Discrimination algorithms for the remediation of unexploded ordnance

by

Laurens Sander Beran

B.Sc., University of Victoria, 2001
M.Sc., University of British Columbia, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Geophysics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)
February 2010
© Laurens Sander Beran 2010
Abstract

This thesis considers analysis of magnetic and electromagnetic data for the purpose of discriminating between buried unexploded ordnance (UXO) and non-hazardous metallic clutter. Magnetic data acquired over a ferrous object are modelled as a magnetostatic dipole. For time-domain electromagnetic data, a linear combination of decaying, orthogonal dipoles represents the secondary magnetic field radiated by a conductor.

Model parameters estimated with inversion can be input into a discrimination algorithm whose output is a ranked diglist of targets. Algorithms that have been applied to UXO discrimination can be broadly categorized as library or statistical methods. Library methods assume that for each ordnance type there is a single set of parameters which is representative of intrinsic target properties. Statistical methods try to formulate a decision rule with a training set of models for targets with known ground truth.

Observed data can sometimes have non-normally distributed noise and consequently parameter estimates obtained via least squares inversion may be biased. Robust misfit functions provide improved estimates of model parameters when there are outliers in the data. I also investigate propagation of uncertainties from data to model parameters. For inversion of electromagnetic data I find that parameters derived from the dipole model are approximately normally distributed. However, when data coverage is poor or SNR is low, the posterior distribution of these parameters may be multimodal. I develop a statistical classification algorithm that incorporates parameter uncertainty by integrating over the posterior probability distribution. Simulations and applications to real data indicate that this technique can detect outliers to the distribution of ordnance sooner than conventional classifiers.

I quantify the performance of a discrimination algorithm using metrics derived from the receiver operating characteristic (ROC) curve. I use a bootstrapping algorithm to estimate mean values of performance metrics from limited training data and to identify the discrimination algorithm that is best suited to a particular problem. Finally, I consider the problem of determining the stop digging (or operating) point on the ROC. I derive an approximate probability distribution for the point on the ROC at which all ordnance are found and develop methods for estimating this point.
# Table of Contents

Abstract ............................................................................................................. ii

Table of Contents ............................................................................................... iii

List of Tables ....................................................................................................... v

List of Figures ...................................................................................................... vi

Acknowledgements .............................................................................................. viii

Dedication ............................................................................................................. ix

Co-authorship Statement ...................................................................................... x

1 Introduction ....................................................................................................... 1
   1.1 Research motivation .................................................................................. 1
   1.2 UXO detection and discrimination: the state of the art ......................... 2
      1.2.1 UXO detection .................................................................................. 2
      1.2.2 UXO discrimination ......................................................................... 7
   1.3 Case study: Camp Lejeune ....................................................................... 16
   1.4 Outline of the thesis ................................................................................ 19

Bibliography ......................................................................................................... 26

2 Robust inversion for UXO discrimination ........................................................ 29
   2.1 Introduction ............................................................................................... 29
   2.2 Forward modelling .................................................................................... 30
   2.3 Positional errors ....................................................................................... 31
   2.4 Robust norms ............................................................................................ 33
   2.5 Simulations ............................................................................................... 36
   2.6 Application to real data ........................................................................... 39
   2.7 Conclusions ............................................................................................... 43

Bibliography ......................................................................................................... 47

3 Uncertainty in UXO discrimination ................................................................. 48
   3.1 Introduction ............................................................................................... 48
   3.2 Model uncertainty ..................................................................................... 51
   3.3 Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters .................................................................................................................. 53
   3.4 Incorporating uncertainties in discrimination: local uncertainties ........... 61
      3.4.1 Simulations ........................................................................................ 66
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.2</td>
<td>Application to real data</td>
<td>67</td>
</tr>
<tr>
<td>3.5</td>
<td>Incorporating uncertainties in discrimination: global uncertainties</td>
<td>70</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusions</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>Selecting a discrimination algorithm</td>
<td>78</td>
</tr>
<tr>
<td>4.1</td>
<td>Advanced discrimination: a review</td>
<td>79</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Statistical classification</td>
<td>80</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Library methods</td>
<td>83</td>
</tr>
<tr>
<td>4.2</td>
<td>Selecting a discrimination strategy</td>
<td>84</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Measuring discrimination performance</td>
<td>85</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Bootstrap estimation of discrimination performance</td>
<td>87</td>
</tr>
<tr>
<td>4.3</td>
<td>Application to electromagnetic and magnetic data sets</td>
<td>90</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>97</td>
</tr>
<tr>
<td>4.5</td>
<td>Conclusion</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>101</td>
</tr>
<tr>
<td>5</td>
<td>Selecting a receiver operating point</td>
<td>104</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>104</td>
</tr>
<tr>
<td>5.2</td>
<td>Optimal operating points</td>
<td>105</td>
</tr>
<tr>
<td>5.3</td>
<td>Optimal operating points for samples</td>
<td>109</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Simulations</td>
<td>116</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Application to real data</td>
<td>116</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>121</td>
</tr>
<tr>
<td>6</td>
<td>Conclusions</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>Appendices</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>A simplified fingerprinting method</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>134</td>
</tr>
</tbody>
</table>
List of Tables

1.1 EM sensors used for UXO discrimination ........................................ 6
2.1 Example norms used for robust inversion ......................................... 35
2.2 Comparison of performance metrics for robust vs. L2 inversion .......... 43
3.1 Comparison of QDA and GP classifiers ............................................. 67
3.2 Performance of classifiers on EM61 test data .................................. 72
4.1 Bootstrap estimates of AUC and FAR .............................................. 93
4.2 AUC and FAR for algorithms applied to Guthrie road data ............... 94
5.1 Example costs for risk minimization .............................................. 108
5.2 Performance of operating point criteria for simulations ................. 117
# List of Figures

1.1 The UXO remediation process ........................................... 3
1.2 Surface clutter at Camp Lejeune ...................................... 3
1.3 Data acquisition ......................................................... 4
1.4 Flow chart for advanced discrimination of UXO. .................. 7
1.5 EM63 data with overlapping anomalies from two targets .......... 7
1.6 Frequency and time-domain EM responses of a sphere .............. 9
1.7 Dependence of frequency and time-domain responses upon physical properties 10
1.8 Camp Lejeune adapter .................................................. 16
1.9 Failed fit report for EM63 target, Camp Lejeune .................... 21
1.10 Passed fit report for EM63 target, Camp Lejeune ................... 22
1.11 Model parameters estimated from EM63 data collected at Camp Lejeune 23
1.12 Canonical analysis of Camp Lejeune EM63 data .................... 24
1.13 Two-dimensional QDA decision surface .............................. 24
1.14 Bootstrapped classifier performance, Camp Lejeune EM63 data ... 25
2.1 Moments of data residuals .............................................. 32
2.2 Distributions of relative residuals .................................... 33
2.3 Robust norms .......................................................... 35
2.4 Fit to synthetic data with independent errors ...................... 36
2.5 Simulated parameter estimates with independent errors .......... 37
2.6 Sensitivity curve for estimation of mean of a sample ............. 38
2.7 Simulated parameter estimates with correlated errors .......... 40
2.8 Fit to synthetic data with correlated errors ....................... 41
2.9 Camp Sibert target classes ........................................... 41
2.10 Camp Sibert EM63 features ........................................... 44
2.11 Comparison of L2 and robust fits for a single target ............ 45
2.12 Camp Sibert MTADS features ....................................... 46
3.1 Fit of a dipole model to TEM data acquired over an ordnance ... 49
3.2 Estimated model parameters for ordnance and non-ordnance targets 50
3.4 Principle axes of the covariance estimated by linearized and nonlinear appraisals 58
3.5 Accepted models for nonlinear appraisal of a single target, EM61 58
3.6 Misfit versus depth curve for EM61 target .......................... 59
3.7 Camp Sibert EM63 features ........................................... 59
3.8 Accepted models for nonlinear appraisal of a single target, EM63 60
3.9 Misfit versus depth curve for EM63 target .......................... 60
3.10 Estimating normal distributions in the presence of local uncertainty 62
3.11 Dependence of Gaussian product on feature vector variance .... 64
3.12 Classification with the Gaussian product in one dimension .... 65
3.13 Classification with the Gaussian product in two dimensions .... 66
List of Figures

3.14 Simulating discrimination under uncertainty ........................................ 68
3.15 Application of GP and QDA to EM61 data ........................................... 70
3.16 Application of GP and QDA to EM63 data ........................................... 70
3.17 Estimated data uncertainties and residuals from inversion .................... 71
3.18 Comparison of discrimination with EM61 data using multiple or single feature vectors. 74
3.19 Discrimination with EM63 data using multiple feature vectors. ............... 75

4.1 Feature estimation for magnetic data ....................................................... 80
4.2 Support vector machine formulation ...................................................... 82
4.3 Estimation of false alarm rate (FAR) and area under curve (AUC) for finite samples. 88
4.4 Training data for 20 mm Range Fan discrimination ................................ 92
4.5 Sounding at anomaly maximum for EM-63 anomalies .......................... 93
4.6 Application of statistical classifiers to Guthrie road magnetics data .......... 95
4.7 Selection of a discrimination algorithm using bootstrap estimates of the FAR and AUC. 100

5.1 Generating the ROC ................................................................................. 106
5.2 Methods for selecting the operating point on an ROC curve .................. 109
5.3 Effect of sample size on ROC and false alarm rate ............................... 110
5.4 Steps in computing the probability \(p(T|x(i))\) ....................................... 112
5.5 Computing the probability \(P(T|x(i))\) ....................................................... 113
5.6 Computing the distribution of the operating point ............................... 113
5.7 Dependence of the optimal operating point on sample size and prior probability .... 114
5.8 Probability of observing \(N_L F\) instances sequentially ............................ 115
5.9 Operating points applied to Guthrie road data ..................................... 118
5.10 Operating points applied to San Luis Obispo data ............................... 119

A.1 EM63 polarization coefficients ............................................................... 132
A.2 Library polarizations and fingerprinting fits ........................................ 133
Acknowledgements

I am grateful to my supervisor, Dr. Doug Oldenburg, for his motivation, encouragement and support. His positivity and enthusiasm are a continued inspiration. Thank you as well to the members of my committee, Dr. Stephen Billings and Dr. Mark Jellinek, for helpful feedback and guidance.

This work has only been possible with the help of Sky Research, which has given me opportunities to work on real projects. Thank you to everyone in the Vancouver office for their collaboration, assistance and careful selection of lottery numbers. Thank you also to members of the UBC-GIF group over the years.

Data from Guthrie road was provided by Cliff Yeomans of the Montana National Guard. This work was funded by an NSERC postgraduate scholarship and by SERDP-ESTCP.
Dedication

To my parents.
Co-authorship Statement

Chapters 2 through 5 of this thesis are versions of research articles accepted or in preparation to be submitted for publication in peer-reviewed journals. As my supervisor, Doug Oldenburg oversaw the research and provided advice for all chapters. I was principally responsible for manuscript preparation, although all co-authors had editorial roles for the chapters to which they are attached.

Chapter 2: Robust inversion for UXO discrimination

Authors: Laurens Beran, Stephen Billings and Doug Oldenburg

This work was carried out as part of SERDP project MM-1629: Robust Statistics and Regularization for Feature Extraction and UXO Discrimination. Dr. Stephen Billings is principal investigator and performed initial investigations with robust inversions for the project proposal. Dr. Lin Ping Song of UBC-GIF also worked on this project and pointed out that the non-normal distributions of the data residuals could be fit by a GEV distribution. I implemented robust inversion by extending existing forward modelling and inversion codes in the UXOLab software package. All simulations and applications to Camp Sibert data sets were done by me. Initial processing (e.g. masking) of real data sets was done by colleagues at Sky Research. I wrote the manuscript, with editorial input from my co-authors.

Chapter 3: Uncertainty in UXO discrimination

Authors: Laurens Beran and Doug Oldenburg.

This work in this chapter was also carried out under SERDP MM-1629. Again, initial processing of Camp Sibert data sets was done by Sky Research. Code to estimate the misfit versus depth curve was developed by Dr. Len Pasion and Dr. Nicolas Lhomme of Sky Research. I implemented the Metropolis-Hastings sampling algorithm, numerical calculation of linearized uncertainties, and the statistical classifiers used in this chapter. I wrote the manuscript, with editorial input from my co-author.

Chapter 4: Selecting a discrimination algorithm for unexploded ordnance remediation

Authors: Laurens Beran and Doug Oldenburg.

I implemented the bootstrapping approach and all statistical classifiers in this chapter. The Guthrie road data set (courtesy of the Montana National Guard) was inverted by Dr. Stephen Billings and the Former Lowry Bombing and Gunnery Range data set was inverted by Sky Research. I wrote the manuscript, with editorial input from my co-author.
Chapter 5: Selecting a receiver operating point

Authors: Laurens Beran and Doug Oldenburg.

This work was motivated by results of the Camp Sibert demonstration and further investigation was suggested by Dr. Herb Nelson of SERDP-ESTCP and by Dr. Stephen Billings. All of the theoretical and computational work was done by me. I wrote the manuscript, with editorial input from my co-author.
Chapter 1

Introduction

1.1 Research motivation

As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don’t know
We don’t know.

-Donald Rumsfeld

Applied geophysicists make inferences about the Earth’s subsurface using uncertain and limited data. In this process our “known knowns” are the data acquired with a geophysical experiment and a model of the physics which underlies the experiment. The mathematical machinery of inverse theory allows us to generate models which fit the observed data and so to draw objective conclusions regarding the scientific questions at hand. However, we must always contend with “known unknowns” which create discrepancies between our observations and predictions and thereby introduce uncertainty in our model and in the inferences we make with this model. Such discrepancies may be stochastic noise arising from myriad sources during data acquisition and processing. Errors may also have a deterministic component caused by approximations in the physical model.

This thesis is concerned with handling uncertainty in a specific application of near-surface geophysics: the detection and discrimination of unexploded ordnance (UXO). The primary scientific question is whether a buried object that we have detected with a geophysical survey is a hazardous ordnance. In answering this basic question, I address a number of secondary questions related to uncertainty:

• How do particular stochastic error sources translate to errors in the geophysical data? How can we account for these errors in an inversion to ensure that our estimated model is useful?
1.2 UXO detection and discrimination: the state of the art

Unexploded ordnances (UXOs) are defined as “explosive, propellant or chemical containing munitions that were armed, fired and remain unexploded through malfunction” (Delaney and Etter, 2003). The extent of UXO contamination within the United States and abroad has motivated research into improved technologies for detection and discrimination of UXO. This work is part of a broader effort to remediate “munitions and explosives of concern” (MEC). This encompasses UXO as well as discarded military munitions (often disposed by dumping at sea) and potentially toxic munitions constituents (Grone, 2003). Advanced discrimination methods are expected to greatly reduce false alarm rates, improving both the safety and cost of remediation efforts of UXO and other MEC (figure 1.1). These methods require the acquisition of digital geophysical data and subsequent signal processing.

1.2.1 UXO detection

Prior to detailed geophysical surveys a wide area assessment of the contaminated site must often be carried out to identify areas which are of interest (e.g. former bombing targets). At military sites this can sometimes be accomplished with historical records of training activities. However, these records are not always accurate, and so wide-area assessment with aerial photography or, increasingly, with LIDAR
1.2. UXO detection and discrimination: the state of the art

Figure 1.1: The UXO remediation process (after Delaney and Etter (2003)).

Figure 1.1 depicts two paradigms for acquisition of geophysical data. The first, “mag-and-flag”, uses metal detectors operated by expert technicians to identify targets which are then flagged for subsequent digging. No data are recorded and detection is usually indicated by changes in an audio tone. This method is inefficient because it depends upon the skill of the operator and offers limited possibility for discrimination between hazardous ordnance and clutter. Consequently, the projected cost for clearance of contaminated sites within the United States using this standard approach is prohibitively high. However, there will always be a role for analog detectors in UXO clearance. Many digital geophysical sensors are quite cumbersome and so man-portable analog sensors may be preferable in
areas where there is difficult topography or vegetation. Secondly, metal detectors are required when ordnance technicians return to relocate and excavate targets.

The second mode of UXO detection uses geophysical sensors connected to a data acquisition system to record digitized data acquired over a survey grid. These data are subsequently processed to identify high priority targets which are likely to be buried ordnance. As indicated in figure 1.1, this mode of operation is expected to greatly reduce the number of false alarms (non-UXO) which must be excavated.

Magnetic and electromagnetic data are the most common geophysical data types which are acquired for UXO detection and discrimination. Most magnetic instruments measure the total magnitude of the geomagnetic field, which may be locally perturbed by anomalous fields produced by magnetically susceptible materials (e.g. steel). Magnetic sensor arrays have been deployed for helicopter-borne surveys in wide area assessments. Multiple magnetometers can also be arranged in arrays for ground-based surveying, with the increased swath decreasing the number of passes required to cover a grid (figure 1.3(a)). Sensor position is usually measured with differential GPS, with an accuracy of 5-10 cm.

Alternatively, a robotic total station (RTS), which uses a laser to track the position of a reflector attached to the sensor array, can provide position accuracies of less than a centimeter. Ground based magnetic data are typically acquired with an along-line spacing of approximately 5 cm and a between-line spacing of 25-50 cm.

Electromagnetic instruments actively transmit a time-varying primary magnetic field which illuminates the earth. The variation of the primary field induces currents in the ground and these currents in turn produce a secondary field which can be measured by a receiver at the surface. EM sensors operating in the time domain measure the decay of secondary fields after a primary field is switched off. These secondary fields can provide information regarding the distribution of electrical conductivity and magnetic permeability in the ground.

Electromagnetic sensors come in a wide variety of geometries ranging from cart systems with multiple transmitters and receivers to single loop, man-portable systems. A ubiquitous time-domain instrument in use today is the Geonics EM61, which transmits from a single horizontal coil and receives
1.2. UXO detection and discrimination: the state of the art

with one or more horizontal loops at four time channels. This instrument is robust and easy to use and so is popular for UXO and other environmental applications. However, the range of time-channels is relatively short and the lack of late-time information limits the discrimination capability of this instrument. A more advanced horizontal loop system is the Geonics EM63 (figure 1.3(b)), which has a horizontal transmitter coil and multiple receiver coils which measure the secondary field at 26 time channels extending much later in time than the EM61. Table 1.1 summarizes the electromagnetic sensors which have been applied to UXO detection problems, either in production or research settings. This is not a comprehensive list of electromagnetic sensors, but is intended to illustrate the evolution of sensors from monostatic configurations (concentric transmitters and receivers) with a few channels towards multistatic (multiple transmitters and receivers) arrays with many channels over a wide time range.

Accurate positioning is crucial to the viability of discrimination with electromagnetic data. This is in part because the models used to fit EM data (section 2.2) are more complicated than those required for magnetic data, and data quality controls must therefore be much more stringent to ensure reliable EM parameter estimates. Consequently, for EM data there has been recent emphasis on accurate positioning with RTS systems as well as measurement of sensor orientation with inertial motion units.

Two types of surveys are common with electromagnetic instruments. A discrimination mode survey passes the sensor over an area in parallel lines spaced approximately 50 cm apart. Sometimes perpendicular lines are also acquired to maximize data coverage over targets and ensure that they are illuminated from multiple directions. Along line data spacing is on the order of 10 cm. Towed arrays of EM sensors can quickly cover large areas, while single sensor pushcart systems are much slower. Pushcart, or man-portable, EM systems are therefore better suited to the cued-interrogation mode of surveying (figure 1.3(b)). In this mode, we revisit targets which have been identified in a discrimination mode survey and acquire finely spaced, high SNR data in a small area (e.g. 5 x 5 m) about the target. Some recently developed multistatic systems for cued-interrogation illuminate the target with multiple transmitters and receivers from a single observation location and thereby avoid the requirement for accurate positioning of moving sensors. Smith and Morrison (2005) have shown that multistatic systems operating in a cued interrogation mode produce smaller uncertainties in model parameter estimates relative to simpler systems operating in a discrimination mode survey.

Finally, we mention ground penetrating radar (GPR) as a possible sensor technology for UXO detection. GPR operates at much higher frequencies than TEM systems, such that the electromagnetic pulse propagates as a wave into the earth and is scattered when it encounters discontinuities in electromagnetic properties, particularly changes in dielectric permittivity. While the technique can provide a high resolution image of scatterers in the earth, its sensitivity to variations in dielectric permittivity makes it prone to false alarms and so it is often deployed with another sensor (commonly EM) to facilitate joint interpretation. Experimental applications of GPR to UXO detection are described in O’Neill (2001) and Chen et al. (2001), the technology is now commonly applied to the related problem of landmine detection (e.g. Zhu and Collins (2005)).
## 1.2. UXO detection and discrimination: the state of the art

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Channels</th>
<th>Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM61 MKII</td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td>GEM3</td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td>EM63</td>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td>MTADS</td>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
</tr>
<tr>
<td>TEMTADS</td>
<td><img src="image9.png" alt="Graph" /></td>
<td><img src="image10.png" alt="Graph" /></td>
</tr>
<tr>
<td>BUD</td>
<td><img src="image11.png" alt="Graph" /></td>
<td><img src="image12.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

Table 1.1: Electromagnetic sensors which have been used for UXO discrimination. In right hand column dashed lines are transmitters, solid lines are receivers.
1.2. UXO detection and discrimination: the state of the art

1.2.2 UXO discrimination

Given digital geophysical data acquired with a sensor, a number of processing steps are required to produce an ordered list of targets for excavation. Figure 1.4 shows the typical processing involved in advanced discrimination. Some data pre-processing is usually required (merging position and sensor data, filtering, levelling, etc.), but will not be discussed in detail here. In a discrimination mode survey, the first task following data pre-processing is to identify targets of interest in the observed data. Target picking in magnetic data is complicated by the dipolar nature of anomalies, which can make manual picking somewhat complicated because the source location does not generally correspond to minima or maxima of an anomaly. An automatic target picker for magnetics data which follows target maxima and minima as they are upward continued is developed in Billings and Herrmann (2003).

Target picking with electromagnetic data identifies local maxima in an interpolated plan-view image of the first time channel. If targets are isolated and the SNR is high, the procedure is straightforward and can be automated. However, if anomalies are closely spaced or of low amplitude, target picking becomes more challenging and often requires careful judgements by the data analyst (e.g. figure 1.5).

Forward modelling

Once target anomalies have been identified in the observed geophysical data (magnetic or electromagnetic), we can characterize each anomaly by estimating features which will subsequently allow a discrimination algorithm to discern UXO from non-hazardous clutter. These features may be directly related to the observed data (e.g. anomaly amplitude at the first time channel), or they may be the...
parameters of a physical model. The former approach is appealing in its simplicity but may not always work. For example, an ordnance at depth will produce a small anomaly amplitude and might be left in the ground with a diglist generated based solely upon anomaly amplitude. Most discrimination strategies therefore use physical modelling to resolve such ambiguities, as described below.

In magnetics, the anomalous field produced by an buried ferrous ordnance can be approximately represented as a magnetostatic dipole (Billings, 2004). The magnetic field is

$$
B(r,m) = \frac{\mu_0}{4\pi r^3} \left[ \frac{3(m \cdot r)r}{r^2} - m \right]
$$

(1.1)

with $r$ and $m$ the source-observation separation and dipole moment vector, respectively. The constant $\mu_0$ is the permeability of free space. The data observed by a total field magnetometer is approximated as the projection of the dipole field onto the direction of the earth’s field at that location. More general forward models compute the fields produced by a susceptible object by solving a discretized form of the magnetostatic governing equations (e.g. Sanchez et al. (2008)). This approach can capture higher order (quadrupole, octupole) effects which are neglected in the dipole forward model, but is perhaps too computationally intensive for application to the large numbers of targets encountered in UXO data processing.

The electromagnetic scattering problem can be solved analytically for simple shapes. We first consider the frequency-domain spectrum and the corresponding time-domain response for a conductive, permeable sphere in a uniform field (figure 1.6). These forward modellings were computed for the vertical (z) component of the secondary field using the analytic solution derived in Wait and Spies (1969). Notable features of the frequency-domain response are the low and high frequency asymptotes of the in-phase response. At low frequencies, the secondary field is entirely in-phase with the primary field and tends to a magnetostatic induced dipole as frequency decreases. The secondary fields are produced by volume currents which circulate within the body and produce a uniform magnetization. In contrast, at the high frequency end of the spectrum the induced currents circulate only on the surface of the body. By Lenz’s law, currents are initially induced on the surface of the target after the shut-off of the primary field. These currents circulate so as to oppose any change in the primary field. The corresponding time-domain response is characterized by two stages in log-log space: an early time linear decay followed by late time exponential decay. The early time portion of the time-domain response corresponds to the diffusion of currents into the body. In the late time stage, the secondary fields decay exponentially.

Figure 1.7 shows the dependence of frequency and time-domain responses upon physical properties. An increase in relative permeability increases the strength of the induced dipole at low frequencies and so the magnitude of the in-phase component grows. The peak of the quadrature component shifts to higher frequencies with increasing permeability. In the time domain, the effect of increasing permeability is best understood in terms of the time constant

$$
\tau \propto \sigma \mu R^2.
$$

(1.2)
The time constant is the time at which late-stage exponential decay begins. As shown in figure 1.7, increasing either the conductivity, permeability, or radius extends the early time stage. In the case of increased conductivity this effect is easy to understand: surface currents, which produce the early time stage, persist longer in a more conductive medium.

In the frequency domain a change in the conductivity or radius for a fixed value of permeability simply shifts the response with respect to frequency. If the frequency-domain response is plotted as a function of the induction number

$$\xi = \sqrt{\omega \sigma \mu_o R} \quad (1.3)$$

then, for a given relative permeability, a single set of in-phase and quadrature curves are required to define the response for all possible values of $\sigma$ and $R$.

Although considerable insight can be gained by analytic forward modelling for simple shapes, extension of electromagnetic forward modelling to more complex bodies has proven to be a difficult endeavour. Finite difference solutions to Maxwell’s equations in both the time and frequency domain are available, but have difficulty modelling large conductivity contrasts between an ordnance target and surrounding soil. Accordingly, a family of frequency domain models have been developed which consider the scattered secondary field to be radiated in free space by a confined conductor. Analytic solutions for the secondary magnetic field radiated by an ellipsoid have been derived (Grzegorczyk et al., 2008). The numerical method of auxiliary sources (MAS) can be used to compute the response of an arbitrary body of revolution and has been used to investigate the effect of target material inhomogeneity, sharp edges, and wall thickness upon the frequency spectrum (Shubitidze et al., 2004). Wall thickness was found to have a marked effect on the peak of the quadrature response, corresponding to the dominant decay mode in the time domain. In Chilaka et al. (2006), experiments verified that low frequency (late time) measurements provided discrimination capability between visually identical targets with differing

![Figure 1.6: Frequency domain (top) and time-domain (bottom) electromagnetic response of a conductive, permeable sphere.](image-url)
1.2. UXO detection and discrimination: the state of the art

Figure 1.7: Dependence of frequency and time-domain responses upon physical properties.

wall thickness.

Models which are directly parameterized in terms of target material properties and geometry have provided considerable insight into the dependence of observed data upon these parameters. However, for scenarios where hundreds or thousands of targets must be modelled, simpler and faster models are typically used. For example, the spheroidal excitation approach (SEA) represents the secondary field as a superposition of fields radiated by magnetic charge rings arranged around the axis of symmetry
1.2. UXO detection and discrimination: the state of the art

of a spheroid (Sun et al., 2005). An important and popular model for UXO applications is the tri-axial dipole model (see e.g. Zhang et al. (2003), Norton et al. (2005), Pasion (2007)) which further simplifies the response of a UXO as the superposition of three orthogonal dipoles: one oriented along the long axis of the object and the other two transverse to the long axis. The spatial response of the secondary field $B(r, t)$ is given by equation 1.1, with a temporally- and spatially-varying magnetic dipole moment

$$m(r, t) = \frac{1}{\mu_0} M(t) \cdot B_\alpha(r)$$

and with $B_\alpha(r)$ the primary field at the object center. Neglecting higher order moments, the polarization tensor $M(t)$ can be assumed symmetric positive definite and so can be decomposed as

$$M(t) = A^T L(t) A.$$ 

The eigenvector matrix $A$ is an orthonormal matrix which rotates from geographic to body-centered coordinates, and $L(t)$ is a diagonal matrix of non-negative eigenvalues corresponding to the amplitudes of induced dipole moments along the principal axes of the target. The dipoles are assumed to decay independently in the time domain, with each decay parameterized by

$$L_i(t) = k_i t^{-\beta_i} \exp(-t/\gamma_i), \ i = 1, 2, 3.$$ 

By convention, we associate $L_1$ with a dipole aligned with the axis of symmetry, with $L_2$ and $L_3$ transverse. If the item is axisymmetric, then its decay characteristics in the transverse axes can be assumed equal, so that we need only solve for two sets of decay parameters. This is termed a two-dipole model, although the response is comprised of three orthogonal dipole moment vectors. Pasion (2007) showed that combinations of time-decay parameters ($k_i$, $\beta_i$, and $\gamma_i$) are diagnostic of target shape and size and can therefore be used to discriminate between UXO and clutter. An alternative formulation is the instantaneous amplitude model, which specifies the amplitude of the polarizations $L_i$ at each time channel without imposing a parametric form on the time decay. This increases the number of model parameters, but also allows more flexibility when fitting observed data in an inversion. An analogous model in the frequency domain is described by Baum (1998). In this model, the spectral response of the item is parameterized by resonant frequencies which correspond to decay constants in the time domain.

