Classification of Unexploded Ordnance

by

Laurens Sander Beran

B.Sc., The University of Victoria, 2001

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Geophysics)

THE UNIVERSITY OF BRITISH COLUMBIA

March 2005

Abstract

In this thesis I investigate methods for discriminating between unexploded ordnances (UXOs) and clutter items (e.g: shrapnel, geology). I first describe a numerical forward model, the method of auxiliary sources (MAS), which can be used to model the magnetic and electromagnetic response of a conductive, permeable body. I use this model to validate the connection between the parameters of approximate forward models and target properties (i.e target shape). I also examine how model parameters can be estimated from observed data using inversion.

I then describe algorithms for discriminating between UXO and clutter. In the statistical classification framework, model parameters are basis vectors within a multi-dimensional feature space. I prioritize features based upon their ability to separate UXO and clutter using canonical analysis. I describe two approaches for partitioning the feature space: modelling the underlying distributions from which the observed feature data are drawn, or directly defining a decision boundary. A suite of statistical classifiers are then applied to magnetics data acquired at three field sites. Finally, I propose an algorithm for selecting a classifier as target excavation proceeds.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Doug Oldenburg for his kindness, patience, and encouragement throughout my studies. His tireless energy and willingness to let me research in sometimes unfamiliar terrain were a great support.

This thesis owes a great deal to work done by Dr. Stephen Billings and Leonard Pasion. Thanks also to office-mates and the wonderful and brilliant members of the UBC-GIF group for their help over the last two years.

Dr. Fridon Shubitidze was incredibly helpful in his explanations of MAS. His willingness to share code and his lightning-fast email responses are remarkable, especially since we have never met.

Thank you to the Montana National Guard for providing data sets from Guthrie Road and Limestone Hills, Montana. I would also like to thank Dr. Larry Carin of Duke University for allowing me to visit and for providing data from Badlands Bombing Range.

Thanks to the many good friends at St. John's and beyond who have entertained me. Finally, special thanks to my brother and my parents for their love and support from too far away.

This research was funded in part by an NSERC post-graduate scholarship and the Egil H. Lorntzsen scholarship.

Chapter 1

Introduction

Unexploded ordnances (UXOs) are defined as "explosive, propellant or chemical containing munitions that were armed, fired and remain unexploded through malfunction [1]." It is estimated that up to 10 million acres are contaminated by UXO within the United States. UXO contamination is a growing problem: the Defense Science Board estimates that 10000 UXOs are produced each year in live-firing exercises. The UXO problem is also prevalent in past and present conflict zones around the world. For example, approximately $1 billion (US) has been spent on clearance of mines and UXOs in Kuwait [2].

Figure 1.1: Ordnance items recovered at Limestone hills, Montana.

The extent of the UXO contamination within the United States and abroad has motivated intensive research into improved technologies for detection and discrimination of UXO. Remediating UXO-contaminated sites in the United States could cost up to $52 billion if metal detectors and pin flags are used to locate potential UXO items (the "mag and flag" technique). This cost is distributed throughout the remediation process, but the majority of the expense is spent on digging non-ordnance items [1]. Mag and flag can require as many as 100 clutter items (geology, shrapnel) to be excavated for each UXO item found. This 100:1 rate is called a False Alarm Rate (FAR) and is a crucial measure for comparing discrimination techniques. "Advanced" discrimination methods are expected to reduce false alarm rates to 10:1, cutting remediation costs to approximately $16 billion. These methods require the acquisition of digital geophysical data and subsequent signal processing.

This thesis examines the signal processing which can be used to discriminate between UXO and clutter. In particular, I apply classification algorithms from the field of pattern recognition to discrimination between UXO and clutter.

Chapter 2 discusses methods for forward modelling the magnetic and electromagnetic response of ordnance items. Models used in UXO discrimination are typically approximate forward models whose parameters serve as proxies for physical properties. I describe a

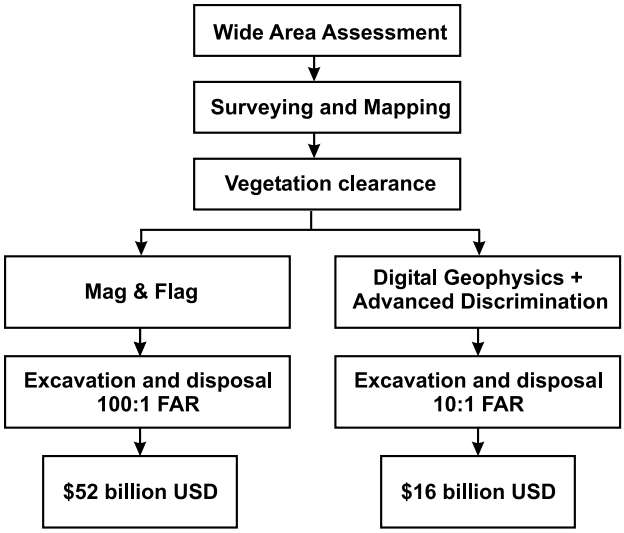


Figure 1.2: The UXO remediation process (after [1]).

numerical model, the Method of Auxiliary Sources (MAS), which can be used to model the response of a conductive, permeable bodies with arbitrary shape. This model is parameterized in terms of physical properties.

In chapter 3, I demonstrate how the parameters of a forward model can be estimated from observed data using inversion. I implement the Levenberg-Marquardt inversion algorithm for a simple synthetic problem. This example uses the analytic solution for the frequency-domain response of a sphere. The parameterization of this model in terms of physical properties makes it a useful analogue for future inversions using MAS. In particular, I show how the correlation between conductivity and permeability can preclude unique estimation of these parameters.

In chapter 4, I use MAS as a benchmark forward model for validating approximate forward models in the time domain. I demonstrate that there is a nonlinear relationship between the estimated parameters of the Pasion-Oldenburg forward model and the aspect ratio of a spheroid.

Chapter 5 gives an overview of pattern recognition. In this framework, model parameters estimated by inversion are basis vectors within a multi-dimensional feature space. I discuss how to prioritize features and reduce the dimensionality of a feature space with canonical analysis. I describe two approaches to defining a classification rule: generative classifiers which estimate the underlying distributions from which the observed feature data are drawn, and discriminative classifiers which directly define decision boundaries to partition the feature space. A crucial step in the classification process is optimization of classifier performance through training.

In chapter 6, I show the application of classification algorithms to real data sets. I apply canonical analysis to prioritize the features obtained from magnetic and electromagnetic data. I then apply a suite of classifiers to real feature data. At two sites (Guthrie road and Limestone hills) statistical classifiers are outperformed by a "rule-based" remanence classifier. However, at a third site (Badlands bombing range), some statistical classifiers provide a modest improvement in classification performance. Statistical classifiers rely strongly upon the available ground truth, or training data. I show how retaining a classifier

as the training data set grows can improve its performance.

I also propose an algorithm for selecting a classifier from a suite of available classifiers. The algorithm estimates the performance of classifiers using cross-validation on the training data. By evaluating performance as digging proceeds, the classifier selection algorithm provides near-optimal performance for the synthetic and real data sets considered.

Finally, I briefly examine the effect of feature uncertainty upon classification with a Monte Carlo simulation. Accounting for uncertainty in the feature vectors can have a strong effect upon classification of electromagnetic data.

Chapter 7 reviews the work carried out to date and suggests directions for future research within the field of UXO classification.

Chapter 2

Forward modelling

2.1 Introduction

In this chapter I review forward modelling of magnetic and electromagnetic data. These data types are most commonly acquired for UXO detection and discrimination. Extracting relevant features from a geophysical data set requires a model which can predict the measured response of a UXO. Given a model vector $\mathbf{m}$, we seek a forward modelling operator F which predicts the data $\mathbf{d}$

$$\mathbf{d} = F\{\mathbf{m}\}. \tag{2.1}$$

In the UXO problem, the model parameters are typically proxies for the relevant physical properties of a target. While these approximate forward models have had success in real applications, recent research has focussed upon parameterizations in terms of physical properties. I consider a numerical forward model, the method of auxiliary sources (MAS), for modelling both magnetic and electromagnetic data. This model provides a single parameterization in terms of target properties (conductivity, susceptibility, shape and size) and geometry (location and orientation). In chapter 4, I will use MAS to validate the connection between approximate forward model parameters and physical properties.

2.2 Magnetics

2.2.1 Governing equations

The governing equations for the study of magnetism are Maxwell's equations for the magnetic field $\mathbf{h}$ and magnetic flux density $\mathbf{b}$

$$\begin{aligned} \nabla \times \mathbf{h} &= \mathbf{j} + \frac{\partial \mathbf{d}}{\partial t} \\ \nabla \cdot \mathbf{b} &= 0 \end{aligned} \tag{2.2}$$

where $\mathbf{j}$ is the current density and $\mathbf{d}$ is the displacement current. Under a quasistatic approximation, displacement currents are assumed negligible [3]. If measurements of the magnetic flux density are made in air, no currents are present and so $\mathbf{j}$ is zero. The first expression in equation 2.2 is then

$$\nabla \times \mathbf{h} = 0. \tag{2.3}$$

The magnetic field is irrotational and can therefore be expressed as the gradient of a scalar potential

$$\mathbf{h} = -\nabla\phi. \tag{2.4}$$

The potential of any arbitrary body occupying a region $\mathbf{R}$ can be expressed as the superposition of potentials due to elementary magnetic dipoles

$$\phi(\mathbf{r}) = \frac{1}{4\pi} \int_{\mathbf{R}} \mathbf{M} \cdot \nabla(\frac{1}{r}) dV \tag{2.5}$$

where the magnetization $\mathbf{M}$ is the magnetic dipole moment per unit volume. The SI unit for magnetization and the magnetic field is A/m. Magnetic flux density is measured in tesla (T).

There are several ways to compute the scalar potential (and flux density) of an arbitrary magnetized body using equation 2.5. If the body is assumed to have uniform magnetization and a simple shape, the volume integral can often be evaluated analytically. For example, the magnetic potential of a permeable sphere is that of a dipole.

Alternatively, the integrand in the above equation can be approximated by expanding $\frac{1}{r}$ as a MacLaurin series. This yields a moment expansion with each term in the series corresponding to the potential of progressively higher order moments (dipole, quadrupole, octupole, etc.) [4].

For bodies of more complicated shape (eg geological structures), a more appropriate approach to computing the magnetic field is to discretize the volume into cells of constant magnetization and numerically integrate equation 2.5. This is a proven method of forward modelling magnetics data for three-dimensional imaging problems encountered in exploration geophysics [5].

Both the moment expansion and discretized methods have been used to forward model magnetics data for UXOs. The most common approach is to represent the induced field of a UXO as a dipole. Billings et al. justify this by approximating a UXO by an equivalent spheroid (most ordnance items can be approximated as prolate spheroids) [4]. The first nonzero components in the moment expansion of a spheroid are the dipole and octupole moments. The octupole moment falls off very rapidly and so the dipole moment provides a reasonable approximation to the field of a spheroid.

Lelievre used a discretized approach to solve the governing equation for the magnetic scalar potential in high susceptibility materials [6]. His implementation of a finite volume discretization was able to account for demagnetization effects observed in spheroidal bodies such as UXO, as described in section 2.2.4.

2.2.2 Magnetization

Magnetization can be modelled as the sum of induced $\mathbf{M_i}$ and remanent $\mathbf{M_r}$ magnetizations

$$\mathbf{M} = \mathbf{M_i} + \mathbf{M_r}. \tag{2.6}$$

Remanent magnetization is a permanent magnetization which does not depend on the inducing magnetic field. This effect is observed in ferromagnetic (eg. iron) materials commonly used to construct ordnance casings as well as in geologic materials.

Two types of remanent magnetization are thought to be relevant to the study of UXOs: thermoremanent magnetization and viscous remanent magnetization. Thermoremanent magnetization is acquired when magnetizable material is cooled from a high temperature. Below the material-dependent Curie temperature, magnetic moments are locked into permanent domains, producing a net remanent magnetization. Viscous remanent magnetization also involves alignment of moments, but can be acquired at lower temperatures by prolonged exposure to the inducing field [7].

Induced magnetization is produced by the instantaneous alignment of magnetic moments with an inducing magnetic field. In geophysics, the primary inducing field is generally the earth's magnetic field. When the magnetic susceptibility is small ($\chi < 0.1$), as is

the case for most earth materials, the induced magnetization is modelled as

$$\mathbf{M} = \chi \mathbf{h}$$

where $\mathbf{h}$ is the magnetic field. The susceptibility is in general a tensor quantity, but in most applications it can be assumed to be isotropic (ie a scalar). The total magnetic flux density $\mathbf{b}$ is then the superposition of the inducing magnetic field plus the magnetization

$$\mathbf{b} = \mu_o(\mathbf{h} + \mathbf{M}) \tag{2.7}$$

In the absence of remanent magnetization (ie $\mathbf{M} = \mathbf{M_i}$), this yields the constitutive relation

$$\mathbf{b} = \mu \mathbf{h} \tag{2.8}$$

with the permeability $\mu = \mu_o(1 + \chi)$.

2.2.3 Geophysical measurements of b

Equation 2.7 shows that the measured flux density $\mathbf{b}$ is the superposition of an inducing field $\mathbf{b_o} = \mu_o \mathbf{h_o}$ and an anomalous field $\mathbf{b_a} = \mu_o \mathbf{M}$. Total field magnetometers measure the magnitude of the flux density

$$\|\mathbf{b}\|^2 = \mathbf{b}_o^T \mathbf{b}_o + 2\mathbf{b}_o^T \mathbf{b}_a + \mathbf{b}_a^T \mathbf{b}_a. \tag{2.9}$$

In geophysical applications the anomalous magnetic field is typically far smaller than the earth's magnetic field and so the last term in the above expression can be neglected. Then the magnitude of the total field can be approximated as

$$\begin{aligned} \|\mathbf{b}\| &= (\mathbf{b}_o^T \mathbf{b}_o + 2\mathbf{b}_o^T \mathbf{b}_a)^{1/2} \\ &\approx \|\mathbf{b}_o\| + \mathbf{b}_a \cdot \hat{\mathbf{b}}_o \end{aligned}$$

The total field anomaly is then computed by subtracting the magnitude of the earth's magnetic field to give

$$b_a \approx \mathbf{b}_a \cdot \hat{\mathbf{b}}_o.$$

The total field anomaly measured by a magnetometer is the projection of the anomalous field on to the direction of the earth's magnetic field.

2.2.4 Demagnetization

In highly susceptible materials the induced magnetization is reduced by self-demagnetization. This effect can be understood by representing the induced dipole moment of a spheroid as a surface distribution of (fictitious) magnetic charges, as shown in figure 2.1. These charges produce an internal magnetic field which opposes the induced dipole moment and thereby reduces the net moment. If the induced moment is parallel to the long axis of the spheroid, then the demagnetizing field is weaker than if the moment is perpendicular to the long axis, as illustrated in figure 2.1(b).

For a given spheroid a range of induced moments can therefore be generated as the spheroid is rotated about the inducing field. This "feasibility" curve provides a template against which an observed moment can be matched. The minimum distance between the

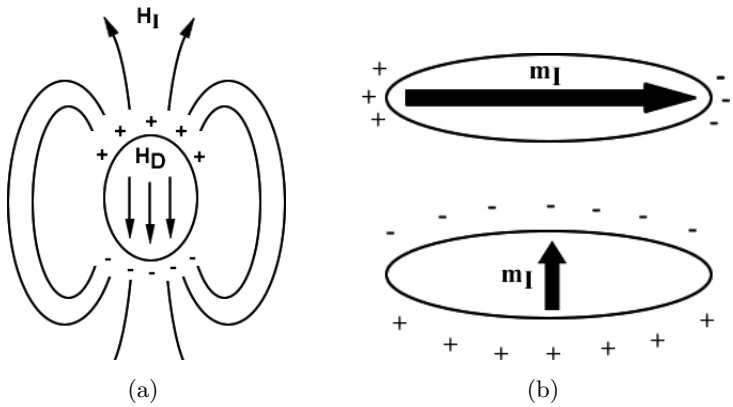


Figure 2.1: (a) Demagnetization of a spheroid. Magnetic charges which produce an induced field $\mathbf{H_I}$ also produce an internal demagnetizing field $\mathbf{H_D}$ which reduces the net induced moment. (b) The demagnetizing field is weakest when the magnetic charges are at their maximum separation, ie when the induced moment $\mathbf{m_I}$ is parallel to the long axis of the spheroid (after [7]).

observed moment and a feasibility curve for a given ordnance type then provides a feature which can be used for UXO discrimination [8].

Unfortunately, there is ambiguity in the interpretation of magnetics data, even for relatively simple objects. Billings showed that there is an infinite family of spheroids which can predict an observed dipole anomaly [8]. This nonuniqueness prevents us from drawing any inferences about shape and orientation of a target; we can only extract information about the induced moment.

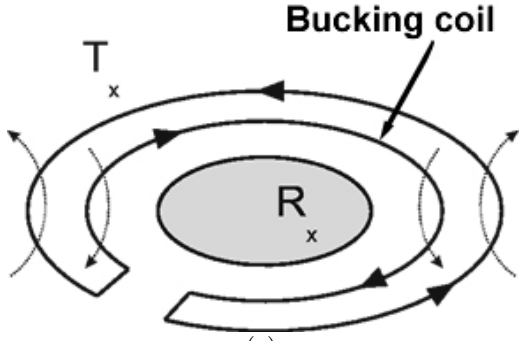
2.3 Time and frequency-domain electromagnetics

Electromagnetic instruments actively transmit a time-varying primary magnetic field which illuminates a buried target. The variation of the primary field induces currents in the target and these currents in turn produce a secondary field which can be measured by a receiver at the surface.

EM data can be acquired in the time-domain or in the frequency-domain. In the time-domain the decay of secondary fields is measured after a primary field is switched off. The frequency-domain mode measures the secondary field during transmission of a primary field. Frequency-domain instrumentation tries to cancel (or "buck out") the primary field at the receiver so that only secondary fields are measured.

While the two modes of operation theoretically provide the same information about a target, in practice there are advantages and disadvantages to both methods. Because time-domain instruments only "listen" for the secondary field after the primary is shut off, the complications of bucking out the primary field are circumvented. However, self-induction in the transmitter and receiver makes a true step-off waveform difficult to achieve, and so it has historically been difficult to sample the secondary field at very early times after the primary field is shut off.

In the frequency-domain mode of operation there are no complications from the step-off of the primary field and information at high frequencies (equivalent to early times in the time domain) can be obtained. However, careful alignment of transmitter, receiver



(a)

Figure 2.2: Operation of a frequency domain EM sensor with concentric transmitter ($\mathbf{T_x}$), bucking and receiver ($\mathbf{R_x}$) coils. The bucking coil is designed to produce a field which cancels the primary field at the receiver.

and bucking coils is required if the secondary field, which is typically orders of magnitude smaller than the primary, is to be accurately measured.

2.3.1 The time and frequency domain responses of a sphere

Figure 2.3 shows the frequency-domain spectrum and the corresponding time-domain response for a conductive, permeable sphere in a uniform field. These forward modellings were computed for the vertical (z) component of the secondary field using the analytic solution derived by Ward [9]. 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 the 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 2.4 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

$$\tau \propto \sigma \mu R^2. \tag{2.10}$$

The time constant is the time at which late-stage exponential decay begins. As shown in figure 2.4, 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.

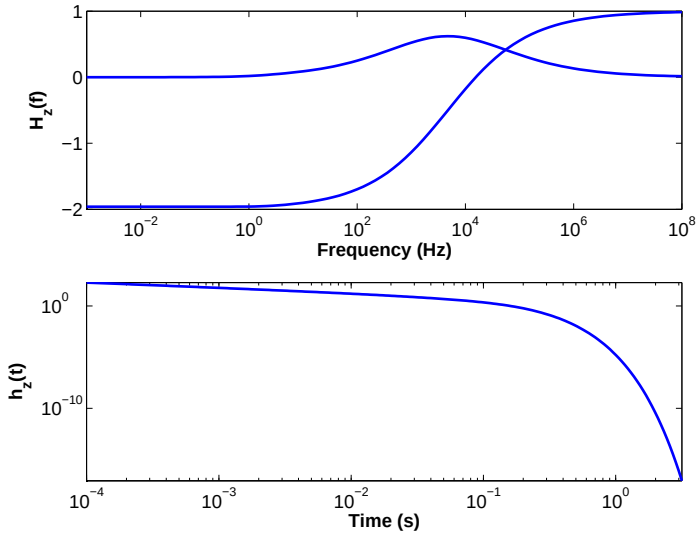


Figure 2.3: Frequency domain (top) and time-domain (bottom) electromagnetic response of a conductive, permeable sphere.

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

$$\xi = \sqrt{\omega\sigma\mu_o}R \tag{2.11}$$

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 σ and R.

2.3.2 Approximate forward models for EM

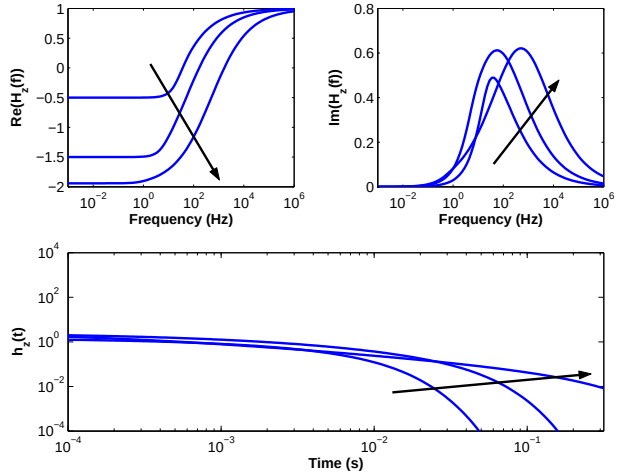
Although considerable insight can be gained by analytic forward modelling for simple shapes, extension of electromagnetic forward modelling to more complex bodies has proven to be a difficult endeavour. Finite difference solutions to Maxwell's equations in both the time and frequency domain are available. However, these methods are typically too computationally intensive for use in UXO applications, where hundreds, if not thousands, of anomalies must be modelled. Rigorous solutions to the electromagnetic scattering problem can be regarded as benchmarks against which simpler models can be validated.

Laboratory and field measurements have shown that many UXOs can be approximated as axisymmetric bodies of revolution. Though many UXOs are heterogenous objects with asymmetries such as fins and nose cones, numerical and analytic models which use approximate spheroids have generated good agreement between observed and predicted data.

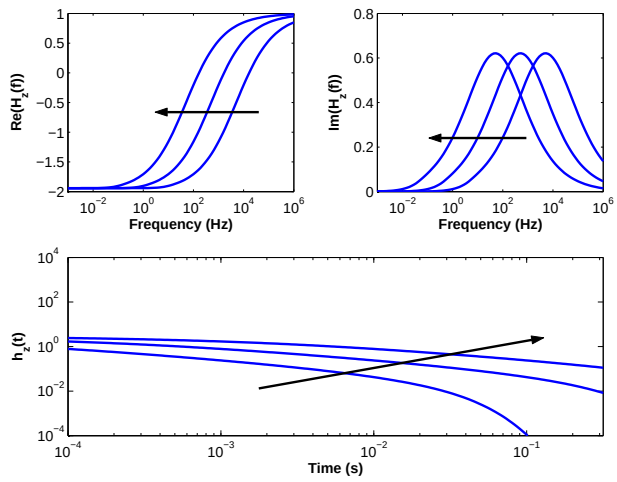
These results have motivated the development of approximate forward models which parameterize the response of an elongated body as a superposition of excitations along axial and transverse directions.

