). Strategic Arms Limitation Treaty II Treaty. Online at: http://www.atomicarchive.com/Treaties/Treaty13.shtml.

United States and Union of Soviet Socialist Republics (2003). Strategic Offensive Reduction Treaty. Online: http://www.nti.org/treaties-and-regimes/strategic-offensive-reductions-treaty-sort/.

U.S. Department of State (2011). New START. Online: http://www.state.gov/t/avc/newstart/index.htm.

Watt, B. (1952). Energy Spectrum of Neutrons from Thermal Fission of U 235. *Physical Review*, **87**(6), p. 1037.

Whitaker, M. K., E. Clarkson, and H. H. Barrett (2008). Estimating Random Signal Parameters from Noisy Images with Nuisance Parameters: Linear and Scanning-Linear Methods. *Optics express*, **16**(11), pp. 8150–8173.

Williams, R. C. (1989). *Klaus Fuchs, Atom Spy*. Harvard University Press.

Willman, C., A. Håkansson, O. Osifo, A. Bäcklin, and S. J. Svärd (2006). Nondestructive Assay of Spent Nuclear Fuel with Gamma-Ray Spectroscopy. *Annals of Nuclear Energy*, **33**(5), pp. 427–438.

Wollenweber, S., B. Tsui, D. Lalush, E. Frey, K. LaCroix, and G. Gullberg (1999). Comparison of Hotelling Observer Models and Human Observers in Defect Detection from Myocardial SPECT Imaging. *Nuclear Science, IEEE Transactions on*, **46**(6), pp. 2098–2103.

Woodbury, M. A. (1950). Inverting Modified Matrices. *Memorandum report*, **42**, p. 106.

Wright, D. (2015). Physics Simulation Packages. Online: http://nuclear.llnl.gov/simulation/main.html.

Yamashita, M., L. D. Stephens, and H. W. Patterson (1966). Cosmic-Ray-Produced Neutrons at Ground Level: Neutron Production Rate and Flux Distribution. *Journal of Geophysical Research*, **71**(16), pp. 3817–3834.

Yao, J. and H. H. Barrett (1992). Predicting Human Performance by a Channelized Hotelling Observer Model. In *San Diego'92*, pp. 161–168. International Society for Optics and Photonics.

Ye, J., X. Song, Z. Zhao, A. J. Da Silva, J. S. Wiener, and L. Shao (2006). Iterative SPECT reconstruction using matched filtering for improved image quality. In *Nuclear Science Symposium Conference Record, 2006. IEEE*, volume 4, pp. 2285–2287. IEEE.

Young, R. (1960). Ike Slashes Sugar Quotas.

Zeng, G. L. (2012). A Filtered Backprojection MAP Algorithm with Nonuniform Sampling and Noise Modeling. *Medical physics*, **39**(4), pp. 2170–2178.

Ziock, K., J. Collins, L. Fabris, S. Gallagher, B. Horn, R. Lanza, and N. Madden (2006). Source-Search Sensitivity of a Large-Area, Coded-Aperture, Gamma-Ray Imager. *Nuclear Science, IEEE Transactions on*, **53**(3), pp. 1614–1621.

Zuhoski, P. B., J. P. Indusi, and P. E. Vanier (1999). Building a Dedicated Information Barrier System for Warhead and Sensitive Item Verification. Technical Report BNL-66214, Brookhaven National Lab.