Inversions using the dipole model have produced good agreement with observed field data and so can be used to estimate extrinsic (location and orientation) and intrinsic (polarization amplitudes) target properties. The dipolar form of the secondary field can be a poor approximation if a target is observed in the near-field, and more sophisticated models (e.g. MAS) have been shown to produce better fits to the data in these cases (Shubitidze et al., 2005). However, in practice this shortcoming is more than compensated by the implicit regularization (stabilization) of the inverse problem afforded by the dipole model. A simpler model has fewer degrees of freedom with which to overfit a noisy data set, and this property is desirable when dealing with real data sets.
1.2. UXO detection and discrimination: the state of the art

Inversion

The forward models described in the previous sections are examples of the forward modelling operation

$$d = F\{m\}.$$  

The data vector $d$ is generated by a forward modelling operator $F$ operating on the model vector $m$. When real data are acquired, the related inverse problem is to estimate model parameters which produced the observed data. In the presence of noise, the inverse problem can be written as

$$\hat{m} = F^{-1}\{d^{\text{obs}}\},$$  

where the observed data $d^{\text{obs}}$ are the true data plus noise $\epsilon$

$$d^{\text{obs}} = d + \epsilon.$$  

For magnetic and electromagnetic data the number of observations typically outnumbers the number of model parameters in a parametric forward model. The inverse problem is therefore overdetermined and the solution involves minimizing an objective function which quantifies the misfit between observed and predicted data. A common choice is the least squares (L2) misfit function

$$\phi_d = \|W_d(d^{\text{obs}} - F\{m\})\|^2. \quad (1.7)$$

The diagonal data weighting matrix $W_d$ weights the contribution of a datum based on its estimated standard deviation $\sigma_i$

$$W_{d\,ii} = \frac{1}{\sigma_i}. \quad (1.8)$$

Minimization of the L2 norm is equivalent to maximizing the likelihood function of the data given the model (Menke, 1989). This assumes that

$$d^{\text{obs}}_i = d^{\text{pred}}_i + \epsilon_i,$$  

the noise on the data is independent and Gaussian distributed ($\epsilon_i \sim N(0, \sigma_i)$). While the central limit theorem can be employed to justify the assumption of Gaussian noise, it is often difficult in practice to characterize the uncertainties on the data. Data uncertainty is usually estimated as a percentage of each observed datum plus a noise floor. This weighting is particularly important for inversion of time-domain electromagnetic data, which can decay over several orders of magnitude in the range of measured channels. Weighting the data by an estimated standard deviation ensures that early time, large amplitude data do not dominate the misfit. In addition, an appropriate floor value ensures that small amplitude data do not dominate the misfit after scaling by a percentage. The choice of data standard deviations remains something of an educated guess which can be informed by data pre-processing. For example, a noise floor can be estimated for each time channel by windowing regions.
where no significant signal is observed. In contrast, magnetic data have much less dynamic range and it is often sufficient to specify a noise floor of a few nanotesla when inverting for dipole model parameters.

If the forward modelling operator is linear, then there is a single minimum to the misfit function and the best-fitting model can be obtained in one step by solving a linear system of equations. For a non-linear forward model there may be multiple minima of the misfit function and the solution of the inverse problem cannot be obtained in one step. This is usually the case in UXO applications: all forward models described above are nonlinear functions of the input model parameters. Iterative approaches to the nonlinear inverse problem involve minimizing a quadratic approximation to the objective function with respect to the model perturbation \( \delta m \) at each iteration. For example, the Gauss-Newton method solves

\[
J^T W_d^T W_d J \delta m = -J^T W_d^T W_d (d^{obs} - F\{m\})
\]  

(1.10)

with \( J \) the Jacobian matrix of sensitivities. Given an initial guess for the model parameters, we can solve the above equation for a model perturbation which will reduce the misfit. We then update our model with this perturbation and repeat the procedure until a convergence criterion is achieved (e.g. \( \| \delta m \| < \epsilon \)). Iterative methods can converge to local, suboptimal minima and so it is common practice to initialize these algorithms from multiple starting models.

Local minima of the misfit function have been identified as a particular problem with electromagnetic data. We often observe local minima at incorrect depths (Lhomme et al., 2008). A robust solution to this problem is to cooperatively invert magnetic data: we use a depth estimated from magnetic data to constrain target depth when inverting the corresponding electromagnetic data set (Pasion, 2007). The choice of an inversion algorithm may also help to avoid convergence to local minima: Benavides and Everett (2006) implemented a continuation inversion algorithm to invert TEM data and found that it generally achieved a lower misfit than the Levenberg-Marquardt method (a variant of the Gauss-Newton algorithm).

Another concern when inverting time-domain electromagnetic data is the effect of sensor positional error on the uncertainty in the data. Bell (2005) showed that reliable parameter estimation and classification from TEM data requires a relative positional accuracy on the order of 2 cm. This accuracy can be nominally achieved with laser positioning systems, or by deploying sensors in a cued-interrogation mode. For data which do not meet these requirements, algorithms have been developed to account for positional error in the inversion process. For example, Tarokh and Miller (2007) impose box constraints on positional errors and then estimate model parameters with a min-max approach: they maximize the misfit function with respect to positional errors, and then minimize this same function with respect to the model parameters. They show that this former maximization can be determined by evaluating the misfit function at the extrema (corners) of the positional box constraints. Tantum et al. (2008) adopt a Bayesian approach, treating positional errors as nuisance parameters and marginalizing these parameters from the posterior distribution of the model parameters. Both these approaches provide improvements over conventional least squares minimization, though some additional computational cost is incurred. For example, the Bayesian approach requires Monte Carlo integration of the posterior
A complication which will occur when inverting any data type is overlap between anomalies from two or more adjacent targets. Multi-target scenarios are likely at bombing targets where there is a high anomaly density. Resolution of multiple targets using independent components analysis (ICA) is tested in Hu et al. (2004) and Throckmorton et al. (2007) on synthetic data. ICA assumes that the observed data are linear combinations of polarization spectra and estimates these spectra as transformations of the data which maximize some assumed statistical property (e.g. the spectra should be dissimilar to a Gaussian random variable). Another approach which directly employs the problem physics is to invert the observed data for a postulated number of targets. This technique was tested for TEM on real data sets in Song et al. (2009) and successfully recovered estimates of multiple polarizations.

Soil response can also be an important consideration when processing TEM data. Ferrimagnetic grains of maghaemite or magnetite will undergo a viscous magnetic relaxation when a primary field is applied, and this relaxation can appear as a characteristic time decay which is superposed on the measured decay of an ordnance. Methods for removing a magnetic soil response prior to inversion or incorporating it as a parameter in inversion are described in Pasion (2007).

Finally, we emphasize that quality control (QC) of fits to observed data is a necessary and important step. Because we often have a poor handle on the noise, metrics such as the final data misfit and correlation coefficient may not always be reliable for deciding whether a fit is successful. QC’ing magnetic data is relatively quick, as there is only one channel of data to consider, but TEM data often requires visual inspection of multiple channels in plan view, lines, and individual soundings to determine whether a fit is adequate. Quality control is presently a major bottleneck in UXO data processing.

**Discrimination algorithms**

The end product of geophysical data processing is a diglist which ranks targets from most to least likely to be ordnance, as well as a “stop dig” point (or operating point) on the diglist where digging can be safely terminated. In some cases targets beyond the operating point will be left in the ground. Some sites, however, will require total clearance in order to satisfy environmental regulations. In this case the template for field operations is excavation of high risk targets (as identified on the diglist) by expert disposal teams, with low risk targets excavated by labor under expert supervision. The cost savings of advanced discrimination is realized in the reduction in the number of targets dug by EOD (explosive ordnance disposal) technicians and the choice of operating point is less critical because we are guaranteed to find all detected ordnance.

To rank targets for digging, we use the information in our observed geophysical data. Features of the observed data, estimated without resorting to inversion with a physics-based model, can sometimes suffice as criteria to discriminate between ordnance and non-ordnance targets. For example, in Williams et al. (2007) a bivariate Gaussian distribution is fit to observed EM61 data at each time channel and the average width of the anomaly, as measured by the estimated covariance matrix, is then used as a criterion to rank ordnance (wide anomaly) ahead of clutter (small anomaly). This approach significantly outperformed a statistical classification approach employing features estimated
1.2. UXO detection and discrimination: the state of the art

with the dipole model. This can work when ordnance is significantly larger than clutter, but may fail if there are large, deep clutter which can generate broad anomalies. Furthermore, a horizontal target can sometimes produce an anomaly which is better described by a bimodal distribution (i.e. two Gaussians, see Pasion (2007)). Data features are nonetheless useful when data quality is not sufficient to support estimation of useful parameters in an inversion or when time constraints preclude processing with inversion.¹

Parameters of models estimated from inversion can resolve some of the ambiguities of data features because model parameters can be related to intrinsic target properties. An intuitive template matching approach to discrimination compares estimated model parameters with those previously derived from a library of known targets. For magnetics data template matching has been used to compare the estimated magnetostatic dipole with library responses (Billings, 2004). An ordnance can produce a range of induced moments depending upon its orientation relative to the earth’s magnetic field, and the deviation of estimated dipole moments from this type curve provides a ranking of targets. Discrimination with TEM data is often performed by comparing estimated polarization decays with library responses and then ranking a target based on some measure of closeness between observed and expected responses. Care must be taken here to use parameters which can be reliably estimated: late time polarizations are more susceptible to noise and poor polarization estimates may unduly affect the discrimination decision. Pasion et al. (2007) solve this problem with a fingerprinting algorithm that inverts for target location and orientation while holding polarizations fixed at their library values. Reducing the model’s degrees of freedom in this way makes the inversion less susceptible to fitting the noise. Targets are then dug based upon the proposed library item which produces the best fit to the observed data. We can regard this method as incorporating information from our target library directly into the inversion, whereas conventional template matching uses library information in the discrimination stage.

Library methods assume that there is a true set of model parameters that, under ideal circumstances, can be perfectly reconstructed from an observed data set. Statistical classification algorithms which have been applied to UXO discrimination can be regarded as Bayesian solutions to the discrimination problem: we treat the parameters of interest as fundamentally uncertain random variables which are characterized by probability distributions. We then try to learn these probability distributions from a sample of labelled targets for which ground truth is known (the training data), and then formulate a decision rule that tries to minimize the probability of making an incorrect decision for unlabelled targets (the test data). One approach to formulating the decision rule is to fit some assumed parametric distributions to each class of targets in the training data, and then assign a test target to the class distribution which is most likely. The class distributions are defined in a multidimensional feature space spanned by some subset of estimated model parameters, or transformations thereof. The success of a statistical classifier is measured by its ability to generalize to the unseen test data (i.e. correctly classify), and having a training data set which is representative of class variability in the test data set is crucial. In Aliamiri et al. (2007), for example, class distributions are generated by simulating data for each target class in a range of orientations and depths, and then inverting these synthetic data. This assumes that

¹Periodic clearance of active ranges must often be carried out within a short time.
simulations can capture the noise conditions which are encountered in experimental data. Alternatively, training data can be generated by full clearance of selected grids in a geophysical prove-out. Active learning techniques for iteratively selecting targets to build the training data set, based upon reducing uncertainties in the resulting classifier, are developed in Zhang et al. (2004).

### 1.3 Case study: Camp Lejeune

As an example of UXO data processing, in this section I present results of a discrimination study carried out at Marine Corps Base Camp Lejeune, North Carolina in 2006. The goal of this study was to identify pre-existing and emplaced ordnance items from ubiquitous clutter on a range used for training exercises. Only one “native” ordnance was encountered during surveying (an anti-tank rocket), all other ordnance targets were emplaced items taken from the ATC (Aberdeen test center) standardized set or from a set of training mortars excavated in a previous remediation project. Clutter items were predominantly non-ferrous (aluminum) “adapters” used as casings for armor-piercing sabot rounds (figure 1.8).

![Figure 1.8: Adapter excavated at Camp Lejeune](image)

Four sensors were deployed for this study: a magnetometer array (figure 1.3(a)), a towed array of EM61 sensors, an EM63 cart (figure 1.3(b)) and ground penetrating radar. Because the majority of emplaced ordnance were ferrous, discrimination between ordnance and non-ordnance with magnetic data was relatively easy (non-ferrous adapters should have no magnetic anomaly). Here we focus on analysis of the EM63 data set. Inversions were carried out for all identified targets using a tri-axial dipole model with polarization decays parameterized in terms of equation 1.6.

Figures 1.9 and 1.10 show fits obtained for two targets in this data set. These images were generated as part of a fit report in the UXOLab software package. The report summarizes the spatial and temporal observed and predicted data, as well as the estimated model parameters. This facilitates the decision made by a data analyst to fail (figure 1.9) or pass (figure 1.10) an inversion result. The inversion in figure 1.9 is failed because there is poor data coverage over the anomaly and the recovered polarizations are poorly constrained. Only passed inversions should be input into a discrimination algorithm which uses model parameters as features. Ideally, failed inversions can be re-inverted with a
new data mask if an initial mask fails to exclude data from neighbouring targets. Targets which cannot be successfully modelled after this additional quality control can sometimes be analysed on the basis of data features alone. In the worst case, we must label a target as “can’t analyze” and designate it for excavation.

Figure 1.11 shows a subset of polarization parameters \((k_i, \beta_i, \gamma_i)\) estimated from the observed EM63 data. For the \(k_i\) parameters there is considerable overlap between distributions of adapters and ordnance. This is expected because \(k_i\) parameters are determined by the early time polarization response. As shown in figure 1.7, the early time, high frequency limit of the induced response for a sphere is insensitive to material composition and so we cannot distinguish between ferrous and non-ferrous targets using the \(k_i\) parameters. The difference between secondary and tertiary parameters \((k_2\) and \(k_3\)) is sometimes used as a criterion for discriminating between ordnance and clutter on the basis that ordnance are axisymmetric so that \(k_2 = k_3\), whereas clutter are generally irregularly shaped and non-axisymmetric \((k_2 \neq k_3)\), (Bell and Barrow, 2001). In figure 1.11(b), the secondary and tertiary polarizations, while positively correlated, are not sufficiently constrained by the data to make inferences regarding target shape. Regardless, if our goal is to discriminate between adapters and ordnance, then both target classes are axisymmetric (see figure 1.8). Parameters sensitive to late time response \((\beta_i, \gamma_i)\) are more useful for discrimination between ordnance and adapters. This result can be understood by noting that the low frequency, late time TEM response is fundamentally sensitive to a target’s magnetic composition (figure 1.7(a)).

To discriminate between adapters and ordnance on the basis of estimated model parameters, we must select some subset of these parameters. Figure 1.11 suggests that the parameters \(\beta_1\) and \(\beta_2\) have the best capability to separate the distributions of adapters and ordnance, with \(\gamma_1\) and \(\gamma_2\) also providing some discrimination information. Statistical criteria can be used to validate these qualitative observations. For example, canonical analysis is an algorithm which tries to find linear combinations of model parameters which maximize the separation between two or more classes (Gittins, 1985). This analysis assumes that class distributions are approximately normally distributed and have equal covariance. While this is certainly not true for these data, canonical analysis of all polarization parameters spanning a nine-dimensional feature space does favor decay parameters \(\beta_i\), and \(\gamma_i\) ahead of amplitude parameters \(k_i\) (figure 1.12). More sophisticated techniques have been proposed for the problem of feature selection in UXO discrimination (e.g. Krishnapuram et al. (2004)). However, our experience has been that features should be chosen based upon the physical properties (e.g. ferrous vs. non-ferrous) which can provide discrimination capability. The task is then to connect these properties to the parameters of the parametric model. Statistical feature selection criteria should then be viewed as tools to support physical intuition.

Because non-ordnance targets have consistent and distinct physical properties at Camp Lejeune, this data set is amenable to analysis with a library-based discrimination method. Pasion et al. (2007) applied their fingerprinting algorithm to these data and were able to eliminate the majority of adapters from the diglist. Here I consider a simplified form of the Pasion fingerprinting technique, similar to methods considered in Tantum and Collins (2001) and Norton et al. (2001). More description of this
1.3. Case study: Camp Lejeune

Figure 1.14 compares the expected performance of quadratic discriminant analysis (QDA) classifiers and the fingerprinting technique applied to the Camp Sibert EM63 data. QDA assumes that the class distributions in the feature space are Gaussian, with individual means and covariances estimated for each class (Hastie et al., 2001). An example of the surface of maximum probability output by this classifier in a two-dimensional feature space is shown in figure 1.13.

In figure 1.14, I quantify the performance of each classifier with the receiver operating characteristic (ROC) curve, which displays the proportion of true positives (ordnance) versus the proportion of false positives (adapters). The curve is generated by labelling targets in the order predicted by a classifier, beginning with targets that are most probably ordnance. This corresponds to the order in which targets will be excavated in the field. An ideal ROC will label all ordnance before any non-ordnance are encountered and so passes through the point (0,1).

For a statistical classifier the ROC curve depends upon the particular realization of training data used to model the class distributions, as well as the test data for which predictions are generated. I therefore employ a bootstrapping scheme involving random resampling of all estimated features to generate the mean expected ROC, as well as best and worst case bounds on expected performance. This procedure is described in more detail in Beran and Oldenburg (2008) (chapter 4 of this thesis).

In figure 1.14(a), I attempt a naive discrimination approach with all polarization parameters spanning a nine-dimensional feature space. While the mean ROC in this case compares favorably to the other scenarios, we see that the bandwidth of the upper and lower bounds is relatively wide. This is indicative of the “curse of dimensionality”: as the dimension of the feature space grows, the number of parameters required to estimate the parameters of the classifier grows exponentially (Hastie et al., 2001). For these data there are simply not enough targets in the ordnance class to obtain a stable estimate of the class covariance, and consequently the output of the QDA classifier is highly variable. Consistent with our physical intuition and canonical analysis, discrimination with the $k_i$ in (b) performs quite poorly on these data. Reducing the dimensionality of the feature space in (c) by using only the $\beta_i$ parameters somewhat reduces the variability of the ROC and improves mean performance. This type of dimensionality reduction is often required for UXO discrimination: we generally have relatively small samples with which to model class distributions and so must restrict our feature space to a small subset of parameters which are useful. Supplementing the training data with synthetics as in Aliamiri et al. (2007) may help stabilize classifier training. Finally, the fingerprinting technique has a narrow bandwidth ROC, with average discrimination performance comparable to QDA on the $\beta_i$.

On the basis of the average performance shown in figure 1.14, I conclude that for this particular data set fingerprinting is the best choice of discrimination algorithm, followed closely by statistical classification with the $\beta$ parameters. This does not imply, however, that fingerprinting will always be the best technique for discrimination with electromagnetic data. Indeed, Aliamiri et al. (2007) found that statistical classifiers outperformed “data-space” discrimination algorithms (analogous to fingerprinting) for discrimination of EM63 data. In this thesis my approach will be to develop a suite of inversion and discrimination tools and to objectively evaluate the expected performance of these tools for the problem.
1.4 Outline of the thesis

This is a manuscript-based thesis with the main body (Chapters 2 through 5) comprised of four research articles. Chapters 2, 3, and 5 will be submitted for publication after final approval of the thesis, and Chapter 4 was published in 2008. As each chapter is intended to be independent there is some repetition of material. I hope that this will reinforce concepts rather than bore the reader.

Chapter 2 examines stochastic noise sources - in particular sensor positional error - and their effect upon errors in electromagnetic data. Inversion with the conventional least squares misfit function assumes that errors are normally-distributed. I show that this assumption does not hold in the presence of positional errors. This result motivates investigation of alternative, robust misfit functions which are designed to reduce the effect of non-normal errors on model estimates obtained with inversion. Application of robust inversion to time-domain electromagnetic data sets indicates that it can improve model estimates and subsequent discrimination performance.

Chapter 3 considers propagation of errors from data to model in the context of inversion of time-domain electromagnetic data with a parametric model. Two methods of uncertainty appraisal are considered: a linearized technique which approximates the uncertainty distribution of the model as a Gaussian, and a nonlinear technique which makes no restrictions on the form of the model probability distribution. For this problem, I find that the linearized analysis is often an approximation to one mode of the multimodal probability distribution obtained with nonlinear appraisal. Conventional discrimination algorithms treat model uncertainty as point estimates, and I develop a generalization of discriminant analysis which accounts for the variability of model parameters represented by a probability distribution. Simulations indicate that this technique can improve discrimination performance by identifying outliers which may be characterized by large uncertainties. In addition, I develop a method which characterizes model uncertainty in the feature space without resorting to numerically intensive nonlinear appraisal. This method also improves outlier detection.

In chapter 4 I investigate the choice of discrimination algorithm. A wide variety of techniques have been proposed for the UXO discrimination problem and while comparisons abound there is no single method which is universally successful. Every data set presents particular challenges and we must adapt our discrimination approach to the individual problem. Here I apply a bootstrapping technique to objectively estimate metrics of discrimination performance and select a discrimination algorithm with the best expected performance.

Finally chapter 5 deals with the question of selecting a cut-off point at which we terminate excavation of targets. This question is relevant to a variety of applications. For example, a medical test may return the level of a specific antigen in the blood. To make a diagnosis of a condition with this test, we must decide at what antigen level we can deem that condition to be present. In UXO discrimination we must similarly deal with numerical proxies which indicate the presence of a buried ordnance. I develop new techniques for selecting a cut-off and compare these approaches with existing methods.
1.4. Outline of the thesis

I conclude in chapter 6 with a discussion of the thesis contributions and directions for future research.
Failed fit report for EM63 target, Camp Lejeune. Gridded images show observed, predicted and residual data, with points indicating observation locations. Magenta circle defines target mask, and red line is shown as a profile at bottom left. Displayed sounding is for the observation location closest to the target pick (white cross in gridded images). Misfit versus depth curve displays the relative misfit of models obtained with nonlinear inversion, as well as all initial models from a scan over a range of proposed target locations.
1.4. Outline of the thesis

- Channel 1 observed
- Channel 1 predicted
- Channel 1 residual
- Channel 10 observed
- Channel 10 predicted
- Channel 10 residual
- Channel 20 observed
- Channel 20 predicted
- Channel 20 residual

Line profile

Sounding

Polarizations

Estimated parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>-0.10</td>
</tr>
<tr>
<td>y</td>
<td>-0.10</td>
</tr>
<tr>
<td>z</td>
<td>-0.15</td>
</tr>
<tr>
<td>$\theta$</td>
<td>-1.47</td>
</tr>
<tr>
<td>$\phi$</td>
<td>1.32</td>
</tr>
<tr>
<td>$\delta$</td>
<td>-0.95</td>
</tr>
<tr>
<td>$k_1$</td>
<td>2.70</td>
</tr>
<tr>
<td>$b_1$</td>
<td>0.41</td>
</tr>
<tr>
<td>$g_1$</td>
<td>1.75</td>
</tr>
<tr>
<td>$k_2$</td>
<td>3.37</td>
</tr>
<tr>
<td>$b_2$</td>
<td>0.70</td>
</tr>
<tr>
<td>$g_2$</td>
<td>2.81</td>
</tr>
<tr>
<td>$k_3$</td>
<td>3.47</td>
</tr>
<tr>
<td>$b_3$</td>
<td>0.59</td>
</tr>
<tr>
<td>$g_3$</td>
<td>5.34</td>
</tr>
</tbody>
</table>

Figure 1.10: Passed fit report for EM63 target, Camp Lejeune. Plots are as described in figure 1.9.
1.4. Outline of the thesis

Figure 1.11: Model parameters estimated from EM63 data collected at Camp Lejeune
Figure 1.12: First canonical eigenvector from canonical analysis of Camp Lejeune EM63 data. A linear combination of the model parameters with weightings $v_i$ is expected to maximize the separation between means of the ordnance and adapter classes.

Figure 1.13: Surface of maximum probability output by quadratic discriminant analysis classifier trained to discriminate between ordnance and adapters.
1.4. Outline of the thesis

Figure 1.14: Receiver operating characteristics (ROC) showing true positive fraction (TPF) versus false positive fraction (FPF) for quadratic discriminant analysis (QDA) classifiers (a-c) and library method (d). Here QDA is applied to discrimination between ordnance and adapters using features estimated from EM63 data. Solid line is the mean bootstrapped ROC, dashed lines are best and worst case ROCs from bootstrapping. Feature spaces are: (a) all polarization parameters $k_i, \beta_i, \gamma_i, i = 1, 2, 3$, (b) polarization amplitude parameters $k_1, k_2, k_3$, (c) decay parameters $\beta_i, i = 1, 2, 3$, (d) fingerprinting.
Bibliography


Chapter 1. Bibliography


Chapter 1. Bibliography


Chapter 2

Robust inversion for UXO discrimination

2.1 Introduction

Estimating model parameters from observed geophysical data is a crucial step in discriminating between hazardous unexploded ordnance and benign clutter items. Inverting data involves minimizing some norm of the discrepancies between observed \(d^{\text{obs}}\) and predicted \(d^{\text{pred}}\) data. The parameters of the best-fitting model can then be used to infer a target’s intrinsic properties (size, shape, material composition, etc.). The least squares (L2) norm is most commonly chosen as a misfit function

\[
\phi = \| W_d (d^{\text{obs}} - d^{\text{pred}}) \|^2
\]  

(2.1)

with \(W_d\) a weighting matrix. When the weightings are selected to be the inverse standard deviations of the data \((W_{ii} = 1/\sigma_i)\), then minimization of the L2 norm is equivalent to maximizing the likelihood function of the data (Menke, 1989). This assumes that

\[
d^{\text{obs}}_i = d^{\text{pred}}_i + \epsilon_i,
\]  

(2.2)

the noise on the data is independent and Gaussian distributed \((\epsilon_i \sim N(0, \sigma_i))\). While the central limit theorem can be employed to justify the assumption of Gaussian noise, it is often difficult in practice to characterize the uncertainties on the data.

Of particular concern when inverting time-domain electromagnetic (TEM) data is the effect of sensor positional error on the uncertainty in the data. Bell (2005) showed that reliable parameter estimation and classification from TEM data requires a relative positional accuracy on the order of 2 cm. This accuracy can be nominally achieved with laser positioning systems, or by deploying sensors in a cued-interrogation mode where measurements are made on a template. For data which do not meet these requirements, algorithms have been developed to account for positional error in the inversion process. For example, Tarokh and Miller (2007) impose box constraints on positional errors and then estimate model parameters with a min-max approach: they maximize the misfit function with respect to positional errors, and then minimize this same function with respect to the model parameters. They show that this former maximization can be determined by evaluating the misfit function at the extrema (corners) of the positional box constraints. Tantum et al. (2008) adopt a Bayesian approach, treating positional errors as nuisance parameters and marginalizing these parameters from the posterior distribution of the

\[2\] A version of this chapter will be submitted for publication. Beran, L., Billings S. and Oldenburg D. Robust inversion for unexploded ordnance discrimination.
2.2. Forward modelling

model parameters. Both these approaches provide improvements over conventional least squares minimization, though some additional computational cost is incurred, especially by the Bayesian approach, which requires Monte Carlo integration of the posterior distribution.

In this study we further examine the role of positional error on inversion of TEM data. We first show that the distributions of data residuals in the presence of positional error do not follow a Gaussian distribution. This motivates consideration of robust norms which are designed to provide reliable parameter estimates when a proportion of the data is contaminated with non-Gaussian noise. We carry out simulations to understand improvements in parameter estimation afforded by robust estimation in the presence of both uncorrelated and correlated noise. Finally we apply robust inversion to real datasets acquired in detection and cued-interrogation surveys.

2.2 Forward modelling

We use the dipole model (Bell (2005), Pasion (2007)) to predict observed TEM data and to investigate the effect of positional error on the data and the model estimated from those data. The secondary magnetic field is computed as

\[ B_s(r, t) = \frac{m(t)}{r^3} \cdot (3(\hat{m}(t) \cdot \hat{r})\hat{r} - \hat{m}(t)) \tag{2.3} \]

with \( r = r\hat{r} \) the separation between target and observation location, and \( m(t) = m(t)\hat{m}(t) \) a time-varying dipole moment

\[ m(t) = \frac{1}{\mu_0} M(t) \cdot B_o. \tag{2.4} \]

The induced dipole is the projection of the primary field \( B_o \) onto the target’s polarization tensor \( M(t) \).