For example, the Pasion-Oldenburg model represents the time-domain response of a UXO as the superposition of two orthogonal dipoles: one oriented along the long axis of the object and the other transverse to the long axis (figure 2.5). The secondary field $\mathbf{B}(\mathbf{r}, t)$ is expressed as

$$\mathbf{B}(\mathbf{r}, t) = \mathbf{B_1}(\mathbf{r})L_1(t) + \mathbf{B_2}(\mathbf{r})L_2(t) \tag{2.12}$$



(a) Frequency domain in-phase response (top left), quadrature response (top right) and time domain response (bottom) for three values of relative permeability: $\mu_r = (2,\ 10,\ 100)$. Conductivity is constant ($\sigma = 10^5$ S/m). Arrows indicate increasing relative permeability.



(b) Frequency domain in-phase response (top left), quadrature response (top right) and time domain response (bottom) for three values of conductivity: $\sigma = \left(1 \times 10^4,\ 1 \times 10^5,\ 1 \times 10^6\right)$ S/m. Relative permeability is constant ($\mu_r = 100$). Arrows indicate increasing conductivity.

Figure 2.4: Dependence of frequency and time-domain responses upon physical properties.

where $\mathbf{B_1}(\mathbf{r})$ and $\mathbf{B_2}(\mathbf{r})$ are the spatial responses of the orthogonal dipoles. The dipoles are assumed to decay independently, with each decay parameterized by

$$L_i(t) = k_i(t + \alpha_i)^{-\beta_i} \exp(-t/\gamma_i),\ \ i = 1, 2. \tag{2.13}$$

Pasion showed that combinations time-decays parameters (k_i, α_i,β_i, and γ_i) are diagnostic of target shape and size and can therefore be used to discriminate between UXO and clutter [10].

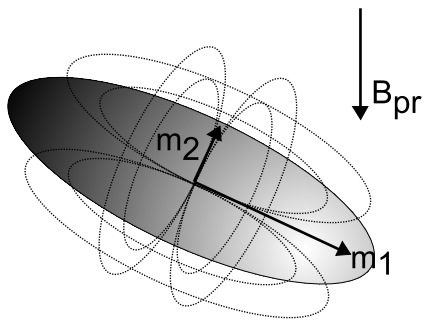


Figure 2.5: The Pasion-Oldenburg model represents the secondary field of an arbitrary object as the sum of fields radiated by two orthogonal induced dipoles (m_1 and m_2).

Inversions using this model have produced good agreement with observed field data and can be used to estimate target depth and orientation. A similar two-dipole model has been developed in the frequency domain by Baum [11]. In this mode, the spectral response of the item is parameterized by resonant frequencies which correspond to decay constants in the time domain.

In magnetics we usually cannot infer orientation or shape information from the observed data. However, with electromagnetic data we can hope to extract orientation and shape information. A simple analogy to explain the extra information available from EM data is the illumination of a spheroid and sphere with a flashlight. If we shine our flashlight along the long axis of the spheroid, it casts a circular shadow, just like the sphere. However, if we illuminate the spheroid and sphere from multiple angles it is clear that they will cast different shadows and we can infer something about the shape of the targets. In a similar manner, the earth's magnetic field provides a static primary field which illuminates the target from one direction. The primary field from a moving EM transmitter illuminates the target from multiple directions and so we can hope to gain some information about target shape with EM data.

Because UXOs are typically prolate spheroids, estimates of target shape can potentially provide a feature for discrimination between UXO and clutter items. A criticism of the approximate forward models described above is the indirect connection between the parameters of the forward models and the physical properties of the target. For example, the relative strengths of the orthogonal dipoles in the Pasion-Oldenburg model provide an indication of target shape, but the link between model parameters and physical properties is qualitative.

I address this criticism in chapter 4 using the method of auxiliary sources (MAS), a numerical forward model which is parameterized in terms of physical properties.

2.3.3 Method of auxiliary sources (MAS)

The method of auxiliary sources can model the secondary magnetic field of an arbitrary body. The body is parameterized by its shape, size, conductivity and relative permeability. MAS formulates the EM forward problem as a boundary value problem. The formulation is related to the method of moments, which requires that the integral equation for the magnetic field be satisfied on the surface of the body. MAS displaces fictitious sources of the secondary field (the auxiliary sources) away from the surface of the conductor and requires that the fields produced by these sources satisfy continuity conditions on the surface

of the body [12].

The first step in implementing MAS is to divide the computational domain into two regions: a region outside the conductor (region 1) and a region inside the conductor (region 2), as shown in figure 2.6. Next consider Maxwell's curl equation for the magnetic field **H**

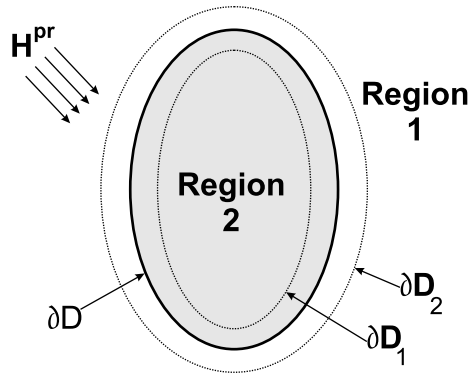


Figure 2.6: Geometry for MAS (after Shubitidze et al.). The boundary of the conductor ∂D defines the boundary between regions 1 and 2. The auxiliary surfaces ∂D_1 and ∂D_2 are concentric with the conductor.

in the frequency domain

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} - i\omega\epsilon \mathbf{E}. \tag{2.14}$$

In region 1 (outside the conductive body), both displacement and conduction currents are negligible and so the magnetic field is irrotational. I can then write the magnetic field as the gradient of a scalar potential Φ

$$\mathbf{H} = -\nabla\Phi. \tag{2.15}$$

By analogy with Gauss' law for the electric field, the scattered magnetic field is

$$\nabla \cdot \mathbf{H}_1^{sc} = \frac{1}{\mu_o}\sigma_m(\mathbf{r}) \tag{2.16}$$

where σ_m is a (fictitious) magnetic charge density situated on the auxiliary surface ∂D_1. So the magnetic scalar potential satisfies Poisson's equation

$$\nabla^2\Phi(\mathbf{r}) = -\frac{1}{\mu_o}\sigma_m(\mathbf{r}). \tag{2.17}$$

The solution is

$$\Phi(\mathbf{r}) = -\int_{\partial D_1} \frac{\sigma_m(\mathbf{r}')}{4\pi\mu_o|\mathbf{r}-\mathbf{r}'|} dS'. \tag{2.18}$$

The scattered magnetic field in region 1 is then obtained by taking the gradient of the above equation. An important feature of MAS is that sources are located outside the region where the fields are calculated. This a major advantage over the method of moments, where integration over regions where source and observation locations coincide must be handled with caution due to the singularity of the Green's function.

Things are not quite so straightforward inside the conductive body (region 2). We can make a quasistatic approximation but the conduction currents are certainly not negligible, so the governing equation is Helmholtz's equation (with sources)

$$\left[\nabla^2 + k^2\right] \mathbf{H_2} = -i\omega\mu\mathbf{P}(\mathbf{r}) \tag{2.19}$$

where the complex wavenumber is $k = \sqrt{-i\omega\sigma\mu}$ and $\mathbf{P(r)}$ is a magnetic current density (A/m^3) situated on the auxiliary surface ∂D_2. The solution is given by

$$\mathbf{H_2}(\mathbf{r}) = [k^2 + \nabla\nabla\cdot]\mathbf{\Pi^m}(\mathbf{r}) \tag{2.20}$$

with the magnetic Hertz potential given by

$$\mathbf{\Pi^m}(\mathbf{r}) = \int\limits_{\partial D_2} \frac{\mathbf{P}(\mathbf{r}')}{4\pi\mu} \frac{e^{-ik|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} dS' \tag{2.21}$$

Having obtained expressions for the magnetic field inside and outside the conductive body, we must impose continuity of the tangential $\mathbf{H}$ field and normal $\mathbf{B}$ field at the interface between the ground and conductor

$$\begin{aligned} \hat{\mathbf{n}} \times (\mathbf{H_1^{sc}} + \mathbf{H^{pr}}) &= \hat{\mathbf{n}} \times \mathbf{H_2} \\ \hat{\mathbf{n}} \cdot (\mathbf{H_1^{sc}} + \mathbf{H^{pr}}) &= \hat{\mathbf{n}} \cdot \mu_\mathbf{r}\mathbf{H_2} \end{aligned} \tag{2.22}$$

where $\mathbf{H}^{pr}$ is the primary field, $\mathbf{H}^{sc}$ is the scattered field and $\mathbf{H_2}$ is the total field inside the body. The problem is now to determine a distribution of auxiliary sources which will satisfy equation 2.22. The source distributions are conveniently expressed by an expansion in terms of orthogonal basis functions $\{\chi_n\}$

$$\begin{aligned} \sigma_m(\mathbf{r}') &= \sum_{n=1}^{N} Q_n \chi_n(\mathbf{r}') \\ \mathbf{P}(\mathbf{r}') &= \sum_{n=1}^{N} \mathbf{P_n} \chi_n(\mathbf{r}') \end{aligned} \tag{2.23}$$

The coefficients Q_n and $\mathbf{P_n}$ can be found by integrating the continuity conditions (equation 2.22) over the physical surface

$$\int\limits_{\partial D} [\xi \cdot (\mathbf{H}_1^{sc} - \mathbf{H}_2)] w_m dS = -\int\limits_{\partial D} [\xi \cdot \mathbf{H}^{pr}] w_m dS \tag{2.24}$$

where ξ is a normal or tangential unit vector on ∂D. A convenient choice of the weighting functions $\{w_m\}$ makes evaluation of these integrals straightforward, producing an linear system of equations. For example, if both the expansion and weighting functions are chosen to be delta functions, then the above expression reduces to point matching. Continuity of the fields produced by auxiliary point sources is enforced at testing points on the surface of the conductor. In this formulation, Shubitdize et al. call the sources positioned on the outer auxiliary surface ∂D_2 magnetic dipoles [12]. While these sources are parameterized in terms of the vector-valued coefficients $\mathbf{P_n}$, inspection of equation 2.21 shows they do not decay as $1/r^3$, as we might expect for dipole sources. To avoid this confusion I call the sources on surface ∂D_2 vector sources.

Point-matching implementation of MAS

As an example of MAS, I implement a point-matching solution for a sphere under axial excitation. Figure 2.7 shows the distribution of points on the surface of the sphere and on concentric auxiliary surfaces. The vector and charge sources are distributed in 50 bands of

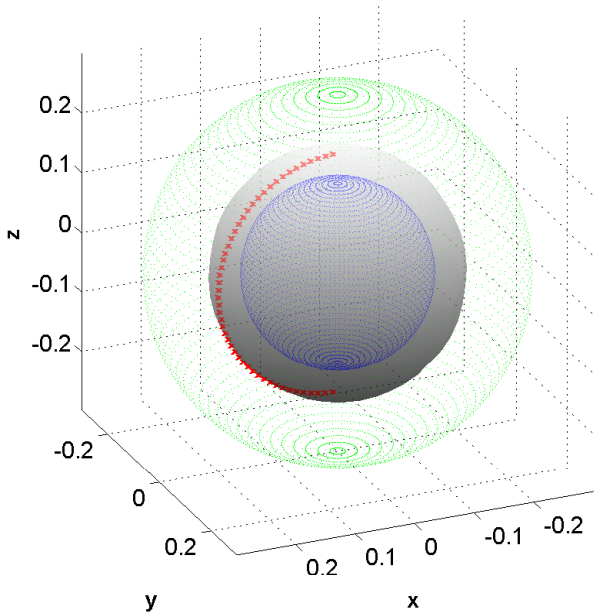


Figure 2.7: Distribution of auxiliary sources (points) and testing points (crosses) on a sphere of radius 0.2 m

constant polar angle, with each band comprised of 120 auxiliary sources (a total of 6000 sources on each auxiliary surface).

The rotational symmetry of the sphere, or indeed any body of revolution, dictates that under a uniform axial excitation the continuity conditions at any two testing points with the same polar angle must be the same. This implies that the strength of auxiliary sources must be constant in a given auxiliary source band of constant polar angle. This greatly reduces the computational burden: in this example we need only solve for the charge strengths of 50 bands, rather than the charge strengths of 6000 individual auxiliary charges. A further simplification is achieved by requiring that the vector sources be oriented tangential to the auxiliary surface ∂D_2 [Shubitidze, personal communication]. Under this assumption, the expressions for the secondary fields produced by the vector sources are considerably simplified, so that we need solve for one coefficient for each band of vector sources (see [12] for details). Figure 2.8 shows the point-matching solution for the z-component of the secondary magnetic field at an observation location 1 m above the centre of the sphere. The sphere has radius 0.2 m, conductivity 10^6 S/m, and relative permeability 150. The auxiliary surfaces have radii of 0.15 m and 0.30 m, respectively. Also shown is an MAS implementation provided by Shubitidze which uses the "thin-skin" approximation at high frequencies [13].

Agreement between the analytic and standard MAS solutions is observed up to approximately 10^3 Hz. At higher frequencies, the linear system becomes ill-conditioned and

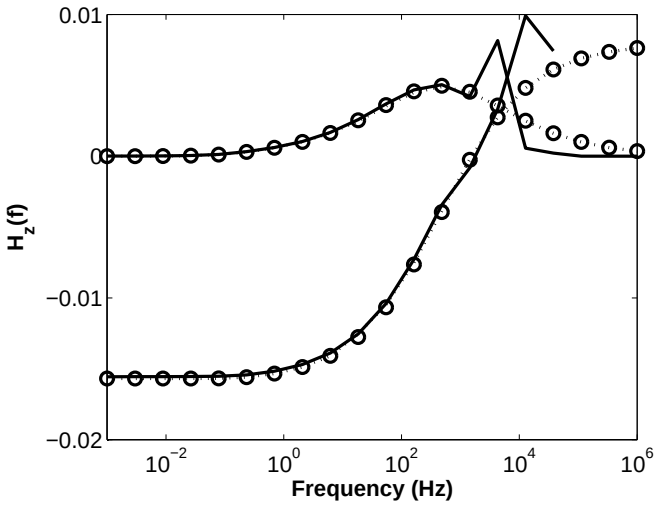
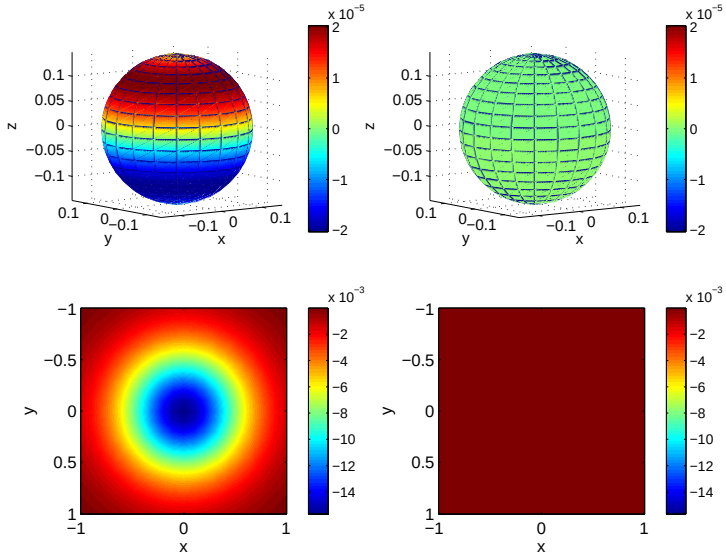
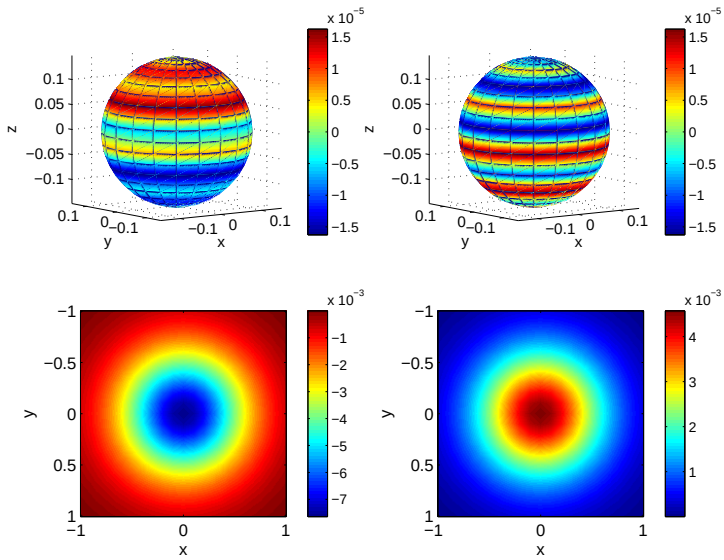


Figure 2.8: MAS point-matching solution (solid line) and analytic solution (circles) for the frequency domain response of a conductive, permeable sphere. Dashed line is MAS point-matching solution with thin-skin approximation at high frequencies provided by Shubitidze.

an accurate solution cannot be obtained. This ill-conditioning is due to the rapid decay of vector sources at high frequencies. This is a weakness of the standard MAS formulation: by displacing sources from the physical surface we avoid singularities in the Green's functions but problems can arise from the rapid falloff of the vector sources. One possible solution is to move the vector auxiliary sources closer to the object surface and to increase the number of sources. However, the increased resolution required at these frequencies quickly makes the linear system intractable, even for numerical solvers (to quote one comedian in the field: "Fugheddaboudit!"). The alternative "thin skin approximation" circumvents this problem by solving only for the exterior (scattered) fields together with a modified boundary condition which accounts for material properties [13].



(a) 10^{-3} Hz



(b) 162 Hz

Figure 2.9: Top left: real auxiliary point sources. Top right: imaginary auxiliary point sources. Bottom left: real response at a height of 1 m above the center of the sphere, plan view. Bottom right: imaginary response, plan view.

The distributions of auxiliary point sources at two frequencies are shown in figure 2.9. At low frequencies, the imaginary component of the charges is very small and so there is no quadrature response. The distribution of real charges divides into a positively charged upper hemisphere and a negatively charged lower hemisphere. Hence the fictitious magnetic charges behave like electric charges to produce a (downward-directed) vertical dipole moment.

At higher frequencies (162 Hz in figure 2.9), the quadrature response is opposite in sign to the in-phase response. In this case the real charge bands have opposite sign to the corresponding imaginary charge bands and produce a dipole moment which is oppositely directed to the imaginary component.

The distribution of auxiliary sources at 162 Hz does not neatly divide into oppositely charged hemispheres. How is a dipolar field produced by this distribution of auxiliary sources? Closer inspection of the charge strengths at higher frequencies reveals that for each band with a given charge there is a corresponding band of equal, but opposite, charge in the other hemisphere (figure 2.11). The corresponding bands are vertically displaced by the same distance from the centre of the sphere.

An equivalent dipole moment can be computed from the charge bands by considering bands of equal but opposite charge situated in the upper and lower hemispheres of the sphere (figure 2.10). Each band is vertically displaced by a distance $d/2$ from the centre of the sphere, so that the total vertical separation between the two bands is d. By analogy with the electric dipole moment, the bands of opposite magnetic charges constitute a vertical dipole moment with magnitude

$$m = Q_{eq} d \tag{2.25}$$

with Q_{eq} an equivalent charge strength for a point charge at the centre of each band (figure 2.10). To compute the equivalent strength Q_{eq} of a point charge, consider an observation

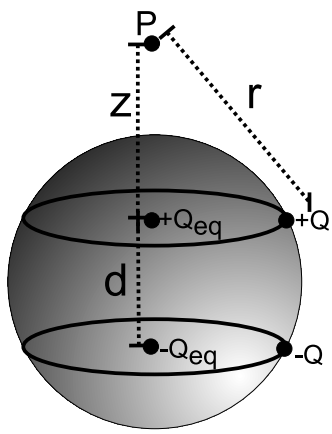


Figure 2.10: Computing an equivalent dipole moment from the MAS solution. Two oppositely charged bands of magnitude Q with vertical separation d are equivalent to two point charges with magnitude Q_{eq}.

location P along the axis of symmetry (figure 2.10). For P far from the centre of the sphere (i.e for z much greater than the radius of the sphere), we require the same secondary field

from charge Q_{eq} and the total charge on a band, so that

$$\begin{aligned} \frac{NQ}{r^2} &= \frac{Q_{eq}}{z^2} \\ \Rightarrow Q_{eq} &= NQ\left(\frac{z}{r}\right)^2. \end{aligned} \tag{2.26}$$

with N the number of charges on a band. As we move from the equator towards both poles we encounter pairs of bands with increasing separation. Hence the total dipole moment produced by the auxiliary charge sources can be represented as a superposition of vertical dipole moments, all located at the centre of the sphere (figure 2.11). The equivalent dipole

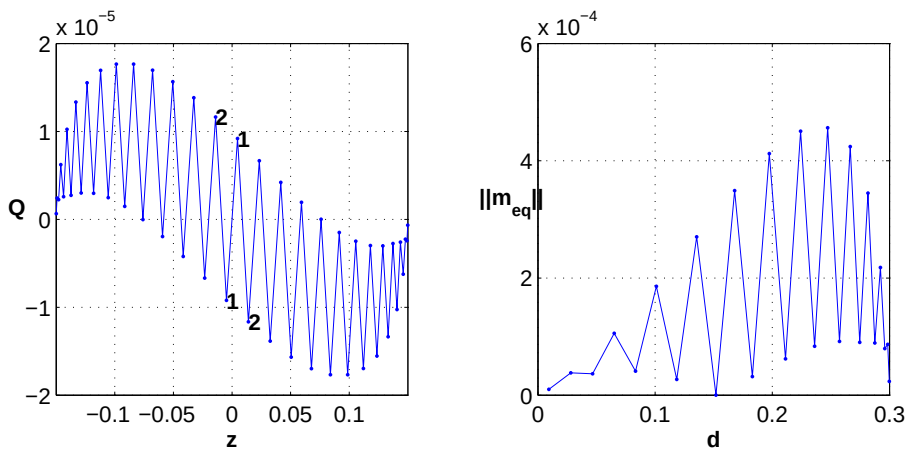


Figure 2.11: Left: Real charge band strengths Q at 162 Hz as a function of vertical location z. The centre of the auxiliary surface is at $z = 0$. The first two pairs of oppositely charged bands, constituting vertical dipole moments, are labelled 1 and 2. Right: Magnitude of the equivalent dipole moments as a function of band separation d.

anomaly computed in this manner is identical to the superposition of fields from individual charges (figure 2.12). The agreement between analytic, MAS and equivalent dipole models shown in figure 2.12 is observed at other frequencies. However, as will be discussed in section 4.3, the scattered fields radiated by non-spherical bodies are not wholly dipolar, and so the equivalent dipole is, in general, an approximation.

The MAS solution also provides the total magnetic field inside the body (figure 2.14). The field is computed as the superposition of magnetic fields produced by vector sources. At low frequencies, the real part of the induced magnetic field is constant within the sphere (figure 2.14(a)). The sphere is illuminated by primary fields which are constant throughout the sphere and the secondary fields are produced by volume currents, resulting in a uniform induced magnetization. The imaginary component of the field is small relative to the real component at this frequency and quickly falls off to zero with increasing distance from the sphere. In contrast, at higher frequencies the skin depth decreases and so the induced magnetic field is nonzero only in a small region near the surface. In this case, the secondary fields are produced by surface currents. For the observation locations selected in figure 2.14, the z-component of the magnetic field is the tangential component of the field and so it is continuous at the boundary between sphere and air.