If the primary field is itself modelled as a dipole, then the above expressions indicate that the observed data will have a \( 1/r^6 \) dependence. The polarization tensor is assumed to be symmetric and positive definite and so can be decomposed as

\[ M(t) = A^T L(t) A \tag{2.5} \]

with \( A \) an orthogonal matrix which rotates the coordinate system from geographic coordinates to a local, body centered coordinate system.

\[ L(t) = \begin{bmatrix} L_1(t) & 0 & 0 \\ 0 & L_2(t) & 0 \\ 0 & 0 & L_3(t) \end{bmatrix} \tag{2.6} \]

with the eigenvalues ordered by convention so that \( L_1(t) \geq L_2(t) \geq L_3(t) \). The eigenvalues decay independently in time and can be parameterized according to

\[ L_i(t) = k_i t^{-\beta_i} \exp(-t/\gamma_i), \quad i = 1, 2, 3. \tag{2.7} \]
The parameters $k_i$, $\beta_i$, and $\gamma_i$ are then related to the intrinsic properties of the target (Pasion, 2007). Another option is to parameterize the model at each time channel directly in terms of the instantaneous polarization amplitudes $L_i(t)$, with no functional dependence between polarizations at different times.

Decomposing the polarization tensor with equation 3.15 parameterizes the model in an orthogonal coordinate system which is assumed to correspond to axial and transverse coordinates of a target. However, this parameterization introduces additional nonlinearity into the forward model: the rotation matrix $A$ is a nonlinear function of target orientation. Parameterizing the polarization decay with an expression such as 2.7 also produces a nonlinear dependence between data and model. A final source of nonlinearity in the forward model is the $1/r^6$ dependence arising from the dipolar primary and secondary fields. All of these nonlinearities complicate the corresponding inverse problem. Iterative algorithms may converge to local minima of the chosen objective function and thereby produce model estimates which are far from the true parameters.

One way to address these complications is to repeatedly solve a related linear problem. If the target location is known, then the forward model for the data at a single time channel is linear in terms of the elements of $M(t)$. Solving this linear problem over a range of proposed target locations provides a preliminary search of model space and can help identify local minima of the misfit and starting models for subsequent nonlinear inversion.

## 2.3 Positional errors

In this section we provide an analysis of the distribution of the data when the received signal follows a $1/r^6$ decay. We show that the distribution of data residuals about the true datum has positive skew and so we might expect to see large positive outliers and non-normal distributions of residuals.

Consider a datum

$$d = f(r) = \frac{1}{r^6}$$

with $r$ the position vector between sensor and target, and $r$ its norm. Assume that positional error is a normally distributed random variable $\Delta r = N(0, S)$, with $S$ the covariance matrix. For simplicity assume independent positional errors with identical variance ($\sigma^2$), so that $S = \sigma^2 I$. Addition of positional error is approximated by a Taylor series

$$f(r + \Delta r) = f(r) + J\Delta r + \frac{1}{2}(\Delta r)^T H\Delta r + \ldots$$

The distribution of a datum about its true value is then given by $p(\Delta d) = p(f(r + \Delta r) - f(r))$. We can
2.3. Positional errors

characterize this distribution by its moments. The first three moments are, to leading order,

\[ E(\Delta d) = \frac{1}{2} \sum_{i=1}^{3} \frac{\partial^2 f}{\partial x_i} \sigma_i^2 \propto \frac{\sigma^2}{r^8} \tag{2.10} \]

\[ E((\Delta d)^2) = \sum_{i=1}^{3} \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2 \propto \frac{\sigma^2}{r^{14}} \tag{2.11} \]

\[ E((\Delta d)^3) = 3 \sum_{i=1}^{3} \left( \frac{\partial f}{\partial x_i} \right)^2 \frac{\partial^2 f}{\partial x_i} \sigma_i^4 + 3 \sum_{i=1}^{3} \sum_{j \neq i} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \frac{\partial^2 f}{\partial x_i \partial x_j} \sigma_i^4 \propto \frac{\sigma^4}{r^{22}}. \tag{2.12} \]

Here \( x_i \) is the \( i^{th} \) basis vector in the chosen coordinate system. Now for \( f(r) \) given by equation 2.8, the third moment is strictly positive. This implies that the distribution of residuals has positive skew, i.e. there are larger magnitude positive outliers than negative outliers. This is in contrast to the generating distribution of positional errors, which has zero skew.

Figure 2.1 shows the agreement of the moments of the distribution of residuals computed by simulation and using the above expressions. It is important to note that the distribution of residuals at a given location is determined by the dimensionless ratio of \( r/\sigma \). There is some disagreement between simulations and the predicted second and third moments for small \( r/\sigma \), likely due to contributions of higher order terms. We see that the data residuals have nonzero mean, with the mean proportional to the variance of the positional error (\( \sigma^2 \)) and inversely proportional to target sensor separation. Similarly, the variance and skew of the residuals decreases nonlinearly as we increase \( r \) or decrease \( \sigma \). This result suggests that the closer the sensor is to a target, or the larger the positional error, then the larger the variance and the more likely we are to encounter outlying data. In figure 2.2 we show the distribu-

![Figure 2.1](image_url)

Figure 2.1: Dependence of the first three normalized moments of the data residuals on the ratio of target-sensor separation (\( r \)) and positional standard deviation (\( \sigma \)). Moments are normalized by the true datum (\( d \)). Points indicate Monte Carlo simulations, solid lines are predictions for moments given by equations 2.10 to 2.12.
tions of residuals generated by simulation for a few cases, together with maximum likelihood estimates of generalized extreme value (GEV) distributions (Gumbel, 1958). There is a good correspondence between the empirical and fitted distributions, however any distribution which can have nonzero skew will likely provide an adequate fit to the data residuals. We see that there is significant skew in the data residual distribution for $r/\sigma = 30$, but the distribution is effectively normal for $r/\sigma = 100$. The former case can be interpreted as the distribution of residuals we might expect for a target 30 cm directly below the sensor, with a 1 cm standard deviation on positional error.

![Figure 2.2: Distributions of relative residuals $(\Delta d)/d$ for increasing $r/\sigma$](image)

Noise in observed data is not limited to positional error, other sources such as orientation error, cultural noise, and instrument noise will also contribute additively to the observed data. While many of these noise sources are likely non-normally distributed as well, their superposition will tend towards a normal distribution. However, individual noise sources may dominate a particular datum and so cause large deviations. For example, if the sensor quickly traverses uneven topography which cannot be adequately sampled by a positioning system, then these data will have relatively large deviations from their true values and may have undue influence on inversion with a least squares norm. In the following section we therefore investigate inversion with robust norms.

### 2.4 Robust norms

Robust inversion algorithms in the geophysical and statistical literature modify the misfit function so that outlying data have downweighted contributions to the total misfit. In the context of geophysical inversion, Farquharson and Oldenburg (1998) have demonstrated the use of various robust norms for measuring both data misfit and model norm. Here we follow their description of an iterative approach to minimizing these norms for linear and nonlinear inverse problems. Consider the norm

$$\phi = \rho(x)$$

with

$$x = W_d (d^{obs} - d^{pred})$$

(2.13)

(2.14)
2.4. Robust norms

the weighted discrepancy between observed and predicted data, and \( \rho(x) \) defining some norm on this
vector (e.g. \( \rho(x) = x^T x \) for the L2 norm). For a linear problem, we can write \( d^{pred} = Gm \). Now to
minimize \( \phi \) with respect to the model vector \( m \) we have

\[
\frac{\partial \rho}{\partial m} = \frac{\partial \rho}{\partial x} \frac{\partial x}{\partial m} = B^T Rx
\]

(2.15)

with

\[
R = \text{diag}\left(\frac{\partial \rho/\partial x}{x}\right)
\]

(2.16)

and

\[
B_{ij} = \frac{\partial x_i}{\partial m_j}
\]

(2.17)

For a linear forward problem then

\[
\frac{\partial \phi}{\partial m} = B^T Rx = G^T W_d^T R W_d (d^{obs} - Gm)
\]

(2.18)

Setting the above expression equal to zero and solving for \( m \) yields an expression identical to that
in a standard least squares problem, except for the presence of \( R \). This matrix depends upon \( x \) the
weighted residual, and so is a function of the model \( m \). Hence minimization of an arbitrary norm
becomes a nonlinear problem even when the forward problem is linear. This nonlinearity can be cir-
cumvented with the "iteratively reweighted least squares" (IRLS) algorithm, which iteratively updates
the model with the following procedure

1. Set \( R_{ii} = 1 \) and solve for \( m \) with equation 2.18.

2. Update the elements of \( R \) (equation 2.16) using the current estimate of \( m \)

3. Recompute \( m \) and iterate from step 2 until convergence.

For a nonlinear forward problem, we can use the linearization

\[
F(m + \delta m) \approx F(m) + J \delta m
\]

(2.19)

to derive an expression for the model perturbation \( \delta m \) at each iteration, with the sensitivity matrix \( J \)
taking the place of the forward modelling operator \( G \) in equation 2.18. IRLS provides a general pro-
cedure for minimizing a norm, table 2.1 summarizes a number of norms which appear in the literature.
Figure 2.3 compares these norms as a function of \( x \) and also shows the resulting weightings \( R_{ii} \). The
bisquare norm is unique amongst the norms considered here in that it has the capability to completely
disregard outlying data (\( R_{ii} = 0 \) for \( |x_i| > k \)). However, some care must be taken in initializing IRLS
for the bisquare norm, since the algorithm is not guaranteed to converge for this norm (Marrona et al.,
2006). This is evident in figure 2.3: if the starting model generates predicted data that is far from the
2.4. Robust norms

<table>
<thead>
<tr>
<th>Norm</th>
<th>$\rho(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huber</td>
<td>$\begin{cases} x^2, &amp;</td>
</tr>
<tr>
<td>Eckblom</td>
<td>$(x^2 + k^2)^{p/2}$</td>
</tr>
<tr>
<td>Bisquare (Tukey)</td>
<td>$\begin{cases} \frac{k^2}{6} \left(1 - \left[1 - \left(\frac{x}{k}\right)^2\right]^3\right), &amp;</td>
</tr>
</tbody>
</table>

Table 2.1: Example norms used for robust inversion

observed data, then the bisquare misfit function has no curvature and cannot converge to the global minimum of the misfit. We find that initializing IRLS with the least squares solution works well in this application.

Figure 2.3: Left: comparison of norms as a function of $x$, the weighted discrepancy between observed and predicted data. Right: weightings ($R_{ii}$) applied to data for IRLS minimization of robust norms. For all norms $k = 1$, and $p = 1$ for the Eckblom $p$ norm.

A practical question when applying robust norms is specification of the parameters of the norm (e.g. parameter $k$ in table 2.1). In overdetermined inversion, norm parameters are usually set based upon consideration of the variance of the model estimate when the data are normally distributed. When solving a linear inverse problem, the model covariance at the IRLS solution is

$$cov(m) = (G^T W_d^T R W_d G)^{-1}. \quad (2.20)$$

Referring to figure 2.3, we see that the weightings $R$ for robust norms are always less than or equal to the weightings for the L2 norm, for which $R$ is the identity. This implies that the model variance will always be greater when carrying out robust versus L2 estimation if the data are normally distributed. The ratio of L2 estimator variance to robust estimator variance is called estimator efficiency. Recommended values of norm parameters are then set to achieve 95% efficiency (Marrona et al., 2006).
2.5 Simulations

To investigate the effect of robust norms on parameter estimation from EM data, we first consider solution of the inverse problem by repeatedly solving the linear problem for the non-diagonalized polarization tensor over a range of target locations (see section 2.2). We simulate data sets over two targets with polarizations of comparable magnitude, in both horizontal and vertical orientations. One target is axisymmetric and is approximately the size of a mortar baseplate from the Camp Sibert, Alabama ESTCP demonstration study (see figure 2.9). The second target is slightly smaller and non-axisymmetric and is representative of clutter encountered at the same site. In this example we simulate positional errors as independent Gaussian perturbations to the sensor locations with a standard deviation of 5 cm. We select a relatively large positional error here to investigate the performance of various norms when positional accuracy does not meet the data quality requirements in Bell (2005). Background noise is simulated as 1 percent plus a 5 mV floor Gaussian noise. Stations are spaced uniformly at 20 cm between lines and 20 cm along lines, and only data which are twice the noise floor are used in each inversion. We choose a relatively sparse along line sampling in this case to allow comparison with the results for independent positional error in Tarokh and Miller (2007).

Figure 2.4: Fits (dashed line) to synthetic data (dots) for a single realization of noise. True data are shown as a solid line.

Figure 2.4 shows an example realization of data over a single target, as well as fits predicted by each norm. As expected from the analysis in section 2.3, large positive outliers correspond to large magnitude data close to the target. Figure 2.5 shows the results of these simulations for 50 realizations of positional and background noise, with each point representing the minimum misfit model identified in the scan of target positions. We present the L2 and bisquare norms only, other robust norms in table 2.1 had comparable performance to the bisquare norm. We see that there is no significant improvement in our ability to discriminate between targets provided by the robust norm relative to conventional least squares when errors are independent.

The difference in the distributions of polarizations for horizontal versus vertical targets in figure 2.5 is a result of the positive correlation between polarization strength and target depth. When a target
is oriented horizontally, the data are more sensitive to the transverse (secondary and tertiary) polarizations. The difficulty in constraining target depth is therefore manifested as a positive correlation between depth and transverse polarizations for horizontal targets. For a vertical target, the data are more sensitive to the primary, axial polarization, so that the correlation between target depth and transverse polarizations is reduced. We also note that the clustering of parameter estimates in figure 2.5 is a function of scanning over a discrete set of target depths; scanning over a finer range of target locations will produce more continuous parameter distributions.

Figure 2.5: Effect of independent positional error upon L2 and bisquare secondary polarization estimates ($L_2$, $L_3$) for two targets (dots and crosses). Dots are parameter estimates for an axisymmetric target with true transverse polarizations $L_2 = 10$, $L_3 = 10$, and crosses are parameter estimates for non-axisymmetric target with $L_2 = 10$, $L_3 = 5$. True target parameters are shown as open circles. Top row is for targets with primary polarization oriented horizontally, bottom row is for targets with primary polarization oriented vertically.

These results seem inconsistent with previous research in Tarokh and Miller (2007) which found that a robust minimax inversion technique provides marked improvement of conventional least squares for independent positional error. We note that in Tarokh and Miller (2007) the authors did not weight their L2 inversion with an estimated standard deviation. In contrast, here we estimate a standard devia-
2.5. Simulations

tion for each datum as a percentage of the observed datum (5 percent in this case) plus a floor. This is especially crucial for a inversion of TEM data, where the exponential decay of data requires a weighting which ensures that large amplitude early time data do not dominate the misfit. Hence, even when minimizing the L2 norm we essentially carry out a single iteration of IRLS, so that we expect the weighted L2 to be robust to outliers which might have a significant impact if no reweighting is employed. To illustrate this effect, we compute the sensitivity curves of the unweighted L2, weighted L2 and bisquare estimators for the simple problem of estimating the mean of a random data set \( (x_1, x_2, \ldots, x_N) \) contaminated by a single outlying datum \( x_o \). The sensitivity curve shows the bias in the estimated parameter

\[
\text{bias}(\hat{\mu}) = \hat{\mu}(x_1, x_2, \ldots, x_N, x_o) - \hat{\mu}(x_1, x_2, \ldots, x_N)
\]  

(2.21)
as a function of the location of the outlying datum (Marrona et al., 2006). Figure 2.6 shows sensitivity curves estimated from 100 realizations of Gaussian data, with \( N = 100 \) data in each realization. We see that while the bias in the L2 estimator increases linearly with outlier location, the weighted L2 estimate is able to control the bias by downweighting outliers. As expected, the bisquare estimator ignores large outliers so that there is zero bias in the estimate when \( x_o \) is far from \( \mu \).

![Figure 2.6: Sensitivity curves of estimates of the sample mean obtained from realizations of \( N = 100 \) Gaussian data contaminated with a single outlier \( x_o \).](image)

In the above simulations, we considered positional errors to be independent. In practice this assumption may not necessarily hold. For example, transient noise sources may cause neighbouring data to deviate systematically from their actual position, so that positional error must be treated as correlated. To understand the effect of correlated noise on parameter estimation, we repeat the above simulations with correlated noise added as follows

1. Background Gaussian noise is added to the true data for each target.
2. Along each line, data exceeding the noise floor are identified.
3. A positive shift is then added to a number, $n_\epsilon$, of adjacent points exceeding the noise threshold. This simulates a correlated error, possibly due to correlated positional error, though in this case the shift is generated deterministically as a percentage of the largest perturbed datum, rather than as a random perturbation on position. An additional random Gaussian error is added to each shifted datum, with standard deviation defined as 5 percent of the shift (we do not expect all points to be shifted by the same amount in the presence of a correlated error).

4. for subsequent realizations, correlated errors are then added to all sets of adjacent $n_\epsilon$ points along all lines (the larger $n_\epsilon$ the fewer perturbations along each line)

Given a noise realization computed in this manner, we then re-estimate the standard deviation of each datum as a percentage of the noisy datum, as discussed above. These standard deviations then define the elements of $W_d$, the data weighting matrix for inversions carried out with L2 and robust norms. These simulations are motivated by the observation that the robustness of an estimator can be measured as a function of the proportion of data $\epsilon$ which is contaminated by non-Gaussian noise (Marrona et al., 2006). Figure 2.7 shows the resulting distributions of target parameters when $n_\epsilon$ is 5 and 10. These correspond to correlated error applied to 2 and 4 percent of the data used in each inversion, respectively. The two cases for $n_\epsilon$ are not shown separately here, but increasing the proportion of data contaminated with correlated noise increases the spread in L2 estimates, while robust norms are able to maintain a relatively constant performance. The bisquare norm provides an appreciable improvement over the L2 norm in terms of overlap between parameters for the targets considered and so we might expect an improvement in parameter estimation and discrimination when inverting real data contaminated with correlated noise. Consistent with the recommendation in Marrona et al. (2006), we also find that the bisquare provided tighter parameter distributions than the other robust norms considered in table 2.1. Figure 2.8 shows an analogous result to figure 2.4, illustrating how correlated noise skews the L2 estimate while the robust estimate can effectively ignore these contaminated data.

2.6 Application to real data

To study the effect of robust parameter estimation upon discrimination, we consider inversion of three TEM data sets acquired at Camp Sibert, Alabama (Billings, 2008). Data were collected with a Geonics EM61 sensor, the MTADS towed array of three EM61 sensors, and a Geonics EM63 sensor deployed in a cued-interrogation mode. Figure 2.9 shows representative items from target classes encountered at this site. In this case target location is not known, and so we solve a nonlinear inverse problem for target location, orientation, and polarization parameters. For the EM63 data, we parameterize each polarization decay according to equation 2.7, while for EM61 data sets we recover instantaneous polarization amplitudes at each time channel. We find that the nonlinearity of the forward problem produces local, suboptimal minima of the misfit function, and so we select a set of starting models which adequately explore the misfit surface as follows:
2.6. Application to real data

Figure 2.7: Effect of correlated noise upon L2 and bisquare secondary polarization estimates \((L_2, L_3)\) for two targets (dots and crosses). Dots are parameter estimates for an axisymmetric target with true transverse polarizations \(L_2 = 10, \ L_3 = 10\), and crosses are parameter estimates for non-axisymmetric target with \(L_2 = 10, \ L_3 = 5\). True target parameters are shown as open circles. Top row is for targets with primary polarization oriented horizontally, bottom row is for targets with primary polarization oriented vertically.

1. We perform a scan over target location and depth, solving the linear problem for the polarization tensor, as in section 2.5

2. To define \(N\) (e.g. 10) starting models for subsequent nonlinear inversion, we identify the minimum misfit model in each of \(N - 1\) depth intervals, ranging from \(z = 0\) m to \(z = z_{max}\), a maximum target depth (here \(z_{max} = 1.2\) m). An additional \(N^{th}\) starting model is specified as the best fitting model at the surface (\(z = 0\) m).

3. We then solve each of the \(N\) nonlinear inverse problems using the set of starting models, and select the minimum misfit model. The final model is parameterized in terms of target location, orientation and the eigenvalues of the polarization tensor at each time channel.
2.6. Application to real data

Application of a robust norm in this procedure can occur at several stages. For example, we might apply a robust norm to the linear problem in step 1 as well as for the nonlinear problem at step 2. We have tried several approaches and found that the most computationally efficient is to solve both the linear and nonlinear problems first with the L2 norm, and then subsequently apply the bisquare norm at 95% efficiency, using the final L2 models as starting models for the robust inversion. We do not repeat all $N$ inversions with the robust norm, rather we identify the unique models obtained in the initial L2 inversion, where uniqueness is determined by the depth of the target. That is, we identify all final models with similar depths (say, within 5 cm of each other), and then use the best fitting model from each subset to initialize a refit with the robust norm.

Based upon recommended practice in Marrona et al. (2006) for robust linear regression, for each robust refit we also re-scale the uncertainties at each time channel based upon the residuals predicted by the L2 inversions. The scaling $\gamma_i$ on the $i^{th}$ time channel is the median absolute deviation (MADN) of the residuals from the L2 inversion

$$\gamma_i = \text{MADN}(x_i)$$

with $x_i$ as defined in equation 2.14 and

$$\text{MADN}(x) = \frac{1}{0.675} \text{Median}(|x - \text{Median}(x)|).$$

(2.23)

Figure 2.8: Fits (dashed line) to synthetic data (dots) for a single realization of correlated noise. True data is shown as a solid line.

Figure 2.9: Representative items from Camp Sibert target classes. Left to right: 4.2" mortar (UXO class), partial mortar (Partial class), mortar base-plate (Base class), mortar fragment (OE class), cultural debris (Culture class).
The MADN is a robust estimate of the standard deviation, with the normalization factor ensuring that MADN equals one for $x \sim N(0, 1)$ (Marrona et al., 2006). In this application the rescaling ensures that all time channels have approximately equal contributions to the bisquare misfit, at least for the data predicted by the starting model taken from the weighted L2 inversion. Without the scaling we found that later time channels can sometimes be overemphasized by the bisquare norm, leading to an inadequate fit at early times.

Figure 2.10 compares features extracted from Camp Sibert EM63 data using L2 and robust inversions. The feature space selected for discrimination is spanned by parameters related to the amplitude and decay of the induced polarization, both normalized to zero mean and unit standard deviation. The feature space for the robust inversion is qualitatively very similar to that of the L2 inversion. This is perhaps unsurprising, since the data are acquired along tightly spaced lines with robotic total station positioning and so can be expected to yield reliable parameter estimates regardless of the misfit function. To understand the benefit of robust inversion, we highlight three targets in 2.10. Target 1 is a 4.2" mortar which was identified in subsequent processing to have a single line of data with large positional errors (figure 2.11(a)). When these data are included in the L2 inversion, the best fitting model places the target at $z = 0$ m, and the estimated feature vector is an outlier to its class. Robust inversion determines that a better fit exists at a depth of approximately $z = 0.4$ m and the resulting feature vector lies well within the class distribution. In this example robust inversion correctly reweights the data so that the minimum at depth becomes the global minimum (figure 2.11(b)), and a better fit to the observed data is obtained (figure 2.11(c)). Targets 2 and 3 are both scenarios where robust inversion remains in the same neighbourhood as the optimal L2 model. For target 2 (a base-plate), robust inversion moves the feature vector closer to the class mean, whereas for target 3 (a mortar) the feature vector moves slightly away from the class mean. In the case of ordnance related scrap and culture, we do not expect robust inversion to provide an improvement in clustering in the feature space, since these targets are highly physically variable. For targets with consistent physical properties (UXO and base-plates), we find that robust inversion decreases the variance of parameter distributions. Figure 2.10 shows a contour of a quadratic discriminant analysis classifier trained in each feature space to discriminate between 4.2" ordnance items and all other target classes (partials, base plates, etc.). The contour corresponds to the location in the feature space where all ordnance items are identified by the classifier.

A second application of robust inversion is shown in figure 2.12, which considers inversion of MTADS EM61 data acquired in a detection mode survey during the Camp Sibert geophysical prove-out. We again see that robust inversion brings the highlighted outliers closer to their respective class distributions.

Table 2.2 summarizes the discrimination performance improvements gained by applying robust inversion to Camp Sibert TEM data sets. We use a bootstrapping algorithm to estimate metrics of discrimination performance for quadratic discriminant analysis (Beran and Oldenburg, 2008). We have also included results for the EM61 cart data in this table, though for brevity we have not shown the corresponding feature spaces from L2 and robust inversions of these sensor data. We see an improvement in discrimination performance as we move from a detection mode survey with a single sensor
2.7. Conclusions

In this work we have applied robust norms to estimation of parameters from time-domain electromagnetic data. When positional errors are modelled as independent Gaussian perturbations, we find that least squares and robust inversion have comparable performance. Both inversion techniques estimate data uncertainties from observed data, and this has the effect of “robustifying” the least squares inversion. However, when simulated errors are correlated, robust inversion with the bisquare norm provides a marked improvement over L2 inversion. Application of robust inversion to real data sets produced an incremental improvement to the initial L2 inversion, identifying outlying ordnance items and generally tightening parameter distributions in the feature space.

When applying robust norms we assume that the errors on the data are uncorrelated and non-normally distributed. In simulations, however, we found that robust norms can perform well when errors are correlated. An alternative approach to handling correlated errors is to introduce off-diagonal terms in the data covariance. For example, Dosso et al. (2006) iteratively update the data covariance using the data residuals predicted by the current model estimate. This seems very similar to the IRLS procedure employed here, and comparison of these techniques is an avenue for future research.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>L2</th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>FAR</td>
</tr>
<tr>
<td>EM61</td>
<td>0.973</td>
<td>0.297</td>
</tr>
<tr>
<td>MTADS</td>
<td>0.997</td>
<td>0.111</td>
</tr>
<tr>
<td>EM63</td>
<td>0.999</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 2.2: Comparison of bootstrapped performance metrics for robust vs. L2 inversion applied to Camp Sibert TEM datasets.

(EM61) to an array of sensors (MTADS) to a cued-interrogation survey (EM63). In all cases robust inversion with the bisquare norm improves expected discrimination performance versus weighted L2 inversion.

2.7 Conclusions
Figure 2.10: Comparison of features estimated from Camp Sibert EM63 data with L2 and robust norms. Contour indicates output of quadratic discriminant analysis classifier required to find all ordnance items in the feature space.
2.7. Conclusions

(a) Gridded EM63 data. Data locations are shown as points. Line 1 has probable positional errors.

(b) Relative misfit as a function of depth for inversion of observed EM63 data with L2 and robust norms. The relative misfit is computed so that the minimum misfit model has relative misfit equal to zero.

(c) Predicted data from L2 and robust inversions. Circles and squares indicate observed data for lines 1 and 2, respectively.

Figure 2.11: Comparison of fits for ordnance target from Camp Sibert EM63 data.
2.7. Conclusions

Figure 2.12: Comparison of features estimated from Camp Sibert MTADS data with L2 and robust norms. Contour indicates output of quadratic discriminant analysis classifier required to find all ordnance items in the feature space.
Bibliography


Chapter 3

Uncertainty in UXO discrimination

3.1 Introduction

Unexploded ordnance (UXO) discrimination with digital geophysical data is a challenging problem which requires careful application of inversion and decision theory. In the inversion stage, we try to estimate the parameters of a physical model such that the data predicted by a model adequately reproduce the observed data acquired over a buried target. Magnetic and electromagnetic sensors are most commonly deployed for ordnance detection, and for these data types simple dipole models are usually fit to the data. A magnetostatic dipole, parameterized in terms of a cartesian location and dipole moment vector, can be used to predict the anomalous magnetic field observed over a susceptible target. For electromagnetic (EM) data the induced secondary field radiated by a conductive object is often modelled as a superposition of orthogonal dipoles with characteristic decays (or spectra, in the frequency domain). To obtain estimates of these model parameters in an inversion, we minimize a function quantifying the misfit between the observed ($d^{obs}$) and predicted ($d^{pred}$) data. A common choice of misfit function is the least squares data misfit

$$\phi = \frac{1}{2} ||W_{d} (d^{obs} - d^{pred})||^2. \quad (3.1)$$

The data weighting matrix $W_{d}$ quantifies the uncertainty in the data. A typical choice when dealing with geophysical data is to assign a weighting to each datum as a percentage of the observed datum plus a floor value. The noise floor represents a baseline background noise level, and can be estimated from regions of a survey where there are no detected targets.

Parameters of empirical models used for UXO discrimination are proxies for physical properties such as target size and shape. For example, the strength and decay rate of the induced dipole moment are diagnostic of target size and wall-thickness (Shubitidze et al., 2004). As such, the parameters estimated in an inversion can be used to decide whether a target is likely to be an ordnance item. Rule-based approaches can perform well when the discrimination task is relatively simple. For example, when there is a single ordnance type which is much larger than the metallic clutter, shrapnel etc. that we wish to leave in the ground, then the EM induced moment amplitude can be an effective metric for distinguishing ordnance from clutter.

Statistical classification algorithms have also been employed for UXO discrimination. In this approach a subset of model parameters is selected to span a feature space. We also have a training

---

3 A version of this chapter will be submitted for publication. Beran L. and Oldenburg D. Uncertainty in UXO discrimination.
3.1. Introduction

data set comprised of feature vectors for which ground truth is known. Using this training data set we try to formulate a decision rule in the feature space which we then apply to the remaining, unlabelled feature vectors (the test data). A common approach is to fit probability distributions to the training data and then use these distributions to predict the probability that a test feature vector is an ordnance item. Detailed descriptions of the inversion and statistical classification procedures for UXO discrimination can be found in Zhang et al. (2003) and Beran and Oldenburg (2008).

In this article we focus upon the role of parameter uncertainty in inversion and discrimination. Figure 3.1 shows a motivating example for the problem of discrimination with features extracted from time-domain electromagnetic (TEM) data. More details of the forward modelling, inversion and discrimination algorithms used in this example will be provided in section 3.3. The fit between observed

![](image)

Figure 3.1: Fit of a parametric dipole model to time-domain electromagnetic data acquired over a buried ordnance. Top row shows gridded images of observed predicted and residual (observed - predicted) data. Points indicate observation locations, highlighted line is plotted at bottom left, open circle near anomaly maximum is the location of single sounding shown at bottom right.

and predicted data for this target appears quite good, though there are some relatively large residuals for large amplitude data. However, the model parameters obtained in the inversion are far removed from the typical values we obtain for this type of ordnance (figure 3.2). This result highlights that the
3.1. Introduction

TEM parameter estimation problem is fundamentally ill-posed: a small perturbation in the data caused by stochastic noise can produce a large change in the estimated model (ill-conditioning), and multiple models can fit the data equally well (non-uniqueness). We can address ill-posedness in a number of ways. TEM sensors developed specifically for UXO detection and discrimination can better constrain the inverse problem by providing measurements of orthogonal components of the received secondary magnetic field. With single-component TEM data acquired with a production sensor (as in figure 3.1), non-uniqueness can be somewhat alleviated by careful parameterization of the forward model (especially with respect to decay parameters) and by incorporating prior information (e.g. bound constraints) into the inversion (Pasion, 2007). However, there will always be uncertainty in our parameter estimates arising from the approximate nature of the forward model and the noise on the data. In this work we therefore investigate how model uncertainty can be propagated through inversion and into a discrimination algorithm.

![Figure 3.2: Estimated model parameters for ordnance and non-ordnance targets for Camp Sibert EM63 data. Feature vector indicated by arrow corresponds to fit shown in figure 3.1. Crosshairs show the location of the mean of the UXO class.](image)

In section 3.2 we explore two methods of uncertainty appraisal: a local, linearized analysis which approximates the model parameters as Gaussian distributions, and a global, nonlinear appraisal which makes no assumptions regarding the parametric form of the model probability density function (PDF). We find that the two techniques agree for well-constructed experiments where the data quality is high. In section 3.4, we develop a discrimination algorithm which incorporates local uncertainties in the discrimination procedure. Simulations indicate that when parameter uncertainties are unimodal the technique can improve performance relative to conventional algorithms where parameters are treated as point
3.2 Model uncertainty

For a nonlinear forward modelling \( \mathbf{d}^{\text{pred}} = F(\mathbf{m}) \), the misfit can be minimized iteratively by solving for the model perturbation \( \delta \mathbf{m} \)

\[
\mathbf{H} \delta \mathbf{m} = \mathbf{J}^T \mathbf{W}_d^T \mathbf{W}_d \delta \mathbf{d}
\]

with \( \mathbf{H} = \nabla^2 \phi \) the Hessian of the misfit, \( \mathbf{J} \) the Jacobian matrix of sensitivities, and \( \delta \mathbf{d} = (\mathbf{d}^{\text{obs}} - F(\mathbf{m})) \).

At the minimizer \( \mathbf{m} \) we can approximate the model covariance as

\[
\text{cov}(\mathbf{m}) = E(\delta \mathbf{m} \delta \mathbf{m}^T)
\]

\[
= E(\mathbf{H}^{-1} \mathbf{J}^T \mathbf{W}_d^T \mathbf{W}_d \delta \mathbf{d} \delta \mathbf{d}^T \mathbf{W}_d^T \mathbf{W}_d \mathbf{J}^T \mathbf{H}^{-1})
\]

\[
= \mathbf{H}^{-1} \mathbf{J}^T \mathbf{W}_d^T E((\mathbf{W}_d \delta \mathbf{d})(\mathbf{W}_d \delta \mathbf{d})^T) \mathbf{W}_d \mathbf{J}^T \mathbf{H}^{-1}
\]

(3.3)

If the errors on the data are independent and Gaussian and the weightings in \( \mathbf{W}_d \) are the inverse standard deviations of the noise, then the expectation of the terms in the last expression is the identity, so that

\[
\text{cov}(\mathbf{m}) = \mathbf{H}^{-1} \mathbf{J}^T \mathbf{W}_d^T \mathbf{W}_d \mathbf{J}^T \mathbf{H}^{-1}.
\]

(3.4)

The expression can be further simplified by noting that \( \mathbf{H} \approx \mathbf{J}^T \mathbf{W}_d^T \mathbf{W}_d \mathbf{J}^T \), giving us an approximate expression for the covariance of the model parameters

\[
\text{cov}(\mathbf{m}) \approx \mathbf{H}^{-1}.
\]

(3.5)

From this result we see that if there is a large curvature to the misfit function at \( \mathbf{m} \), then the model is well constrained and the variance of the model parameters is small. The probability distribution of the model parameters may then be approximated as a normal PDF with mean \( \mathbf{m} \) and covariance computed with 3.5 (Menke, 1989). The above expression for the model covariance is the Cramer-Rao lower bound for the model uncertainty: the covariance of any unbiased estimator is at least that of equation 3.5 (van Trees, 2001).

Features used in discrimination algorithms are often functions of the model \( \mathbf{m} \). For a transformation \( \hat{\mathbf{m}} = g(\mathbf{m}) \) operating on the model vector, a first order approximation to the transformed model
3.2. Model uncertainty

covariance $S$ is

$$
cov(\hat{m}) = G^T \text{cov}(m) G, $$

with the elements of the sensitivity matrix of the transformation computed as

$$
G_{ij} = \frac{\partial g_i(m)}{\partial m_j}. $$

A linearized uncertainty analysis may not be valid if the objective function is nonconvex. In this case the local quadratic approximation provides a poor approximation to the actual objective function and uncertainties estimated with equation 3.5 may not be reflective of the actual uncertainties in the model. An alternative approach to estimating uncertainties is to use a Bayesian framework to estimate the model posterior probability distribution (PPD)

$$
p(m|d^{\text{obs}}) \propto p(d^{\text{obs}}|m)p(m). $$

The posterior $p(m|d^{\text{obs}})$ is the product of a likelihood function $p(d^{\text{obs}}|m)$ and a prior probability $p(m)$. For continuous models, the likelihood function is often assumed to have the form

$$
p(d^{\text{obs}}|m) \propto \exp[-\phi_d] $$

with $\phi_d$ the least squares misfit function. The posterior probability is then computed as

$$
p(m|d^{\text{obs}}) = \frac{\exp[-\phi_d] p(m)}{\int \exp[-\phi_d] p(m) dm}. $$

Here the normalizing integral is over all of model space. For a nonlinear forward problem, the normalizing integral is often difficult to evaluate analytically, especially in high-dimensional model spaces. However, the PPD for a nonlinear problem can be estimated numerically using the Metropolis-Hastings algorithm. This algorithm works by randomly perturbing the current model $m_{\text{current}}$ to a proposed model $m_{\text{proposed}}$ and accepting the proposed model according to the Metropolis criterion

$$
\eta \leq \exp[-\Delta \phi] = \exp[-(\phi(m_{\text{proposed}}) - \phi(m_{\text{current}}))], $$

If $\exp[-\Delta \phi]$ is less than or equal to $\eta$, then the perturbation is accepted and the proposed model becomes the current model. Perturbations that decrease the misfit function are always accepted, while perturbations that increase $\phi$ are accepted randomly according to the above criterion. At each model perturbation $\eta$ is drawn from a uniform random distribution on the interval $[0, 1]$.

This scheme is a Markov chain; acceptance of the proposed model depends only on the current model. After a sufficient number of samples the chain of accepted models will converge to a stationary distribution which is the posterior distribution. We adopt the fast sampler algorithm developed in Dosso (2003) to estimate the posterior distribution of the model parameters. A key feature of this algorithm
3.3. Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

is the use of two independent samplers. Convergence of these samplers to the same distribution, as measured by the maximum difference in their cumulative distributions, ensures that the sample provides a reasonable estimate of the PPD.

The posterior probability density is a function in an \( N \)-dimensional model space, with \( N \) the number of model parameters. It is therefore useful to consider the one-dimensional marginal distribution of each parameter.

\[
p(m_i|d^{\text{obs}}) = \int p(m|d^{\text{obs}}) dm_1 dm_2 \ldots dm_{i-1} dm_{i+1} \ldots dm_N \quad (3.12)
\]

3.3 Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

As an example of feature vector estimation and uncertainty propagation in the context of unexploded ordnance discrimination, we first consider time-domain electromagnetic data acquired with Geonics EM61 and EM63 sensors at Camp Sibert, Alabama. The discrimination task for this demonstration project was to distinguish between emplaced 4.2" mortars and non-hazardous clutter (munitions related scrap, cultural items, etc.). In this study we model observed data using a dipole forward model (Pasion, 2007). The secondary magnetic field is computed as

\[
B^s(r, t) = \frac{m(t)}{r^3} \left( 3(\mathbf{m}(t) \cdot \hat{r})\hat{r} - \mathbf{m}(t) \right) \quad (3.13)
\]

with \( r \) the separation between target and observation location, and \( \mathbf{m}(t) \) a time-varying dipole moment

\[
\mathbf{m}(t) = \frac{1}{\mu_0} \mathbf{M}(t) \cdot \mathbf{B}_o. \quad (3.14)
\]

The induced dipole is the projection of the primary field \( \mathbf{B}_o \) onto the target’s polarization tensor \( \mathbf{M}(t) \). The polarization tensor can be decomposed as

\[
\mathbf{M}(t) = \mathbf{A}^T \mathbf{L}(t) \mathbf{A} \quad (3.15)
\]

with \( \mathbf{A} \) an orthogonal matrix which rotates the coordinate system from geographic coordinates to a local, body centered coordinate system. In body-centered coordinates the tensor is diagonal

\[
\mathbf{L}(t) = \begin{bmatrix}
L_1(t) & 0 & 0 \\
0 & L_2(t) & 0 \\
0 & 0 & L_3(t)
\end{bmatrix} \quad (3.16)
\]

with the eigenvalues ordered by convention so that \( L_1(t) \geq L_2(t) \geq L_3(t) \). When inverting EM61 data, we estimate a 15 element model vector comprised of target location \((x, y, z)\), orientation \((\phi, \theta, \psi)\), and the instantaneous amplitudes of the polarization tensor at three time channels \((L_i(t_j), i = 1, 2, 3, j = 1, 2, 3)\). In this example we have selected a two-dimensional feature space spanned by the polarization
Polarization amplitude = \( \left( \sum_{i=1}^{3} L_i^2(t_1) \right)^{1/2} \), \( (3.17) \)

Polarization decay = \( \left( \sum_{i=1}^{3} L_i^2(t_m) \right)^{1/2} / \left( \sum_{i=1}^{3} L_i^2(t_1) \right)^{1/2} \). \( (3.18) \)

The polarization amplitude is then proportional to the magnitude of the induced dipole moment at the first time channel, and the polarization decay is the ratio of the magnitude of the induced dipole at the first time channel and some later time \( t_m \). For the EM61, \( t_1 = 0.22 \) ms and we use \( t_m = t_3 = 0.66 \) ms. The EM63 measures the decay of the secondary field at 26 time channels spanning a wider range of off-times and so for these data we can parameterize the decays of the individual polarizations as

\[ L_i(t_j) = k_i t_j^{-b_i} \exp(-t_j/g_i), \quad i = 1, 2, 3. \] \( (3.19) \)

The model vector for inverting EM63 data is then target location, orientation, and polarization parameters \((k_i, b_i, g_i, i = 1, 2, 3)\). The recovered parameters can then be used to compute polarization amplitude and decay using equations 3.17 and 3.18, with \( t_1 = 0.18 \) ms and \( t_m = t_{15} = 2.17 \) ms for the EM63.

Figure 3.3(a) shows the distributions of ordnance and non-ordnance feature vectors for the EM61 data set. In 3.3(b) we zoom in on the region of the feature space populated with ordnance feature vectors and show the linearized uncertainties in the model parameters. By inspection of equations 3.17 and 3.18, we expect that the polarization decay will vary inversely with the polarization amplitude, and this negative correlation appears in both the overall distribution of ordnance items and the linearized uncertainties in many of the feature vectors. This trend is not obvious in the distribution of non-ordnance items because this class is comprised of a number of different target types with variable physical properties. We note also that the variance of individual ordnance targets is generally much smaller than the overall variance of the ordnance class, with little correlation between parameter uncertainty and the deviation of a feature vector from the class mean.

To validate the computation of uncertainties shown in figure 3.3, figure 3.4 compares the variances of model parameters estimated by linearized and nonlinear uncertainty appraisal for a single ordnance item in the Camp Sibert EM61 data. This target is indicated by an arrow in figure 3.3(b). The nonlinear appraisal is run with uniform priors on the model, so that agreement between linearized and nonlinear appraisals is expected. Consistent with results in Pasion (2007) obtained with the neighbourhood sampling algorithm (Sambridge, 1999), parameter uncertainties propagated to the feature vectors using equation 3.6 agree well with those estimated by nonlinear appraisal. The nonlinear marginal PDF of the polarization amplitude has a slight positive skew, resulting in a shift of the mean of this distribution to a slightly larger value than is obtained with the linearized appraisal.

Figure 3.5 shows a second example of model uncertainty for an ordnance item. In this case the distribution of model parameters is a multimodal distribution. Linearized uncertainty analysis is a good
3.3. Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

approximation to one mode of this distribution centered about the minimum misfit model. However, the second mode of the nonlinear model PDF is in better agreement with the expected features for targets belonging to this ordnance class. To further understand this result, in figure 3.6 we show the misfit versus depth (MVD) curve for this inversion. This curve is generated by carrying out multiple inversions of the same data set with each inversion constrained to lie within a narrow interval of target depth (Lhomme et al., 2008). In 3.6(a) we display the relative misfit, i.e. the misfit relative to the global minimum value on the MVD curve. Two minima are identified on the MVD curve and correspond to the modes of the model PDF in figure 3.5. The shallower, local minimum of the MVD curve is a better estimate of true target depth for this case. We see in 3.6(b) that polarization amplitude increases monotonically with depth, so that the optimal model at depth overestimates polarization amplitude. Figure 3.6(c) shows that for models on the MVD curve the polarization decay varies nonlinearly with polarization amplitude. This follows from the nonlinear relationship between these features (equations 3.17 and 3.18) and the nonlinear dependence of both of these features on target depth.

Figure 3.7 shows estimated feature vectors and associated linearized uncertainties for the Camp Sibert EM63 data. Parameter uncertainties are generally smaller for the EM63 than for the EM61. The EM63 data was acquired in high quality cued interrogation surveys over previously-identified targets, and so we expect smaller errors on both the data and the model than for the EM61 detection mode data set.

A similar result to that in figure 3.5 is obtained in figure 3.8 for appraisal of the EM63 target highlighted in 3.7: local minima of the misfit function produce a multimodal PDF. Again, the linearized model PDF is a good approximation to one mode of the nonlinear model PDF. The MVD curve for this target has a false global minimum at a shallow depth, resulting in an underestimated polarization amplitude at the best-fitting model (figure 3.9).

When data are acquired in a controlled experiment (e.g. test stand measurements), then we expect that the global minimum of the misfit function will correspond to the true model. However, when inverting field measurements we generally have only naive estimates of the noise on the observed data, and the weightings (\(W_d\)) we choose have a direct effect on determining the minimum misfit model through equation 3.1. For this study we have assigned errors as 10 percent of each observed datum plus a floor value estimated at each time channel. While this choice generally allows us to recover useful estimates of model parameters from the model parameters, in these last two examples the data are weighted so that a “false” minimum at an incorrect depth becomes the global minimum. From these examples we therefore conclude that

1. the minimum misfit model is not always the best model to use for discrimination,

2. linearized uncertainty appraisal cannot always account for the deviation of an estimated feature vector from the expected value of that target class.

We can address this problem in a number of ways. First, we might investigate different choices of data weightings or data norm. For example in Beran et al. (2009), robust inversion techniques were applied to the same TEM data sets presented here. We found that robust norms can sometimes reweight min-
3.3. Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

ima of the misfit so that the desired minimum (i.e. closest to the true target depth) becomes the global minimum. Robustness, however, does not imply a foolproof inversion procedure and may not always unambiguously eliminate local minima as possible solutions to the inverse problem. Similarly, different choices of noise standard deviations can correctly reweight minima of the misfit, but again no noise estimation procedure is likely to be universally successful. In light of this ambiguity, here we instead investigate how our estimates of model uncertainty (linearized or nonlinear) can be incorporated into the discrimination process. In the next section we develop a technique for incorporating local (linearized) uncertainties in discrimination via a generalization of discriminant analysis. While this approach has the potential to detect outlying feature vectors, for this particular problem linearized error estimates are often a poor characterization of model uncertainty. In section 3.5 we therefore investigate a simple technique for incorporating nonlinear (global) uncertainties in discrimination.
3.3. Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

(a) Estimated feature vectors. Dashed area indicates plotted region in (b) and crosshairs show the location of the mean of the UXO class.

(b) Estimated feature vectors with local standard deviations along principle axes. Principle axes do not appear orthogonal because of unequal scalings of plot axes. Arrows indicate targets appraised in figures 3.4 and 3.5.

Figure 3.3: Feature vectors estimated from Camp Sibert EM61 data.
Figure 3.4: Left: Principle axes of the covariance matrix estimated by linearized and nonlinear uncertainty appraisals of EM61 data for highlighted target in figure 3.3(b). Also shown are all accepted models from nonlinear sampling. Right: marginal parameter distributions from linearized and nonlinear appraisals.

Figure 3.5: Left: Accepted models for nonlinear appraisal of a single target in EM61 data, highlighted in 3.3(b). Crosshairs intersect at the mean of the UXO class. Right: marginal parameter distributions from linearized and nonlinear appraisals.
3.3. Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

Figure 3.6: (a) Misfit versus depth (MVD) curve for EM61 target considered in figure 3.5. Vertical line indicates minimum of misfit which best agrees with actual target depth. (b) Dependence of MVD curve model polarization amplitude on estimated target depth. (c) Trajectory of MVD curve models in the feature space. In all plots markers correspond to models at minima of the MVD curve.

Figure 3.7: Feature vectors and linearized uncertainties estimated from Camp Sibert EM63 data. Arrow indicates target appraised in figure 3.8.
3.3. Comparison of linearized and nonlinear uncertainty appraisals for TEM dipole model parameters

Figure 3.8: Accepted models for nonlinear appraisal of a single target in EM63 data, highlighted in figure 3.7. Crosshairs intersect at the mean of the UXO class. Right: marginal parameter distributions from nonlinear appraisal.

Figure 3.9: (a) Misfit versus depth (MVD) curve for EM63 target considered in figure 3.8. Vertical line indicates minimum of misfit which best agrees with actual target depth (b) Dependence of MVD curve model polarization amplitude on estimated target depth. (c) Trajectory of MVD curve models in the feature space. In all plots markers correspond to models at minima of the MVD curve.
3.4 Incorporating uncertainties in discrimination: local uncertainties

As mentioned previously, feature vectors are normally treated as point estimates when they are input into a discrimination algorithm. Shivaswamy et al. (2006) develop a support vector machine algorithm which accounts for uncertainty in training feature vectors. Uncertainty is characterized by an ellipsoid about each training feature vector and the algorithm tries to find a separating hyperplane such that all points in an ellipsoid lie on the correct side of the hyperplane. This has the effect of rotating the decision boundary to be orthogonal to the direction of maximum uncertainty in the training vectors. However, this method does not account for the uncertainty in test vectors in the prediction stage.

Here we try to account for uncertainties in both test and training data. We adopt a generative approach to classification where we model the probability distributions of classes and then use these probability distributions to make predictions for the test data. In a conventional generative classifier we typically fit a parametric distribution such as a Gaussian to the empirical distribution of feature vectors in each class. Consider a sample of \( N \) feature vectors belonging to the same class. In the absence of uncertainty, the empirical distribution of these data can be represented as a superposition of delta functions

\[
p(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_{i}^{\text{train}})
\]

The mean and variance of this distribution are the sample mean and variance

\[
\mu = E(x) = \frac{1}{N} \sum_{i=1}^{N} x_{i}^{\text{train}}
\]

\[
\sigma^2 = E((x - \mu)^2) = \frac{1}{N-1} \sum_{i=1}^{N} (x_{i}^{\text{train}} - \mu)^2
\]

For the sample variance the normalization must be changed to \( N - 1 \) to obtain an unbiased estimator. When feature vectors are uncertain, the empirical distribution becomes

\[
p(x) = \frac{1}{N} \sum_{i=1}^{N} p_{i}^{\text{train}}(x)
\]

where \( p_{i}^{\text{train}} \) is the uncertainty distribution of the feature vector \( x_{i}^{\text{train}} \). Assuming the \( p_{i}^{\text{train}} \) are normally distributed with means \( x_{i}^{\text{train}} \) and variances \( (\sigma_{i}^{\text{train}})^2 \), the mean and standard deviation of \( p(x) \) are

\[
\mu = E(x) = \frac{1}{N} \sum_{i=1}^{N} x_{i}^{\text{train}},
\]

\[
\sigma^2 = E((x - \mu)^2) = \frac{1}{N-1} \sum_{i=1}^{N} (x_{i}^{\text{train}} - \mu)^2 + \frac{1}{N} \sum (\sigma_{i}^{\text{train}})^2
\]

\[
= \sigma_B^2 + \sigma_L^2,
\]
3.4. Incorporating uncertainties in discrimination: local uncertainties

with the second expression following from the identity $E(x^2) = \sigma^2 + \mu^2$. Again, the normalization on the first term of the variance is modified to obtain an unbiased estimator of the variance of the $x_i^{\text{train}}$. We see that the total variance of the class distribution can be decomposed into contributions arising from the variability between feature vectors ($\sigma_B$) and the local uncertainties about feature vectors ($\sigma_L$). When there is no uncertainty in the feature vectors, we again obtain the expressions in 3.21. This is analogous to the variance decomposition which is employed in analysis of variance (ANOVA) or canonical analysis (Walpole et al., 2007). The estimates of the mean and standard deviation from the above expressions can then be used as the moments of the class distribution. When dealing with multivariate data we can replace the variances in 3.23 with covariances. Figure 3.10 shows estimation of a normal class distribution for varying values of $\sigma_L/\sigma_B$. The estimated normal distribution from equation 3.23 is a good fit to the empirical distribution functions, computed by integrating equation 3.22.

![Figure 3.10: Estimating normal class probability distributions in the presence of local uncertainty $\sigma_L$. Points are empirical cumulative distributions generated from a sample of 100 feature vectors from a univariate normal distribution with zero mean and standard deviation $\sigma_B = 1$. A local uncertainty $\sigma_L$ (the same for all feature vectors in the sample) is used to compute the empirical cdf by integrating equation 3.22 and solid lines are normal cumulative distributions $P(x)$ estimated with equation 3.23.](image)

We now turn to classification of uncertain test feature vectors. When there is no uncertainty in the test feature vectors, a binary discrimination algorithm computes the probabilities of membership of a test vector $x^{\text{test}}$ in the classes (e.g. $T$, $F$) using Bayes rule

$$p(T|x^{\text{test}}) = \frac{p(x^{\text{test}}|T)p(T)}{p(x^{\text{test}}|T)p(T) + p(x^{\text{test}}|F)p(F)}$$

where $p(T|x)$ is the $T$ class distribution estimated from the training data and $p(T)$ is the prior probability of the $T$ class (and similarly for class $F$). When $x^{\text{test}}$ is uncertain with probability distribution $p^{\text{test}}(x)$, we simply evaluate the above expressions over all possible values of $x^{\text{test}}$, weighted by the probability
3.4. Incorporating uncertainties in discrimination: local uncertainties

of each respective value

\[ p(T|x^{test}) = \frac{\int p(x|T)p^{test}(x)p(T)dx}{\int p(x|T)p^{test}(x)p(T)dx + \int p(x|F)p^{test}(x)p(F)dx} \]  

(3.25)

\[ p(F|x^{test}) = 1 - p(T|x^{test}). \]

When there is no uncertainty \( (p^{test}(x) = \delta(x - x^{test})) \), the above expressions reduce to equation 3.24. The required integrals must be evaluated over the entire feature space. For multivariate normal distributions the integral can be solved analytically. A full derivation is given in Brookes (2005), here we summarize the key steps. Consider two normal distributions in an \( M \)-dimensional feature space with means \( \mu_i \) and covariances \( S_i \), i.e.

\[ p_i(x) = \frac{1}{(2\pi)^{M/2}|S_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T S_i^{-1}(x-\mu_i)}, \]  

(3.26)

with \( |S_i| \) denoting the determinant of the covariance. The product is

\[ p_1(x)p_2(x) = \frac{1}{(2\pi)^{M}|S_1|^{1/2}|S_2|^{1/2}} e^{-\frac{1}{2}((x-\mu_1)^T S_1^{-1}(x-\mu_1)+(x-\mu_2)^T S_2^{-1}(x-\mu_2))} \]  

(3.27)

Completing the square gives

\[ p_1(x)p_2(x) = \frac{1}{(2\pi)^{M}|S_1|^{1/2}|S_2|^{1/2}} e^{-\frac{1}{2}(x-\mu_p)^T S_p^{-1}(x-\mu_p)} e^{-\frac{1}{2}(\mu_1-\mu_2)^T (S_1+S_2)^{-1}(\mu_1-\mu_2)} \]  

(3.28)

with

\[ \mu_p = S_1^{-1}\mu_1 + S_2^{-1}\mu_2 \]  

(3.29)

and

\[ S_p = (S_1^{-1} + S_2^{-1})^{-1}. \]  

(3.30)

Integrating equation 3.28 with respect to \( x \) yields

\[ I = \int_{-\infty}^{\infty} p_1(x)p_2(x)dx \]  

(3.31)

\[ = \frac{1}{(2\pi)^{M/2}|S_1 + S_2|^{1/2}} e^{-\frac{1}{2}(\mu_1-\mu_2)^T (S_1+S_2)^{-1}(\mu_1-\mu_2)} \]

The integral of two Gaussians is itself a Gaussian distribution. For this reason, we term a classifier employing equation 3.25 with Gaussian distributions for both classes and feature vectors a “Gaussian product” (GP) classifier. Figure 3.11 shows the integral \( I \) from the above equation for two probability distributions at a fixed separation. As the variance of one of the distributions is increased, \( I \) increases until it reaches a maximum value. This maximum occurs when the squared distance between means equals the total variance of the feature vectors.

The nonlinearity of the Gaussian product produces some interesting behaviour when classifying
3.4. Incorporating uncertainties in discrimination: local uncertainties

Figure 3.11: Top: two normal distributions with fixed separation. The variance of one of the vectors (solid line) is successively increased to produce the bottom curve, with points a, b, and c showing the pdfs which generate corresponding values of \( I = \int p_1(x)p_2(x)dx \). The second distribution (dashed line) has constant variance. Bottom: dependence of \( I \) on the separation between means relative to variances.

uncertain feature vectors. Figure 3.12 shows three scenarios for classification in one dimension. In example (a) the test vector is equidistant from two identical class distributions. Regardless of the standard deviation of the test vector, there is an equal probability of membership in the two classes. In example (b), the test vector is again equidistant between the two classes, but \( p(x|F) \) has a standard deviation which is five times larger than that of \( p(x|T) \). In this case an \( x^{test} \) with no uncertainty will have \( p(T|x) < 0.5 \). This is true for the Gaussian product for \( \sigma(x^{test}) \) small (less than one), but as the uncertainty in \( x^{test} \) is increased \( I(p(x|T),p^{test}(x)) \) exceeds \( I(p(x|F),p^{test}(x)) \) such that \( p(T|x) > 0.5 \). This suggests that if there is an outlying test vector belonging to the \( T \) class with an uncertainty such that \( I(p(x|T),p^{test}(x)) \) is maximized (or nearly maximized), then the Gaussian product will detect that outlier sooner than a conventional classifier. However if \( \sigma(x^{test}) \) is large relative to the separation between \( x^{test} \) and the class means, then the Gaussian product becomes uncertain about the class of \( x^{test} \) and so \( p(T|x) \) asymptotes to 0.5. This is demonstrated in example (c), where the test vector mean is shifted slightly closer to the \( T \) distribution. Again \( p(T|x) \) attains a maximum and thereafter asymptotes to 0.5.