Note that the MAS solution in figure 2.14 is for the *total* magnetic field, both inside and outside the sphere. The total field outside the sphere is the superposition of the scattered

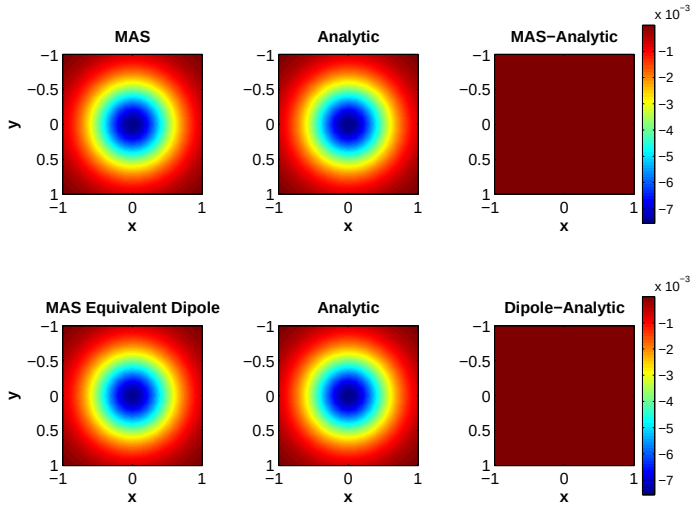


Figure 2.12: Real component of scattered fields at 162 Hz predicted by MAS, analytic forward model, and an equivalent dipole computed from the MAS solution. Observations are in plan view at 1m above the sphere centre.

and primary fields. For observation locations lying along the x-axis, the vertical scattered field is opposite to the primary field, and so in figure 2.14 the total field increases as the magnitude of the scattered field decreases.

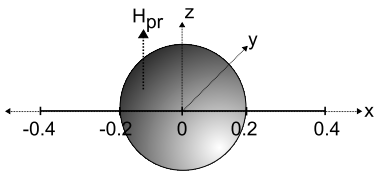
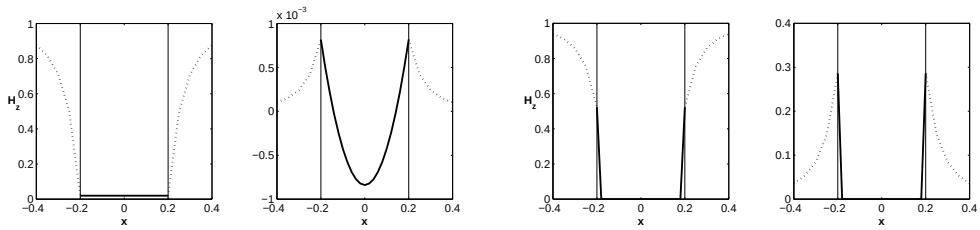


Figure 2.13: Solid horizontal line indicates extent of observation locations used to generate figure 2.14.



(a) 10^{-3} Hz (b) 162 Hz

Figure 2.14: MAS solution for vertical (z) fields at observation locations inside the sphere (heavy solid lines) and outside the sphere (heavy dashed lines). Left plots in (a) and (b) are the real components, right plots are the imaginary components. Vertical lines indicate the boundary between sphere and air.

2.3.4 Forward modelling of sensor data with MAS

The method of auxiliary sources described above provides a frequency-domain solution for the secondary magnetic field produced by a body for axial and transverse excitations. This is not a full solution to the forward problem: for most EM sensors we must compute the z-component of the secondary field with an arbitrary excitation and for a target with arbitrary orientation. For an implementation of MAS which uses magnetic charges, the secondary field $\mathbf{B^{sc}}$ at observation location $\mathbf{r}$ is

$$\mathbf{B^{sc}}(\mathbf{r}) = \sum_{i=1}^{N} Q_i \mathbf{G}(\mathbf{r}, \mathbf{r_i}) \tag{2.27}$$

with Q_i the (complex) magnetic charge strength for the ith magnetic charge. The summation is over all magnetic charges on the auxiliary surface. The Green's function for a magnetic charge with position vector $\mathbf{r_i}$ is

$$\mathbf{G}(\mathbf{r}, \mathbf{r_i}) = \frac{\mathbf{r} - \mathbf{r_i}}{|\mathbf{r} - \mathbf{r_i}|^3} \tag{2.28}$$

The primary field produced by the transmitter loop can be computed with the Biot-Savart law [10]. However, if the target is more than a transmitter radius away from the transmitter, the primary field can be approximated as a dipolar field. I therefore compute the primary field to be that of a dipole at the centre of the transmitter loop. Moreover, if the target is small compared to the distance to the transmitter, then the magnetic field is assumed to be uniform over the surface of the target (ie the MAS problem is solved for a uniform primary excitation).The uniform primary field excitation is then the value of the dipole field at the centre of the target.

The secondary field $\mathbf{B^{sc}}$ at observation location $\mathbf{r}$ is calculated as

$$\mathbf{B^{sc}}(\mathbf{r}) = \mathbf{A}^T \hat{\mathbf{B}}^{\mathbf{sc}} \mathbf{A} \mathbf{B^{pr}} \tag{2.29}$$

where the secondary magnetic field for unit axial and transverse excitations in target-centred coordinates is expressed as

$$\hat{\mathbf{B}}^{\mathbf{sc}} = \begin{bmatrix} \hat{B}^a_a & \hat{B}^{t1}_a & \hat{B}^{t2}_a \\ \hat{B}^a_{t1} & \hat{B}^{t1}_{t1} & \hat{B}^{t2}_{t1} \\ \hat{B}^a_{t2} & \hat{B}^{t1}_{t2} & \hat{B}^{t2}_{t2} \end{bmatrix}. \tag{2.30}$$

Here the superscript denotes unit axial (a) or transverse ($t1,\ t2$) excitations and the subscript denotes the component of the secondary field in target-centred coordinates.

The matrix $\mathbf{A}$ is an orthogonal rotation matrix which rotates the primary field $\mathbf{B^{pr}}$ into target centered coordinates. I adopt the coordinate system shown in Figure 2.15 with z-positive up. The declination angle ϕ defines the angle clockwise from the y axis. The inclination angle θ measures the angle from the horizontal plane (positive upwards). With this convention, the Euler rotation matrix is

$$\mathbf{A} = \begin{bmatrix} \cos(\phi) & -\sin(\phi) & 0 \\ \sin(\theta)\sin(\phi) & \sin(\theta)\cos(\phi) & \cos(\theta) \\ \cos(\theta)\sin(\phi) & \cos(\theta)\cos(\phi) & -\sin(\theta) \end{bmatrix}. \tag{2.31}$$

In target-centered coordinates the MAS solution for the secondary field is computed for unit axial and transverse excitations. These excitations are scaled by the primary field and rotated back into geographic coordinates by the inverse rotation matrix.

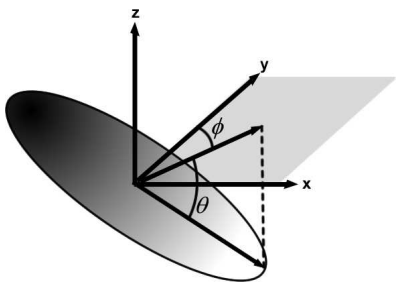


Figure 2.15: Coordinate system for computation of EM data.

2.3.5 Forward modelling of magnetics data with MAS

At low frequencies the EM response of a magnetically permeable target is entirely in-phase with the primary field. A low-frequency MAS forward modelling can therefore be used to solve for the anomalous magnetic field produced by a spheroid for a static primary field (ie the earth's magnetic field). Figure 2.16 shows a forward modelling for the anomalous magnetic field produced by a sphere and spheroid. The analytic solution is computed using demagnetization factors, as explained in [4]. The observations are modelled using the earth's magnetic field at Yuma, Arizona (58° inclination, 12° declination and 48000 nT field strength) The MAS solution is computed at a frequency of 10^{-6} Hz. The agreement between analytic and MAS solutions indicates that we can model both magnetics and EM data with a single forward model. This makes MAS a promising candidate as a forward model in joint inversion of these data types.

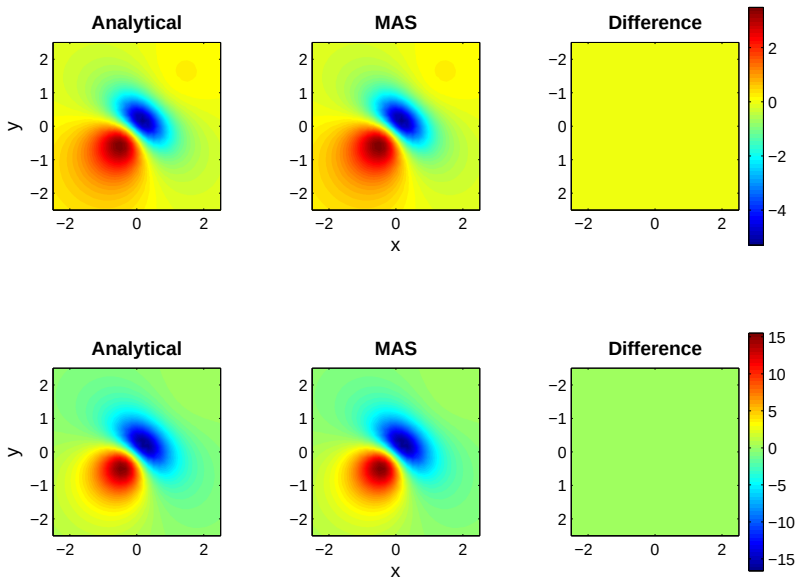


Figure 2.16: Analytic and MAS forward modellings of the anomalous magnetic field (nT) produced by a sphere (top row) and a spheroid (bottom row).

2.4 Discussion and conclusions

In this chapter, I have reviewed forward modelling of magnetic and electromagnetic data in UXO applications. Magnetics data are typically handled using a dipole forward model. Similarly, time and frequency domain EM data can be accurately modelled with parametric models which represent the spatial response as two orthogonal dipoles. In the time domain, the Pasion-Oldenburg model represents the decay of the secondary field in terms of independent decay functions for the two dipoles. In chapter 4, I examine the dependencies of the Pasion-Oldenburg model parameters upon target shape using MAS as a benchmark forward model.

Chapter 3

Inversion

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

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

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

$$\hat{\mathbf{m}} = F^{-1}\{\mathbf{d^{obs}}\}.$$

where the observed data $\mathbf{d^{obs}}$ are the true data plus noise ϵ

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

For magnetic and electromagnetic data the number of observations typically outnumbers the number of model parameters in an approximate forward model such as MAS. 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 L2 misfit function

$$\phi_d = \|\mathbf{W_d}(\mathbf{d}^{obs} - F\{\mathbf{m}\})\|^2. \tag{3.1}$$

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

$$\mathbf{W_d}_{ii} = \frac{1}{\sigma_i}. \tag{3.2}$$

If the forward modelling operator is linear, then there is a unique global minimum to the misfit function. In this case, the solution can be obtained in one step by solving a linear system of equations. The minimum of ϕ_d corresponds to a single element of model space. This is in contrast to the underdetermined case, where an infinite number of models can predict the observed data to a specified degree.

3.1 Linearized inversion

If the forward modelling operator is not linear, then 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 in chapter 2 are nonlinear functions of the input model parameters.

A deterministic approach to solving nonlinear inverse problems is to solve the problem iteratively. The first step is to linearize the misfit function by expanding equation 3.1 as a

Taylor series with a model perturbation $\delta\mathbf{m}$

$$\phi_d(\mathbf{m}+\delta\mathbf{m}) = \phi_d(\mathbf{m}) + (\nabla_m\phi_d)|_{\mathbf{m}}\delta\mathbf{m} + \frac{1}{2}\delta\mathbf{m}^T(\nabla_m^2\phi_d)|_{\mathbf{m}}\delta\mathbf{m} + \dots \tag{3.3}$$

The gradient of the misfit function (equation 3.1) with respect to the model parameters is

$$(\nabla_m\phi_d)|_{\mathbf{m}} = -\mathbf{J}^T\mathbf{W_d}^T\mathbf{W_d}(\mathbf{d^{obs}} - F\{\mathbf{m}\}) \tag{3.4}$$

where the sensitivity matrix $\mathbf{J}$, evaluated at the model $\mathbf{m}$, has elements

$$J_{ij}|_{\mathbf{m}} = \frac{\partial F\{\mathbf{m}\}}{\partial m_j}. \tag{3.5}$$

The operation $(\nabla_m^2\phi_d)$ defines the symmetric Hessian matrix $\mathbf{H}$ with elements

$$H_{ij} = \frac{\partial^2\phi_d}{\partial m_i \partial m_j}. \tag{3.6}$$

For an L2 data misfit function the Hessian is

$$\mathbf{H} = \mathbf{J}^T\mathbf{W_d}^T\mathbf{W_d}\mathbf{J} - (\nabla_m\mathbf{J}^T)\mathbf{W_d}^T\mathbf{W_d}(\mathbf{d^{obs}} - F\{\mathbf{m}\}). \tag{3.7}$$

The second term in this expression contains second-order information about the curvature of the objective function and is often negligible in comparison to the first term [14].

Minimizing a quadratic approximation to the objective function (equation 3.3) with respect to the model perturbation yields the following expression

$$\mathbf{J}^T\mathbf{W_d}^T\mathbf{W_d}\mathbf{J}\delta\mathbf{m} = -\mathbf{J}^T\mathbf{W_d}^T\mathbf{W_d}(\mathbf{d^{obs}} - F\{\mathbf{m}\}). \tag{3.8}$$

This expression is the Gauss-Newton method with the Hessian matrix approximated by $\mathbf{J}^T\mathbf{W}_d^T\mathbf{W}_d\mathbf{J}$. 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 no further reduction in misfit is achieved.

Difficulties can arise with the Gauss-Newton method if the approximate Hessian is ill-conditioned. This can be circumvented by applying a singular value decomposition to the approximate Hessian and discarding small singular values (truncated SVD).

Alternatively, a trust region method can be used to modify the search direction and regularize the approximate Hessian. The Levenberg-Marquardt method solves the following expression for the model perturbation at each iteration

$$(\mathbf{J}^T\mathbf{W}_d^T\mathbf{W}_d\mathbf{J} + \lambda\mathbf{I})\delta\mathbf{m} = -\mathbf{J}^T\mathbf{W}_d^T\mathbf{W}_d(\mathbf{d^{obs}} - F\{\mathbf{m}\}) \tag{3.9}$$

with λ a non-negative Lagrange multiplier. This parameter controls the step direction: for $\lambda \to \infty$ the model perturbation is in the steepest-descent direction, while for $\lambda \to 0$ the model perturbation is in the Gauss-Newton direction. Trust region methods therefore adjust the step direction within a trusted region of model space where the quadratic approximation to the objective function is considered valid.

Various authors propose different methods of determining a value of λ at each iteration. Dennis and Schnabel outline a method to calculate this parameter so that the resulting step size is approximately the radius of the trust region [14]. I adopt the simpler heuristic

originally suggested by Marquardt [15]. If the misfit is decreased by a model step using the current value of lambda, I update $\lambda \rightarrow \lambda/\nu$, where ν is a constant factor which I set equal to 10. If the misfit is increased by a model step using the current value of λ, I discard this step and increase λ by a factor of ν until the misfit decreases. I also monitor the condition number of the approximate Hessian to guard against ill-conditioning. If the condition number exceeds a maximum $maxcond = 10^3$, then I increase λ until the approximate Hessian is regularized by the addition of this parameter to its diagonal elements.

Convergence can be evaluated with the misfit function. Under the assumption of independent identically-distributed Gaussian noise on the data, the data misfit function is a χ^2 random variable with expected value $\langle \chi^2 \rangle = N$, where N is the number of data. As in an underdetermined problem, this expected value defines a target misfit which must be achieved by the solution. Since the misfit is a random variable, we do not expect to exactly obtain the target misfit, but a model which produces a misfit "close" to this target misfit (ie within about 10%) can be said to adequately reproduce the data. The solution obtained with this approach is nonunique, since different models may satisfy the convergence criterion. However, barring convergence to a local minimum, I find that the final models obtained with this criterion are not significantly different.

If the algorithm is given a poor starting model it may converge to a local minimum of the misfit function. It is therefore important to have a second convergence criterion based on relative changes in the model parameters. Dennis and Schnabel suggest termination when

$$\frac{max(|m_i^j - m_i^{j-1}|)}{max(|m_i|, 1)} < \epsilon_i \tag{3.10}$$

with m_i^j and m_i^{j-1} the values of the ith model parameter at successive iterations $j-1$ and j [14]. The presence of $max(|m_i|, 1)$ in the denominator safeguards against blowup when m_i is close to zero.

As an example of linearized inversion using the Marquardt-Levenberg algorithm, I consider the simple case of inverting frequency-domain data using the analytic forward model for a sphere derived in [16]. The data are measured in a plane above the target at 121 regularly-spaced observation locations and at 10 logarithmically-spaced frequencies ranging between 10^3 Hz and 10^8 Hz. I treat the real and imaginary parts of the data as independent observations, so that there are a total of 2420 data. To generate synthetic data, I add Gaussian random noise to each datum d_i with standard deviation

$$\sigma_i = 5\% \ of \ datum + floor.$$

The parameter $floor$ is a minimum error which ensures that data with small absolute values have realistic errors and do not dominate the misfit.

The model vector $\mathbf{m}$ is

$$\mathbf{m} = [x, y, z, \mu_r, \sigma, R]^T. \tag{3.11}$$

with $x,\ y,\ z$ the target location, μ_r the relative permeability, σ the conductivity, and R the radius. All model parameters but the $[x, y]$ position are required to be positive (z is positive down in this model). This condition is enforced by making a nonlinear transformation of the model parameters

$$\tilde{\mathbf{m}} = [x, y, \sqrt{z}, \sqrt{\mu_r}, \log_{10}(\sigma), \sqrt{R}]^T. \tag{3.12}$$

The conductivity is transformed using a logarithm because this transformation ensures positivity and scales this parameter such that it varies over approximately the same order of magnitude as the other rescaled parameters.

Table 3.1 shows the progression of the Levenberg-Marquardt algorithm for inversion of synthetic data. The estimated model agrees well with the true model in this simulation.

Iteration	Misfit	x	y	μ_r	σ	R	z
0	4585.811	0.500	0.500	15.000	1000.000	0.500	1.000
2	3233.974	0.127	0.116	17.098	64.240	1.028	1.974
4	2467.132	0.024	0.000	17.869	16.207	0.985	1.518
6	2456.011	0.025	0.006	9.667	9.467	1.002	1.514
8	2455.254	0.026	0.005	9.763	9.201	1.006	1.514
True Model		0	0	10	10	1	1.5

Table 3.1: Iterations of Levenberg-Marquardt algorithm for inversion of frequency domain data using an analytic forward model. The expected misfit is $\phi_d* = 2420$ and the misfit of the true model is $\phi_d^{true} = 2458$.

Figure 3.1 shows the path of the inversion algorithm through cross sections of model space. The cross section with respect to x and y target coordinates is a bullseye shape characteristic of uncorrelated parameters. Following the gradient of the objective function leads directly to the minimum. However, the cross section with respect to relative

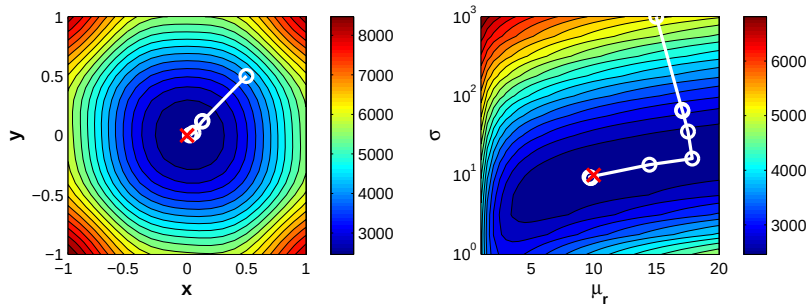


Figure 3.1: Cross sections of the model objective function for an inversion of EM data using an analytic forward model. The path of the Levenberg-Marquardt algorithm through model space is plotted with circles and the true model parameters are shown with a cross.

permeability and conductivity is elongated, indicating a strong correlation between these parameters. For the first two iterations of the algorithm, the approximate Hessian matrix is regularized with a large value of λ so that these steps are in a steepest descent direction (figure 3.2). Once the algorithm reaches the bottom of the misfit "valley", less regularization is required and λ is decreased accordingly. Near convergence, Gauss-Newton type steps allow the algorithm to move efficiently to the global minimum of the misfit. If a steepest descent algorithm were used in this problem, we would expect slow convergence due to the correlation between these parameters.

The positive correlation between conductivity and permeability indicated by figure 3.1 seems, at first inspection, inconsistent with the analytic expression for the frequency-domain transfer function (Fourier transform of the impulse response) derived by Wait [16]

$$\psi(\omega) = -\frac{(2\mu_r + 1)(\tanh(\alpha) - \alpha) + \alpha^2 \tanh(\alpha)}{(\mu_r - 1)(\tanh(\alpha) - \alpha) - \alpha^2 \tanh(\alpha)} \tag{3.13}$$

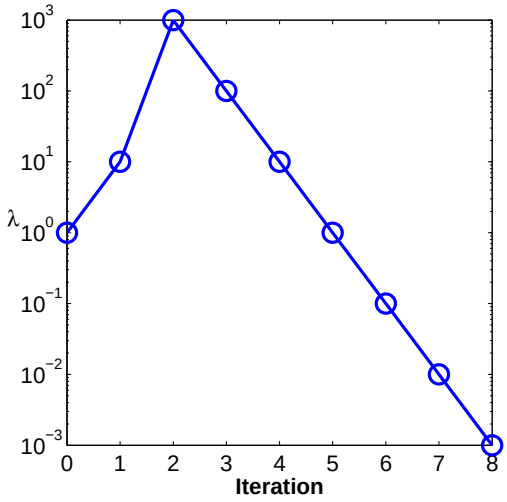


Figure 3.2: Regularization parameter λ as a function of iteration for the inversion in table 3.1

with

$$\begin{aligned} \alpha &= \sqrt{\frac{\mu\sigma\omega}{2}}R(1+i) \\ &= \beta(1+i). \end{aligned} \tag{3.14}$$

and $\beta = \sqrt{\frac{\mu\sigma\omega}{2}}R$. Conductivity and permeability appear as a product in the above expression, and so we might expect these parameters to be negatively correlated. However, for ω sufficiently high that $\beta \gg 1$ we can make the approximation $\tanh(\alpha) \approx 1$ so that

$$\psi(\omega) \approx -\frac{(2\mu_r+1)(1-\alpha)+\alpha^2}{(\mu_r-1)(1-\alpha)-\alpha^2} \tag{3.15}$$

For moderately permeable objects (say $\mu_r \geq 10$), we can make the further approximations $(2\mu_r+1) \approx 2\mu_r$ and $(\mu_r-1) \approx \mu_r$. Some further manipulation leads to the expression

$$\psi(\omega) \approx -\frac{2-\beta/\mu_r(1+i)}{1+\beta/\mu_r(1+i)} \tag{3.16}$$

with

$$\frac{\beta}{\mu_r} = \sqrt{\frac{\mu_o\sigma\omega}{2\mu_r}}R. \tag{3.17}$$

The impulse response for moderately permeable objects depends upon the *ratio* of conductivity and relative permeability and these parameters are therefore positively correlated. Figure 3.3 shows the exact and approximate impulse responses for the sphere used in the synthetic inversion in this section. The two expressions agree at higher frequencies where the assumption $\beta \gg 1$ is valid.