In higher dimensional feature spaces, the covariance of a test feature vector must be considered. The relative sizes of the variances along principal directions of the covariance affect the orientation of
3.4. Incorporating uncertainties in discrimination: local uncertainties

Figure 3.12: Classification with the Gaussian product in one dimension. Top row: class distributions $p(x|T)$ and $p(x|F)$ and location of test feature vector $x^{test}$. Middle row: corresponding values of the integral $I$ as the standard deviation of the test vector $\sigma(x^{test})$ is varied. Bottom row: corresponding probability of membership in $T$ for Gaussian product classifier for varying $\sigma(x^{test})$. Equal priors are used in all examples.

the decision boundary (figure 3.13). The decision surface for a classifier is generated by evaluating the maximum probability output by the classifier

$$p_{max}(x) = \max(p(T|x), p(F|x))$$

over a grid of points. This generates an image of the classifier output in the feature space, with $P(T|x) = P(F|x) = 0.5$ corresponding to the decision boundary. For the Gaussian product classifier, a separate decision surface is generated for each possible test vector covariance. In figure 3.13 we show three decision surfaces corresponding to three different specifications of test vector covariances. In (a), the covariance is isotropic and so a linear decision boundary, identical to that generated when there is no uncertainty, is obtained. In (b) and (c) we successively increase the variance of the test vector along each orthogonal direction in the feature space. This rotates the decision boundary
3.4. Incorporating uncertainties in discrimination: local uncertainties

such that it closer to orthogonal to the direction of smallest test vector uncertainty. This implies that when we make classification decisions for an individual test vector, we rely most upon the element in that vector which is least uncertain. The rotation of the decision boundary shown here is similar to the result in Shivaswamy et al. (2006): their support vector machine formulation rotates the decision boundary so that the covariance ellipses of the training data lie on the correct side of the boundary. In this case, however, the orientation of the boundary is determined by uncertainties in both test and training data.

Figure 3.13: Classification with the Gaussian product in two dimensions. (a) Two class discrimination problem with one standard deviation ellipses of classes shown as solid and dashed lines. Grayscale image is the decision surface of the Gaussian product classifier for a test feature vector with equal standard deviations along principal axes (dotted line). (b) As in (a) but with the decision surface computed for a test feature vector with $\sigma_{x1} = 2\sigma_{x2}$, as indicated by dotted line. (c) Decision surface for $\sigma_{x2} = 2\sigma_{x1}$.

3.4.1 Simulations

We conclude our development of the Gaussian product with simulations comparing its performance with that of conventional classifier. We consider a two-class discrimination problem in a two-dimensional feature space and simulate feature vector uncertainty as follows

1. Each feature vector is drawn as a sample from a normal distribution.

2. A local standard deviation $\sigma_L$ is then computed for each feature vector as a fraction $\gamma$ of the deviation of the feature vector from its class mean, so that feature vectors which are farther from their assigned class are more uncertain. Since the expected deviation of a feature vector from its class mean is $\sigma_B$, the class standard deviation, we have $E(\sigma_L) = \gamma \sigma_B$

3. Additional local uncertainty is then added to the feature vector as a normally distributed random variable with the standard deviation $\sigma_L$ computed in the previous step.
3.4. Incorporating uncertainties in discrimination: local uncertainties

We generate training and test data sets in this manner, and then apply Gaussian product (GP) and quadratic discriminant analysis (QDA) classifiers to each realization, with both classifiers evaluated on the same set of realizations. The QDA classifier represents class distributions as normal probability distributions with means and covariances estimated from the training data (Hastie et al., 2001). In the case that there is no uncertainty in the test or training feature vectors, the GP classifier with class distributions estimated using equation 3.23 is identical to QDA.

To simulate scenarios where there are outlying feature vectors in the test data, we restrict the training distributions such that no training feature vector exceeds two standard deviations from its class mean. The test feature vectors are not restricted in this way, so we expect that there will be larger deviations from the class means in the test data realizations. Table 3.1 summarizes mean performance metrics for the two classifiers for increasing values of the parameter $\gamma$. The area under the receiver operating curve (AUC) is the integral of the ROC, with a larger value indicating better discrimination performance. The false alarm rate (FAR) is the proportion of $F$ instances (e.g. clutter) which must be labelled in order to find all $T$ (e.g. UXO) instances, with better performance correlated with decreased FAR.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>QDA AUC</th>
<th>QDA FAR</th>
<th>GP AUC</th>
<th>GP FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.9753 ± 0.0084</td>
<td>0.3835 ± 0.1452</td>
<td>0.9753 ± 0.0083</td>
<td>0.3843 ± 0.1453</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9742 ± 0.0095</td>
<td>0.4195 ± 0.1614</td>
<td>0.9760 ± 0.0088</td>
<td>0.3790 ± 0.1391</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9709 ± 0.0105</td>
<td>0.4749 ± 0.1836</td>
<td>0.9757 ± 0.0086</td>
<td>0.3685 ± 0.1294</td>
</tr>
</tbody>
</table>

Table 3.1: Comparison of area under receiver operating curve (AUC) and false alarm rate (FAR) for quadratic discriminant analysis (QDA) and Gaussian product (GP) classifiers. Simulations are generated with 100 realizations of test and training data with varying values of $\gamma = E(\sigma_L) / \sigma_B$, where $\sigma_L$ is the expected local standard deviation of feature vectors and $\sigma_B$ is the class standard deviation.

With the exception of the simulations for $\gamma = 0.01$, the mean performance metrics for the classifiers in table 3.1 are all significantly different$^4$. We conclude that when the local uncertainty in feature vectors is small relative to the class uncertainty, then the Gaussian product performs as a conventional generative classifier. However, when the local uncertainty for outlying feature vectors is large relative to the class uncertainty, then incorporation of uncertainties via the GP can significantly improve discrimination performance. Figure 3.14 shows two realizations of data for these simulations, corresponding to the largest differences in false alarm rate observed for the two classifiers over 100 realizations when $\gamma = 0.2$. In the best case, QDA provides a negligible reduction in false alarm rate. In the best case for the GP, however, we see a significant reduction in false alarm rate.

3.4.2 Application to real data

Figures 3.15 and 3.16 show application of the Gaussian product and quadratic discriminant analysis classifiers to the Camp Sibert EM61 and EM63 data sets discussed in section 3.3. To test the ability

---

$^4$paired t-test, $\alpha = 0.01$. This assumes that the AUC and FAR are approximately normally distributed.
of our discrimination algorithms to detect outliers, for the EM61 data we have generated the training data as a random sample of 20 percent of all feature vectors, with training samples from the UXO class restricted to lie within a 2 standard deviation confidence ellipse of the class mean. This ensures that outliers occur only in the test data. For the EM63 data there is only one outlier to the ordnance class, and so here we have used the actual test and training data from the Camp Sibert study (Billings, 2008).

The Gaussian product provides a reduction in false alarm rate relative to QDA for the EM61 data, with one outlier detected much earlier by the GP. This outlier is the test feature vector with the largest polarization decay in figure 3.15 and is a deep, lower SNR ordnance item. While incorporation of uncertainty is beneficial in this case, we remark that this particular outlier is also far from the distribution of clutter. More sophisticated classifiers (e.g. a neural network or support vector machine) have
3.4. Incorporating uncertainties in discrimination: local uncertainties

Both GP and QDA classifiers have identical performance when applied to the EM63 data: incorporating feature vector uncertainty does not help detect the one outlying ordnance target in this example. This result is consistent with our simulations in the previous section: when the uncertainty in individual feature vectors is small relative to their mean deviation from the class mean, then the GP classifier should give the same result as QDA. In two or more dimensions, we can generalize the ratio of local uncertainty to class uncertainty \( \gamma = \sigma_L/\sigma_B \) as

\[
\gamma = \frac{|S_L|^{1/2}}{|S_B|^{1/2}}
\]  

(3.33)

with \( S_L \) the mean covariance of all feature vectors, and \( S_B \) the mean of the class covariances. For the EM61 and EM63 data sets, \( \gamma \) is approximately 0.02 and 0.001, respectively. This indicates that local feature vector uncertainty has a negligible contribution to the total variability of feature vectors within a class.

A possible explanation for this result is that we have underestimated data uncertainties, so that model uncertainty, which varies linearly with data variance (equation 3.5), is also underestimated. If errors are correctly estimated, then the expected value of the \( i^{th} \) data residual \( (|d_{i}^{obs} - d_{i}^{pred}|) \) is the estimated standard deviation \( \sigma_i \). Furthermore, the data misfit \( \phi \) is a \( \chi^2 \) random variable with \( N \) degrees of freedom, where \( N \) is the number of observed data used in the inversion. Figure 3.17 compares data residuals and estimated standard deviations for the outlying EM63 target highlighted in figure 3.2. For low amplitude data the data residuals are typically smaller than the estimated standard deviations, suggesting that in this case we have slightly overestimated the floor noise level. Because the majority of the data are low amplitude data, the obtained misfit \( \phi \approx 5575 \) is significantly less\(^5\) than the expected misfit of \( E(\phi) = 6000 \) for this data set. However, we note that for some large amplitude data the data residuals are consistently larger than their estimated standard deviations. This suggests that for these data we may have underestimated data errors. We can update the estimated standard deviations as

\[
\sigma_i = \max(\sigma_i, |d_{i}^{obs} - d_{i}^{pred}|).
\]  

(3.34)

We take the maximum of the residuals and the initial estimate to obtain updated error estimates which will tend to err conservatively, i.e. overestimate model uncertainty. Recomputing model uncertainty with the updated data uncertainties, we find that the overall variance of feature vectors, as quantified by the determinant of the model covariance, increases by approximately ten percent for this target. This is not sufficient to account for the deviation of this feature vector from its class: application of the GP classifier with uncertainties updated with equation 3.34 produces no change in classification performance to that shown in figure 3.16. A similar result is obtained for the EM61 data. We conclude that while initial error estimates are at best only a rough estimate, they are not greatly underestimated so as to produce model uncertainties which are a negligible proportion of feature vector variability. As demonstrated in

---

\(^5\) \( \alpha = 0.01 \), one-tailed \( \chi^2 \) test
section 3.3, linearized uncertainties are sometimes only an approximation to one mode of the nonlinear model PDF.

![Figure 3.15: Application of Gaussian product (GP) and quadratic discriminant analysis (QDA) classifiers to Camp Sibert EM61 data. Decision surface is for the GP classifier generated with the mean test vector covariance.](image1)

![Figure 3.16: Application of Gaussian product (GP) and quadratic discriminant analysis (QDA) classifiers to Camp Sibert EM63 data. Decision surface is for the GP classifier generated with the mean test vector covariance.](image2)

### 3.5 Incorporating uncertainties in discrimination: global uncertainties

How can global model uncertainty, characterized by a potentially multimodal PDF, be incorporated into discrimination? One approach is to carry out nonlinear appraisals on all targets and then apply the
3.5. Incorporating uncertainties in discrimination: global uncertainties

Gaussian product classifier to the resulting distributions. If the resulting distributions are multimodal, then they can be approximated as mixtures of Gaussian kernels and the Gaussian product can be evaluated on each element in the mixture. In this work we propose a simpler solution: we identify all minima of the misfit function for a given target, and use all models at minima to make discrimination decisions for that target. That is, when prioritizing targets in a diglist, we dig a target on the basis of the feature vector which is most likely to be an ordnance item, regardless of whether that feature vector achieves the minimum misfit. This avoids the significant computational burden posed by nonlinear appraisal for large numbers of targets.

To obtain an ensemble of possible models for each target, we repeatedly minimize the misfit with an iterative algorithm. We initialize these inversions with a number of models selected over a range of possible target depths. Figure 3.18(a) shows EM61 test data generated with this approach. To define the feature vectors for test items we consider the set of converged models from our repeated iterative inversions. If a subset of these models is within a range of 5 cm in depth, then we select the minimum misfit model from this subset to use for discrimination. Most targets have at least two unique feature vectors identified in this manner, from a total of 10 converged inversions per target. In figure 3.18(b) we consider the performance of four classifiers

1. QDA single: QDA applied to minimum misfit feature vectors
2. GP single: GP applied to minimum misfit feature vectors
3. QDA multiple: QDA applied to multiple feature vectors
4. GP multiple: GP applied to multiple feature vectors. For this classifier the Gaussian product is evaluated using the linearized uncertainty estimated about each recovered model in the ensemble of models for a target. The target is then classified on the basis of the feature vector which is most probably a UXO (i.e. has maximal Gaussian product with the UXO class).

Figure 3.17: Initial estimates of data uncertainties (left) and absolute data residuals (right) for EM63 data acquired over an ordnance.
As seen already in figure 3.15, the GP is able to detect an outlying ordnance feature vector earlier than QDA, and so has a reduced false alarm rate. Using multiple feature vectors improves the performance of both QDA and GP in the early part of the ROC curve, but also requires QDA to dig a slightly larger proportion of clutter. However, applying the GP multiple technique again detects outlying feature vectors and produces the best classification performance for this test data set. Table 3.2 summarizes performance metrics obtained with the various techniques for these data.

Table 3.2: Performance of classifiers on EM61 test data.

<table>
<thead>
<tr>
<th></th>
<th>QDA single</th>
<th>GP single</th>
<th>QDA multiple</th>
<th>GP multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.986</td>
<td>0.988</td>
<td>0.990</td>
<td>0.994</td>
</tr>
<tr>
<td>FAR</td>
<td>0.350</td>
<td>0.131</td>
<td>0.442</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Figure 3.19(a) shows EM63 test feature vectors employing multiple feature vectors. In this case, classification with multiple feature vectors using either QDA or GP classifiers yields perfect classification performance (figure 3.19(b)).

### 3.6 Conclusions

In this work we have investigated uncertainty in dipole model parameters estimated from time-domain electromagnetic data. We found that linearized uncertainty estimates are a good approximation to the distribution of model parameters about local minima. However, in many cases the data misfit has multiple minima, and a linearized uncertainty appraisal about the global minimum cannot fully capture the variability of the model. A nonlinear appraisal with Markov sampling algorithms was used then to explore model space and to estimate a multimodal model pdf. The non-Gaussian form of the model probability density may have implications for sensor and survey design. For example, Smith and Morrison (2005) minimized local uncertainty of dipole model parameters as a function of receiver placement and orientation. This analysis could be repeated with an aim to eliminating local minima of the misfit versus depth curve, thereby improving parameter estimation and subsequent discrimination. Initial results with multistatic systems which can measure multiple components of the secondary magnetic field already suggest that these systems overcome many of the complications encountered here with monostatic sensors.

To account for model uncertainty in discrimination, we developed the Gaussian product classifier. This algorithm is a generalization of discriminant analysis which incorporates uncertainty in both test and training feature vectors. Simulations showed that the Gaussian product classifier can improve discrimination performance by finding outlying test feature vectors which have large uncertainties. Application of this technique to Camp Sibert EM61 data did provide a slight improvement in discrimination performance relative to quadratic discriminant analysis. In addition, repeated iterative inversions with different initializations can identify modes of the model probability density function. We used this latter strategy to generate an ensemble of possible models for each target, and then input these models into a discrimination algorithm. For the Camp Sibert EM61 data, the multiple feature vector approach
in conjunction with the Gaussian product classifier yielded the best classification performance for all techniques considered on this data set. Inputting multiple feature vectors into either discriminant analysis or the Gaussian product classifier found a previously undetectable outlier and produced perfect discrimination performance for the Camp Sibert EM63 data set.

Aliamiri et al. (2007) investigated dipole parameter uncertainty by forward modelling data for a range of target locations and orientations, and then inverted these synthetic data to obtain expected parameter distributions for a given target class in the feature space. These class distributions were non-Gaussian, and so a non-parametric model of class distributions was used for discrimination. In this study we have used discriminant analysis (or its Gaussian product generalization). This algorithm assumes that class distributions are normally-distributed. While this assumption is certainly not true for the distribution of clutter in the feature space, discriminant analysis can often be successfully employed on non-Gaussian data (Hastie et al., 2001) and gave good results for the data sets considered here. There is no difficulty, however, in generalizing the Gaussian product classifier to other distributions which we may represent as mixtures of Gaussian kernels. The concept of using multiple feature vectors when classifying test data can be regarded as an extension of the technique in Aliamiri et al. (2007) to the test data. While synthetics can augment training data and help to gain a sense of the variability of features, it is impossible to fully anticipate the complications (i.e. noise) which will be encountered in field data. By allowing a test target to be represented by a number of possible models, we can characterize the model conditional upon the observed data and possibly detect outliers which are not captured in the training data.
3.6. Conclusions

(a) EM61 test data, showing multiple feature vectors per target. Feature vectors for a given target are connected by a line, with the largest marker indicating the minimum misfit model.

(b) Receiver operating characteristics for classifiers applied to EM61 test data.

Figure 3.18: Left: ROC curves for Gaussian product (GP) and quadratic discriminant analysis (QDA) applied to EM61 test data with a single feature vector for each target. Right: ROC curves for classifiers with multiple feature vectors for each target, as shown in (a).
3.6. Conclusions

(a) Feature vectors estimated from Camp Sibert EM63 data. Feature vectors for a given target are connected by a line, with the largest marker indicating the minimum misfit model. Highlighted feature vector is an outlier to the ordnance class.

(b) Receiver operating characteristics for classifiers applied to EM63 test data.

Figure 3.19: Discrimination with EM63 data using multiple feature vectors.
Bibliography


Chapter 4

Selecting a discrimination algorithm

The extent of unexploded ordnance (UXO) contamination within the United States and abroad has motivated research into improved technologies for detection and discrimination of UXO. Discrimination algorithms are expected to lower remediation costs by reducing the number of clutter items (geology, shrapnel, etc.) which must be excavated, while ensuring that all munitions of concern are identified. Classification of ordnance type may also be a priority when particular items must be excavated and disposed of with extra caution (e.g. chemical munitions).

Advanced discrimination requires acquisition of digital geophysical data. The current industry standard is time-domain electromagnetic data, typically acquired with a towed array of sensors. Simple criteria, such as signal amplitude, are often used in production settings to prioritize detected targets for digging (SERDP-ESTCP, 2006). Anomaly amplitude is easily extracted from the observed data, and can be quite effective when the site only requires identification of large ordnance (e.g. 100 lb bombs). However, when ordnance and clutter are of comparable size, anomaly amplitude is not a particularly robust parameter for discrimination. For example, UXO items at depth may produce a comparable anomaly amplitude to shallower clutter items.

In contrast, a wide variety of discrimination algorithms has been proposed by researchers and these algorithms have been shown to outperform simple amplitude thresholding in many cases. However, many of these algorithms require intensive data processing, specialized knowledge and user experience. For example, training a neural network to discriminate between UXO and clutter is by no means an automatic process. The design of the network and its inputs must be specified by the user, and the results can vary widely depending on the particulars of the training process.

While comparisons of discrimination algorithms abound (e.g. Zhang et al. (2004), Hart et al. (2001)), no single algorithm is a general panacea for UXO discrimination. A given discrimination algorithm might be suitable in a certain context, but it is unlikely to generalize well to all situations. Indeed, one abiding lesson which can be drawn from reviews of pattern recognition research is that there is no single best algorithm for discerning patterns amongst data (Jain et al., 2000).

The remainder of this paper is organized as follows. First, we provide a review of algorithms which have been used for UXO discrimination. Secondly, in light of the uncertainty over which discrimination approach will work best at a given site, we compare two metrics of classifier performance: the area under the receiver operating curve (AUC) and the false alarm rate (FAR). We propose a bootstrapping approach for estimating these metrics as digging proceeds and show how these estimates can be used

---

to select a discrimination algorithm when there is limited groundtruth. Finally, we demonstrate applica-
tions of this approach to UXO discrimination with both time-domain electromagnetic and magnetic data
sets.

4.1 Advanced discrimination: a review

In this section we describe the processing required to apply advanced discrimination algorithms to UXO
problems. Typically magnetic and/or electromagnetic (either time or frequency-domain) data are col-
lected. Once anomalies have been identified in the observed data, we can characterize each anomaly
by estimating features which will allow our discrimination algorithm to discern UXO from clutter. These
features may be directly related to the observed data (e.g. anomaly amplitude at the first time channel),
or they may be model parameters which must be estimated via inversion.

The inverse problem in this context is overdetermined: its solution requires minimization of a data
misfit function. This is not to say that parameter estimation is straightforward; most problems are
nonlinear and so can present difficulties with local minima. While the focus of this paper is on using
model parameters for discrimination, we emphasize here that parameter estimation, with the attendant
complications presented by real data (noise, overlapping targets, etc.), is crucial to the success of any
discrimination algorithm. A discrimination algorithm can only be as good as our ability to extract useful
parameter estimates from the observed data.

In the case of magnetics data, the forward model is a dipole parameterized by its location, orienta-
tion and strength. Figure 4.1 shows an example fit to observed magnetics data obtained using a dipole
forward model. For electromagnetic data, empirical models are used to forward model the secondary
fields produced by an arbitrary conductive body. For example, the Pasion-Oldenburg model represents
a conductor as a superposition of orthogonal dipoles which decay independently in time (Pasion and
Oldenburg, 2001). This model can faithfully reproduce observed time-domain data in many situations.
The parameters of this model include location, orientation and polarization parameters. The polariza-
tion parameters serve as proxies for target size, shape and material composition and hence can be
used for discrimination. A number of electromagnetic models which account for non-dipolar effects
and material heterogeneity are also being investigated. For the purposes of discrimination, a model is
useful if its parameters allow us to reliably discriminate between UXO and clutter.

Given the model parameters obtained from inversion, we must decide whether a target is likely
to be a UXO. A common approach is to use the model parameters estimated via inversion as basis
vectors in an \( M \)-dimensional feature space. A discrimination algorithm is then a function with a domain
spanning the feature space. The value of the function may signify the probability that a given feature
vector is a UXO. However, in general the particular value of this function is of secondary importance
to the ordering of targets provided by the function. In a production setting we are primarily interested
in providing explosive ordnance disposal technicians in the field with a prioritized diglist. In the next
section, we describe some important methodologies for generating this diglist.
4.1. Advanced discrimination: a review

Figure 4.1: Feature estimation for magnetic data showing observed data in plan view (left), data predicted with a dipole model (middle), and residual (observed minus predicted, right). Points indicate data locations used in inversion.

4.1.1 Statistical classification

UXO discrimination has been treated as a supervised learning problem. Supervised statistical classification makes discrimination decisions for a test data set for which labels are unknown. The classifier performance is optimized using a training data set for which labels are known. Given training and test data sets, the goal of a statistical classifier is then to find an optimal partition of the feature space. Here optimality can be defined by minimizing the probability of misclassifying a test feature vector (Hastie et al., 2001).

In this framework, the problem of discriminating between UXO and clutter is in fact a classification problem. This means that we assume there are two classes (UXO and clutter) and assign all test feature vectors to one of these two classes. This implies that the clutter class is comprised of items with consistent physical properties. For example, at some sites small munitions (e.g. 20mm) may be safely left in the ground and can be considered as clutter. In this case we would expect the clutter to have a distribution of size-related features which is distinct from that of larger munitions comprising the UXO class. However, in other cases clutter may encompass geology, garbage (metal cans, barbed wire, etc.), or shrapnel. Lumping these targets into one class can be detrimental to the discrimination task, as there is likely no consistency in the model parameters which are used as proxies for physical properties. In particular, the overlap between UXO and clutter classes in the training data can result in very poor generalization to the test data.

With this important caveat in mind, we can take two approaches to formulating a statistical decision rule. A generative algorithm seeks to model the underlying distributions which produced the observed
data, often assuming a parametric distribution such as the Gaussian. A discriminative algorithm is not concerned with underlying distributions but rather seeks to identify decision boundaries which provide an optimal separation of classes (Ng and Jordan, 2002).

**Generative classifiers**

The starting point for any generative classifier is Bayes’ rule

\[ P(\omega_i|x) \propto P(x|\omega_i)P(\omega_i). \]  

(4.1)

The likelihood \( P(x|\omega_i) \) is the probability of observing the feature vector \( x \) given the class \( \omega_i \). The prior probability \( P(\omega_i) \) quantifies our expectation of how likely we are to observe class \( \omega_i \) before (ie prior to) observing any feature vector data. Bayes’ rule translates the prior probability into a posterior probability \( P(\omega_i|x) \). The posterior is the probability that we have observed class \( \omega_i \) given the observed feature vector.

Application of Bayes’ rule to classification requires knowledge of the prior probabilities and the form of the likelihood function. The likelihood function can take either a parametric or nonparametric form. The parametric approach assumes a probability distribution for each class and tries to estimate the parameters of these distributions from the training data. The most common parametric classifier is discriminant analysis, which assumes a Gaussian form for the likelihood function. To implement this classifier, we estimate the mean and covariance of each class (UXO and clutter) in the feature space. Quadratic discriminant analysis computes a separate covariance for each class. In this case the decision boundary is a quadratic function in the feature space. Alternatively, if the same class covariance is assumed equal for all classes, then the classifier produces a linear decision boundary in the feature space (linear discriminant analysis).

A parametric classifier which was used in Zhang et al. (2003) for UXO discrimination is the Gaussian likelihood ratio. This classifier considers the ratio of posterior probabilities

\[ \lambda = \frac{P(x|\omega_1)P(\omega_1)}{P(x|\omega_2)P(\omega_2)} \]  

(4.2)

so that \( \lambda = 1 \) corresponds to a feature vector \( x \) on the decision boundary between classes \( \omega_1 \) and \( \omega_2 \). This classifier is a reformulation of discriminant analysis (either linear or quadratic). In the Bayesian framework, prior distributions play a central role: they quantify our subjective expectations. When Bayes’ rule is used in the form given in equation 4.1, the prior probabilities weight the relative importance of classes. However, in equation 4.2 we see that the ratio of prior probabilities is a constant multiplicative factor and so the ordering of test vectors from most to least probably UXO is independent of the particular choice of prior probabilities.

In Barrow and Nelson (2001), the authors define a class distribution for 81 mm mortars in a feature space spanned by instantaneous polarizations. These polarizations are estimated from MTADS electromagnetic data acquired at a single time channel. The recovered distribution of polarizations follows a
4.1. Advanced discrimination: a review

Figure 4.2: Support vector machine formulation for linearly separable feature data belonging to two classes (squares and circles) in a two-dimensional feature space. The classifier tries to maximize the margin between support planes, subject to the constraint that the feature data are classified correctly.

lognormal distribution, and so the class distribution is Gaussian with respect to the logarithm of the polarizations. By thresholding on contours of equal standard deviation of this distribution, a prioritized dig list is generated which significantly improves upon the performance of simple amplitude thresholding.

Assuming a parametric form for the likelihood function greatly simplifies the problem of estimating class distributions. However, this assumption may be difficult to justify if limited training data are available. In this situation, we may turn to nonparametric methods, which define a likelihood function directly from the training data. A representative nonparametric classifier which has been applied to UXO discrimination is the probabilistic neural network (PNN). This classifier represents the class distributions as superpositions of Gaussian "kernels", with each kernel centered on a feature vector in the training data (Hart et al., 2001).

**Discriminative classifiers**

Instead of estimating posterior probability distributions, discriminative classifiers directly define a decision boundary to classify test data. Finding a decision boundary which separates the training data and generalizes well to the test data can be approached as a constrained optimization problem. A commonly-used classifier of this form is the support vector machine (SVM). The basic idea of this classifier is illustrated in figure 4.2. We maximize the margin between classes, subject to the constraint that the training data are classified correctly (Burges, 1998). The margin is defined as the perpendicular distance between support planes. As shown in figure 4.2, a support plane is a line (or plane in higher dimensions) such that all feature vectors in a class fall to one side of that line.

A more general formulation of the SVM allows for nonlinear decision boundaries with overlapping
classes. The idea is to map the feature data to a higher-dimensional space where the training data become separable. We then construct the decision boundary in this space (Hastie et al., 2001).

In UXO applications, the SVM was used by Zhang et al. (2003) to discriminate with features derived from both electromagnetic and magnetic data. In this case, a nonlinear support vector machine trained on EM features outperformed discriminant analysis applied to the same feature space. In Zhang et al. (submitted), an SVM and a neural network had comparable performance when classifying targets with EM features estimated from synthetic data.

### 4.1.2 Library methods

As mentioned previously, statistical classifiers require training data for both UXO and clutter in order to make discrimination decisions. If clutter is highly variable, defining a clutter class may not be sensible. In this case, we can make discrimination decisions using UXO feature vectors in the training data, with no assertions made about the distribution of clutter items in the feature space. This approach encompasses library methods, which try to match estimated features to a predefined library of features for ordnance items. An example of a library-based method for magnetics data is the remanence classifier developed by Billings (2004). Here the estimated dipole moment is matched to a “feasibility curve” which defines the range of dipole moments a particular ordnance type can produce. The degree to which an estimated dipole moment matches the feasibility curve is an indication of how likely the target is to be an ordnance item. In this case the distribution of UXO can be computed analytically.

A library of features for various UXO types can also be defined using previously-acquired training data. Norton and Won (2001) applied this idea to the discrimination of ordnance using GEM-3 frequency domain data. They recover estimates of “orientation invariants” (eigenvalues of the polarization tensor) from observed data at each frequency. The spectrum of eigenvalues is then compared to a library of eigenvalues for known UXO using an L2 norm.

A similar approach was implemented by Pasion et al. (2007) for discrimination with time-domain EM data. They used EM measurements made on a test stand to determine polarization parameters for UXOs. To classify an anomaly as a particular ordnance type, the observed data are fit with the polarization parameters fixed at their library values (i.e. only target location and orientation are estimated in the inversion). The data misfit is then a feature which can be used to discriminate between UXO (low misfit) and clutter (high misfit).

Hu et al. (2004) apply a library method for the difficult problem of discrimination in multi-object scenarios. They represent observed frequency domain electromagnetic data as a linear combination of the data predicted by a specified number of unknown subsurface sources. They then apply independent components analysis (ICA) to determine the mixture of sources from a predefined library which can best reproduce the observed data.
4.2 Selecting a discrimination strategy

As in the broader statistical classification literature, there is no “magic bullet” algorithm in the field of UXO discrimination. Comparisons of statistical discrimination algorithms (e.g. Zhang et al. (2003), Hart et al. (2001)) demonstrate the performance of particular algorithms on particular data sets, but it is often difficult to gauge how dependent an algorithm is upon the authors’ expertise or the difficulty of the classification task. For example, Hart et al. (2001) demonstrate the application of a PNN for discrimination of UXO using magnetics data. They showed that the PNN outperformed linear discriminant analysis at two of three sites. However, a subsequent application of the PNN at a different site produced very poor results compared to all other discrimination methods, including interpretation by a human expert Nelson et al. (2003). These results highlight the need to carefully select a discrimination strategy which is appropriate to the remediation task.

One of the key weaknesses of statistical classifiers is their reliance upon a representative sample of training items. Recent work by Zhang et al. (2004) addresses this shortcoming with an active learning approach to building a training set. The authors develop a metric to identify the test feature vectors which provide the most “information gain,” as quantified by the Fisher information matrix. A greedy search for the target with the most information gain is implemented to iteratively build an informative training set. This training set can then be used to train a statistical classifier and generate predictions for the remaining test vectors. For application to UXO discrimination, the authors consider magnetometer and GEM-3 electromagnetic data from a Jefferson Proving Ground demonstration. The performance of the active learning algorithm, measured by the receiver operating characteristic, is better than that of a classifier which is trained using multiple realizations of random training data.

The active learning approach is an important contribution to UXO research and is especially useful when the training data set is difficult to generate. However, our practical experience at a number of field sites has indicated that there are often situations where a sizeable and representative training set can be obtained both safely and quickly. The sites most in need of remediation have often been used for intensive training and have considerable clutter and ordnance on the surface. In some cases it may be possible to generate a large number of labelled feature vectors by full clearance of selected areas. For example, generating a training set was comparatively easy at one of the Lowry range presented later in this paper: clearance of some 200 targets by explosive ordnance technicians required only two days work. Furthermore, a component of random digging for quality assurance should always proceed in parallel with digging directed by a discrimination algorithm. Finally, any anomalies which cannot be confidently modelled (failed inversions) should be excavated. For these reasons, our focus here is not on the efficient generation of a training set but rather on iterative identification of an optimal discrimination algorithm as digging proceeds.

Given our set of statistical or library discrimination algorithms, we must choose an approach which is appropriate to the discrimination task at hand. One option is to combine the outputs of all available discrimination algorithms into a single decision. A wide variety of classifier combination schemes exist in the statistical literature, ranging from simple schemes (e.g. voting or averaging) to more sophisticated...
4.2. Selecting a discrimination strategy

approaches which seek to optimize some weighted combination of classifier outputs using the training data. As with classification algorithms, comparisons of combination algorithms yield no single best method. Combiners generally outperform the input classifiers, but not always. A thorough comparison of combination schemes in Jain et al. (2000) showed that combining a set of statistical classifiers (both generative and discriminative) trained on the same feature set can, at best, provide only marginal improvement over the best single classifier. Combination is most appropriate when classifiers are trained using independent feature sets.

While platforms allowing joint acquisition of magnetic and EM data exist for UXO applications, in this paper we consider the most common scenario in UXO remediation: given features extracted from a single data type (EM or magnetics) how can we identify the discrimination strategy which provides the lowest false alarm rate? Although classifiers may use different features, the training data are drawn from the same observed data, and so it is preferable to try to identify a single algorithm which can provide the best available discrimination performance. However, we also advocate continual re-evaluation of the discrimination strategy as digging proceeds and the training set grows. An advantage of statistical classifiers is their ability to learn from the data and improve their performance as more information becomes available, and we hope to exploit this ability if it is advantageous.

4.2.1 Measuring discrimination performance

In UXO applications, the performance of a discrimination strategy is often displayed using the receiver operating characteristic (ROC), which shows the true positive fraction ($TPF$) as a function of the false positive fraction ($FPF$). Here, the $TPF$ is the proportion of UXOs found and the $FPF$ is the proportion of clutter found. The ordinate is sometimes also displayed as the number of false alarms per acre, or simply the total number of clutter items which are dug. To generate an ROC curve for a discrimination algorithm, we compute the output of the algorithm for each unlabelled test vector and then sort the outputs from highest to lowest rank, where a higher rank indicates that a test vector is more likely to be a UXO. We then proceed to label (i.e. dig) the test vectors according to their rank, generating an ROC curve which indicates the proportion of UXO found as a function of the proportion of clutter found throughout the digging process.

A metric of classifier performance which is derived from the ROC is the area under the curve (AUC). The AUC is defined as the integral of the true positive fraction with respect to the false positive fraction

$$\text{AUC} = \int_0^1 TPF \, d(FPF).$$

(4.3)

If the $FPF$ is the fraction of all test clutter items which are dug, then an ideal discrimination algorithm will have an $\text{AUC} = 1$ (i.e. all UXOs are found before a single clutter item is dug). Conversely, the worst possible classifier will require us to dig all clutter items before finding any UXO, producing an $\text{AUC} = 0$. The AUC statistic is commonly used to assess medical diagnostic tests (Hanley and McNeil, 1982) and machine learning algorithms (Bradley, 1997). In these contexts, there is an extensive
4.2. Selecting a discrimination strategy

literature describing the equivalence of an algorithm’s AUC and the probability that the algorithm will correctly rank a randomly selected pair of “abnormal” and “normal” (e.g. UXO and clutter) test feature vectors (Hanley and McNeil, 1982). Specifically, let us denote an algorithm’s output as \( \lambda \), such that \( \lambda(a) \) is the output for test vector \( a \). Further, let \( a \) be ranked as more likely to be a UXO than test vector \( b \) if \( \lambda(a) > \lambda(b) \). For example, with a generative classifier we define \( \lambda \) as the predicted probability that a test vector is a UXO. If \( a \in \text{UXO} \), the set of all test vectors belonging to the UXO class, and \( b \in \text{Clutter} \), the set of all test vectors belonging to the clutter class, then

\[
\text{AUC} = P(\lambda(a) > \lambda(b)),
\]

(4.4)

the probability that \( a \) is ranked ahead of \( b \) (Hanley and McNeil, 1982). For finite samples, the AUC provides an unbiased estimate of this probability. If the generating distributions for UXO and clutter are Gaussian, as shown in figure 4.3(a), then the true and false positive fractions are

\[
TPF(\lambda) = \int_{\lambda}^{\infty} N(x|TP) dx = 1 - C(\lambda|TP)
\]

(4.5)

\[
FPF(\lambda) = \int_{\lambda}^{\infty} N(x|FP) dx = 1 - C(\lambda|FP)
\]

where \( N(x|TP) \) and \( C(x|TP) \) denote the normal and cumulative probability density functions for true positives, respectively (with the corresponding distributions for false positives denoted by \( FP \)). Then from equation 4.3 the AUC can be computed as

\[
\text{AUC} = \int_{-\infty}^{\infty} TPF(\lambda) \frac{d(FPF(\lambda))}{d\lambda} d\lambda
\]

(4.6)

\[
= 1 - \int_{-\infty}^{\infty} N(\lambda|FP)C(\lambda|TP) d\lambda
\]

An analogous result holds for any form of the generating distributions: the integral will always involve the product of the false positive pdf and the true positive cdf.

If assuming a parametric form for the AUC is too restrictive, then we can simply estimate it by numerical integration of the empirical ROC.

An alternative metric for measuring discrimination performance is the false alarm rate (FAR), which we define as the proportion of clutter which must be dug in order to find all UXOs. The FAR is defined graphically by the point at which the TPF first attains a value of one (see figure 4.6). Intuitively, we can also regard the FAR as an estimate of a probability

\[
FAR = P(\lambda(b) > \min_{\text{UXO}} \lambda(a))).
\]

(4.7)
4.2. Selecting a discrimination strategy

This is the probability that a randomly drawn scrap item is ranked ahead of the worst case (i.e. the minimum) prediction for all feature vectors belonging to the UXO class. If, as shown in figure 4.3(a), we know the generating distributions for true and false positives, then we can compute a lower bound for the FAR for a given sample size

\[ \text{FAR}_{\text{min}} = 1 - C(z_N|FP), \]  

(4.8)

where \( C(z_N|FP) \) is the cumulative distribution of false positives evaluated at the characteristic smallest value \( z_N \) of the true positive distribution. This value is defined so that if we draw \( N \) samples from the true positive distribution, then we expect the smallest of these samples to have a value of \( z_N \) or less (Gumbel, 1958). The lower bound on the FAR is then the integral of the false positive distribution from \( z_N \) to infinity (hatched area in figure 4.3(a)).

From figure 4.3(a), it is easy to see that as sample size increases \( (N \to \infty) \), the characteristic minimum value \( z_N \) goes to \(-\infty\), so that

\[ \lim_{N \to \infty} \text{FAR} = 1. \]  

(4.9)

This limiting behavior is demonstrated in figure 4.3(b), which shows the dependence of estimates of AUC and FAR on sample size when the generating distributions are Gaussian. As the number of samples increases, both the mean estimated FAR and its lower bound tend towards one. In practice, however, our samples are of limited size and so the FAR may provide a useful statistic for comparing discrimination performance.

In contrast, the expected AUC is independent of sample size, and mean estimates converge to the expected value computed with equation 4.6. Furthermore, the variance of AUC estimates is much smaller than those of the FAR for corresponding sample sizes. These simulations suggest that the AUC is a more suitable parameter for measuring discrimination performance. However, we will demonstrate in the next sections that bootstrap estimation of the FAR can produce a more robust parameter for comparing discrimination performance than does the AUC.

Estimation of the AUC and FAR metrics given an ordered list of true and false positives is straightforward. We can construct the receiver operating curve from our ordered list, and then estimate the AUC via numerical integration of this curve. In the following examples, we estimate the AUC by trapezoidal integration of the empirical ROC. The FAR is estimated as the point at which the ROC first attains \( TPF = 1 \).

4.2.2 Bootstrap estimation of discrimination performance

Any potential application of the performance metrics discussed in the previous section to UXO discrimination requires their estimation, ideally with an independent test data set. Generating such a test set may be possible when all anomalies in selected areas are cleared in an initial digging stage. In this case, we may estimate performance using our previously-trained classifiers on the newly labelled test feature vectors.

In many real situations, however, we have limited labelled data with which to train our discrimination
4.2. Selecting a discrimination strategy

(a) Derivation of the minimum FAR from the distributions of true positives (dashed line) and false positives (solid line) for a sample of size $N$. The minimum false alarm rate (FAR) is the integral of the false positive distribution from the characteristic minimum value $z_N$ to infinity (hatched area).

(b) In each of 1000 trials, an equal number $N$ of true and false positive samples were drawn from Gaussian generating distributions, as shown in (a). The resulting distributions of estimated AUC and FAR are shown for varying sample size $N$. Vertical dashed lines show the expected AUC (upper plot, independent of $N$) and the expected minimum FAR for a sample of size $N$ (lower plot).

Figure 4.3: Estimation of false alarm rate (FAR) and area under curve (AUC) for finite samples.
4.2. Selecting a discrimination strategy

algorithms and validate their performance. We would like to train and assess our discrimination algorithms as digging proceeds and identify the best algorithm using the available information. Arbitrary division of labelled data into training and test sets is undesirable in this case: there is potential to learn from all feature vectors and so we want to include all labelled data in the training set. However, estimating performance is problematic if there are no independent test data. An algorithm which perfectly discriminates the training data may not generalize well to an unseen test set. This is analogous to overfitting the data in regression, where fitting a noisy function too closely can produce poor estimates of the function parameters.

A standard way to estimate discrimination performance when no independent test data are available is with cross-validation. In leave-one-out cross-validation, a single vector is left out of the training set and the algorithm is trained on the remaining vectors. A discrimination prediction can then be made for the hold-out vector and the process is repeated for all training vectors. The AUC or FAR can then be estimated from the set of cross-validation predictions. However, the training samples in this approach are substantially the same, and so if the classifier overfits this training set then we will obtain an overly optimistic estimate of discrimination performance (Hastie et al., 2001). We can address this difficulty with bootstrap estimation. If the full set of labelled data \( L \) comprises \( N \) feature vectors, then we can approximate the true (unknown) class distributions as discrete distributions with all labelled vectors in \( L \) attributed equal weight \( 1/N \). We can estimate any desired statistic by drawing samples from these empirical distributions. In practice, bootstrapping generates a training realization by sampling with replacement \( N \) times from \( L \). This procedure will generate repeated feature vectors in the training realization, so that the expected number of unique feature vectors is then

\[
E(N_{\text{bootstrap}}) = \left[ 1 - \left(1 - \frac{1}{N}\right)^N \right] N 
\approx 0.632 N. \tag{4.10}
\]

The remaining feature vectors (on average \( 1 - 0.632 = 0.368 \) of the vectors in \( L \)) can then be used as a hold-out test set for estimation of the performance metrics. In discrimination problems, the “0.632” bootstrap estimator is the preferred estimator of discrimination performance statistics (Hastie et al., 2001). This estimate is computed by

1. Generating a bootstrap realization of training and test sets by sampling with replacement from the full set of labelled data.
2. Training the discrimination algorithm on the bootstrap training set
3. Generating predictions for both the bootstrap training and test sets
4. Estimating the performance statistic \( \phi \) (e.g. FAR, AUC) of interest, again for both bootstrap training and test sets. For a given bootstrap realization \( B \), this produces the estimates \( \hat{\phi}_B^{\text{train}} \) and \( \hat{\phi}_B^{\text{test}} \).
5. Averaging the bootstrap performance statistics according to

\[ \hat{\phi}_{0.632} = 0.632\hat{\phi}^B_{\text{test}} + 0.368\hat{\phi}^B_{\text{train}}. \]  

(4.11)

6. Repeating steps (1)-(5) to obtain a distribution for \( \hat{\phi}_{0.632}. \)

Intuitively, the weighting of training and test estimates in equation 4.11 corrects for the unequal sizes of bootstrap training and test sets and ensures that all labelled feature vectors are included in each estimate \( \hat{\phi}_{0.632}. \)

To select a discrimination algorithm as digging proceeds, we propose a method which iteratively re-evaluates the available algorithms using the bootstrap estimates of the AUC or FAR. At each iteration, we estimate the mean metric and select the discrimination algorithm with the best expected performance (i.e. largest AUC or smallest FAR) as the active algorithm. We then dig a given number (say, 20) of the highest priority targets identified by this algorithm. These newly-excavated items now become part of an updated training data set, with which we can re-train and re-evaluate our available discrimination algorithms. With this approach we do not explicitly include the uncertainty in our bootstrap estimates. However, by continually evaluating algorithms with the available information, we allow ourselves to correct for errors in the selection of an algorithm which may be caused by the variance in bootstrap estimates.

### 4.3 Application to electromagnetic and magnetic data sets

As a first example of this procedure applied to UXO discrimination, we compare the performance of discrimination algorithms at the 20 mm Range Fan of the Former Lowry Bombing and Gunnery Range, Colorado. At this site our aim was to discriminate 37 mm projectiles from ubiquitous 20 mm projectiles and 50 caliber bullets using Geonics EM-63 time-domain electromagnetic data. The data were acquired along lines spaced at 50 cm with a single sensor mounted on a fibreglass pushcart. A Leica robotic total station and an inertial motion unit were used for positioning and orientation measurements and were merged with the sensor data in post-processing. Digging proceeded in two phases, with the first phase involving excavation of all targets identified in the EM-63 data within a training grid. Twenty five emplaced 37 mm projectiles, twenty two 20 mm projectiles, and seventy three 50 caliber bullets were recovered in this first phase. The emplaced 37 mm projectiles were buried at depths ranging from 5 to 40 cm. The smaller 20 mm and 50 caliber items, which were not emplaced but were left in the ground after live training, were typically found at shallow depths (< 10 cm). The second digging phase produced an independent test set which was used to evaluate the final performance of discrimination strategies. This test set is comprised of seven 37 mm projectiles, twenty nine 20 mm projectiles, and fifty 50 caliber bullets.

Target feature extraction was carried out with a two-dipole model parameterized by instantaneous polarizations \( (L_i(t_j), i = 1, 2) \) estimated at each time channel \( t_j. \) The EM-63 instrument records 26

90
4.3. Application to electromagnetic and magnetic data sets

time channels ranging from 0.18 ms to 25 ms. Inclusion of late time, low SNR channels in the inver-
sion can produce poor fits to the observed data and unreliable polarization estimates. Consequently,
for each target we fit a subset of channels with estimated SNR above an estimated noise floor (see
figure 4.5). The estimated polarizations were then fit with a parametric function of the form

\[ L_i(t) = k_i t^{-\beta_i} \exp(-t/\gamma_i). \]  

(4.12)

As mentioned in section 4.1, the parameters of this function (or others with similar parameterizations)
have been shown to be indicative of target size and shape. A useful diagnostic derived from this
parameterization is the integral of the polarization

\[ (\text{Integrated Polarization})_i = \int_{0.18}^{25} k_i t^{-\beta_i} \exp(-t/\gamma_i) dt. \]  

(4.13)

The integral of the polarization is an approximation to the magnetostatic polarization, which can in
turn be related to target size (McFee, 1989). After careful experimentation with the available training
features we selected a feature space spanned by the largest integrated polarization and the associated
decay parameters \( \beta \) and \( \gamma \). All statistical classifiers considered in the subsequent bootstrap analysis are
trained with this feature space. Figure 4.4 shows the training features, as well as the decision surface
computed for a nonlinear support vector machine. In this feature space we note that the pervasive
50 caliber bullets have very large feature variance relative to the 37 mm and 20 mm classes. This is
somewhat surprising given that the bullets are much smaller, and were generally shallower, than the
other ordnance classes.

The variability of the features for the 50 caliber bullets can be understood by examining maximum
observed decays for the different classes of ordnance (figure 4.5). We see that 50 caliber bullets tend
to have lower SNR than the larger ordnance, and so fewer time channels are available to constrain the
inversion for these items. Furthermore, a multivariate analysis of variance (MANOVA) indicates that
there is not a significant separation between class means for 50 caliber bullets and 37 mm projectiles,
whereas the separation between 37 mm and 20 mm is significant (both at a 95% confidence level). For
these reasons, we decided that the training features for 50 caliber bullets were unreliable and these
items should not be treated as a class in the training data. Although they were not used for training
process, the 50 caliber bullets are still included in estimation of ROC performance metrics (i.e. true
positive denotes 37 mm, and false positive denotes 20 mm and 50 cal.). This result emphasizes our
erlier remark that inversion is a crucial step in the application of advanced discrimination, and careful
assessment of fits is always necessary if we are to make useful inferences with the training features.

Our goal in this retrospective analysis is then to identify the best available algorithm for discriminat-
ing between 37 mm and 20 mm projectiles using bootstrapped performance metrics estimated after the
first phase of digging is complete (and before an independent test set becomes available).

Table 4.1 shows the mean AUC and FAR of 100 bootstrap samples for four discrimination algo-
rithms, as well as the performance metrics computed independently from the test data set. To ensure
4.3. Application to electromagnetic and magnetic data sets

Figure 4.4: Training data for 20 mm Range Fan discrimination. Grayscale image shows the decision surface for a nonlinear support vector classifier trained to discriminate between 20 mm and 37 mm projectiles that the reported differences between performance metrics are significant for the independent data, we test the receiver operating characteristic curves using a two-sample Kolmogorov-Smirnov test (Press et al., 1992). At a 95% confidence level, there is a significant difference between all receiver operating characteristic curves for the discrimination algorithms considered in table 4.1. This ensures that there is a significant difference between the generating distributions of true and false positives so that the reported metrics on the independent test data can be deemed significant.

The resulting ranking of discrimination algorithms provided by the bootstrap analysis is consistent with that obtained from the independent test data set. The support vector machine has the best performance, while thresholding on the amplitude of the target anomaly provides poor performance relative to all statistical classifiers. Although both performance metrics agree on the preferred order of algorithms, the bootstrap estimates overestimate performance relative to the values obtained from the test data. This is consistent with the simulations of Yousef et al., who found that the 0.632 estimator of the AUC was optimistically biased (Yousef et al., 2004). While more elaborate estimators can be employed to correct for this bias, our aim is to prioritize discrimination algorithms and the estimator employed here
4.3. Application to electromagnetic and magnetic data sets

Figure 4.5: Sounding at anomaly maximum for EM-63 anomalies (with ground-truth) in 20 mm Range Fan training data, along with theoretical \((t^{-1/2})\) and calculated (obtained through analysis of a signal-free part of the dataset) noise floors (from Billings et al. (2007)).

Table 4.1: Bootstrap estimates of area under the receiver operating curve \((\text{AUC}_{0.632})\) and false alarm rate \((\text{FAR}_{0.632})\) for discrimination algorithms applied to 20 mm Range Fan training data. AUC and FAR denote performance metrics evaluated on independent test data. Discrimination algorithms are: (SVM): nonlinear support vector machine, (PNN): probabilistic neural network, (LDA) linear discriminant analysis, (Amplitude): thresholding on anomaly amplitude.

<table>
<thead>
<tr>
<th></th>
<th>(\text{AUC}_{0.632})</th>
<th>AUC</th>
<th>(\text{FAR}_{0.632})</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.99</td>
<td>0.96</td>
<td>0.03</td>
<td>0.10</td>
</tr>
<tr>
<td>PNN</td>
<td>0.96</td>
<td>0.92</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>LDA</td>
<td>0.91</td>
<td>0.89</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>Amplitude</td>
<td>0.86</td>
<td>0.73</td>
<td>0.84</td>
<td>0.95</td>
</tr>
</tbody>
</table>

appears adequate for this purpose.

In our second example we consider re-assessment of the discrimination strategy through several digging iterations. Here we compare the performance of three discrimination algorithms applied to magnetics data from Guthrie Road, Montana. Analysis of these data is presented in detail in Billings (2004). The discrimination task at this site was to identify a total of 80 live and emplaced 76-mm projectiles and 81-mm mortars. Magnetostatic dipole fits were successfully estimated for 724 anomalies in the data. The model is parameterized by the location of the magnetic dipole moment and its vector components. For magnetostatic data the induced dipole is time-independent and so a time-varying polarization such as in equation 4.12 is not required. The 76-mm and 81-mm items comprise the UXO class, while the remaining 644 targets (encompassing shrapnel, ferrous scrap, and geologic false alarms) define the clutter class. As shown in figure 4.6, in both training and test data clutter are characterized by small dipole moments oriented at large angles with the earth’s magnetic field, while ordnance have large moments oriented at small angles. It is therefore appropriate in this case to treat
4.3. Application to electromagnetic and magnetic data sets

all non-ordnance items as a single class and train statistical classifiers to discriminate between UXO and clutter.

Following Hart et al. (2001), we train a probabilistic neural network (PNN) in a feature space spanned by the logarithm of the moment magnitude and the angle of the dipole moment with respect to the earth’s magnetic field. Training data were generated as a random sample of 100 feature vectors drawn from the full set of estimated feature vectors. In addition, we train a nonlinear support vector machine with Gaussian kernels using the same realization of training data. Finally, we use the remanence algorithm originally applied to these data in Billings (2004).

Figure 4.6 shows the feature data used to train and test the PNN and SVM algorithms. The grayscale images in these plots are displayed such that darker regions correspond to the decision boundary (e.g. \( P(\text{UXO}) = 0.5 \) for the PNN). Again, the decision threshold is not fixed in the feature space, but is swept through the space to generate the ROC. No feature space is displayed for the remanence discriminant because it does not require training data, and because this algorithm is a nonlinear transformation of the two-dimensional feature space in figure 4.6 into a single feature, remanence.

Figure 4.6 also shows the ROC curves generated by the two statistical classifiers and remanence. The ROC curves in this plot are all significantly different at the 95% confidence level. In this case, remanence requires us to dig the fewest clutter items in order to find all UXOs. This observation can be quantified by the AUC and the FAR shown in table 4.2. As expected, there is a negative correlation between the AUC and the FAR, with both parameters indicating that remanence provides the best discrimination performance for these data.

Table 4.2: Area under the receiver operating curve (AUC) and false alarm rate (FAR) for algorithms applied to Guthrie road data.

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNN</td>
<td>0.90</td>
<td>0.41</td>
</tr>
<tr>
<td>SVM</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td>Remanence</td>
<td>0.93</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Can we identify remanence as the optimal available algorithm using the limited information in the initial training set? Figure 4.7 shows the bootstrap estimates of AUC and FAR as digging proceeds, and the resulting ROCs produced by our iterative selection of the active algorithm. We consider both the mean AUC and the mean FAR as criteria for selecting the active algorithm. In this simulation, each iteration of bootstrapping follows digging the 20 highest priority items predicted by the active algorithm. This interval is fairly small since dig teams can sometimes excavate hundreds of items in a single day. However, this interval is chosen to demonstrate the evolution of the bootstrapped performance metrics as the training data set grows. Simulations with larger intervals produced similar results. We account for the dig interval in figure 4.7(a) by evaluating the TPF and FPF only at the end of each iteration of digging, producing an ROC curve evaluated at only a few operating points. For comparison, the best and worst ROC curves in figure 4.6, produced using remanence and the SVM are shown in figure 4.7(a).
4.3. Application to electromagnetic and magnetic data sets

(a) Receiver operating characteristics for three discrimination algorithms applied to Guthrie road data. Solid circles show the false alarm rate for each algorithm.

(b) Training and test feature vectors used to generate ROCs for statistical classifiers in (a). Squares are UXO and circles are clutter. Greyscale images are classifier outputs for the probabilistic neural network (top row) and the support vector machine (bottom row).

Figure 4.6: Application of statistical classifiers to Guthrie road magnetics data

At the 95% confidence level there is no significant difference between the ROCs generated by the mean FAR algorithm and remanence. However, the ROCs generated by the mean AUC and remanence are
4.3. Application to electromagnetic and magnetic data sets

significantly different, indicating that in this case the AUC metric was unable to reproduce the best observed performance for these data.

At the first iteration we set the active classifier to be the support vector machine. This choice tests the ability of our approach to correct for a poor discrimination strategy. As shown in figure 4.6, this algorithm has the worst performance of all algorithms when applied to the test data using the initial training data. However, the SVM is often a reasonable choice in the initial stages of digging, since it makes no distributional assumptions and is therefore appropriate when there are insufficient training data to estimate the parameters of generative distributions.

Notable in figure 4.7(b) is the result that the mean AUC ranks remanence behind the PNN for a substantial portion of the digging process. It is only towards the end of digging that remanence finally emerges as the optimal algorithm. Although the bootstrapped AUC fails to identify remanence as the best available algorithm, the ROC produced using the AUC as the selection criterion is an improvement on the ROCs for individual statistical classifiers shown in figure 4.6. This is because the statistical classifiers are retrained after each iteration of digging and their performance accordingly improves as they learn from the growing training set.

In contrast, the mean FAR provides a consistent ranking of remanence ahead of the statistical classifiers after the first two iterations. The mean FAR quickly corrects for our poor initial choice of discrimination strategy and switches at the second iteration to the PNN. Thereafter, the optimal discriminant is remanence, and the resulting ROC produced by the selection algorithm using the FAR is not significantly different from the ROC curve for remanence alone.

The failure of the bootstrapped AUC to identify the optimal available algorithm in this example stems from its formulation as an average metric of discrimination performance. In the early stages of digging, we primarily recover UXOs, so that the overlap between UXO and clutter classes in the training data is small. As the training data set grows, we see increased overlap between UXO and clutter. We therefore expect the FAR to grow and the AUC to decrease as the discrimination task becomes more challenging. The limiting behavior of the FAR (equation 4.9) also dictates that this parameter should grow as the bootstrap test sets increase in size. This dependence is evident in figure 4.7(b).

However, the addition of a few difficult UXO items to the training data will not greatly affect the average performance of the algorithm, as quantified by the bootstrapped AUC. Consequently, in the early to intermediate stages of digging, bootstrapping overestimates the AUC and our procedure fails to identify remanence as the best available algorithm.

The FAR, on the other hand, is most sensitive to those items which are most difficult to identify. Although bootstrapping requires that we average over multiple training and test realizations, the FAR is an average of worst case scenarios, and so provides a more reliable ranking of algorithms when the training data is of limited size.
4.4 Discussion

Discrimination algorithms which are specifically designed for a particular application will often outperform more general statistical approaches. In case of magnetics data, the remanence algorithm uses the known dipole responses for expected ordnance items to reliably identify buried items. Statistical algorithms, in contrast, rely upon a representative training sample to generate the decision rule. The initial training data set in the magnetics simulation was generated as a small random sample of the full set of feature vectors and so the particular realization of training data can have important implications for expected performance. However, even with large training sets and continual retraining, our simulations consistently showed that remanence will outperform competing statistical algorithms for these data. This is not to say that remanence is the best algorithm for UXO discrimination with magnetic data; its applicability depends upon the validity of its assumptions for a particular data set. The method proposed here provides a mechanism for evaluating whether an algorithm is appropriate for the data set at hand.