This result suggests that for permeable objects it may be preferable to parameterize the model vector in terms of σ/μ_r rather than solve for these parameters separately. In this synthetic example there is sufficient low frequency information available to recover accurate

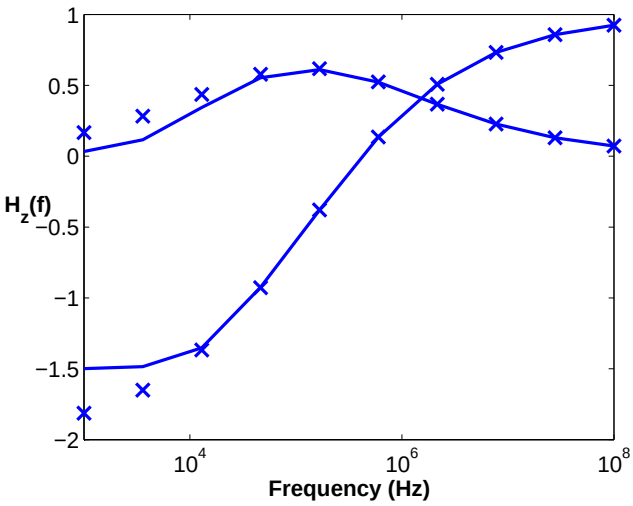


Figure 3.3: True (crosses) and approximate (solid lines) impulse response of the sphere used in the example synthetic inversion.

estimates of the permeability and conductivity. However, the radius (1 m), conductivity (10 S/m), and frequency range used here are not particularly realistic for UXO applications.

To investigate this effect further, I forward model the frequency and time-domain EM response using more reasonable physical properties and the range of time-channels from the EM-63 (figure 3.4). In this figure, the conductivity and permeability are varied such that their ratio remains fixed. For a relatively small sphere (figure 3.4(a)), the impulse response only depends upon permeability and conductivity at low frequencies and late times. Furthermore, there is no distinguishable difference between the responses of moderately ($\mu_r = 10$) and highly permeable ($\mu_r = 100$) spheres. For a larger sphere (figure 3.4(b)), we cannot make any inferences about conductivity and permeability. These results indicate that an inversion of EM data which uses physical parameters may, at best, tell us whether a target is magnetic, but an accurate estimate of the permeability cannot be recovered. This is not a "showstopper" for UXO classification, since permeability and conductivity are not likely to be useful features for discriminating between UXO and clutter.

3.2 Uncertainty appraisal

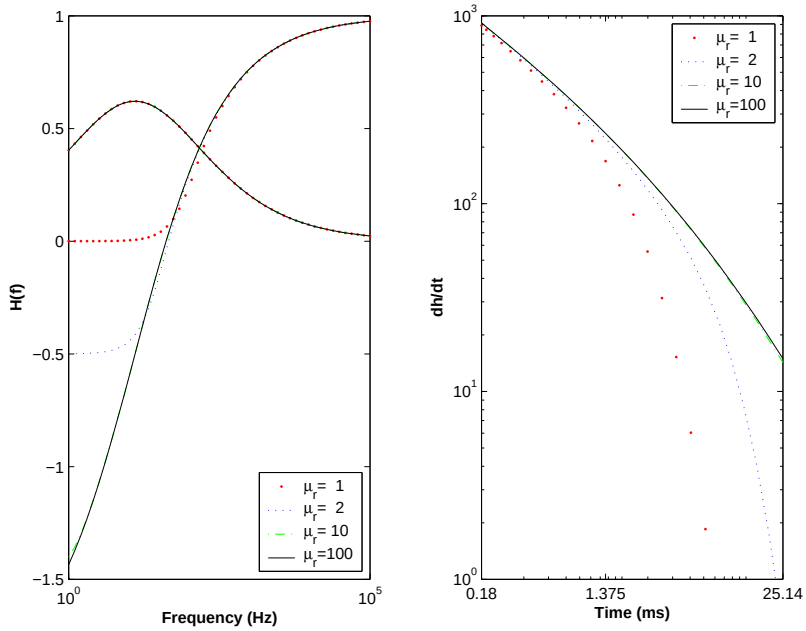
A complete solution to an overdetermined inverse problem must include not only estimates of the model parameters, but also estimates of the parameter uncertainties.

For an overdetermined linearized inverse problem with Gaussian errors on the data, the model parameters are Gaussian distributed with covariance

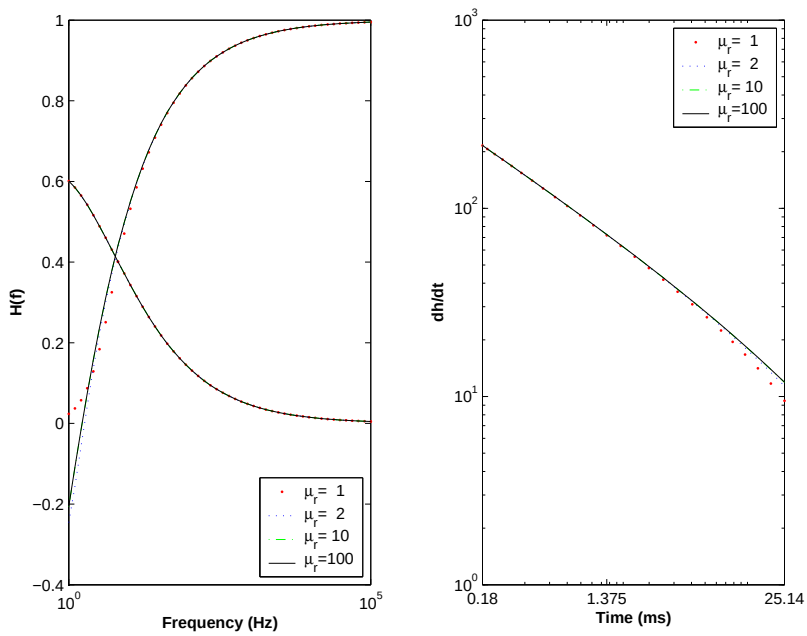
$$cov(\mathbf{m}) = \mathbf{H}^{-1}. \tag{3.18}$$

This result states that if the objective function has a large curvature for a given model (ie the eigenvalues of the Hessian are large) then this model will have a small uncertainty. Conversely, if the model resides in a broad minimum of the objective function then there is a large uncertainty in the model parameters [17].

A linearized uncertainty analysis may not be valid if the objective function is highly nonlinear. In this case the local quadratic approximation provides a poor approximation to



(a) $R = 0.1$ m



(b) $R = 0.5$ m

Figure 3.4: Frequency and time-domain impulse responses for two spheres of different radii. In each plot the conductivity and permeability are varied such that their ratio remains constant.

the actual objective function. Hence uncertainties estimated with equation 3.18 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)

$$p(\mathbf{m}|\mathbf{d^{obs}}) \propto p(\mathbf{d^{obs}}|\mathbf{m})p(\mathbf{m}). \tag{3.19}$$

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

$$p(\mathbf{d^{obs}}|\mathbf{m}) \propto \exp[-\phi_d] \tag{3.20}$$

with ϕ_d the L2 misfit function. The posterior probability is then computed as

$$p(\mathbf{m}|\mathbf{d^{obs}}) = \frac{\exp[-\phi_d]p(\mathbf{m})}{\int \exp[-\phi_d]p(\mathbf{m})d\mathbf{m}}. \tag{3.21}$$

Here the normalizing integral is over all of model space. A special case in the Bayesian framework arises with a uniform prior, Gaussian errors on the data and a linear (or linearized) misfit. In this case the likelihood function is itself a Gaussian distribution and the normalization in the above equation can be evaluated analytically. Maximizing the posterior probability distribution then corresponds to maximizing the likelihood function, or equivalently, minimizing the negative of the log likelihood. By inspection of equation 3.20, the maximum likelihood estimate of the model is obtained by minimizing ϕ_d. Hence the L2 data misfit used to solve most overdetermined problems has a probabilistic justification [17].

For a nonlinear forward problem, the normalizing integral is often difficult to evaluate analytically or numerically, especially in high-dimensional model spaces. However, the PPD for a nonlinear problem can be estimated numerically using the *Gibbs' sampler* algorithm. This algorithm works by randomly perturbing model parameters and accepting these perturbations according to the Metropolis criterion

$$\eta \leq exp[-\Delta\phi_d]. \tag{3.22}$$

If the change in the objective function $\Delta\phi_d$ is less than or equal to η, then the perturbation is accepted. At each model perturbation η is drawn from a uniform random distribution on the interval $[0, 1]$.

This scheme is a Markov chain; acceptance of the perturbed model parameter depends only on the current value of that model parameter. After a sufficient number of samples the chain of accepted models will converge to a stationary distribution which is in fact the posterior distribution [18].

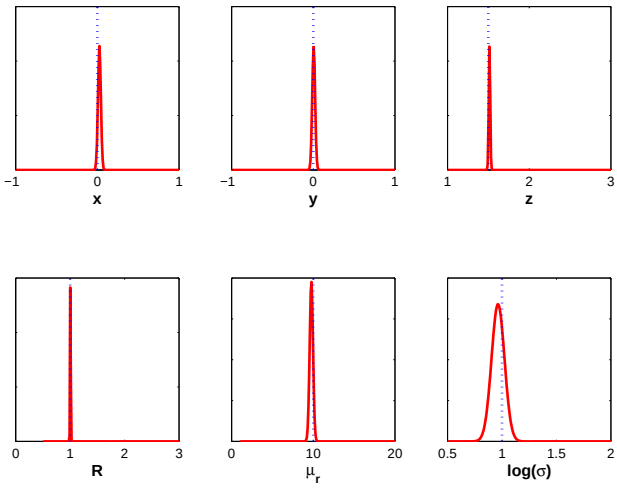
I adopt the fast Gibbs sampler algorithm developed by Dosso to sample the posterior distribution [18]. A key feature of this algorithm is the use of two independent samplers. Convergence of these samplers to the same distribution, as measured by the maximum difference in their cumulative distributions, 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.

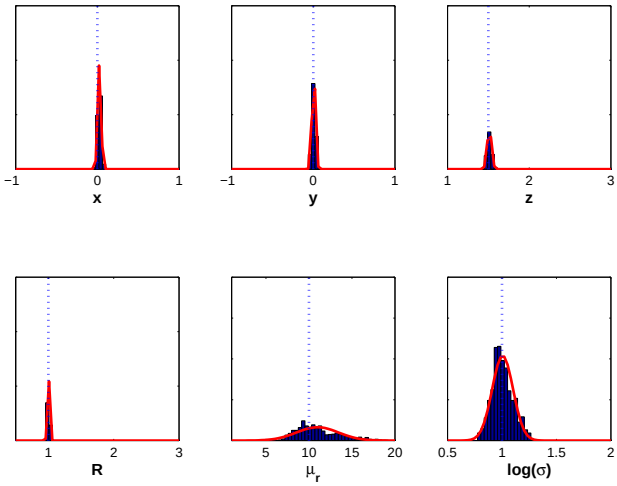
$$p(m_i|\mathbf{d^{obs}}) = \int p(\mathbf{m}|\mathbf{d^{obs}})dm_1 dm_2 \ldots dm_{i-1} dm_{i+1} \ldots dm_N \tag{3.23}$$

Figure 3.5 shows linearized and nonlinear uncertainty appraisals for the synthetic inversion summarized in table 3.1 The nonlinear appraisal generates marginal distributions

which can be reasonably approximated as Gaussians. The nonlinear marginal PPDs are generally broader than those generated by the linearized appraisal, and the two methods produce significantly different distributions for both conductivity and permeability. This discrepancy is due to the strong positive correlation between these parameters. The misfit function has a narrow "valley" when plotted as a function of conductivity and permeability (figure 3.1). Consequently, the nonlinear appraisal, which samples directly from the misfit surface and is sensitive to these correlations, estimates a relatively large uncertainty for these parameters. In contrast, the Hessian used for linearized appraisal is an approximation to the curvature of the misfit function at the solution. This approximation overestimates the curvature at the solution so that the resulting uncertainty is too small.



(a) Linearized uncertainty appraisal.



(b) Nonlinear uncertainty appraisal. Gaussian distributions with the mean and variance of the Gibbs sample are also shown as a solid line.

Figure 3.5: Comparison of uncertainty appraisal methods for inversion of frequency domain EM data with an analytic forward model. The true model parameters are shown with a dashed line.

3.3 Discussion and conclusions

In this chapter, I have applied a linearized inversion algorithm to a simple problem. Although I have used a simple analytic forward model, this example has many of the considerations encountered when using approximate forward models. Parameters are typically estimated with a linearized algorithm and scaling and positivity must often be applied.

Because models recovered by linearized inversion will subsequently be used for classification, accurate estimation of parameter uncertainties is essential. In chapter 6 I will investigate the effect of parameter uncertainties on classification with a Monte Carlo simulation.

Both its parameterization in terms of physical properties and its ability to accurately model magnetic and electromagnetic data make MAS a promising candidate as a forward model for joint inversion. I have implemented an initial MAS inversion for target location and orientation, with physical properties (size, shape, conductivity and permeability) fixed at some initial guess. Under the assumption that the primary field is uniform over the surface of the target, the auxiliary source strengths do not depend on the position and orientation of the target. This inversion therefore required only one solution for the auxiliary sources. Using this method, I successfully recovered estimates of target location and orientation from synthetic magnetic and frequency-domain EM data.

Despite the computational savings offered by this approach, I found that MAS inversions were prohibitively slow for application to real data. However, this is a preliminary result and there are avenues of investigation which may significantly improve the speed of the forward modelling.

A subsequent inversion for target properties will also encounter the strong positive correlation between conductivity and permeability seen in my simple example. Sun et al. have shown that this positive correlation also exists for arbitrary spheroids [19]. If there is sufficient low-frequency or late-time data then it is possible to infer these parameters separately. Alternatively, we might invert for the ratio of these parameters.

Chapter 4

Validating parametric models

In this chapter, I use MAS as a benchmark forward model for investigating the dependence of parametric models on physical properties. Of particular interest is the relationship between the shape of an item and the model parameters estimated with parametric inversion. The dependence of parameters upon other physical properties (conductivity, permeability, size) can be investigated with analytic modelling. However, because no analytic solutions previously existed for the EM response of prolate or oblate spheroids, the connection between shape and model parameters can be verified using modelling with MAS.

A commonly-used forward model for TDEM data is the Pasion-Oldenburg model, which parameterizes the time decay response as

$$L(t) = k_i(t + \alpha_i)^{-\beta_i} \exp(-t/\gamma_i) \tag{4.1}$$

where t is the time since the step-off of the primary field and α_i, β_i, and γ_i determine the shape of the decay curve for the ith dipole source. The response of an item is then given as the superposition two orthogonal dipoles which decay in time according to the above expression. The first dipole represents currents circulating perpendicular to the semi-major axis of the item while the second dipole models currents circulating perpendicular to the semi-minor axis.

Pasion demonstrated via laboratory measurements that there is a connection between the aspect ratio of an object and the ratio of the k parameters. He showed that for nonmagnetic objects the ratio k_1/k_2 decreases as the aspect ratio is increased. An opposite dependence was observed for magnetic ($\mu_r > 1$) objects [10].

To examine the dependence of the EM response on target shape using MAS, I first model a conductive, nonmagnetic spheroid. I vary the spheroid aspect ratio from one (spherical) to eight (prolate spheroid), while maintaining a constant volume, conductivity and permeability. I then compute the impulse response for axial and transverse excitations, as defined in figure 4.1. Figure 4.2 shows the MAS forward modellings of axial and transverse uniform

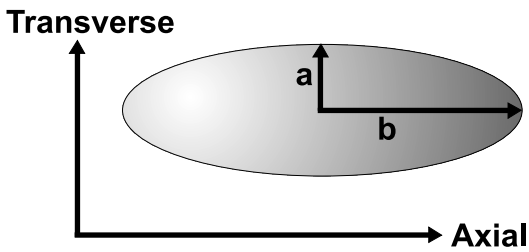
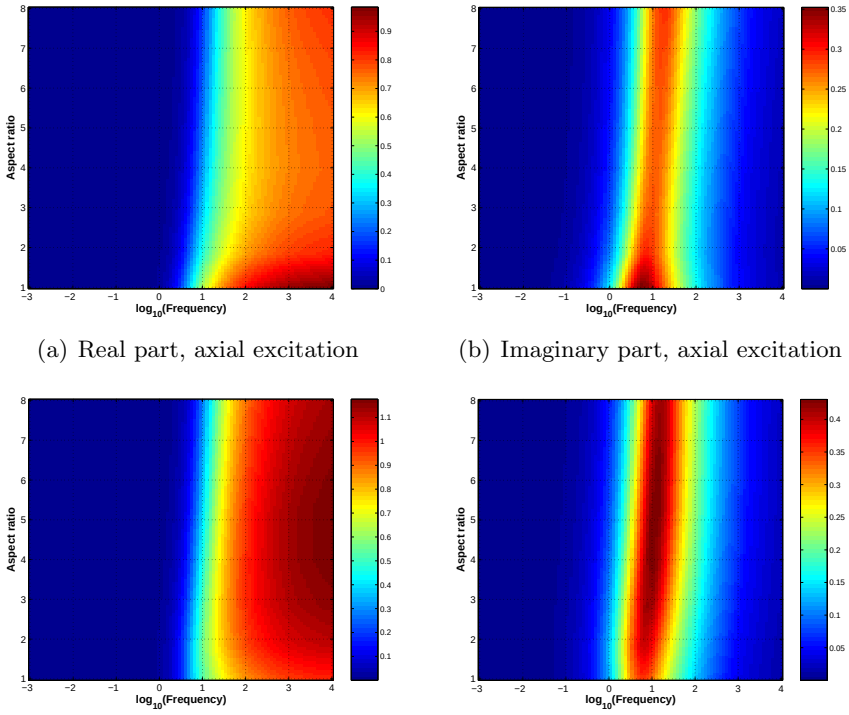


Figure 4.1: Definition of axial and transverse excitation directions used in this chapter. Semi-major (a) and semi-minor (b) axes are also labelled.

field excitations as the aspect ratio is varied. As the spheroid becomes increasingly prolate, the spectra for both transverse and axial excitations are shifted higher in frequency.



(a) Real part, axial excitation

(b) Imaginary part, axial excitation

(c) Real part, transverse excitation

(d) Imaginary part, transverse excitation

Figure 4.2: Real and imaginary parts of the impulse response for a nonmagnetic spheroid as a function of frequency and aspect ratio.

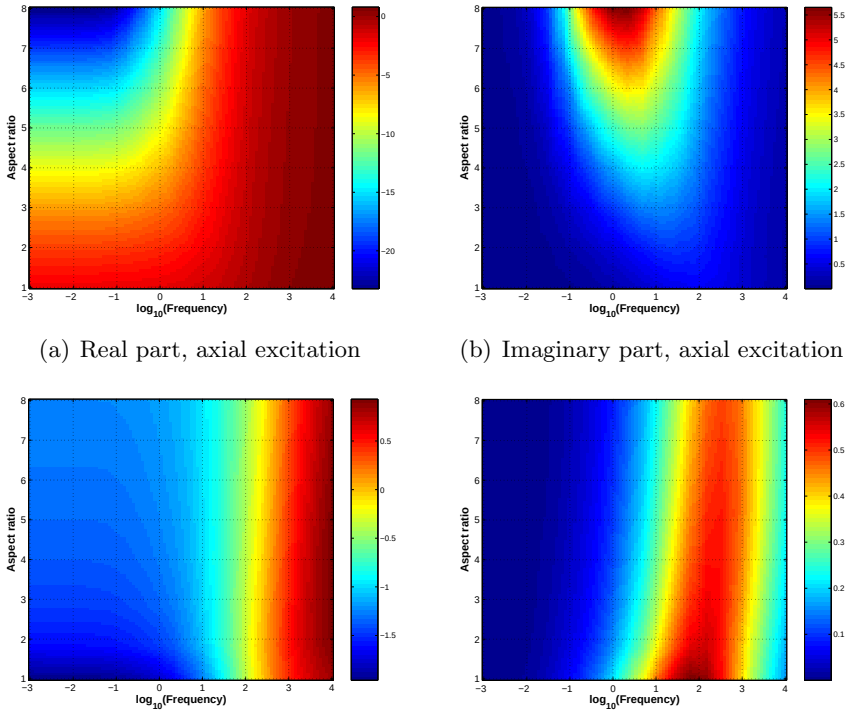
Figure 4.3 shows the same forward modellings for a magnetic spheroid ($\mu_r = 150$). A different dependence is observed for a magnetic spheroid: as the aspect ratio is increased, the peak of the imaginary component of the axial excitation moves lower in frequency, while the peak for the transverse excitation moves higher in frequency.

4.1 Inferring aspect ratio from frequency-domain data

The peak of the imaginary component of the impulse response has important significance in the frequency domain. In the time domain, the decay of the secondary field can be represented as a superposition of exponentials. The dominant decay $\exp(-\omega_p t)$ falls off with time constant ω_p, corresponding to the peak frequency of the imaginary part of the impulse response.

For high-frequency EM induction systems (ie ground penetrating radar), multiple peak frequencies can be estimated from backscattered data. These frequencies correspond to integer multiples of a fundamental resonant frequency and have been successfully used to estimate target length for UXO discrimination [20].

The peak frequencies of axial and transverse excitations can be related to target shape by considering the behaviour of induced currents on the one-dimensional domain [0,L].



(a) Real part, axial excitation

(b) Imaginary part, axial excitation

(c) Real part, transverse excitation

(d) Imaginary part, transverse excitation

Figure 4.3: Real and imaginary parts of the impulse response for a magnetic spheroid as a function of frequency and aspect ratio.

Under the quasistatic assumption, the current density J obeys the diffusion equation

$$\frac{\partial^2 J}{\partial x^2} = k^2 J \tag{4.2}$$

with complex diffusivity

$$k = \sqrt{i\omega\sigma\mu} = \sqrt{\frac{\omega\sigma\mu}{2}}(1+i) \tag{4.3}$$

The solution is then a superposition of decaying harmonic functions

$$J \propto \exp(-ikx) = \exp(-i\gamma x)\exp(-\gamma x) \tag{4.4}$$

where $\gamma = \sqrt{\omega\mu\sigma/2}$ is the reciprocal of the skin depth. The diffusivity k takes on resonant values at the peak frequencies ω_a and ω_t for axial and transverse excitations. By analogy with the wavenumber in the wave equation regime, the diffusivity can be related to wavelength according to

$$k = \frac{2\pi}{\lambda}. \tag{4.5}$$

O'Neill proposed that resonance occurs when the dimension of the target L equals an integer number of half wavelengths [13], so that

$$\Im(k_n) = \frac{n\pi}{L}. \tag{4.6}$$

How does resonance arise in a diffusive regime? If the current density is confined to the surface of a target, then the solution for harmonic functions on the surface is equivalent to a solution on a one-dimensional domain (figure 4.4). In the absence of current sources,

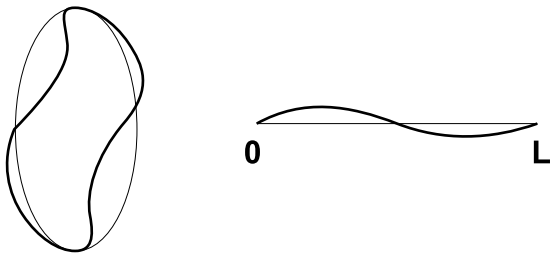


Figure 4.4: Resonant modes circulating on the surface of a spheroid of circumference L are equivalent to resonant modes confined to the one-dimensional domain [0,L] with periodic boundary conditions.

the current density must be continuous and so in one dimension we can impose periodic boundary conditions

$$J(0) = J(L). \tag{4.7}$$

The above boundary conditions lead to eigenvalues of the diffusivity required by equation 4.6. In the high frequency wave equation regime an identical resonance condition arises for the wavenumber k. This is because the spatial components of the diffusion and wave equations have the same form (equation 4.2).