The variable performance of statistical classifiers in these examples highlights the requirement for careful assessment of discrimination algorithms at each site. In the case of discrimination using features extracted from electromagnetic data, the SVM provided the best performance. However, this same classifier produced the worst performance of all algorithms when applied to magnetics data. The discrepancy in SVM performance in these examples can be attributed to the difficulty of the discrimination tasks: EM features provided a very clear separation between 37 mm and 20 mm classes, whilst magnetics features produced more overlap between classes. These differences are attributable to different site characteristics (i.e. different ordnance classes are present), as well as the non-uniqueness inherent to magnetic features described in Billings (2004).

Recent work in Lee et al. (2007) presents an algorithm for optimizing the performance of statistical classifiers using an approximation to the AUC. The method is applicable to algorithms whose predictions are differentiable functions of the classifier parameters (e.g. kernel smoothings in a neural network). By directly optimizing these parameters with respect to the AUC, the authors obtain improved performance (i.e. lower FAR) relative to classifiers optimized with respect to the probability of misclassification. This method has the potential to significantly improve the performance of some statistical classifiers (PNN, SVM) in our examples. However, optimization of the AUC is not possible for library methods, and so the bootstrapping technique described here should still be used to compare the expected performance of all discrimination algorithms.

Our examples are limited to binary hypothesis cases. That is, we have considered discrimination between two classes: UXO and clutter. In general, however, we may be faced with a classification problem which requires us to identify several classes of ordnance, as well as clutter. The metrics presented here can easily be adapted to the multi-class case by combining the classification predictions for all UXO classes into a single UXO class. This can be accomplished by summing the predictions for a test vector over all UXO sub-classes, and then generating an ROC with these merged predictions. If we are concerned with our ability to distinguish between different ordnance classes, we can similarly
pool predictions and generate an ROC for each UXO class separately. AUC or FAR metrics can be averaged over all classes to obtain an estimate of overall classification performance.

The interval at which algorithms are evaluated is dictated by the speed at which field work proceeds and by the need for thorough quality control of all feature vectors. If our discrimination algorithms are at all useful, then much of the initial excavation effort will be spent on digging UXOs. These items are likely dangerous, and so the early stages of remediation will proceed slowly. This gives the data analyst time to carefully evaluate data fits and retrain algorithms. As the training set grows, we have more confidence in the choice of active algorithm and so the retraining interval can be increased.

Bootstrapping is a computationally-intensive operation, and this approach is only suitable for algorithms which can be trained relatively quickly. Bootstrapping with library methods such as remanence is fast, as these algorithms do not require retraining with each realization. The support vector machine, on the other hand, requires that we solve a quadratic programming problem for each bootstrap realization. However, the simulations presented here took a few minutes to run for each iteration of digging on a 3.4 GHz Pentium IV desktop, and so this is a viable procedure for data sets of this size (approximately $10^3$ feature vectors).

When adding new feature vectors to the training data set, it is crucial that parameter estimates are reliable. To this end, it may be necessary to reacquire and re-invert new training items in a controlled test-pit setup to ensure adequate spatial coverage and low noise. This is especially a consideration for electromagnetic data, where the added complexity of the model makes the data requirements more stringent than for magnetic data. Again, we emphasize that parameter estimation is the most important step in the application of advanced discrimination algorithms.

Any application of our proposed method for identifying an optimal algorithm requires close coordination between field technicians and geophysicists. This is consistent with the “Total Quality Management” model of field operations adopted by Billings and Youmans (2007). They emphasize the need for continual performance monitoring and feedback, and our approach explicitly includes this monitoring using statistical performance metrics. This method can be applied for evaluation of any kind of discrimination algorithm, including algorithms derived from different types of geophysical data, or trained on different subsets of features. Estimation of performance metrics in the initial geophysical prove-out may help to select a sensor which can provide the best available performance at a site. Comparison of this approach to sensor management with methods based upon risk minimization (e.g. Williams (2006)) is an avenue for future research.

### 4.5 Conclusion

In this paper we have provided a brief review of algorithms which have been used to discriminate between UXO and clutter. The performance of algorithms can be measured using the area under the curve (AUC) or the false alarm rate (FAR). When independent validation data are available, these performance statistics can be used to identify an optimal algorithm from a suite of available algorithms. However, when the set of labelled data is small, we must resort to bootstrap estimates
of performance. Our simulations indicate that the bootstrapped FAR is a more robust metric for ranking performance than the AUC when sample sizes are small and there is significant overlap between classes in the feature space.
4.5. Conclusion

(a) ROC curves generated by the remanence and SVM algorithms alone, compared with ROCs generated by iterative selection of an active algorithm using AUC and FAR performance metrics. Circle and cross markers on the ROC curves indicate which algorithm (PNN and remanence, respectively) is active at each iteration of digging, as shown in (b).

(b) Left: estimates of the mean FAR as a function of iteration. The algorithm with the smallest FAR at each iteration is considered optimal. Bottom right: estimates of the mean AUC as a function of iteration. The algorithm with the largest AUC at each iteration is considered optimal.

Figure 4.7: Selection of a discrimination algorithm using bootstrap estimates of the FAR and AUC.
Bibliography


Chapter 5

Selecting a receiver operating point

5.1 Introduction

The receiver operating characteristic (ROC) curve is a tool for displaying the performance of a binary decision process and is used in applications such as medical diagnosis, economics, and machine learning. The curve is generated by varying the threshold of the decision algorithm and plotting the resulting proportion of true positives as a function of the proportion of false positives. For medical studies a true positive often denotes correctly predicting that a condition is present, while a false positive denotes incorrectly predicting the presence of the condition. The decision statistic, or score, is a single continuous or discrete value which corresponds to the output of a diagnostic test. A second application, which is our focus in this article, is unexploded ordnance discrimination. Here a true positive indicates that we have successfully predicted that an ordnance belongs to the class of UXO targets \(T\) which must be dug, while a false positive is a non-ordnance item belonging to the class \(F\) which we have predicted belongs to \(T\). The decision statistic is the output of any algorithm which can discriminate between \(T\) and \(F\) classes. For example, the decision statistic might be the probability of membership in the UXO class predicted with a discriminant analysis classifier.

As the decision statistic is varied monotonically for a test data set, we generate a sequence of true and false positives which cumulatively generate points on the ROC. This empirical curve is necessarily an uncertain estimate of the expected performance of a discrimination algorithm, since it is derived from a limited set of observations which is assumed to be representative of the underlying populations of \(T\) and \(F\) classes. Accordingly, methods for computing ROC confidence intervals are increasingly applied in machine learning contexts (Macskassy et al., 2005).

In this article we address selection of the operating point on the ROC curve. The operating point is a cut-off value of the decision statistic used to generate the ROC. For example, any diagnostic test with a value less than a selected operating point will be classified as “disease present.” In the UXO problem the operating point corresponds to the “stop-dig” point, any targets with a decision statistic less than this threshold will be left in the ground. When full clearance of all detected targets is required by a regulator, the operating point might instead represent the division between digging targets with explosive ordnance disposal (EOD) technicians and digging with slightly fewer precautions (e.g. using a backhoe to dig targets which are likely clutter).

In section 5.2 we set up the problem and discuss existing methods for selecting the operating point.

---

7 A version of this paper will be submitted for publication. Beran L. and Oldenburg D. Choosing a receiver operating point in UXO discrimination.
5.2. Optimal operating points

point. These methods use slightly different notions of optimality to balance the trade-off between true and false positive fractions. In unexploded ordnance discrimination we often know a priori how many unlabelled targets must be classified. We assume that this set of unlabelled targets is a sample from hypothetical generating distributions of $T$ and $F$ classes. In section 5.3, we show that the empirical ROC for a sample depends upon the generating and prior distributions of $T$ and $F$ classes, as well as the size of the sample. In particular, as the sample size grows the expected proportion of targets which must be labelled in order to find all $T$ instances in the sample also grows. We then derive the approximate sampling distribution of the point on the ROC at which all $T$ instances are found. In theory, this distribution can be used to select an operating point which attains a specified probability of identifying all $T$ instances in a sample. In practice, however, this depends upon accurate estimation and extrapolation of the tails of the generating distributions from a limited training data set. We find that this approach has limited success when applied to real data sets.

We therefore propose a simple method that does not depend upon estimating distributions, but rather continues labelling instances until a predefined number $N_L$ of false instances occur in sequence. In applications to synthetic and real data this technique has a higher probability of identifying all $T$ instances than other methods, though with a commensurate increase in the proportion of labelled $F$ instances.

5.2 Optimal operating points

In this section we review existing theory for choosing the operating point. Following work in Kanungo and Haralack (1995) and Fawcett (2004), denote a positive (e.g. target is ordnance) by $T$ and a negative (e.g. target is not ordnance) as $F$. At a specified threshold $\Lambda$ of the decision statistic $x$, we define the following probabilities

\[
P(T|T, \Lambda) = \int_{-\infty}^{\Lambda} p(x|T)dx \quad \text{(true positive)} \quad (5.1)
\]

\[
P(F|T, \Lambda) = \int_{\Lambda}^{\infty} p(x|T)dx = 1 - P(T|T, \Lambda) \quad \text{(false negative)} \quad (5.2)
\]

\[
P(T|F, \Lambda) = \int_{-\infty}^{\Lambda} p(x|F)dx \quad \text{(false positive)} \quad (5.3)
\]

\[
P(F|F, \Lambda) = \int_{\Lambda}^{\infty} p(x|F)dx = 1 - P(T|F, \Lambda) \quad \text{(true negative)} \quad (5.4)
\]

where $p(x|T)$ and $p(x|F)$ are the distributions of $T$ and $F$, as shown in figure 5.1(a). The decision statistic $x$ is a number representing the output of a discrimination algorithm, and $p(x|T)$ and $p(x|F)$ are univariate distributions, sometimes referred to as score distributions, which are specific to that
5.2. Optimal operating points

Figure 5.1: (a) The ROC is generated by integrating the distributions of $T$ and $F$ classes. (b) At a given operating point $\Lambda$, we have a confusion matrix of probabilities, and the ROC is a plot of the first row of this matrix as a function of $\Lambda$. (c) The resulting ROC with operating point $\Lambda$.

discrimination algorithm. Statistical classification algorithms (e.g. discriminant analysis, support vector machines) may operate in a high-dimensional feature space to generate the decision statistic; we emphasize that the score distributions are a univariate projection of the distributions of $T$ and $F$ in the feature space. In figure 5.1(a) we have shown the score distributions as normal probability distributions. However, in practice score distributions are rarely normal (Macskassy et al., 2005) and the development which follows is therefore general to any form of $p(x|T)$ and $p(x|F)$. The receiver operating curve shows the dependence of $P(T|T, \Lambda)$ (true positives) upon $P(T|F, \Lambda)$ (false positives) as $\Lambda$ is varied. The probabilities at a given threshold $\Lambda$ can be displayed as a “confusion matrix” (figure 5.1(b)). Hereafter we retain the convention that if the decision statistic $x$ for a test item is less than the threshold $\Lambda$ then that item classified as $T$ by our discrimination algorithm, as shown in figure 5.1(a)8.

In this framework, confusion matrix probabilities alone are not sufficient to select an optimal threshold, we must assign some cost $C$ to each of the decisions we make. We denote the cost of a true positive as $C(T|T)$, and a false positive as $C(T|F)$. The costs of the other elements of the confusion matrix are similarly defined. A notable feature of the ROC is that it is independent of the relative frequencies of the two classes (the prior probabilities $P(T)$ and $P(F)$). However, once we assign a cost to each of the decisions in the confusion matrix, the prior probabilities become crucial because they determine how often we expect to incur the specified costs. Now given our confusion matrix probabilities,

8Many statistical classifiers output a probability $P(T|x)$, so that a large decision statistic indicates that an instance is likely to be $T$. In this case we simply threshold on $-P(T|x)$.
specified costs and prior probabilities, the expected cost (also called risk) at $\Lambda$ will be

$$E(C(\Lambda)) = P(T|T, \Lambda)C(T|T)P(T) + P(F|T, \Lambda)C(F|T)P(T) + P(F|F, \Lambda)C(F|F)P(F) + P(T|F, \Lambda)C(T|F)P(T) + (1 - P(T|T, \Lambda))C(F|T)P(T) + (1 - P(F|T, \Lambda))C(F|F)P(F) + P(T|F, \Lambda)C(T|F)P(F)$$

(5.5)

Minimizing the expected cost with respect to $\Lambda$ gives

$$\frac{\partial P(T|T, \Lambda)}{\partial P(T|F, \Lambda)} = \frac{(C(F|F) - C(T|F))P(F)}{(C(T|T) - C(F|T))P(T)}.$$  

(5.6)

Recalling that the ROC is a plot of $P(T|T, \Lambda)$ versus $P(T|F, \Lambda)$, the optimal operating point is defined by the point where the slope of the curve equals the right hand side of the above expression (Kanungo and Haralack, 1995). The ROC is a non-decreasing function, and a positive slope can be ensured by requiring

$$C(T|F) \geq C(F|F)$$

$$C(F|T) \geq C(T|T)$$

(5.7)

so that misclassification costs more than correct classification. In machine learning it is often assumed that correct classifications do not incur any cost and we must only specify the relative costs of false negatives and false positives. While in some applications (e.g. economics) costs are available, in many cases it is not obvious how to objectively choose these parameters. Of particular concern is the cost of the false negative. What does it cost to leave an ordnance item in the ground? In the short term, not taking action when it is required may save us expenditures on remediation, but in the long term we must account for potential environmental damage, liability, etc. We might encode expert, but ultimately subjective, judgements into costs. Variable costs of remediating targets might be considered: for example deep targets are more time-consuming (and hence more expensive) to excavate than shallow targets. However, deep targets are also likely to pose less risk for accidental detonation. We conclude that there is considerable ambiguity to cost specification and employing expert judgement when specifying costs may not minimize actual risk.

Even if costs can be chosen in a manner that will satisfy regulators, we also require the prior probability for each class to apply equation 5.6. We are thus faced with the problem of specifying two functions (costs and priors) from limited training data. A further problem is that the empirical ROC generated from observed data is a piecewise constant curve with infinite or zero slope, and so finding an operating point with slope satisfying equation 5.6 requires that we fit a smooth function to the empirical ROC. This interpolated function may also achieve the desired slope in more than one location, so that a given specification of costs and priors does not unambiguously define an operating point.

Billings (2004) suggested that an operating point for UXO discrimination be chosen by terminating digging once a predefined number $N_L$ of successive $F$ instances (clutter) are encountered in the or-
5.2. Optimal operating points

dered list of targets. In this case the sequence of labelled targets will end with one \( T \) instance followed by \( N_L \) \( F \) instances. The slope of the ROC at this operating point is approximately

\[
\frac{\partial P(T|T, \Lambda)}{\partial P(T|F, \Lambda)} \approx \frac{1/N_T}{N_L/N_F} = \frac{1}{N_L} \frac{P(F)}{P(T)} \tag{5.8}
\]

Choosing \( N_L \) is equivalent to specifying the right hand side of equation 5.6. Approximating the slope of the ROC in this way circumvents the need to estimate priors and to interpolate the empirical ROC, but still leaves us with the problem of choosing costs via the parameter \( N_L \). In section 5.3 we explore how \( N_L \) is influenced by sample size (the total number of instances) and propose objective criteria for specifying this parameter.

Simpler options for selecting an operating point which do not require costs or priors have also been proposed. Two criteria which have been considered are the point on the ROC closest to the point \((0,1)\), and the point on the ROC which is the largest vertical distance from the 45 degree line passing through \((0,0)\) and \((1,1)\) (the Youden index, Perkins and Shisterman (2006)). Figure 5.2 illustrates these methods. The point closest to \((0,1)\) is also equivalent to choosing the point with the largest perpendicular distance from the 45 degree line. The 45 degree line corresponds to random chance: a discrimination algorithm with this ROC will correctly rank randomly selected items from each class only half of the time (Fawcett, 2004).

Figure 5.2 compares the location of operating points on a theoretical ROC curve with the three methods discussed thus far. This curve is generated for two univariate Gaussian distributions with standard deviations \( \sigma_T = \sigma \) and \( \sigma_F = 2\sigma \), and means separated by \( 4\sigma \), as shown in figure 5.1(a). To illustrate the effect of unequal costs on the operating point, we also show operating points with hypothetical costs displayed in table 5.1. The Youden index and cost minimization with equal priors and equal costs to incorrect classification give the same operating point. We can see this by writing the vertical distance \( d \) between the ROC and the 45 degree line as

\[
d = P(T|T, \Lambda) - P(T|F, \Lambda). \tag{5.9}
\]

Maximizing \( d \) with respect to \( \Lambda \) yields equation 5.6 for equal priors and costs. The operating point for the closest \((0,1)\) method stops slightly before the Youden index, while cost minimization with lower costs for digging clutter has the intuitive result that decreasing the relative costs of digging clutter causes us to dig more clutter.

| Method | \( C(T|T) \) | \( C(T|F) \) | \( C(F|T) \) | \( C(F|F) \) |
|--------|-------------|-------------|-------------|-------------|
| C1     | 0           | 1           | 1           | 0           |
| C2     | 0           | 0.1         | 1           | 0           |

Table 5.1: Example costs for risk minimization used in figure 5.2.
5.3 Optimal operating points for samples

The operating points discussed in the previous section are expected to be optimal (according to slightly different criteria) for the underlying population from which our observed data are assumed to be a random sample. However, in practice we are not applying our discrimination algorithm to the inferred population, but to a sample from that population. How do the expected ROC curves for a population and a sample from that population differ? The empirical ROC for a sample is generated by ordering the sample from smallest to largest decision statistic, and then incrementing the cumulative counts of T and F instances as we move up the ordered list. The true and false positive fractions are therefore discrete quantities which asymptote to the continuous form (equation 5.4) as \( N \to \infty \).

Consider the example of two univariate Gaussian distributions from the previous section. Figure 5.3 shows the theoretical ROC curve for the population, as well as ROC curves for sample sizes ranging from \( N = 200 \) to \( N = 800 \), with \( N/2 \) the number of samples from each distribution (i.e. equal priors). As sample size increases, the empirical ROCs are increasingly better approximations to the true ROC. Also shown in figure 5.3 are the false alarm rates (\( FAR \)) for the random samples. Here we define the \( FAR \) as the proportion of \( F \) items which must be dug in order to find all \( T \) items in the sample. The generating distributions require that we dig every \( F \) item to guarantee that all \( T \) items are found, so that \( FAR = 1 \) for the population, However, the expected value of the false alarm rate for a given sample size is considerably less than one for finite samples and slowly asymptotes to one as sample size is increased.

Based upon these observations, it seems reasonable to define an optimal operating point \( \lambda \) for a sample of size \( N \) as one which finds all \( T \) items in the sample. Processing of unexploded ordnance...
5.3. Optimal operating points for samples

Figure 5.3: Effect of sample size on ROC and false alarm rate. Left column: population ROC (heavy line) and realizations of empirical ROCs for samples of size $N/2$ from each generating distribution. Right column: false alarm rates (FAR) of empirical ROCs in left column. Solid horizontal line is the mean FAR over all realizations.
data typically occurs in batches, so that the sample size is known prior to selecting the operating point. If the prior probability of $T$ is known (or is estimated from a training data set) and the total number of instances in the unlabelled sample (test data) is known, then the expected number of $T$ instances in the test data can be computed. This suggests a straightforward criterion for selecting an operating point: continue labelling until the expected number of $T$ instances has been found. However, if the number of $T$ instances is overestimated by just one, then this approach will lead us to label all instances in a fruitless search for a non-existent occurrence of $T$. This defeats the purpose of discrimination, and so instead we seek to develop methods which have an improved chance of identifying all $T$ instances than the methods previously discussed.

Here $\lambda$ is a discrete random variable which corresponds to the index of the last $T$ instance encountered in the ordered sample. A lower bound on the expected value of $\lambda$ can be derived by finding the critical value $x^\lambda$ such that
\[
1 - P(x^\lambda|T) = \int_{x^\lambda}^{\infty} p(x|T)dx = 1/N_T,
\]
with $N_T = NP(T)$ the expected number of $T$ instances in the sample (and similarly, $N_F = NP(F) = N(1 - P(T))$). At the point $x^\lambda$, we expect that one $T$ instance will occur somewhere in the interval $[x^\lambda, \infty]$. This extreme value is the last $T$ instance we will encounter in the sample and therefore defines the desired operating point. At $x^\lambda$ we expect the false alarm rate to be
\[
P(x^\lambda|F) = \int_{-\infty}^{x^\lambda} p(x|F)dx
\]
so that the expected number of $F$ instances occurring before the point $x^\lambda$ is $N_F P(x^\lambda|F)$. Then a lower bound on the expected number of instances which must be labelled in order to find all $T$ occurrences in a sample of size $N$ is
\[
E(\lambda) \geq N_T + N_F P(x^\lambda|F).
\]
This expression is only a lower bound on the expected value of $\lambda$ because the extreme $T$ instance can occur anywhere in the interval $[x^\lambda, \infty]$. We note that computing this lower bound requires that we integrate the distribution $p(x|T)$ up to a small probability $1/N_T$ (equation 5.10) so that the corresponding value $x^\lambda$ will be in the tail of $p(x|T)$. In practice the generating distributions must usually be estimated from a small set of labelled training data, and extrapolation beyond the range of observations into the extreme tails of an estimated distribution is prone to large errors (Pandey, 2001). We will demonstrate this problem in section 5.3.2 with applications to real data sets.

Now to derive an approximate probability distribution for the discrete random variable $\lambda$, we consider a test sample comprised of the union of independent samples from $p(x|T)$ and $p(x|F)$. The test data have the score distribution
\[
p(x) = p(x|T)P(T) + p(x|F)P(F).
\]
The empirical ROC is generated by ordering the test data from smallest to largest score (decision
5.3. Optimal operating points for samples

The $i^{th}$ item in this ordered list of $N$ samples (the $i^{th}$ order statistic $x(i)$) has the probability distribution (Balakrishnan and Cohen, 1956)

$$p(x|x(i)) = \frac{N!}{(i-1)!(N-i)!} P(x)^{(i-1)}(1 - P(x))^{(N-i)} p(x). \quad (5.14)$$

The probability of observing a sample from the $T$ class at $x$ is given by Bayes rule

$$P(T|x) = \frac{p(x|T)P(T)}{p(x|T)p(T) + p(x|F)p(F)}. \quad (5.15)$$

Figure 5.4 illustrates the computation of equations 5.13 through 5.15. Marginalizing over $x$ we obtain the probability that the $i^{th}$ order statistic in the test data belongs to the $T$ class

$$P(T|x(i)) = \int_{-\infty}^{\infty} P(T|x)p(x|x(i))dx \quad (5.16)$$

with $P(F|x(i)) = 1 - P(T|x(i))$. Figure 5.5 verifies computation of equation 5.16.

Figure 5.4: Steps in computing the probability $p(T|x(i))$ of a $T$ instance given the $i^{th}$ order statistic from a sample. (a) The probability distribution of the test data $p(x)$ is comprised of the distributions $p(x|T)p(T)$ and $p(x|F)p(F)$ (equation 5.13). (b) The probability $P(T|x)$ of a $T$ instance given score $x$ (equation 5.15). (c) The probability distribution $p(x|x(i))$ of obtaining score $x$ given that $x$ is the $i^{th}$ item in an ordered list of $N$ scores sampled from $p(x)$ (here $N = 800$). The probability $P(T|x(i))$ is obtained by multiplying each $p(x|x(i))$ by $p(T|x)$ and integrating over all $x$.

To compute the probability that the $i^{th}$ order statistic is the last true item in the sample ($P(i = \lambda)$), we must enumerate all permutations for which this outcome can occur and compute the probability of each permutation. This quickly becomes prohibitively expensive for even modest sample sizes, and so here we approximate the distribution of $P(i = \lambda)$ as follows. If the $i^{th}$ order statistic is the last $T$ item in the sample, then the remaining $i + 1$ through $N$ order statistics must all be $F$. Therefore the probability
that the $i^{th}$ order statistic is the last $T$ is approximately equal to

$$P(i = \lambda) \approx P(T|x(i))P(F|x(i+1))P(F|x(i+2))...P(F|x(N)).$$

(5.17)

Figure 5.5: Computing the probability $P(T|x(i))$ of a $T$ instance given the $i^{th}$ order statistic. Points are generated by repeated sampling from the distribution $p(x)$, ordering the sample, and then determining the proportion of realizations for which the $i/N^{th}$ item is $T$. Solid line is the prediction from equation 5.16, with the required integrals evaluated numerically.

Figure 5.6 shows computation of equation 5.17 for the generating distributions shown in figure 5.1(a), with equal priors and for sample sizes ranging from $N = 200$ to $800$.

Figure 5.6: Probability that the $i/N^{th}$ order statistic is the maximum $T$ instance ($\lambda$) in samples of size $N$. In all plots dots are the result of simulations, solid lines are theoretical calculations.

The probability mass function $P(i = \lambda)$ can in principle be used to select an operating point $\lambda$. For example, we might adopt the order statistic which is most probably the maximum $T$ instance in a
sample of size $N$ as an optimal operating point (i.e. the mode of the distribution of $P(i = \lambda)$). Figure 5.7 shows the dependence of the mode of $\lambda$ upon sample size and prior probability of the $T$ class. The operating point increases logarithmically with sample size, with the rate of increase weakly dependent upon the prior probabilities. Also shown in figure 5.7 is the lower bound on the expected value of $\lambda$ computed with equation 5.12. We see that the lower bound asymptotes to the most probable value as sample size increases. This makes equation 5.12 a useful approximation for large sample sizes where computing equation 5.16 over all order statistics can become quite slow.

![Figure 5.7: Dependence of the optimal operating point $\lambda$ on sample size $N$ and prior probability $T$ for the generating distributions of figure 5.1(a). Lines with markers are a lower bound on the expected value of $\lambda$ (equation 5.12).](image)

We now return to the idea of specifying a number $(N_L)$ of $F$ instances occurring sequentially in the ordered diglist which will trigger us to stop digging. The approximate distribution of $\lambda$ can help to understand how a particular choice of $N_L$ affects the probability of finding all $T$ instances. Using the probabilities $P(F|x(i))$ (equation 5.16), figure 5.8 shows the probability $P(N_L)$ of observing $N_L$ $F$ instances in sequence immediately prior to the most likely value of $\lambda$ in an ordered list of length $N$. If $P(N_L)$ is small then there is a small probability that we will terminate digging prior to the final $T$ instance. As $N_L$ is increased, $P(N_L)$ will necessarily decrease monotonically since $P(F|x(i)) < 1$. The curves in figure 5.8 have the character of a Tikhonov curve encountered when solving a regularized inverse problem. This suggests using a heuristic from Tikhonov regularization (such as the point of maximum curvature) to select $N_L$. However, in practice it will be difficult to construct these curves from limited training data and so we seek heuristics which depend only on the size of the sample $N$. In figure 5.8 we find that $P(N_L)$ is bounded by the function

$$P(N_L) \leq \exp\left(-\frac{N^2}{N_L}\right), \quad (5.18)$$
5.3. Optimal operating points for samples

Setting $P(N_L) = 1/N$ in the above expression and solving for $N_L$ yields

$$N_L = \lceil \sqrt{N \log N} \rceil$$

(5.19)

with the inequality converted to an equality and the ceiling function ($\lceil \cdot \rceil$) mapping to the next largest integer. As shown in figure 5.8, this approach yields a consistently small value of $P(N_L)$ for the sample sizes considered. Beyond figure 5.8, we have no theoretical basis for justifying the form of equation 5.19, and the result in figure 5.8 is of course particular to the distributions, prior probabilities, and sample sizes in this example. Other choices for the dependence of $N_L$ on sample size can be considered, in the remainder of this paper we demonstrate the efficacy of this particular form with simulations and applications to real data.

![Figure 5.8](image-url)

Figure 5.8: Solid lines: probability $P(N_L)$ of observing $N_L$ instances in sequence, immediately prior to the most likely value of $\lambda$ (the last $T$ instance), for varying sample sizes $N$. Dashed lines: an upper bound on $P(N_L)$ (equation 5.18). Markers indicate $N_L$ computed with equation 5.19.
5.3. Optimal operating points for samples

5.3.1 Simulations

To illustrate the relative merits of criteria for selecting an operating point, we simulate the performance of the following operating points:

(a) the point on the training data ROC closest to (0,1).

(b) the lower bound on the expected value of \( \lambda \) (equation 5.12), assuming normal probability distributions and with required probabilities and priors estimated from the training data. This is an approximation to the most probable value of \( \lambda \), as shown in figure 5.7.

(c) the number \( N_L \) of \( F \) instances which must occur sequentially, computed with equation 5.19.

We estimate all operating points for (a) and (b) in the above list from the same realization of training data, and then apply these operating points to an independent test data set to compute true and false positive fractions (\( TPF, FPF \)) at each operating point. For each realization of training data a single realization of test data is used to evaluate all operating points. This is repeated for 100 independent realizations of training and test data, allowing us to estimate means and standard deviations of \( TPF \) and \( FPF \) at all operating points. Generating distributions for \( T \) and \( F \) classes are as depicted in figure 5.1, and we used equal priors when generating the test data. These simulations are repeated for training data sets ranging from \( N = 200 \) to \( N = 800 \), as in figure 5.3, with test sample size fixed at \( N_{\text{test}} = 1000 \).

In table 5.2 we display the proportion of test realizations for which a given operating point is able to identify all \( T \) instances, \( P(TPF = 1) \), as well as the mean and standard deviations of the false positive fraction at the operating points. Not surprisingly, the closest (0,1) operating point has zero chance of finding all \( T \) instances. The \( \lambda \) operating point asymptotes to finding all \( T \) instances approximately half of the time. This is consistent with the distribution of \( \lambda \) in figure 5.8: at the most probable value of \( \lambda \) there is still approximately half of the area under the distribution greater than this value. To achieve a higher \( P(TPF = 1) \) we can integrate the distribution of \( \lambda \) up to some desired confidence, though of course this will increase the expected number of \( F \) instances which must be labelled. While this approach seems to represent a reasonable compromise between finding all true positives while limiting the number of false positives, it is contingent upon knowing the correct form of the probability distributions. For these simulations normal probability distributions are assumed and so the performance of the \( \lambda \) operating point is quite good. This assumption will not hold in general and so actual performance of this operating point may not be as suggested by these simulations. This is illustrated on a real data example in the next section. Finally, we see that the \( N_L \) operating point has a high probability of finding all \( T \) instances, though with a corresponding increase in false positive fraction.

5.3.2 Application to real data

We conclude with applications of operating point criteria to unexploded ordnance discrimination. The problem of selecting an operating point has received limited attention in this context: Zhang et al. (2004)
5.3. Optimal operating points for samples

<table>
<thead>
<tr>
<th>$N^{train}$</th>
<th>(a) Closest (0,1)</th>
<th>(b) $\lambda$ (Normal)</th>
<th>(c) $N_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P(\text{TPF}=1)$</td>
<td>$\text{FPF}$</td>
<td>$P(\text{TPF}=1)$</td>
</tr>
<tr>
<td>200</td>
<td>0.00</td>
<td>0.10 ± 0.02</td>
<td>0.43</td>
</tr>
<tr>
<td>400</td>
<td>0.00</td>
<td>0.10 ± 0.02</td>
<td>0.45</td>
</tr>
<tr>
<td>800</td>
<td>0.00</td>
<td>0.10 ± 0.01</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 5.2: Performance of operating point criteria for simulations with varying training data size $N^{train}$. $P(\text{TPF}=1)$ is the proportion of test data realizations for which the operating point found all $T$ instances, $\text{FPF}$ is the mean and standard deviation of the false positive fraction at the operating point.

derive an expression for a probability threshold which depends only upon the numbers of true and false instances in the training data

$$P^\lambda = \frac{\sqrt{N_T}}{\sqrt{N_T} + \sqrt{N_F}}.$$  \hspace{1cm} (5.20)

This assumes that the score distributions are normally distributed with equal variance. Operating points selected by this method typically identify 90% of the test ordnance items in the published results, and we found that this technique had comparable performance to the closest (0,1) method for the simulations reported in the previous section. We retain the latter method for application to real data sets. For discrimination of UXO using frequency domain electromagnetic data, Huang et al. (2006) maximize a “degree of discrimination” parameter identical to the maximum vertical distance from the 45 degree line described in section 5.2. This method also found approximately ninety percent of ordnance in the test data.

In figure 5.9 we compare the performance of operating point criteria applied to a data set from Guthrie road, Montana. The discrimination problem at this site was to discriminate 81 mm mortars and 76 mm projectiles from clutter using magnetics data. As described in Billings (2004), for each observed magnetic anomaly we estimate an (apparent) remanence quantifying the deviation of the observed dipole moment from a library of expected moments for ordnance targets. This provides a decision statistic which can be used to discriminate ordnance (small remanence) from clutter (large remanence). We see in figure 5.9(a) that the empirical distributions of the log transformed remanence are approximately normal. However, the empirical distribution of ordnance is not symmetrical about its mean, such that the upper tail of the normal distribution is a poor characterization of the actual distribution of ordnance. Extrapolation of a fitted normal distribution to determine the operating point $\lambda$ therefore overshoots the last $T$ (ordnance) instance and results in a large number of unnecessary digs on the ROC (5.9(b)). In this plot we have not normalized the ROC axes by the numbers of true and false positives, so that the actual numbers of ordnance (80) and clutter (644) are shown. The operating point selected with $N_L = 70$ finds all ordnance in this case, with an acceptable overshoot past the last ordnance item. In comparison, the expected closest (0,1) operating point achieves a $\text{TPF} = 0.925$ (6 ordnance are left in the ground). This is consistent with previously reported results for this method. In practice, the $\lambda$ or closest (0,1) operating points are estimated from a small training set and subsequently applied to the remaining test data. Here we have instead used the full data set to choose these operating points and so the reported performance does not capture the variability of these
5.3. Optimal operating points for samples

operating points introduced by the training procedure. A bootstrap analysis can be used to quantify this effect, but the simulations in the previous section are likely indicative of the relative variability of each technique. An obvious advantage of the $N_L$ operating point is that it depends only upon the size $N$ of the data set and so does not require a training data set in order to determine an operating point.

![Figure 5.9](image_url)

Figure 5.9: Operating points applied to Guthrie road data. Left plot shows the empirical distributions of the decision statistic, apparent remanence, for ordnance and clutter classes. Solid lines are normal distributions fit to these data. Right plot shows the ROC generated by thresholding from smallest to largest decision statistic, with operating points selected by closest (0,1), $\lambda$ (Normal) and $N_L$ methods.

Figure 5.10 shows example ROCs generated in a demonstration study at San Luis Obispo. The discrimination task at this site was to find a variety of ordnance targets ranging in size from 4.2" mortars down to 60 mm mortars. A number of electromagnetic sensors were deployed at the site, here we show results for

(a) EM61 cart operating in a detection mode survey. The decision statistic is the estimated rate of decay of the induced dipole moment.

(b) EM61 with a magnetometer (MSEMS sensor) operating in a detection mode survey. The decision statistic is as in (a).

(c) TEMTADS array operating in a cued interrogation survey. The decision statistic is the maximum correlation coefficient between the observed data and data predicted by a best fitting item in a library of pre-defined ordnance polarizations (fingerprinting method described in Pasion et al. (2007)).

(d) Same TEMTADS data set as in (c), but with the decision statistic the probability of membership in the ordnance class as predicted by a statistical classifier trained on size and decay rate of the primary polarization (similar to the approach employed in chapter 2 of this thesis on Camp Sibert data).
Given the limitations of the $\lambda$ operating point demonstrated above, here we show only the closest (0,1) and $N_L$ operating points. In figure 5.10 (a), (b) and (c) the $N_L$ operating point achieves $TPF = 1$, while in (d) we obtain $TPF = 0.990$ (2 ordnance left in the ground). Again, the closest (0,1) point finds approximately ninety percent of the ordnance in all cases. The slightly different results for (c) and (d), both derived from the same TEMTADS data set, emphasize that the ROC curve depends not only on the geophysical data, but on the discrimination strategy used to process these data. While careful selection of an operating point can improve our chances of finding all ordnance, failures in inversion and quality control can introduce outliers to the ordnance class. For the TEMTADS statistical classification (figure 5.10(d)), the two outliers to the ordnance class are lower SNR 60 mm mortars for which the decay parameter is difficult to estimate. More advanced discrimination techniques (e.g. robust inversion, or incorporating parameter uncertainty in inversion) may reduce the occurrence of outliers and improve the probability that $TPF = 1$.

Figure 5.10: Operating points applied to San Luis Obispo data. Squares are closest (0,1) operating points, circles are $N_L$ operating points. (a) EM61 cart (b) MSEMS EM61 (c) TEMTADS fingerprinting (d) TEMTADS statistical classification.
5.4 Discussion

We have investigated selecting the operating point on a receiver operating curve. Previously published criteria are generally designed to balance the trade-off between true and false positive fractions on the ROC. This type of trade-off occurs in many analogous problems. For example, in Tikhonov regularization of underdetermined inverse problems we seek a model which balances the misfit to the observed data and some measure of model complexity. For the regularization problem a model can be selected with heuristics (e.g. selecting a point of maximum curvature on the Tikhonov curve), cross-validation techniques, or statistical criteria (e.g. achieving a target data misfit).

Here we have developed heuristics and statistical criteria for the operating point problem when the total number of instances which must be labelled is known. We derived an approximate probability distribution for the discrete random variable $\lambda$, which we defined as the order statistic at which all $T$ instances are labelled. This probability distribution depends upon the generating distributions, prior probabilities, and sample size. Given this probability distribution, we can select an operating point which corresponds to the most likely value of $\lambda$. In addition, we have derived a lower bound to the expected value of $\lambda$ which is a useful approximation to the most likely value. However, this approach has limited practical applicability because it depends upon accurate estimation of the generating distributions and extrapolation into the extreme tails of these distributions. To address this shortcoming, we have proposed a heuristic for selecting $N_L$, the number of $F$ instances which must occur sequentially before digging is terminated. This is equivalent to cost minimization, but the proposed heuristic provides an objective means of choosing relative costs based upon sample size. In simulations and applications to real data we find that this technique has an improved probability of finding all ordnance in a test data set, relative to previously published methods. We have limited our investigations to samples on the order of $N = 10^3$, which is representative of the number of detected targets at many sites. Tests on larger data sets should still be carried out.


Chapter 6

Conclusions

This thesis has demonstrated the viability of advanced discrimination of unexploded ordnance. A notable outcome is the practical application of discrimination techniques at demonstration sites (Lowry, Camp Sibert and elsewhere). These demonstrations have consistently shown that discrimination with physics-based models can achieve real improvements in performance and cost savings. In this concluding chapter I revisit the motivating questions presented in chapter 1 and discuss contributions and directions for future work.

How do particular stochastic error sources translate to errors in the geophysical data? How can we account for these errors in an inversion to ensure that our estimated model is useful?

In chapter 2, I showed via a perturbation analysis that positional errors on the data can translate to non-normal distributions of residuals. The deviation of perturbed data increases with proximity of the sensor to the target, so that large relative deviations are expected for large amplitude data. A similar analysis can be carried out for errors in measurements of sensor orientation and will also produce non-Gaussian residual distributions, as demonstrated in Pasion (2007). These errors, as well as other sources such as sensor noise and geology, will contribute additively to the total error on a given datum, such that the total error will usually tend to normally-distributed random variable. If errors on the data are independent, I found that least squares inversion performs as well as robust inversion techniques, provided that the least squares inversion is properly weighted by estimated errors. However, parameter estimates were markedly improved by robust inversion when simulated errors were correlated. Robust inversion also improved model parameter estimates and subsequent discrimination for all TEM data sets acquired at Camp Sibert. While robust norms can provide a reduction in the bias in model estimates when the noise has a non-normal component, there is a commensurate increase in the variance of parameter estimates when the noise is normal (Marrona et al., 2006). Robust inversions may therefore reduce the deviation of extreme outliers to an ordnance class, but this comes at the expense of an increase in the mean deviation within that class.

How do errors in the data propagate to uncertainty in the model, and how can model uncertainty be incorporated into an algorithm which discriminates between ordnance and non-hazardous objects?

In chapter 3 I studied propagation of uncertainties from data to model to discrimination. I found that the uncertainty distribution of TEM dipole model parameters is often multimodal, with each mode
Chapter 6. Conclusions

of the distribution reasonably approximated by a Gaussian about that mode. This latter representation is equivalent to what is obtained with a linearized uncertainty appraisal at local minima of the misfit. I developed a generalization of discriminant analysis, the Gaussian product, which accounts for local uncertainty in the model. This approach incorporates uncertainty in both the test and training feature vectors, with the end result that features with large relative uncertainty are downweighted in discrimination decisions. Improved detection of outliers was demonstrated on real data sets using this technique.

When applying fingerprinting methods to UXO discrimination with electromagnetic data, we assume that each ordnance class can be characterized by a single set of polarizations. Library polarizations are estimated by inverting high SNR data acquired over known targets and typically have negligible parameter uncertainties. However, inversion of multistatic data can sometimes produce polarizations which vary with target orientation. For example, if an ordnance is composed of a ferrous body and an aluminum tail, then recovered polarizations may differ depending upon whether the target is in nose-up or nose-down orientations. This suggests adopting multiple sets of library polarizations for an ordnance class, or characterizing polarizations by a prior probability distribution. In Aliamiri et al. (2007), the data misfit was augmented with a model norm which accounts for the deviation of estimated polarizations from a prior distribution. This Bayesian formulation allows the inversion to incorporate library information without fixing polarizations at potentially biased values. In addition, uncertainty was incorporated in the discrimination process by integrating the posterior probability distribution over model space. This produces a decision statistic which depends upon the data and model norms, as well as the uncertainty in the estimated parameters. This approach was outperformed by statistical classification when applied to data acquired in a controlled teststand experiment. However, their result may be affected by a potentially poor parameterization of the dipole decay as a superposition of only a few decaying exponentials, and this technique should be further explored with field data sets.

The methods presented in chapters 2 and 3 are primarily applicable to data acquired in a detection mode survey with first generation monostatic sensors. Discrimination with such instruments is often complicated by relatively short time windows, poor spatial coverage, geologic noise introduced by motion of the sensor over topography, and positional errors. These issues make extraction of dipole parameters problematic and necessitate the more complicated approaches to inversion and discrimination presented in this thesis. Early results with multistatic sensors such as the TEMTADS array indicate that improved data quality from cued interrogations, measurement of multiple components of the secondary field, and longer time windows combine to dramatically improve the reliability of inverted dipole parameters and simplify the discrimination procedure. A future model of field operations will therefore be detection with a relatively simple array such as the MTADS, followed by interrogation and discrimination of detected targets with a multistatic sensor. However, there may be scenarios where the discrimination task is relatively simple or where multistatic sensors cannot be deployed. In addition, there will always be a small proportion of small and/or deep targets for which parameter estimation with multistatic cued-interrogation data remains ambiguous. For example, at our demonstration at San Luis Obispo (SLO)\textsuperscript{9}, we encountered a small number of low SNR targets in the TEMTADS test data which

\textsuperscript{9} ongoing at the time of writing
were likely ordnance based upon the match of the estimated polarizations to our library. However, these targets appeared as outliers to the distributions of ordnance in our usual size/decay feature space. QC of these fits indicated that there was a large uncertainty in the estimated decay parameter owing to low SNR at late times. In these cases improved discrimination might be achieved by employing robust inversion techniques or by accounting for uncertainty in discrimination, as described in this thesis.

**Given limited ground truth, how can we objectively evaluate the expected performance of a discrimination algorithm and so choose the best technique for the problem at hand?**

Chapter 4 discussed metrics for measuring discrimination performance and investigates bootstrapping methods to estimate these metrics from a limited training data set. In the published literature on UXO discrimination it is standard to show the performance of a number of discrimination algorithms on a single test data set and then to draw conclusions about the relative advantages of these algorithms. However, this outcome may be highly dependent upon the particular test data set. With a resampling scheme such as bootstrapping we can characterize the expected performance of a classifier over multiple realizations. In chapter 4 I also explored iterative selection of a classification algorithm as discrimination proceeds. While this was perhaps overly complicated for the discrimination problem at hand, the concept of changing the discrimination strategy as digging proceeds seems appropriate at sites where there are multiple, distinct classes of ordnance item. For example, at the San Luis Obispo demonstration ordnance ranged from 4.2” mortars down to 60 mm mortars. Lumping these diverse targets together into a single class was not appropriate, and so we investigated training a classifier for each ordnance type, as well as modelling the overall ordnance distribution as a mixture of Gaussians with each mixture component corresponding to an individual ordnance type.

The statistical classification algorithms used in these studies (discriminant analysis, support vector machines, neural networks) are not state of the art at present, but are often used as baselines for comparison with more advanced algorithms (e.g. semi-supervised classifiers (Zhang et al., 2004) or active learning (Liu et al., 2008)). A new ESTCP project is tasked with demonstrating the advantages of these novel classifiers, our aim has been to test more conventional classifiers on real data and to increase the acceptance of these techniques within the UXO industry. While we have demonstrated clear improvements over simple strategies, statistical classification is by no means automatic and still depends upon the expertise of the analyst. This is evidenced by the feature spaces employed for statistical classification in the various demonstrations. We generally used features related to the amplitude of the polarizations and decay of the polarizations, which are in turn approximately related to target properties (size, wall thickness, material composition). However the features we used to represent amplitude and decay information have evolved through the various demonstrations. Our approach has been to experiment with features during initial training or retrospective analyses to determine those features which provide the tightest clusterings of targets in the feature space. Statistical criteria such as canonical analysis (chapter 1) can sometimes be used to justify these choices, but expert judgement is often required because training data sets are often quite small.
Chapter 6. Conclusions

Library methods presented in chapters 1 and 4 performed as well, or better than, statistical classifiers. Again, it remains to be seen whether advancements in statistical classification can surpass library methods, but circumventing feature selection and classifier design with a library method is very appealing. Indeed with multistatic sensors at SLO we could quickly and easily identify most ordnance by visual comparison of estimated polarizations with library polarizations. This was no more labor intensive than QCing a diglist generated by statistical classification. Of course, library methods will perform poorly if library entries are incorrectly estimated, or if there is a class of ordnance target missing from the library. This second criticism has been used to justify the use of statistical classifiers for UXO discrimination, i.e. statistical classifiers can detect test data feature vectors which are far from anything yet encountered in the training data. While this is true for some classifiers, many algorithms (e.g. discriminant analysis) are formulated as binary hypothesis tests: a test feature vector is either ordnance or clutter, the possibility that it is neither is not explicitly accounted for in the discrimination algorithm. While it is easy to modify an algorithm such as discriminant analysis to identify novel test feature vectors as an initial step, it should not be assumed that novelty detection is inherent to all statistical classifiers.

If a library technique does indeed prove to be the best choice of discrimination algorithm for a given problem, then developments in this thesis can still be applied. Robust inversion can always be used if outliers in the data are suspected, and the choice of operating point is still an important question for any discrimination problem.

How can we decide a cut-off value for a discrimination algorithm, beyond which a buried object is considered benign?

Finally, in chapter 5 I examined selection of the operating (stop digging) point. Risk minimization provides a framework for specifying the operating point, but it requires specification of costs and priors from limited training data (Kanungo and Haralack, 1995). Simpler criteria aim to balance the trade-off between true and false positive fractions in a manner analogous to choosing a regularization parameter when solving an underdetermined inverse problem. Given that the goal of UXO remediation should be to find all ordnance, I derived the statistical distribution of the point on the receiver operating curve where all ordnance are found. This depends upon the generating distributions and prior probabilities of ordnance and clutter, and the size of the test data set. The implication of this last parameter is that the larger the data set, the larger the proportion of targets which must be dug in order to attain an equal probability of finding all ordnance (relative to some smaller data set). However, this distribution has limited practical applicability because it requires accurate estimation of the tails of the generating distributions. To address this limitation I proposed an intuitive technique for choosing the operating point: we continue digging until a prescribed number of clutter are encountered sequentially. This method is equivalent to cost minimization but only requires the specification of one parameter. Choosing the operating point in this way is simple and perhaps obvious, but has not yet been put forth in the UXO community. Previously published techniques for choosing an operating point for UXO discrimination in Zhang et al. (2004) and Huang et al. (2006) are more geared toward balancing the true/false positive
Chapter 6. Conclusions

trade-off, whereas the proposed technique seems to provide an improved probability of identifying all ordnance when applied to real data sets.

The field of unexploded ordnance detection and discrimination has advanced to the point that we have an array of proven tools to address remediation at a variety of sites. Specific directions for future research in the UXO field arising from this thesis research include:

Inclusion of additional features in discrimination. In many of our discrimination studies we restrict statistical classifiers to a two-dimensional feature space spanned by polarization amplitude and decay. However, as sensor platforms improve such that our recovered models are better constrained by the data, additional features may be reliably used for discrimination. In particular, target symmetry, which is indicated by equal transverse (secondary and tertiary) polarizations, has been proposed as a feature for discrimination between axisymmetric ordnance and irregularly-shaped clutter (Bell and Barrow, 2001). Initial efforts in Bell and Barrow (2001) to use symmetry information with production data sets proved unsuccessful because data from monostatic, single-component sensors do not adequately constrain transverse polarizations. At the San Luis Obispo demonstration we found that symmetry information can often be reliably extracted from multistatic sensors, but may be problematic for low SNR targets. This suggests a requirement to constrain the transverse polarizations in an inversion, possibly via the fingerprinting method of Pasion et al. (2007) or by augmenting the misfit function with a model norm, as in Aliamiri et al. (2007). This latter approach penalizes the deviation of the recovered model from a reference model. A less restrictive option which should be investigated is a model norm which penalizes large differences between transverse polarizations.

An alternative to discrimination with size, decay and symmetry parameters derived from estimated polarizations is to work directly with the polarizations themselves. Given that newer sensors provide polarizations at a large number of time channels, we must identify a subset of features to input into discrimination. Canonical analysis is one option for feature selection but it makes simple assumptions that are easily violated. More advanced techniques, such as the recursive feature elimination (RFE) developed in Guyon et al. (2002), show promise. A test of RFE on San Luis Obispo TEMTADS data identified a subset of polarizations at 30 time channels for training a nonlinear support vector machine. Application of this classifier to the test data produced an incremental improvement in discrimination performance relative to discrimination with size and decay parameters. This technique and others (e.g. Krishnapuram et al. (2004)), should be further explored with real data.

Confidence bounds on ordnance clearance. While careful selection of the stop-dig point can improve our probability of identifying all ordnance at a site, there is always a chance that some ordnance remain in the ground. Hathaway et al. (2009) propose that once the stop-dig point is reached, a program of random verification sampling can be used to generate confidence bounds on ordnance clearance. They provide an expression for the number of randomly-selected targets which must be excavated with no true positive (ordnance) instances such that with $1 - \beta$ confidence, we can state that no more than $N_T$ of the remaining targets are ordnance. This approach provides confidence bounds for detected targets
only, it does not consider that some targets may have been missed in surveying. Forward modelling tools should therefore be developed to determine whether a given survey has failed to detect buried ordnance due to inadequate data coverage, low SNR, high target density, etc. These tools would support the purely statistical conclusions of verification sampling with an analysis accounting for the physics of the sensor and the particularities of the survey conducted at a field site.

Automating quality control and multitarget inversion. As advanced discrimination is increasingly applied at real field sites, efficient data processing is required to handle large numbers of targets. Quality control of TEM fits remains a challenge, particularly for monostatic sensor data. Pasion (2007) developed a figure of merit to quantify whether line spacing and SNR of the observed data will support reliable parameter estimation, and this has been extended in Lhomme et al. (2008) to account for metrics related to fit quality (e.g. uncertainty in recovered depth). This work on automatic identification of failed fits should be pursued further to minimize the time required for QC. For example, a machine learning algorithm which is trained to recognize identifying features of failed fits could significantly reduce the number of inversion results requiring visual inspection.

Similarly, automated processing of overlapping anomalies will be needed in areas where there is high target density. Significant progress has been made in multitarget inversion and various information theoretic criteria have been tested for selecting the number of targets required to fit observed data (Song et al., 2009). In chapter 3, I handled non-uniqueness in the TEM parameter estimation problem by allowing an individual target to be represented by multiple models. This suggests a similar strategy for handling high target densities: fit every anomaly with increasing numbers of dipoles and then input all estimated dipoles into discrimination. Alternatively, it may be possible to regularize the model cardinality (i.e. number of dipoles) within a single inversion.

While this thesis has been primarily focussed on developing data processing techniques for improved discrimination, sensor developments will drive much UXO research in the coming years. For example, marine detection and discrimination is a current priority that presents unique challenges, including waterproofing sensors and maintaining a constant clearance above the seabed. In the marine environment we must also model the inductive response of a conductive background (seawater). The method of auxiliary sources (MAS) is well-suited to modelling this problem by matching interior and exterior magnetic fields at the boundary of an arbitrary body of revolution (Shubitidze et al., 2002). This solution could then be used to test whether background and target responses are additive. An additional effect which has been observed in TEM data in marine environments is an induced polarization (IP) response (Pasion, personal communication). Here metallic ordnance impede the flow of ions, producing a potential which is discharged once eddy currents attenuate (analogous to electrode polarization in conventional IP surveys). Whether IP effects can be used for ordnance detection and discrimination is an open research question.

Unexploded ordnance have obvious connections with the improvised explosive devices (IEDs) presently plaguing Afghanistan and Iraq. The same technologies can be employed for detection, but
IED discrimination must occur in real time. Discrimination strategies may therefore be limited to identifying large, slow-decaying TEM responses, or to detecting changes in repeated surveys. If more advanced discrimination is required, then the library method discussed in appendix A could be applicable if a target from a pre-defined set of ordnance types must be immediately identified. Operator-guided or autonomous robotic platforms are increasingly being developed and deployed for IED detection, and these technologies may, in future, be applied to UXO detection and discrimination.
Bibliography


Appendix A

A simplified fingerprinting method

The TEM dipole model assumes that the observed data are a linear superposition of a target’s polarizations $L_i(t)$

$$d^{pred}(t_i) = \sum_{j=1}^{N} w_j L_j(t_i). \quad \text{(A.1)}$$

$j = 2$ or $3$ for an axisymmetric or non-axisymmetric target, respectively. The coefficients $w_j$ vary non-linearly from sounding to sounding as a function of target and sensor location and orientation. However, if the target polarizations are assumed known then for a single sounding the coefficients can be estimated by solving a linear inverse problem. At Camp Lejeune high quality test pit measurements were made to recover estimates of the polarization decays for adapters (figure A.2(a)). We use these polarizations to discriminate between adapters and ordnance as follows:

1. Identify all soundings within the target mask above a predefined threshold. Here I use $15 \text{ mV}$ at the first time channel.

2. For each sounding identified in (1), minimize

$$\phi = \| W_d (d^{obs} - d^{pred}) \|^2 \text{ subject to } w_j \geq 0, \quad j = 1, 2. \quad \text{(A.2)}$$

with $d^{pred}$ given by equation A.1. The data weightings $W_d$ are assigned as the inverse data standard deviations, here estimated as $10$ percent of each observed datum plus a $1 \text{ mV}$ floor. I impose a positivity constraint on the coefficients $w_j$ based upon the expression for the measured voltage over an axisymmetric target at depth $z$ illuminated by a vertical primary field with magnitude $B^p$

$$V(t) = 2\kappa \frac{B^p(z)}{z^3} \left[ L_1 \cos^2(\theta) + L_2 \sin^2(\theta) \right] \quad \text{(A.3)}$$

with $\kappa$ a non-negative constant depending upon sensor geometry, and $\theta$ the dip angle of the target (Pasion, 2007). The coefficients of the polarizations $L_1$ and $L_2$ are therefore non-negative. High SNR soundings identified in step (1) correspond to locations where the horizontal displacement of the sensor from the target is small, so that the primary field is approximately vertical at the target and equation A.3 is a good approximation to the measured voltage. Numerical investigation of the coefficients $w_j$ for the EM63 indicates that the coefficients are non-negative regardless of target-sensor separation (figure A.1), but a formal proof of this property is reserved for future work. The optimization problem (equation A.2) is solved with the lsqlin routine in Matlab.
3. Compute the correlation coefficient (CC) between the (weighted) observed and predicted data, and for each target retain the median correlation coefficient over all soundings. Other CC statistics such as the mean or maximum were also investigated, I choose the median because it is a robust statistic and so is less influenced by noisy soundings with low CC.

4. Dig targets based upon a ranking of median CC.

<table>
<thead>
<tr>
<th>L2 (transverse)</th>
<th>L2 (transverse)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L1 (axial)</th>
<th>EM63 data (t1=0.18 ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure A.1: EM63 polarization coefficients for a vertical target at 30 cm depth. The coefficients are non-negative regardless of sensor-target separation. Predicted EM63 data are for a vertical adapter. The predicted data are a weighted sum of these coefficients, with the weightings given by the polarizations $L_i(t_j)$.

Figure A.2 shows observed and predicted data for soundings corresponding to the median CC for representative adapter (clutter) and ordnance targets at Camp Lejeune. A much better fit is obtained for the adapter than for the ordnance target, and so the former will occur much later in the diglist generated by the fingerprinting technique.

The approach presented here is a simplification of the method developed by Pasion et al. (2007) and is similar to methods considered in Tantum and Collins (2001) and Norton et al. (2001). By eliminating extrinsic target properties (location and orientation) from the forward model, we obtain a linear inverse
Appendix A. A simplified fingerprinting method

Figure A.2: (a) Adapter library polarizations. (b) Fingerprinting fits to soundings for ordnance and adapter targets. Solid line is predicted data, dots are observed data, and dashed line indicates the estimated standard deviations.

problem which can be solved very quickly. The cost of this simplification is that the coefficients are no longer constrained by the physics of the dipole model, although the imposed positivity constraint does at least ensure that the predicted decay is physical. However, it is possible that the coefficients estimated for two different soundings cannot be generated by a single target. This ambiguity can only be resolved by considering all soundings simultaneously in a nonlinear inversion.
Bibliography