Assuming that equation 4.6 holds, the semi-major and minor axis lengths (a and b) can be related to the peak frequencies of axial and transverse excitations according to

$$\begin{aligned} a &= \frac{2\pi}{\gamma_a} = \frac{2\pi}{\sqrt{\omega_a \mu\sigma/2}} \\ b &= \frac{2\pi}{\gamma_t} = \frac{2\pi}{\sqrt{\omega_t \mu\sigma/2}}. \end{aligned} \tag{4.8}$$

In general, the conductivity and permeability of a target are not known. However, taking the ratio of the above expressions yields an expression for the aspect ratio in terms of the resonant frequencies

$$\frac{a}{b} = \sqrt{\frac{f_t}{f_a}}. \tag{4.9}$$

The aspect ratio a/b is the ratio of the target length along axial and transverse directions. Figure 4.5 shows the proposed diagnostic computed from MAS forward modellings. The peak frequency for each excitation was estimated by fitting a spline to the imaginary part of the impulse response. While there is certainly a relationship between the "O'Neill diagnostic" and the aspect ratio, the expected direct relationship only appears to hold for aspect ratios up to 3. This result is inconsistent with O'Neill's result (also using MAS), which accurately predicted the shape of spheroids with aspect ratio $a/b = 6$. This inconsistency may be due to an error in the MAS forward modelling, though agreement with analytic modelling and the results obtained by Pasion (shown later in this section) inspires faith in the MAS.

Alternatively, the errors observed for high aspect ratios may be associated with difficulties in accurately estimating the peak frequencies. As the magnetic spheroid becomes

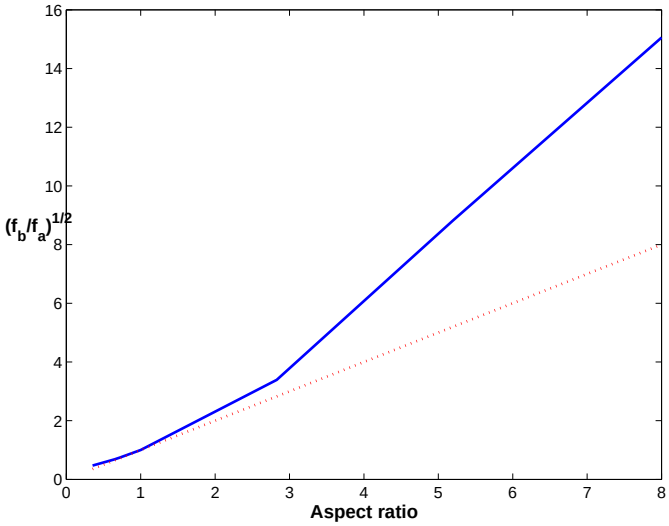


Figure 4.5: O'Neill diagnostic as a function of aspect ratio for a magnetic spheroid (solid line). Dashed line is the expected direct relationship between abscissa and ordinate proposed by O'Neill

increasingly prolate, the axial quadrature response broadens. Consequently, there is greater uncertainty in the peak frequency and so errors in the estimate of the aspect ratio grow. Although a direct relationship between the O'Neill diagnostic and aspect ratio may not hold, the diagnostic does provide a means of discriminating between prolate and oblate magnetic spheroids. Whether an object has an inferred aspect ratio of 6 or 16, it is clearly "rod-like." From a discrimination standpoint, we need only remain mindful that this particular diagnostic may be nonlinearly related to the actual target aspect ratio.

The modelling used to generate figure 4.5 is for a relatively large spheroid (equivalent to a sphere of radius 0.5 m). This raises a concern as to whether resonant frequencies can be observed for actual UXOs for the range of frequencies transmitted by frequency-domain sensors. Figure 4.6 shows frequency domain spectra modelled for spheroids with aspect ratios equivalent to the ordnance listed in table 4.1.

Ordnance Item	**Diameter (cm)**	**Length (cm)**	**a/b**	$\mathbf{f_t/f_a}$
Projectile 14.5 mm	1.45	6.48	4.47	n/a
Montana 76 mm	7.6	27.2	3.58	3.31
Montana 81 mm	8.1	30.375	3.75	4.58
Bomb 1000 lb Mk 83	36.6	191.5	5.23	n/a

Table 4.1: Dimensions of ordnance items used to compute MAS forward modellings. Aspect ratio a/b and O'Neill diagnostic f_t/f_a are also given. n/a indicates that the peak frequency was not observed for that item.

To generate this figure, I assumed a relative permeability of $\mu_r = 150$ and a conductivity of 1×10^6 S/m. The response was computed at 17 frequencies transmitted by a standard GEM-3 instrument (ranging between 30 and 50 kHz). The true and estimated aspect ratios are given in table 4.1. Unfortunately, both axial and transverse peak frequencies are not observed for the small (14.5 mm) and large (1000 lb) ordnance items. This observation has motivated research which extends the GEM-3 frequency range up to 390 kHz [13]. However,

detection of peak frequencies may still be problematic for newer broadband instruments, particularly for very small ordnance items.

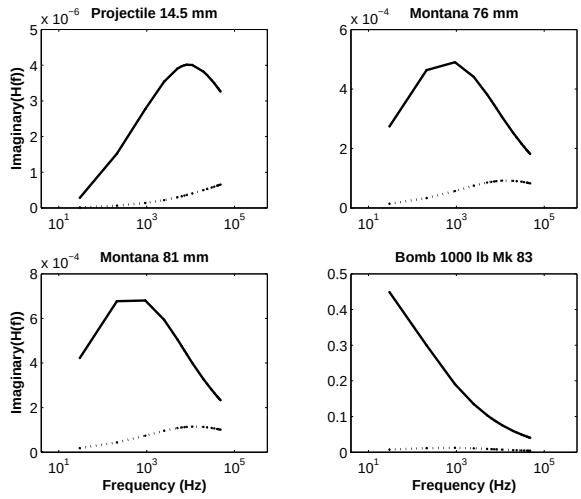


Figure 4.6: Axial (solid line) and transverse (dashed line) quadrature responses of ordnance items modelled with MAS for the GEM-3 range of frequencies.

Another method of inferring aspect ratio from frequency-domain data is suggested by Sun et al [19]. They demonstrate that the ratio of axial and transverse in-phase responses at high frequencies is a diagnostic of target shape. This result is reproduced in figure 4.7 for both magnetic and nonmagnetic spheroids. Both the trend and magnitude of this result are in agreement with that given by Sun et al. However, their result, obtained for perfectly conducting spheroids using the thin skin approximation (TSA), is a monotonically increasing function of aspect ratio. In contrast, my result falls off slightly for the highest aspect ratio. This discrepancy may be due to numerical errors in the MAS for large aspect ratios or possibly due to the finite conductivity of spheroids used in this modelling. Regardless, the agreement of MAS and TSA results for modest elongations ($a/b < 4$) instills some confidence in the validity of the MAS forward modelling.

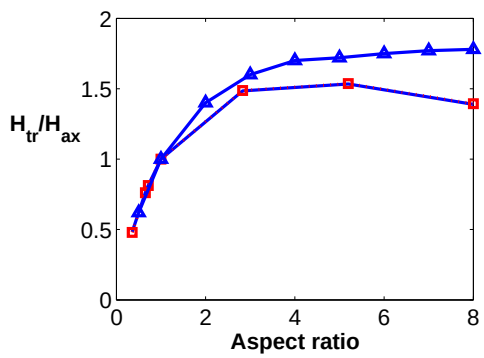


Figure 4.7: Ratio of transverse and axial in-phase response at 1 MHz. Solid line with squares: magnetic and nonmagnetic spheroids, computed with MAS. Solid line with triangles: Result obtained by Sun et al. for perfectly conducting spheroids [19].

A notable feature of this diagnostic is that it appears to be independent of metal type

(magnetic vs. nonmagnetic). This is unique amongst shape diagnostics considered in this chapter. How does this property arise? The ratio of axial and transverse responses is evaluated at 1 MHz, corresponding to skin depths of 3×10^{-5} m and 4×10^{-4} m for magnetic and nonmagnetic spheroids, respectively. Although no physical justification is offered for this result in [19], it is clear that the relationship involves the circulation of surface currents. At the high frequency limit, the impulse response is entirely in-phase and corresponds to eddy currents circulating on the surface. The relative magnitudes of eddy current loops for axial and transverse excitations are then indicative of target aspect ratio.

Is the high-frequency asymptotic response observable in practice? The forward modellings used to generate figure 4.7 are for a relatively large spheroid. In this case, the in-phase response approaches its asymptotic value for the range of frequencies considered. However, in practice most UXOs are much smaller than the spheroid used for this modelling and so the frequency spectrum will be shifted to higher frequencies. This implies that for realistic targets and existing sensors used in production surveys, the asymptotic response will typically not be observable in the frequency domain (figure 4.8).

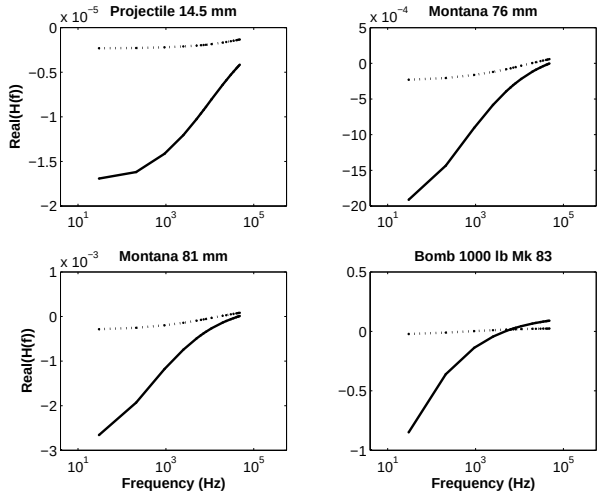
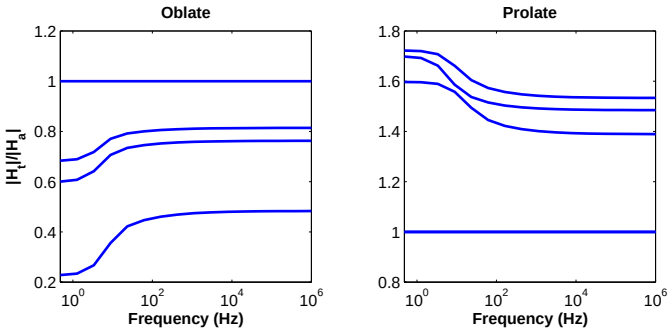


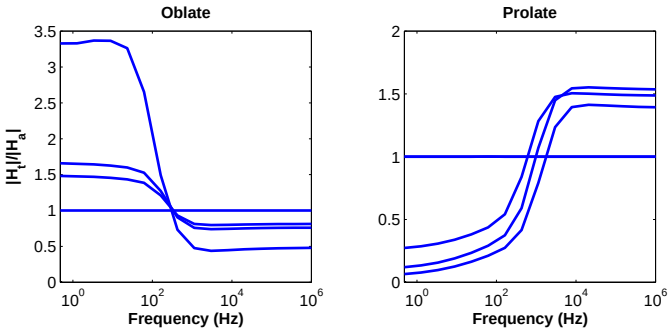
Figure 4.8: Axial (solid line) and transverse (dashed line) in-phase responses of ordnance items modelled with MAS for the GEM-3 range of frequencies.

Finally, an alternative frequency-domain shape diagnostic proposed by Sun et al. is the ratio of the magnitude of transverse and axial responses over the measured frequency-domain spectrum [19]. By considering the entire frequency-domain spectrum this feature circumvents the requirement to measure (or approximate) the peak frequencies or asymptotic in-phase response. Figure 4.9 shows this feature for magnetic and nonmagnetic spheroids. The result obtained here using MAS is in agreement with that presented in [19] using the thin skin approximation method. As shown in figure 4.9, this ratio equals one for a sphere. This is because the axial and transverse responses are identical at all frequencies. O'Neill claims that this feature is more robust than other methods since it uses the entire frequency spectrum to decide whether a target is prolate or oblate. However, an implicit step in this analysis is the resolution of axial and transverse responses from the observed data. This is necessary for all shape diagnostics and so this feature has no particular advantage over other methods presented in this chapter.

Figure 4.9 also suggests a diagnostic for the magnetic properties of a target: if the ratio of responses passes through unity, then the item is magnetic. A more direct indicator of



(a) Nonmagnetic spheroid



(b) Magnetic spheroid

Figure 4.9: Ratio of transverse to axial frequency-domain response for magnetic and non-magnetic spheroids.

magnetic properties is the low-frequency in-phase response. A nonzero in-phase response at low frequencies (either axial or transverse) indicates that the target is magnetic.

4.2 Inferring aspect ratio from time-domain data

To obtain the corresponding time-domain impulse response from frequency-domain MAS data, I use the digital filters of Anderson [21]. The sine transform of the imaginary part of the impulse response $H(\omega)$ produces the time derivative of the secondary magnetic field for a step-off primary field.

$$\frac{\partial h(t)}{\partial t} = \frac{2}{\pi} \int_0^\infty \Im(H(\omega)) \sin(\omega t) d\omega. \tag{4.10}$$

Similarly, the secondary magnetic field is obtained by a cosine transform of the imaginary part of the impulse response

$$h(t) = -\frac{2}{\pi} \int_0^\infty \frac{\Im(H(\omega))}{\omega} \cos(\omega t) d\omega. \tag{4.11}$$

Good agreement is observed between the MAS forward modelling for a sphere and an analytic solution derived by Wait ([16]) in the time domain (figure 4.10). The analytic

solution requires an infinite sum involving the roots of a transcendental equation. The analytic solution shown in figure 4.10 uses the first 5000 roots to compute the time-domain impulse response. Figure 4.11 and 4.12 show time decays obtained by transforming the

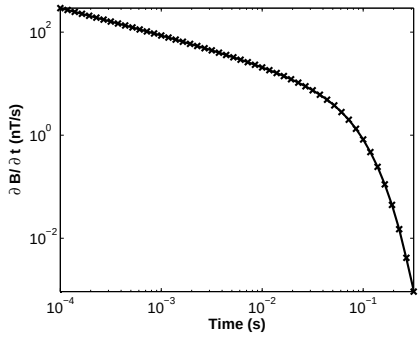


Figure 4.10: MAS forward modelling in the time domain for a conductive sphere (solid line) and analytic solution (crosses)

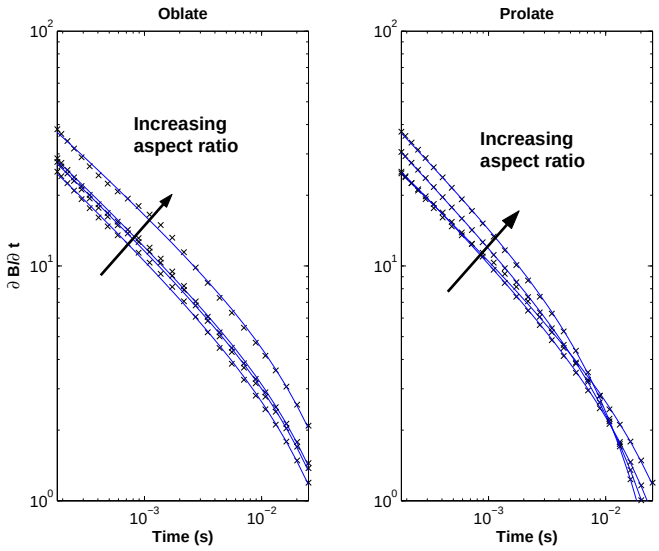
impulse responses from figures 4.2 and 4.3, respectively. The synthetic data were generated for 26 EM63 time channels ranging between 0.18 and 25.14 ms. Numerical errors were produced by the frequency to time domain conversion when the secondary field decayed by more than approximately 4 orders of magnitude for the range of times considered. A large spheroid with volume equivalent to a sphere of radius 0.5 m was therefore used in these simulations. The decays shown in figures 4.11 and 4.12 are characteristic of the early time (linear) decay, with the late time exponential decay only evident for a few cases.

Given the time decays for a spheroid of variable aspect ratio, we can examine the dependence of approximate forward model parameters (k, α, β, γ) upon shape. For each decay, I minimize the difference between the logarithms of the observed and predicted data. Taking the logarithm makes all data the same order of magnitude, so that early time channels do not dominate the misfit. The goal of these inversions is to map from physical properties to model parameters and so no noise is added to the synthetic data. When computing linearized errors in the model parameters I therefore estimate the data uncertainty as the standard deviation of the residuals.

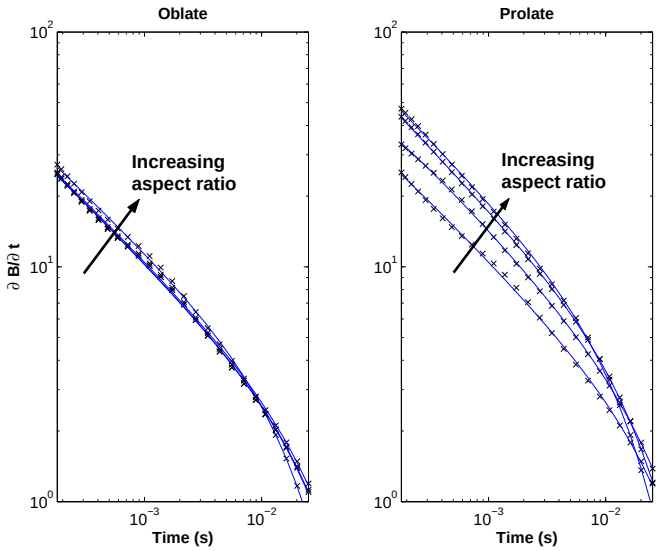
Figures 4.11 and 4.12 show the fits obtained using the Pasion-Oldenburg parameterization. The dependence of model parameters upon aspect ratio is shown in figure 4.13. The magnitude of the time-domain impulse response, governed by the k parameters in the Pasion-Oldenburg model, is highly dependent upon target shape. Figures 4.13 and 4.14 show that the ratio k_1/k_2 provides a robust diagnostic of target shape.

The parameter α controls the asymptotic early time response. For both magnetic and nonmagnetic spheroids considered here, the estimates of this parameter are several orders of magnitude smaller than the earliest time channel. This indicates that the early time response is linear with slope governed primarily by the parameter β. As observed by Pasion, the ratio of β_1/β_2 is a shape diagnostic for magnetic spheroids. However, for nonmagnetic spheroids there is little differentiation between the inferred rates of early time decay for axial and transverse excitations. Similarly, the exponential portion of the decay curve (represented by γ) cannot be reliably used to infer shape for either magnetic or nonmagnetic spheroids.

Figure 4.14 shows the dependence of the ratio k_1/k_2 on target aspect ratio. These results are consistent with laboratory measurements made by Pasion. No synthetic noise



(a) Axial excitation



(b) Transverse excitation

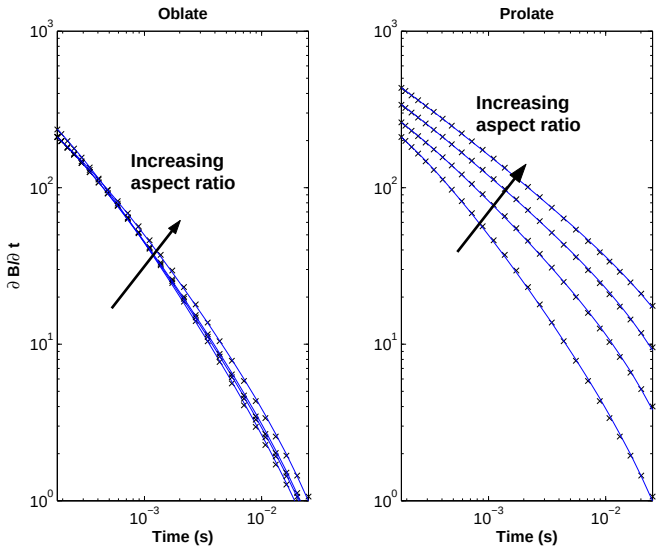
Figure 4.11: MAS forward modellings for a nonmagnetic spheroid in the time domain (crosses). Solid lines are the fits obtained using the Pasion-Oldenburg parameterization.

was added to the time decay data, and so the diagnostic is a smoother function of aspect ratio in these examples than was previously observed for laboratory measurements.

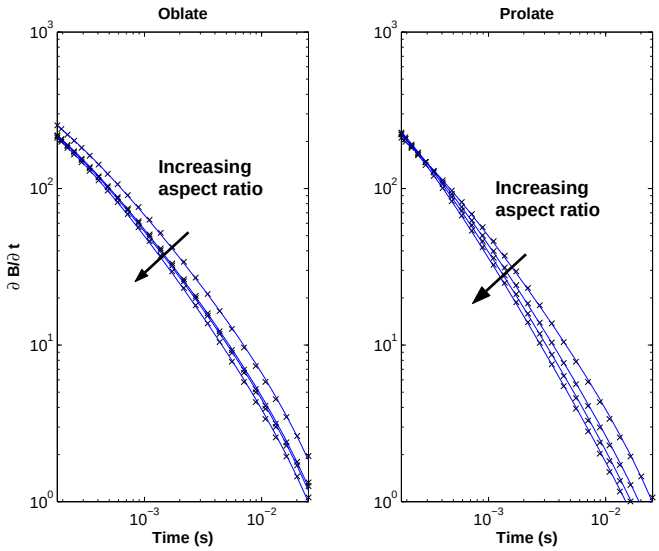
Also shown is the same diagnostic from a power law model fit to four early times (ranging between 0.216 and 1.266 ms) used by the EM61 MKII instrument. The parameterization gives a linear decay in log-log space

$$L_i(t) = k_i t^{-\beta_i}. \tag{4.12}$$

A similar dependence is observed for the power law diagnostic, though the result is less smooth since fewer time channels were used to fit the decays. These results confirm that



(a) Axial excitation

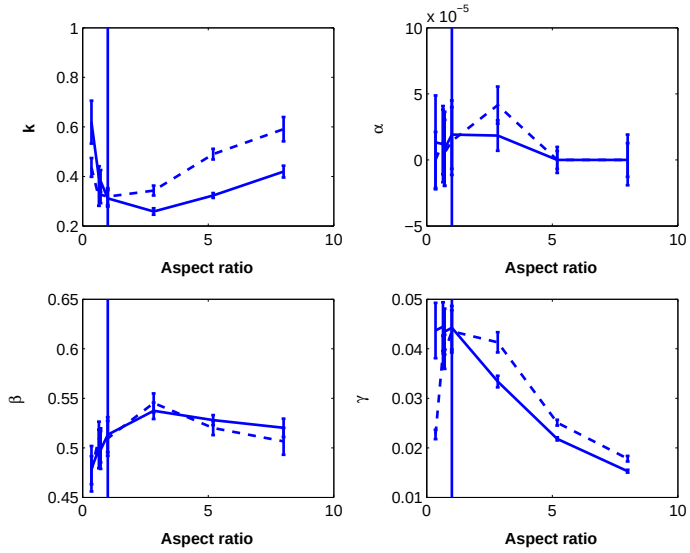


(b) Transverse excitation

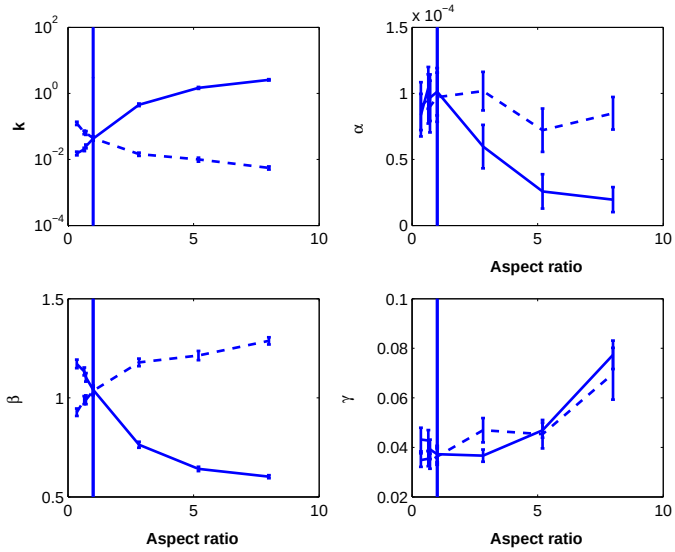
Figure 4.12: MAS forward modellings for a magnetic spheroid in the time domain (crosses). Solid lines are the fits obtained using the Pasion-Oldenburg parameterization.

the diagnostic $k1/k2$ provides information about the aspect ratio of a target.

Figure 4.15 shows the ratio β_1/β_2 as a function of aspect ratio for a magnetic spheroid. As was previously shown by Pasion, fitting the decay of the secondary B-field directly provides a similar dependence of the diagnostics k_1/k_2 and β_1/β_2 upon aspect ratio.



(a) Nonmagnetic spheroid

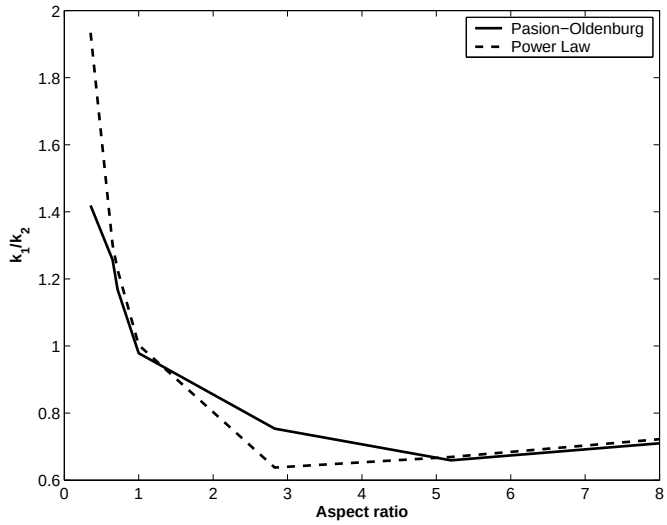


(b) Magnetic spheroid

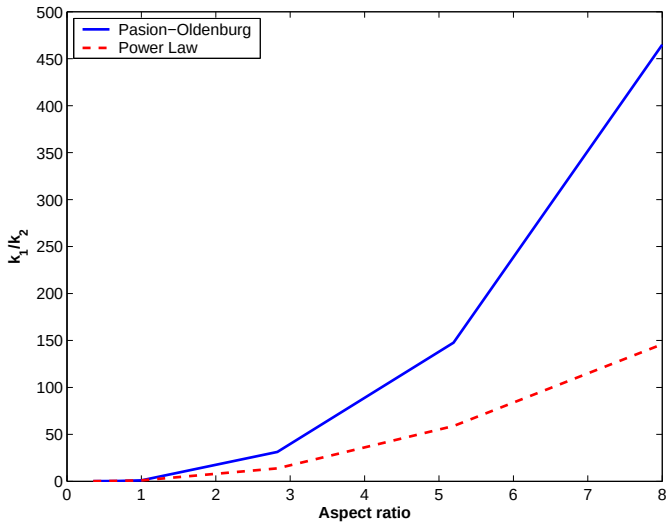
Figure 4.13: Dependence of Pasion-Oldenburg model parameters upon aspect ratio. Solid line is axial excitation and dashed line is transverse excitation. Vertical dashed line indicates an aspect ratio of one.

4.3 Validity of the dipole approximation

Many time-domain (Pasion-Oldenburg, power law) and frequency-domain parametric models assume that the spatial response of an axisymmetric body can be represented by two orthogonal dipoles. These models are motivated by analogy with the analytic solution for a sphere in the time domain, which can be represented as the product of a magnetostatic dipole (the spatial response) and a time or frequency-domain impulse response [10]. Al-



(a) Nonmagnetic spheroid



(b) Magnetic spheroid

Figure 4.14: Dependence of the diagnostic k_1/k_2 upon aspect ratio for magnetic and nonmagnetic spheroids. Solid line is the diagnostic estimated using a Pasion-Oldenburg parameterization and dashed line is the diagnostic estimated using a power law parameterization.

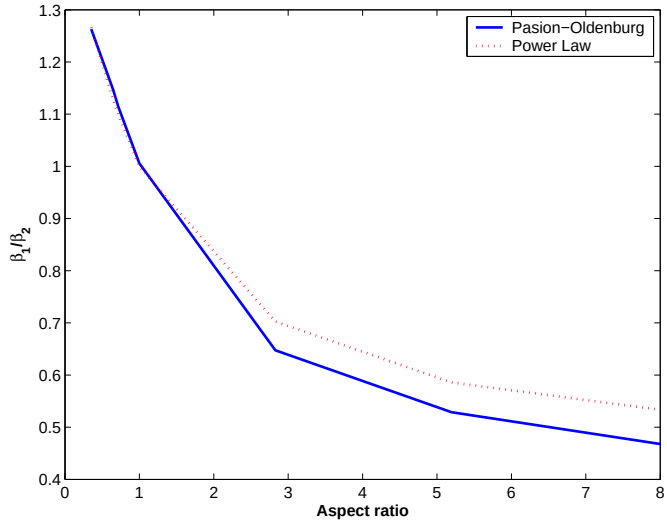
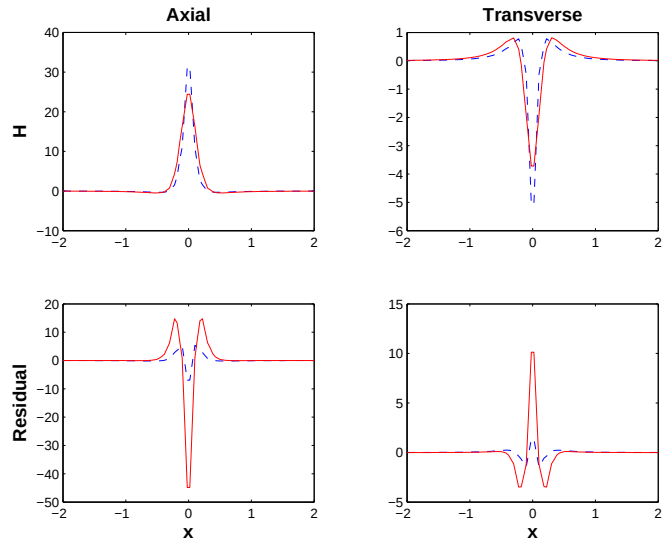


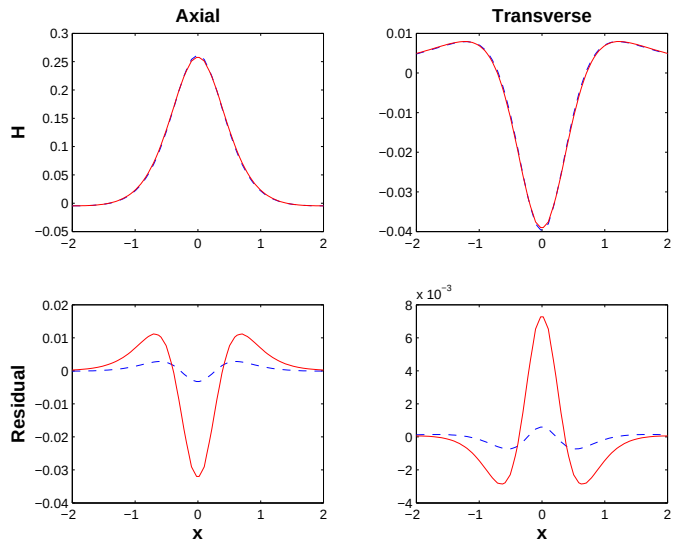
Figure 4.15: The ratio β_1/β_2 as a function of aspect ratio for a magnetic spheroid.

though parametric models can generally reproduce observed data, discrepancies have been observed in the spatial response, especially for laboratory measurements. For example, Das et al. attributed the misfit between observed laboratory data and predicted data from a two-dipole model to modelling error. They speculated that the misfit was due to the presence of higher order poles in the observed data which were not accounted for in the model [22].

The accuracy of the dipole approximation of the spatial response can be investigated using MAS. This is because MAS does not represent the secondary field as a moment expansion but rather computes the secondary magnetic field as a superposition of fields produced by magnetic charges. Figure 4.16 shows the axial and transverse spatial response for a magnetic spheroid with semi major axis $a = 0.2$ m and semi-minor axis $b = 0.1$ m. Best-fitting dipoles are also shown. To generate these fits it was assumed that the spatial response for an axial excitation is represented by a vertical dipole located at the center of the target. In this case, only the magnitude of the dipole moment needs to be estimated to fit the synthetic MAS data (a linear problem). Similarly, the transverse excitation is represented by a horizontal dipole. The residual between synthetic MAS data and data predicted for a dipole shows a systematic variation which indicates the presence of higher order moments. The next nonzero moment for an axisymmetric body with uniform magnetization is the octupole moment [23]. The octupole response for this spheroid can be computed analytically given the estimated dipole moment. The lower panels of figure 4.16 show the residual between MAS and dipole models modellings as well as the predicted octupole response. Although the residual has the character of an octupole response, the fit of octupole moment is poor. A possible explanation for this is the trade-off between octupole and dipole moments described by Billings [24]. The best-fitting dipole is not necessarily *the* induced dipole and overfitting the dipole moment degrades the fit to the octupole. Regardless, the implication of figure 4.16 is that the dipole approximation produces a modelling error which is caused by neglecting higher order moments. This error falls off with distance between receiver and scatterer so that at a distance of 1 m the modelling error of the dipole moment is negligible (figure 4.17).



(a) Depth=0.25 m



(b) Depth=1 m

Figure 4.16: Top row: MAS forward modellings (dashed line) for axial and transverse excitations. Best fit dipoles are shown as solid lines. Bottom row: Residual of fit between MAS and dipole models (dashed lines) The octupole moment corresponding to the magnetization computed for the best fit dipole is shown as a solid line. Observations are at an elevation of 0 m directly above the center of the spheroid.

4.4 Discussion and conclusions

In this chapter, I have examined the connection between parametric model parameters and target shape using MAS as a benchmark forward model. I verified the relationships previously established by Pasion for parametric models in the time-domain. I also investigated the frequency-domain shape diagnostics suggested by O'Neill and collaborators.

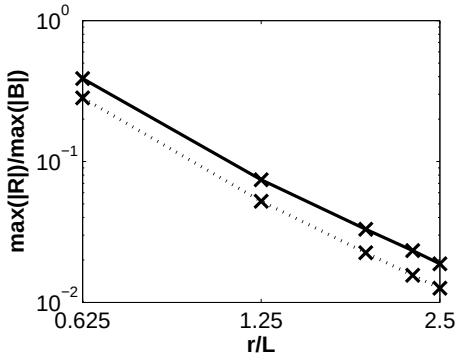


Figure 4.17: Abscissa: Maximum residual of best-fit dipole (max($|R|$)) normalized by maximum amplitude of predicted data (max($|B|$)). Ordinate: Distance from observer to spheroid center (r) normalized by characteristic length $L = 2a$. Solid line is axial excitation and dashed line is transverse excitation.

All shape diagnostics discussed in this chapter involve a ratio of some feature of the axial and transverse response. Implicit in this discussion has been the unique resolution of axial and transverse components of the impulse response. While it is usually possible to resolve the components of the polarization tensor via inversion, difficulties with local minima can arise. For example, the measured response of a vertical prolate spheroid will be dominated by the axial response at observation locations close the target. As the transmitter moves farther from the spheroid, the primary field will increasingly illuminate the target from the side, thereby exciting the transverse response. However, the secondary field will fall off as $1/r^6$ (for a dipole source) so that the SNR at more distant locations may be quite low. This implies that in certain situations the transverse response may be difficult to resolve (ie parameters estimated for the transverse response via inversion may be far from their "true" values)

These considerations will of course have an impact on the performance of the subsequent discrimination. Because we rely on both axial and transverse excitations to make inferences about shape, inaccuracies in shape diagnostics may degrade classifier performance.

It is also important to note that most shape features are strongly nonlinear functions of aspect ratio. In many classifiers, a Euclidean metric is used to measure the "distance" between points in the feature space. However, because of the nonlinear relationship between features and shape, the same distance will have a different physical meaning in different regions of the feature space. This is not necessarily a bad thing: nonlinearity of the features may provide a large separation between items which would appear close together based solely upon their physical properties. Of course, the converse may also be true, in which case statistical classification will likely have a hard time.

With these considerations in mind, I will now discuss statistical methods for partitioning the feature space.

Chapter 5

Statistical classification

5.1 Introduction

As outlined in chapter 1, improved classification algorithms are crucial to achieving cost-effective UXO remediation. A focus of this thesis is therefore a systematic application of existing pattern recognition algorithms to the UXO problem. A full discussion of pattern recognition can be found in many textbooks (e.g. [25]). In this chapter, I give a survey of considerations and methods relevant to UXO discrimination.

Pattern recognition has been defined as

> the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of patterns [26].

Pattern recognition algorithms have found application in a wide variety of fields, including optical character recognition, voice recognition, and DNA sequence analysis. The current state of pattern recognition research within the UXO field is discussed in the next chapter.

Within the field of pattern recognition there is an important dichotomy between *supervised* and *unsupervised* classification. Supervised classification makes classification decisions on a *test* or *prediction* data set for which labels are unknown. The classifier performance is optimized using a *training* data set for which labels are known. In unsupervised classification there is only a test data set; labels are unknown for all observed patterns. Unsupervised classification is often a more difficult problem than supervised classification.

Template matching is a simple solution to supervised learning problems. This approach compares observed data with pre-defined templates; a pattern is classified based upon the template which best matches the observed data. A typical measure of similarity is the correlation between the data and the template. While template matching is an intuitive approach to classification, it is often difficult to generalize templates so that they can match all possible permutations of the data. For example, in biometrics (face recognition) a change in the orientation of the face can be difficult to model within the template framework.

In the statistical classification framework, patterns are represented by vectors within an M-dimensional *feature space*. The basis vectors in this space, the features, are computed from observed data and are chosen so that feature vectors belonging to different classes occupy different regions of the feature space. 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 new feature vector [25]. Statistical classification is the most widely applied framework for pattern recognition and this thesis focuses primarily on statistical methods for discriminating between UXO and scrap.

5.2 Statistical classification

Figure 5.1 summarizes the classification process within the statistical framework. Given test and training data sets, we extract features from the data, select a relevant subset of these features, and optimize the classifier using the available training data. Because the predicted performance of the classifier is dependent upon the feature space, the learning stage can involve further experimentation with feature extraction and selection before adequate performance is achieved.

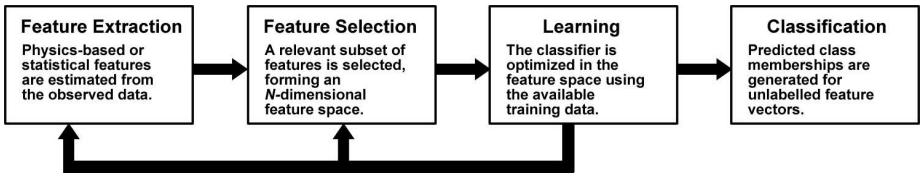


Figure 5.1: A framework for statistical pattern recognition.

Within statistical classification there are two approaches to generating a 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 [27].

5.3 Feature selection and the curse of dimensionality

Before pattern recognition algorithms can be applied to the problem of discriminating between UXO and clutter items, a feature space must be defined. As discussed in the previous chapter, features used in UXO classification are commonly parameters in physics-based models. Because multisensor platforms are used increasingly in the UXO field, available features may be extracted from different data types. For example, the Multi-Sensor Towed Array Detection System (MTADS) combines electromagnetic and magnetic sensors on a single platform [28].

A consequence of multisensor platforms has been joint and cooperative inversion of data sets. Pasion has shown that the additional constraints afforded by combining magnetic and electromagnetic data can provide improved estimates of target parameters [10].

These results suggest that statistical classification will perform better when many features from multiple sensors are used to define a feature space. However, if the true underlying class distributions are unknown and parameters are estimated from a limited training data set, increasing the dimensionality of the feature space will eventually degrade classifier performance. Unbiased estimation of class distributions requires that the number of feature vectors in the training data set must grow exponentially as the dimensionality of the feature space increases. This effect is often referred to as the "curse of dimensionality".

The lesson to be drawn from the curse of dimensionality is that a small set of relevant features should be chosen as basis vectors for a feature space. If features have physical significance, they can be chosen based upon a physical understanding of which features provide the best separation between classes. Alternatively, feature selection or extraction algorithms can be used to define the feature space. Feature selection algorithms choose a subset of existing features to define the feature space, whereas feature extraction algorithms

use linear or nonlinear combinations of features to produce new feature vectors. The process of inverting geophysical data to obtain parameter estimates can be regarded as a (usually nonlinear) feature extraction algorithm.

The simplest transformation of raw features is rescaling to ensure that features have the same range. This is especially important with classifiers which require an estimate of class covariances. If the variability of features differs by orders of magnitude then the covariance matrix may be ill-conditioned and classifier predictions may be numerically unstable. An easy way to prevent this is to standardize the features so that they are zero mean and unit variance.

Principal components analysis (PCA) can be used to project the feature data onto the eigenvectors of the feature data covariance matrix. PCA outputs linear combinations of the input features which are standardized (zero mean and unit variance) and uncorrelated. However, PCA assumes that the feature data follow a multivariate Gaussian distribution. This assumption will certainly be violated if the classes in the training data occupy discrete regions of the feature space. Furthermore, the principal components corresponding to directions of maximum variance in the feature data do not necessarily correspond to directions which provide maximum class separation. In the next section, I describe a more suitable feature extraction algorithm for supervised learning.

5.3.1 Canonical analysis

Canonical analysis is a linear feature extraction algorithm which provides low-dimensional projections of high-dimensional feature spaces. The algorithm differs from principal component analysis in that it projects the data onto non-orthogonal directions which are designed to maximize the separation between classes.

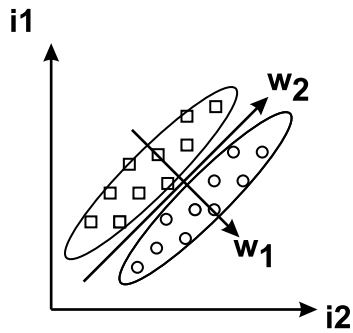


Figure 5.2: Motivation for canonical analysis. Projection of feature vectors (open circles and squares) onto direction $\mathbf{w}_1$ separates the two classes, whereas projection onto direction $\mathbf{w}_2$ does not. Projection onto $\mathbf{w}_1$ provides a maximum separation between the classes relative to the variance of the classes in that direction.

The problem is summarized in figure 5.2 for a simple two dimensional case. We are given a set of N labelled feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in an M-dimensional feature space. We wish to find the optimal set of directions which maximize the variance between K classes, *relative to the variance within classes*, i.e. we want to maximize the number of standard deviations between class means. The $M \times M$ total scatter matrix of the feature vectors is defined as

$$\mathbf{S}_T = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \tag{5.1}$$

There are N feature vectors and the estimate of the mean is

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i. \tag{5.2}$$

The scatter matrix provides an unbiased estimate of the covariance Σ of the feature data

$$\hat{\Sigma} = \frac{1}{N-1}\mathbf{S}_T. \tag{5.3}$$

The estimate is divided by $N-1$ because one degree of freedom is lost in estimating the sample mean with equation 5.2.

Consider a given feature vector belonging to class ω_j with class mean $\boldsymbol{\mu}_j$. For any class, we can decompose the deviation from the mean into

$$(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu}_j) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}). \tag{5.4}$$

If K is the number of classes, then substituting the above expression into equation 5.1 yields

$$\begin{aligned}\mathbf{S}_T = &\sum_{j=1}^{K}\Big[\sum_{\mathbf{x}\in\omega_j}(\mathbf{x}-\boldsymbol{\mu}_j)(\mathbf{x}-\boldsymbol{\mu}_j)^T + \\ &\sum_{\mathbf{x}\in\omega_j}(\mathbf{x}-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_j-\boldsymbol{\mu})^T + \sum_{\mathbf{x}\in\omega_j}(\boldsymbol{\mu}_j-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu}_j)^T\Big] + \\ &\sum_{j=1}^{K}N_{\omega_j}(\boldsymbol{\mu}_j-\boldsymbol{\mu})(\boldsymbol{\mu}_j-\boldsymbol{\mu})^T.\end{aligned} \tag{5.5}$$

Here $\mathbf{x}\in\omega_j$ denotes summation over the N_{ω_j} points belonging to class ω_j. Taking the expectation of the above expression yields

$$\begin{aligned}E\left[\mathbf{S}_T\right] =& E\left[\sum_{j=1}^{K}\sum_{\mathbf{x}\in\omega_j}(\mathbf{x}-\boldsymbol{\mu}_j)(\mathbf{x}-\boldsymbol{\mu}_j)^T\right] + E\left[\sum_{j=1}^{K}N_{\omega_j}(\boldsymbol{\mu}_j-\boldsymbol{\mu})(\boldsymbol{\mu}_j-\boldsymbol{\mu})^T\right] \\ =& E\left[\mathbf{S}_W\right] + E\left[\mathbf{S}_B\right].\end{aligned} \tag{5.6}$$

The first term is the pooled *within-class* scatter $\mathbf{S}_W$, it is a measure of the total scatter of feature vectors about their class means. The second term, the *between-class* scatter $\mathbf{S}_B$, represents the total scatter of class means about mean $\boldsymbol{\mu}$. Normalizing the terms on the right hand side by their respective degrees of freedom, we obtain unbiased estimates of the within-class and between-class covariances

$$\begin{aligned}\hat{\Sigma}_W &= \frac{1}{N-K}\mathbf{S}_W \\ \hat{\Sigma}_B &= \frac{1}{K-1}\mathbf{S}_B.\end{aligned} \tag{5.7}$$

The estimate of the within-class covariance matrix is then identical to the pooled covariance matrix used in linear discriminant analysis.

Now that we have estimates of the within-class and between-class covariances, we can try to find an optimal projection. After projection of means and feature vectors onto the direction $\mathbf{w}$, the estimate of the within-class variance is

$$\begin{aligned}\hat{\sigma}_W^2 =& \frac{1}{N-K}\sum_{j=1}^{K}\sum_{\mathbf{x}\in\omega_j}(\mathbf{w}^T\mathbf{x}-\mathbf{w}^T\boldsymbol{\mu}_j)(\mathbf{w}^T\mathbf{x}-\mathbf{w}^T\boldsymbol{\mu}_j)^T \\ =& \mathbf{w}^T\hat{\Sigma}_W\mathbf{w}\end{aligned} \tag{5.8}$$

Similarly, the estimated between-class variance is

$$\hat{\sigma}_B^2 = \mathbf{w}^T\hat{\Sigma}_B\mathbf{w}. \tag{5.9}$$

To maximize the between-class variance relative to the within-class variance, we maximize

$$\lambda = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2} = \frac{\mathbf{w}^T\hat{\Sigma}_B\mathbf{w}}{\mathbf{w}^T\hat{\Sigma}_W\mathbf{w}}. \tag{5.10}$$

Taking the derivative with respect to $\mathbf{w}$ gives

$$\begin{aligned}\frac{\partial\lambda}{\partial\mathbf{w}} =& \frac{2\hat{\Sigma}_B\mathbf{w}}{\mathbf{w}^T\hat{\Sigma}_W\mathbf{w}} - \left[\frac{\mathbf{w}^T\hat{\Sigma}_B\mathbf{w}}{\left(\mathbf{w}^T\hat{\Sigma}_W\mathbf{w}\right)^2}\right]2\hat{\Sigma}_W\mathbf{w} = 0 \\ \Rightarrow& \hat{\Sigma}_W^{-1}\hat{\Sigma}_B\mathbf{w} = \lambda\mathbf{w}.\end{aligned} \tag{5.11}$$

The solutions for $\mathbf{w}$ are therefore the eigenvectors of the matrix $\mathbf{D} = \hat{\Sigma}_W^{-1}\hat{\Sigma}_B$. The covariance matrices are symmetric and positive definite and so $\hat{\Sigma}_W$ will always be invertible.

Recall that the eigenvectors of a symmetric matrix are orthogonal and that the inverse of a symmetric matrix is itself symmetric. Assuming that neither $\hat{\Sigma}_W$ nor $\hat{\Sigma}_B$ are the identity and that $\hat{\Sigma}_B \neq \hat{\Sigma}_W^{-1}$ we have

$$\begin{aligned}\mathbf{D}^T &= (\hat{\Sigma}_W^{-1}\hat{\Sigma}_B)^T \\ &= (\hat{\Sigma}_B\hat{\Sigma}_W^{-1}) \\ &\neq \mathbf{D}.\end{aligned} \tag{5.12}$$

so that the matrix $\mathbf{D}$ is not symmetric. This implies that the eigenvectors of $\mathbf{D}$ are, in general, not orthogonal. This is a fundamental difference with principal component analysis, which projects the data onto orthogonal eigenvectors of the covariance.

The projection of feature vector $\mathbf{x}$ into the M-dimensional space of canonical variables is

$$\mathbf{x}* = \boldsymbol{\Omega}^T\mathbf{x} \tag{5.13}$$

where the ith column of $\boldsymbol{\Omega}$ is the normalized eigenvector $\hat{\mathbf{w}}_i$. The eigenvectors are normalized so that the projected classes will have unit standard deviation

$$\hat{\mathbf{w}}_\mathbf{i} = \mathbf{w}_i/(\mathbf{w}_i^T\hat{\Sigma}_W\mathbf{w}_i^T)^{1/2}. \tag{5.14}$$

Projection onto eigenvectors ordered by decreasing eigenvalues produces canonical axes with decreasing separation between classes.

The above derivation requires that there are no outliers which will skew the estimated means and covariances. While it is convenient if the observed data follow a well-behaved distribution (eg Gaussian), class distributions which are "reasonably symmetric and not too long-tailed" can provide useful estimates of canonical eigenvectors [29].

As an example of canonical analysis, I consider a synthetic three-dimensional feature space with two normally-distributed classes (figure 5.3). This figure should be interpreted as a matrix of plots, ordered so that the plot at the intersection of jth row and k^{th} column shows the projection of the feature vectors onto basis vectors $\hat{i}_j$ and $\hat{i}_k$. Plots above and below the main diagonal contain redundant information.

The feature data projected onto the first two basis vectors have relatively large variances but provide little separation between the classes. The classes have a small variance when projected onto the third basis vector, but their separation is large relative to the variance in this dimension. The eigenvalues and eigenvectors of **D** are shown in table 5.1. There is

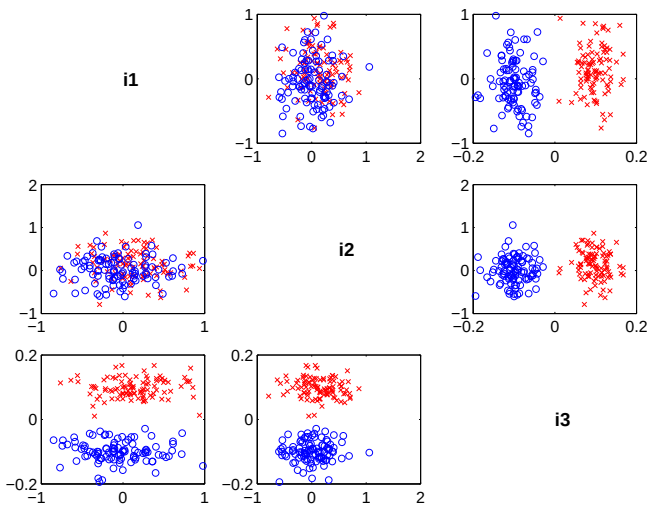


Figure 5.3: Projections of synthetic feature data from two normally-distributed classes (crosses and circles) onto basis vectors $\{\hat{i}_1, \hat{i}_2, \hat{i}_3\}$. The plot at the intersection of jth row and k^{th} column shows the projection of the feature vectors onto basis vectors $\hat{i}_j$ and $\hat{i}_k$.

Index	**1**	**2**	**3**
Eigenvalue	2038	2×10^{-13}	1×10^{-13}
Eigenvector	0.2619	-2.7789	-0.0396
	0.2960	0.0256	-3.1374
	31.9319	2.7606	2.1249

Table 5.1: Eigenvalues and eigenvectors of **D** for the synthetic example shown in figure 5.3.

only one canonical variable whose associated eigenvalue is significantly different from zero. Hence canonical analysis of this simple example identifies a single feature which provides the maximum separation between the classes. The third dimension of the feature space contributes 99.99 percent to the norm of the corresponding vector. Figure 5.4 shows the data projected onto the first two canonical features. As expected, the first canonical variable

$ic1$ provides all of the separation between the classes. Note also that the projected classes have approximately unit within-class variance.

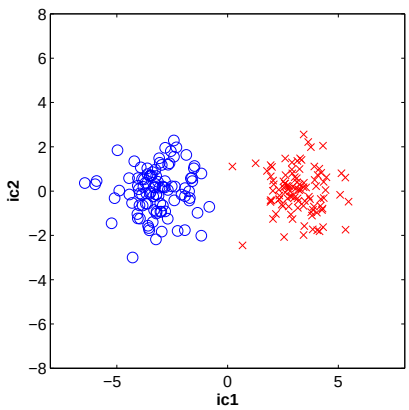


Figure 5.4: Feature data projected onto the first two canonical variables ($ic1$, $ic2$)

5.4 Generative classifiers

Once the feature space is defined by a judicious choice of available features, the goal is to find an optimal partition of the feature space. In this section I discuss *generative* methods: classifiers which seek to model the underlying distributions from which the observed training and test data sets are drawn.

5.4.1 Bayes Decision Theory

The starting point for any generative classifier is Bayes rule

$$P(\omega_i|\mathbf{x}) \propto P(\mathbf{x}|\omega_i)P(\omega_i). \tag{5.15}$$

The *likelihood* $P(\mathbf{x}|\omega_i)$ is the probability of observing the feature vector $\mathbf{x}$ given the class ω_i. The prior probability $P(\omega_i)$ quantifies our expectation of how likely we are to observe class ω_i before (ie prior to) observing any feature vector data. Bayes rule translates the prior probability into a posterior probability $P(\omega_i|\mathbf{x})$. The posterior is the probability that we have observed class ω_i given the observed feature vector. Equality in the above equation can be obtained by normalizing the right hand side by the data prior

$$\begin{aligned} P(\mathbf{x}) &= \sum_{i=1}^{K} P(\mathbf{x}, \omega_i) \\ &= \sum_{i=1}^{K} P(\mathbf{x}|\omega_i)P(\omega_i) \end{aligned} \tag{5.16}$$

where $P(\mathbf{x}, \omega)$ denotes the joint distribution of the feature data and the discrete random variable ω. In classification algorithms the posterior probability is assumed to be

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{k=1}^{K} p(\mathbf{x}|\omega_k)P(\omega_k)} \tag{5.17}$$

where $p(\mathbf{x}|\omega_i)$ is the likelihood function for the class ω_i evaluated at the point $\mathbf{x}$. I use the standard notation that $P(\cdot)$ is a probability and $p(\cdot)$ is a probability density function. If all classes have equal prior probabilities ($P(\omega_k) = 1/K$, $\forall K$), then the likelihood is

$$P(\mathbf{x}|\omega_i) = \frac{p(\mathbf{x}|\omega_i)}{\sum_{k=1}^{K} p(\mathbf{x}|\omega_k)}. \tag{5.18}$$

Equation 5.17 is not a trivial application of Bayes rule: it assumes that the posterior probability is computed by evaluating a density function $p(\mathbf{x}|\omega_i)$ at a point. Strictly speaking, a probability is an integral of a probability density function and the probability of observing any given value of the feature vector $\mathbf{x}$ is zero.

How can we justify equation 5.17? The assumed form of the posterior probability produces generative classifiers which can predict the class of an observed feature vector. However, the predicted probability output by a classifier is not necessarily representative of the probability we might expect from the estimated likelihood function of a test vector given the training data. This effect is summarized in figure 5.5. The posterior probabilities are approximate step functions which vary only in the region where $p(\mathbf{x}|\omega_1) \approx p(\mathbf{x}|\omega_2)$ (i.e. near the decision boundary). Therefore the probabilities output by a generative classifier are insensitive to the likelihood function away from the decision boundary.

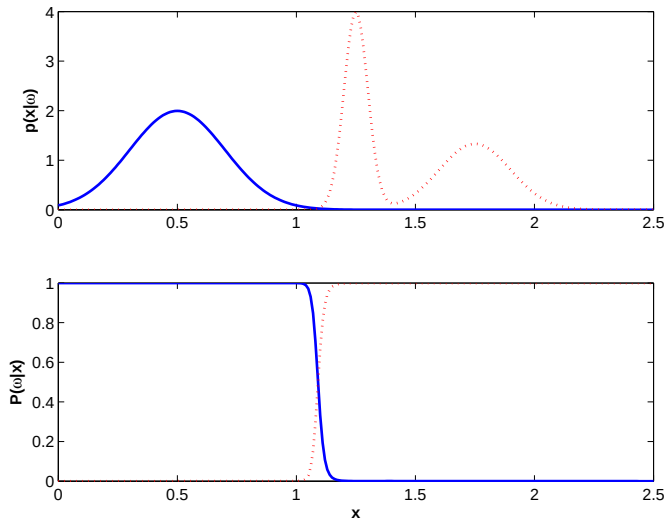


Figure 5.5: Top: Arbitrary likelihood functions $p(\mathbf{x}|\omega)$ for two classes. Bottom: posterior probabilities estimated using Bayes rule (after [25]).

An important implication of figure 5.5 is that the probabilities output by a generative classifier are nearly binary. By assuming equation 5.17, we are designing a classifier which can answer the question: to what class does the observed feature vector belong? This is a reasonable question to pose in many applications. For example, in character recognition we are primarily concerned with assigning each test vector to a class (ie identifying it as a particular character). The value of the posterior probability is of secondary interest. Ripley terms this formulation a *diagnostic* classifier: we are making a definite diagnosis as to the class of a feature vector [30].

Generative classifiers which use equation 5.17 are also insensitive to outliers to the training data. In figure 5.5, a feature vector at $\mathbf{x} = 2.5$ is unlikely to be drawn from either

likelihood function. However, the posterior probability of an outlying feature vector at this point is nearly one. This is not egregious in typical problems: all test data can usually be assumed to fall within one of the classes defined in the training data.

Bayes rule provides a mechanism for classifying feature vectors: assign feature vector $\mathbf{x}$ to the class with the largest posterior probability.

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \ \forall \ i \neq j \tag{5.19}$$

The decision boundary for this rule is a contour along which the posterior probabilities are equal. Is the decision rule described by the above equation optimal? As in an inverse problem, we need some kind of misfit, or loss, function which defines the "best" decision rule. An appropriate misfit for classification is

$$\phi_n(\omega_j, \hat{\omega}) = \begin{cases} 0 & \hat{\omega} = \omega_j, \\ 1 & \hat{\omega} \neq \omega_j. \end{cases} \tag{5.20}$$

where $\hat{\omega}$ is the predicted class membership of the feature vector $\mathbf{x}_n$ belonging to the class ω_j. According to this "0-1" loss function we are only penalized for misclassified feature vectors. The total expected loss for a set of N observed feature vectors is

$$\begin{aligned} \Phi &= \sum_{n=1}^{N} E[\phi_n] \\ &= \sum_{j=1}^{K} \sum_{i=1}^{N_{\omega_j}} P(\hat{\omega} \neq \omega_j|\mathbf{x}_i) \\ &= \sum_{j=1}^{K} \sum_{i=1}^{N_{\omega_j}} (1 - P(\hat{\omega} = \omega_j|\mathbf{x}_i)) \end{aligned} \tag{5.21}$$

where the expectation is taken over realizations of the discrete random variable ω. Here N_{ω_j} denotes the number of feature vectors belonging to class ω_j. This equation states that if we assign each feature vector to the most probable a posteriori class, then we expect the total loss to be minimized. Therefore under the assumption of the "0-1" loss, the decision rule is optimal. The minimum of Φ is obtained when the true posterior probabilities are known for all feature vectors, i.e. when the true distributions for all classes are known.

Application of Bayes rule to classification requires knowledge of the prior probabilities and the form of the likelihood function. 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 5.17, the prior probabilities weight the relative importance of classes. If no prior information is available we may choose to assign equal prior probabilities to each class. In this situation, we must rely on the likelihood function to make classification decisions. The prior probability $P(\omega_i)$ can also be estimated from the training data as the proportion of feature vectors belonging to that class:

$$P(\omega_i) = N_i/N \tag{5.22}$$

with N_i the number of points belonging to class ω_i and N the number of points in the training data set.

The likelihood function can take either a *parametric* or *nonparametric* form. The parametric approach assumes a probability distribution (most commonly the normal distribution) and tries to estimate the parameters of this distribution (eg. mean, covariance) from the training data. This methodology is described in section 5.4.2. This approach can work well if the data follow a simple distribution, but unfortunately this is not always the case in classification problems. If the training data cannot be described by a parametric distribution, we may turn to nonparametric methods (section 5.4.3) which try to model the likelihood function as a superposition of kernel functions.

5.4.2 Parametric methods

The most common parametric classifier is discriminant analysis, which assumes a Gaussian form for the likelihood function

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right) \tag{5.23}$$

with $\boldsymbol{\mu}_i$ and Σ_i the mean and covariance of the ith class. Here $|\cdot|$ denotes the determinant. The decision boundary between classes ω_i and ω_j is found by equating posterior probabilities

$$P(\omega_i|\mathbf{x}) = P(\omega_j|\mathbf{x}). \tag{5.24}$$

The discriminant function f_{ij} is defined as

$$f_{ij} = \log(P(\omega_i|\mathbf{x})) - \log(P(\omega_j|\mathbf{x})) \tag{5.25}$$

so that $f_{ij} = 0$ corresponds to the decision boundary between classes ω_i and ω_j. 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 covariances are assumed equal for all classes ($\Sigma_i = \Sigma_p,\ \forall i$), then the discriminant function is

$$f_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\Sigma_p^{-1}\mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_j^T\Sigma_p^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i^T\Sigma_p^{-1}\boldsymbol{\mu}_i) + \frac{1}{2}\log\left(\frac{P(\omega_i)}{P(\omega_j)}\right). \tag{5.26}$$

The above equation defines a linear decision boundary in the feature space, and so discriminant analysis with a pooled covariance for all classes is called linear discriminant analysis. The pooled covariance Σ_p is

$$\Sigma_p = \frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{N_{\omega_j}}(\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T \tag{5.27}$$

with N the number of feature vectors, K the number of classes, and N_{ω_j} the number of training vectors belonging to class ω_j. We normalize by $N-K$ (as opposed to $N-1$ in the usual covariance estimate) because this estimator has $N-K$ degrees of freedom [31].

Choosing a linear or quadratic form for the decision boundary can be guided by statistical tests of the training data class distributions [31]. The complexity of the decision boundary can also be governed with regularized discriminant analysis, which computes the regularized class covariance matrix for the ith class Σ_i^r as

$$\Sigma_i^r = \alpha\Sigma_i + (1-\alpha)\Sigma_p. \tag{5.28}$$

For $\alpha = 0$ we have the pooled covariance for all classes (linear discriminants) whereas for $\alpha = 1$ we have separate covariances (quadratic discriminants) [25]. The value of the regularization parameter α can be estimated by cross-validation (section 5.6).

An example application of discriminant analysis for UXO classification is shown in figure 5.6. This feature space was generated by cooperative inversion of EM-63 data acquired at the Yuma, Arizona test site. A Pasion-Oldenburg parameterization was used to fit the observed data for each target. As discussed in chapter 4, these parameters can be used to make inferences about target size and shape. Unfortunately, no other ground truth is available and so the performance of trained classifiers cannot be evaluated for this site.

To generate a discriminant analysis classifier for this feature space, I compute the mean of each class (ie ordnance type) and the pooled covariance of the training vectors. I then apply equation 5.17 to compute posterior probabilities of membership in each class on a grid of points in the feature space. Plotting the maximum posterior probability of membership, evaluated over all classes, provides an image of the trained classifier in the feature space. Minima of this function then correspond to decision boundaries.

This choice of feature space produces good separation between ordnance types and so the boundaries delineated by both linear and quadratic discriminant analysis provide very low misclassification rates on the training data. The assumption of pooled or separate covariance matrices has a dramatic effect on the resulting decision boundaries. For example, the class covariance for M75 ordnance (black crosses in figure 5.6) has a large variance along its first principal axis, producing a thin region of the feature space for this class. This is very different from the piecewise linear decision boundary provided by linear discriminant analysis.

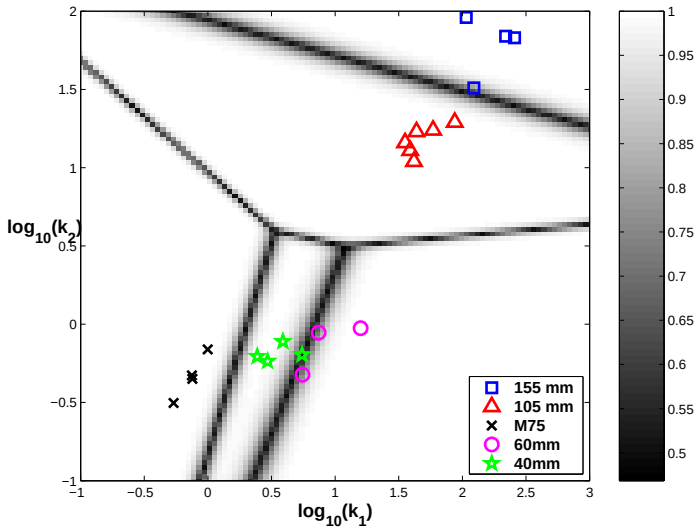
Which form of the decision boundary (linear or quadratic) will provide the best generalization to the test data? For this data set no ground truth was available from other areas of the test site, and so we must evaluate the classifiers based solely upon the training data. A subjective ranking of the classifiers in figure 5.6 might prefer linear discriminant analysis. There does not appear to be sufficient structure in the training data to justify the more complex partition of the feature space offered by quadratic discriminant analysis. This is confirmed by regularized discriminant analysis: leave-one-out cross-validation of the training data reproduced linear discriminant analysis ($\alpha = 0$ in equation 5.28).

5.4.3 Nonparametric methods

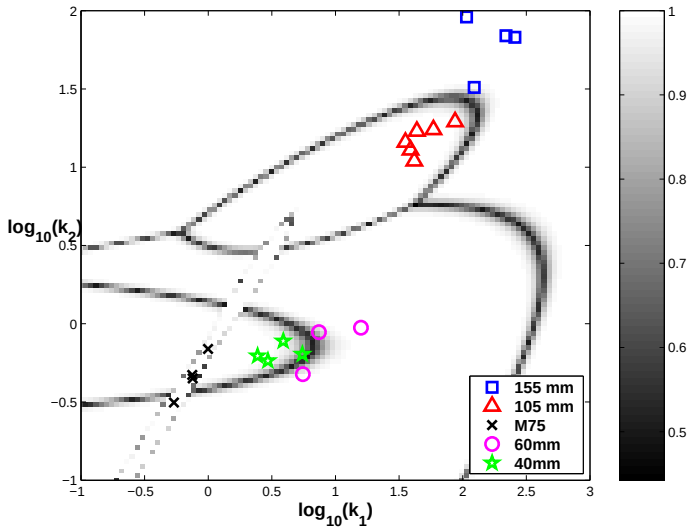
Assuming a parametric form for the likelihood function greatly simplifies the problem of density estimation. However, this assumption may be difficult to justify if limited training data are available (ie statistical tests may not provide sufficient confidence that the assumed distribution is valid). In this situation, we may turn to nonparametric methods, which estimate a likelihood function $p(\boldsymbol{\nu}|\omega_i)$ from the training data

$$\hat{p}(\boldsymbol{\nu}|\omega_j) = \frac{1}{N_{\omega_j}} \sum_{\mathbf{x} \in \omega_j} K(\boldsymbol{\nu}, \mathbf{x_i}). \tag{5.29}$$

where $K(\boldsymbol{\nu}, \mathbf{x_i})$ is a kernel (or windowing) function centered on the training vector $\mathbf{x_i}$. The summation in the above equation is over the N_{ω_j} training vectors belonging to class ω_j. The kernel is commonly a density function (eg a Gaussian) with some parameter σ controlling its width. If the kernel function is chosen to be a delta function, then as the size of the training data set goes to infinity the expected density estimate will approach the true



(a) Linear Discriminant Analysis



(b) Quadratic Discriminant Analysis

Figure 5.6: Discriminant analysis applied to the $k1$ and $k2$ parameters from cooperative inversion. Grayscale images show the maximum posterior probability evaluated over all possible classes. Decision boundaries correspond to minima of this function.

density. However, with limited training data delta functions will produce a high variance, "spiky", estimate of the underlying distribution which will provide poor generalization to the test data (ie an estimate which overfits the training data). Here the choice of kernel width acts as a regularization parameter: a small kernel width will produce a spiky distribution whereas a large kernel width will produce a smooth distribution (figure 5.7).

A representative nonparametric classifier is the probabilistic neural network (PNN). This name is something of a misnomer, the classifier is an application of Bayes rule using a nonparametric estimate of the class likelihood functions. The PNN is often implemented using a single smoothing parameter for all classes. I have also implemented a PNN which

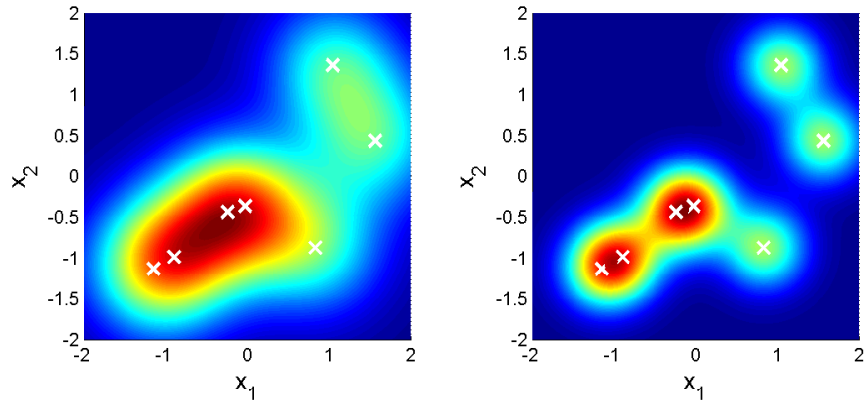


Figure 5.7: Nonparametric density estimation. Kernel centers are shown as crosses. A large kernel width will produce a smooth distribution (left) compared with a small kernel width (right)

allows a separate smoothing parameter for each class. This is analogous to allowing a separate covariance for each class in quadratic discriminant analysis and produces a more complex decision boundary.

Figure 5.8 shows PNNs applied to the Yuma data. The single-smoothing PNN produces decision boundaries which are quite similar to linear discriminant analysis. By the central limit theorem, the superposition of Gaussian kernels will converge to a Gaussian distribution as the number of kernels and the smoothing parameter are increased. The regularized PNN therefore tends to linear discriminant analysis. As was seen with quadratic discriminant analysis, a PNN with separate smoothings can overfit the training data and produce decision boundaries which are overly complex.

5.5 Discriminative Classifiers

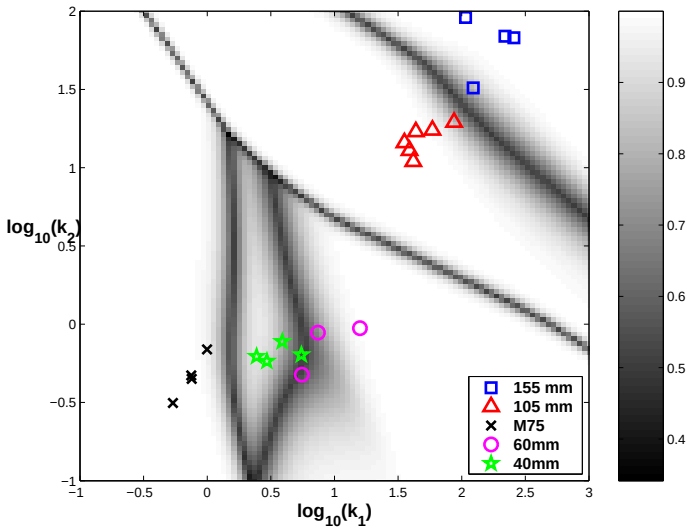
Instead of estimating posterior probability distributions, discriminative classifiers directly define a decision boundary to classify test data. As illustrated in figure 5.9(a), the learning problem is nonunique in this formulation. Given linearly separable feature data there is an infinite number of decision boundaries which can perfectly separate the training 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 is to maximize the *margin* between classes, subject to the constraint that the training data are classified correctly [32]. This concept is illustrated in figure 5.9(b). Referring to figure 5.9(b), the decision boundary satisfies

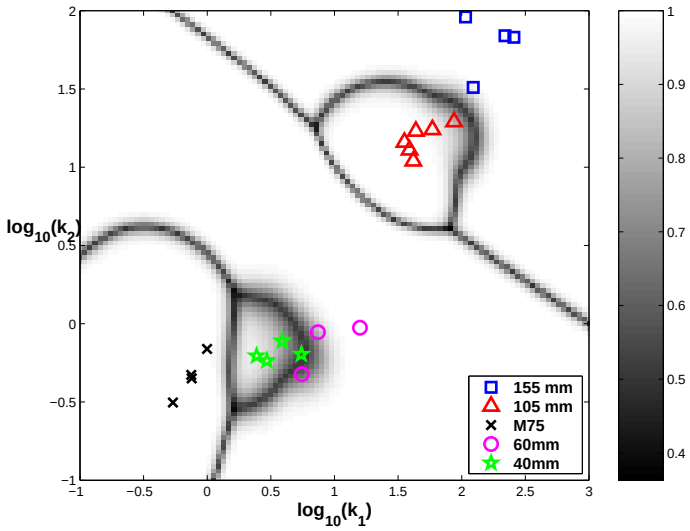
$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{5.30}$$

With this formulation we can derive the constrained optimization problem

$$\min_{\mathbf{w},\, b} \left[\frac{\|\mathbf{w}\|^2}{2} - \sum_{i=1}^{N} \alpha_i((\mathbf{w} \cdot \mathbf{x_i} + b)y_i - 1) \right] \tag{5.31}$$



(a) Single smoothing parameter



(b) Separate smoothing parameters

Figure 5.8: Probabilistic neural networks applied to the $k1$ and $k2$ parameters from cooperative inversion. Grayscale images show the maximum posterior probability evaluated over all possible classes. Decision boundaries correspond to minima of this function.

with α_i the Lagrange multipliers. The optimization problem can also be expressed in terms of its *dual* problem

$$\max_{\alpha \geq 0} \left[\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i y_i (\mathbf{x_i}^T \mathbf{x_j}) y_j \alpha_j \right] \tag{5.32}$$

where N is the number of points in the training data. The dual problem has the constraints

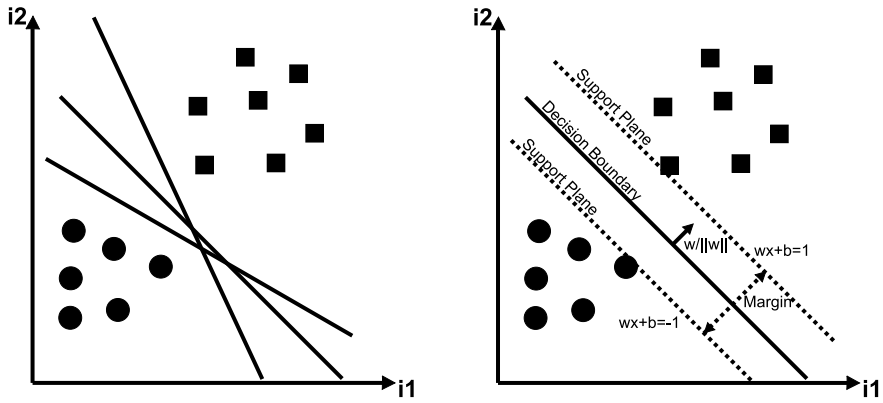


Figure 5.9: (a) Nonuniqueness for discriminative classifiers: an infinite number of decision boundaries can separate the training data. (b) Support vector machine formulation for linearly separable feature data.

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{5.33}$$

and

$$\sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1] = 0, \ \forall i \tag{5.34}$$

with $y_i \in \{-1, 1\}$ defining the class labels for the two classes. Inspection of equation 5.34 shows that only those feature vectors satisfying

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) = 1 \tag{5.35}$$

will have nonzero coefficients α_i. Feature vectors in the training data with nonzero coefficients are called *support vectors*; in a sense they hold up, or support, the decision boundary.

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 optimal separating hyperplane in this space [25].

Mapping the features to a higher-dimensional space seemingly risks the curse of dimensionality. However, there is an implicit dimensionality reduction included with support vector machines. The prediction for a test vector is generated by projection of that test vector onto a small number of (nonorthogonal) support vectors. This is analogous to the dimensionality reduction in canonical analysis where the test vectors are projected onto a small number of nonorthogonal directions.

Extension of the SVM to problems with more than two classes is non-trivial and beyond the scope of this thesis. Instead of the five-class Yuma feature data used earlier in this chapter, I therefore show linear and nonlinear SVM solutions for a synthetic two-class problem in a two-dimensional feature space (figure 5.10). The nonlinear SVM with Gaussian kernels produces a more complicated decision boundary, similar to what we might expect from a PNN. The output of this classifier ranges on $[-\infty, \infty]$. While it is possible to convert

the output of an SVM to a probability, this step is unnecessary if our aim is to prioritize feature vectors based upon their predicted membership in a given class.

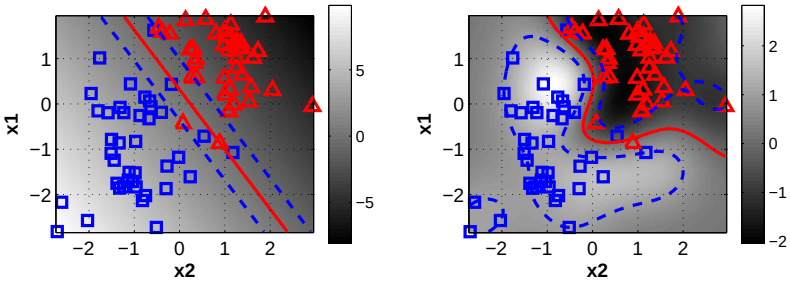


Figure 5.10: Linear (left) and nonlinear (right) support vector machines applied to a synthetic two-class problem. Solid lines indicate the decision boundary and dashed lines are the support planes.

5.6 Classifier training

Before a classifier can generate predictions of class memberships for test data, its performance must be optimized using the available training data. It should be apparent from the previous sections that most classifiers have adjustable parameters which govern the complexity of the decision boundary. The training stage of the learning process requires that we estimate values for these parameters which will minimize the probability of misclassification for the test data.

If a large set of labelled data is available, we can divide it into training and validation sets. We use the training set to estimate parameters of the classifier and then evaluate the classifier using the (independent) validation data. With no regularization, we can often obtain a perfect "fit" to the training data (i.e. no misclassifications). As in an inverse problem, overfitting the feature data results in a model for the decision boundary which is overly complex. On the other hand, if the decision boundary is too simple it may produce a higher misclassification rate on both the training and validation data. The idea in training is to balance our fit to the training data with our desire to have good predictions made on the validation data. In geophysical inverse problems the same "bias-variance" balance must be struck between fitting the observed data and producing a model which reflects our a priori knowledge.

In many situations (including UXO classification) the training data set is relatively small and so we cannot divide it into training and validation sets. This leads to the cross-validation approach, which sets aside a randomly selected portion of the training data for validation. In "leave-one-out" (LOO) cross-validation, a training vector is set aside as a test vector, the classifier is trained on the remaining training set, and a prediction is then generated for the test vector. By repeating this process for all training vectors, an estimate of the misclassification rate for a given value of the regularization parameters is obtained [25]. The estimated misclassification rate obtained with LOO cross-validation is

$$\Phi = \sum_{j=1}^{K} \sum_{\mathbf{x} \in \omega_j} (1 - \hat{P}(\hat{\omega} = \omega_j | \mathbf{x})) \tag{5.36}$$

where $\hat{P}(\hat{\omega} = \omega_j|\mathbf{x})$ is the estimated posterior probability of membership of the vector $\mathbf{x}$ in its true class ω_j.

Figure 5.11 shows Φ as a function of the regularization parameter σ for a single-smoothing probabilistic neural network applied to the Yuma training data. The estimated error function Φ is a smooth function of the regularization parameter and has a global minimum on the range of the smoothing parameter considered here. I therefore apply the Levenberg-Marquardt algorithm discussed in chapter 3 to minimize this function. The optimization routine successfully estimates of the optimal smoothing parameter (indicated by the vertical dashed line in figure 5.11). The error is a nonlinear function of the regular-

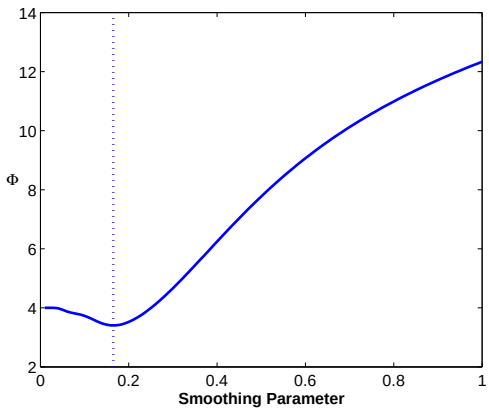


Figure 5.11: Estimated misclassification rate (Φ) as a function of smoothing parameter for a PNN. Vertical dashed line indicates the minimum of this function estimated by a linearized inversion algorithm. The training data used to generate this figure are from the Yuma test site

ization parameter and local minima may in general be a concern with linearized inversion. However, in experimentation with both real and synthetic data sets I have not encountered problems with local minima (ie multiple inversions with random starting models converged to the same parameters).

Using cross-validation as a means to estimate regularization parameters is appropriate for probabilistic neural networks and regularized discriminant analysis. Cross-validation works well if the classes have approximately the same numbers of feature vectors (eg as in the Yuma data). In the next chapter I will demonstrate that minimization of Φ for real data sets with significant skew (unequal class populations) can overfit the training data. I propose a simple heuristic which produces better classification performance than minimization of the LOO error.

Regularization of a support vector machine can also be done with cross-validation. However, each realization of the training data in the cross-validation process requires solution of an optimization problem. This computational burden quickly becomes onerous for even moderately-sized training sets (hundreds of feature vectors). Efficient methods for regularizing SVMs have been developed. However, I found it difficult to obtain reliable performance with these approaches. Instead, I adopt a heuristic for regularizing nonlinear SVMs with Gaussian kernels suggested by [33]. I set the kernel smoothing σ to be the median distance from each training vector to the nearest oppositely-labelled training vector (figure 5.12). I found this heuristic to be reliable for both synthetic data and the real data sets considered in chapter 6.

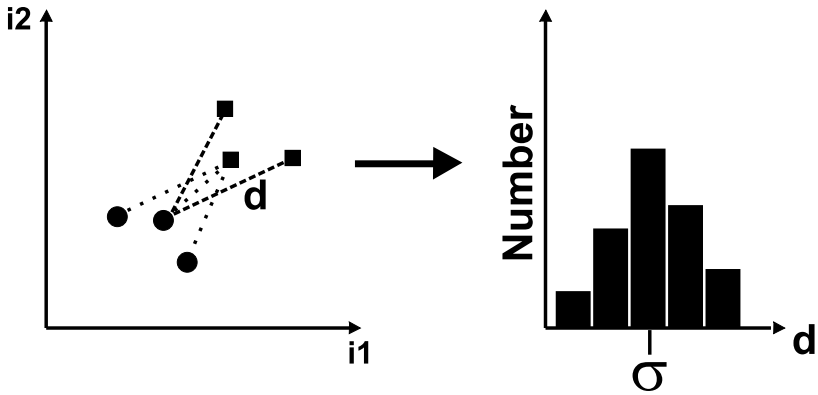


Figure 5.12: Heuristic for regularizing an SVM. The kernel smoothing σ is estimated as the median distance d from each training vector to the nearest oppositely-labelled point.

5.7 Classifier validation

When labels are available for a validation data set, we can compare the performance of various classifiers using these data. A useful tool for evaluating classifier performance is the receiver operating characteristic (ROC) curve. The ROC curve plots the proportion of true positives as a function of the proportion of false positives. In the UXO context this corresponds to plotting the proportion of UXOs correctly identified as a function of the proportion of clutter items incorrectly identified. An optimal classifier is one which finds all UXO items with a minimum number of false positives.

In determining which classifier is best for a given set of training and validation data, we must pose the question: when is the performance of classifiers significantly different so that we may prefer one classifier over all others? For a given realization of training and validation data, the ROC curve for a classifier can be regarded as a cumulative distribution function. We can then test to see whether the ROC curves are significantly different. An appropriate statistical test in this situation is the two-sample Kolmogorov-Smirnov (KS) test, which tests to see whether two random samples $(S1, S2)$ are drawn from the same distribution [34]. The test statistic T is the maximum difference between estimated cumulative distribution functions

$$T = \max(|\hat{P}_c(S1) - \hat{P}_c(S2)|) \tag{5.37}$$

where $\hat{P}_c$ is the estimated cumulative distribution function for a given sample. If the test statistic exceeds a tabulated critical value, then we reject the null hypothesis that the two samples are drawn from the same distribution. Equivalently, if the P-value computed from the test statistic is less than the desired significance level α, then we reject the null hypothesis. The P-value is computed as

$$\text{P-value} = 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2\lambda^2 j^2) \tag{5.38}$$

with

$$\lambda = (\sqrt{N} + 0.12 + 0.11/\sqrt{N})T. \tag{5.39}$$

The infinite sum in equation 5.38 can be accurately approximated with its first hundred terms. The P-value depends upon the degrees of freedom N of the test statistic T. The

number of degrees of freedom in turn depends upon the number of items N_1 and N_2 in our two samples

$$N = \frac{N_1 N_2}{N_1 + N_2}. \tag{5.40}$$

If we are treating ROC curves as cumulative distributions, then it is reasonable to use the numbers of UXO found by each classifier in our KS test to compute the degrees of freedom N.

Figure 5.13 demonstrates the KS test for two hypothetical ROC curves. The right plot shows the computed P-value as a function of the number of UXOs found by the classifiers. Here I assume that both classifiers find the same number of UXOs, so that $N_1 = N_2$. According to figure 5.13, if the ROC curves were generated by more than 160 UXO items, then at a 95 % confidence level we would reject the null hypothesis that the two classifiers provide the same performance. To reject the null hypothesis at a higher confidence level we require more evidence: the ROC curves are significantly different at a 99 % confidence level if they are generated by more than 230 UXO items.

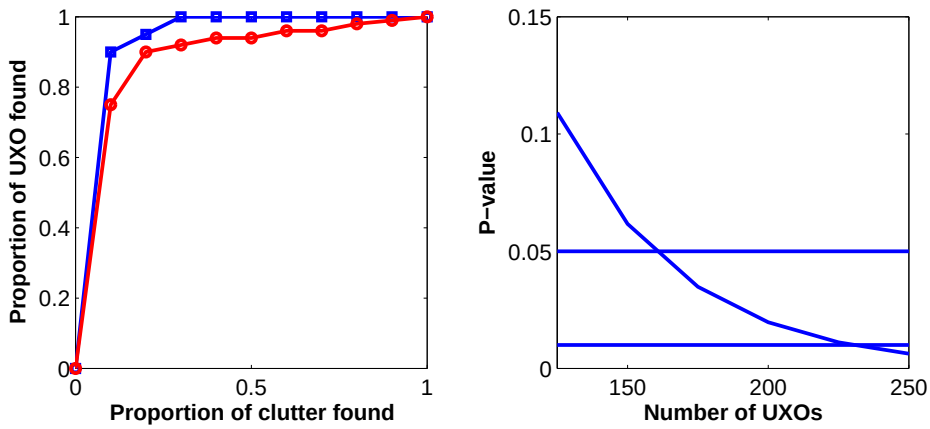


Figure 5.13: Left: two hypothetical ROC curves. Right: Computed P-value for the corresponding Kolmogorov-Smirnov statistic as a function of number of UXOs. The horizontal lines show the cut-off for the $\alpha = 0.05$ and $\alpha = 0.01$ significance levels.

As an example of ROC analysis, I generate synthetic test data for the Yuma feature data. I assume that the generative model used in linear discriminant analysis is the true model (i.e. I use a pooled covariance matrix for all classes). I then generate a total of 200 UXO items in the test data as random samples from normal distributions with the means and covariances of each class. The relative frequencies of each ordnance type in the synthetic test data are equal to the frequencies in the training data. Figure 5.14 shows the observed training data and synthetic test data for this simulation. This example is only intended to illustrate ROC analysis and so comes with the following disclaimer: in the absence of ground truth about the distribution of clutter at this site, the performance of classifiers in this simulation is not necessarily representative of their actual performance.

To generate a synthetic ROC for each classifier in figure 5.15, I first compute the predicted probabilities of class membership for all test data. I then identify all test data with a maximum probability of membership above a given threshold P_{thesh}, starting with $P_{thesh} = 1$. These items are then labelled and the numbers of found UXO and false positives (items whose class was incorrectly predicted by the classifier) are added to running totals.

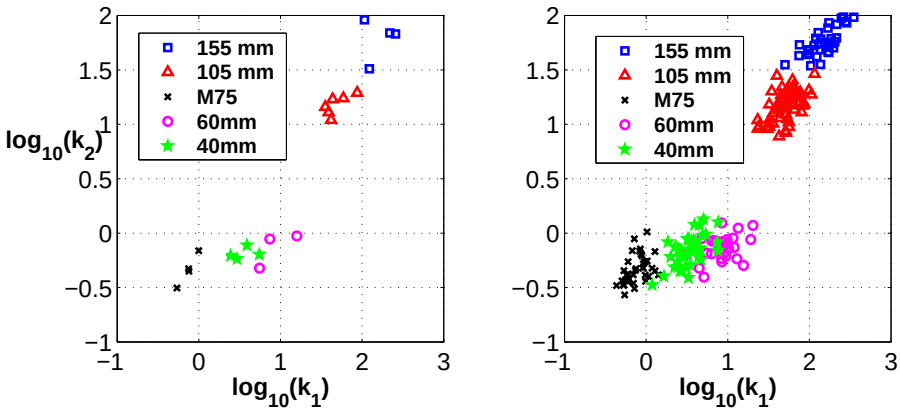


Figure 5.14: Left: Training data for Yuma calibration grid. Right: Synthetic test data for Yuma calibration grid, assuming a pooled covariance for all classes.

In this example all targets are UXO and so the number of UXO found at a given threshold is simply the number of targets which are labelled. This process is repeated by lowering P_{thesh} until all UXO have been labelled.

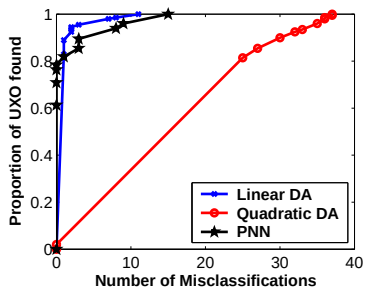


Figure 5.15: Receiver operating characteristic curves for classifiers applied to the feature data of figure 5.14. DA denotes discriminant analysis.

Linear discriminant analysis provides the best performance of all classifiers for this example. This is because this classifier uses the true distributions of the test data and so it will, on average, have an optimal error rate (i.e. it will have the smallest number of misclassifications to find all UXOs). The assumption of separate or pooled covariance matrices has a strong effect on classifier performance: quadratic discriminant analysis performs quite poorly. Finally, the probabilistic neural network has a comparable ROC to that of linear discriminant analysis. This is consistent with the earlier observation that the PNN produces decision boundaries which are quite similar to linear discriminant analysis.

These observations can be quantified with KS tests of the ROC curves in figure 5.15. I test whether the ROCs are significantly different from that of linear discriminant analysis. To compute the tests statistic T, I linearly interpolate the ROC curves at discrete values of the misclassification rate. The ROCs in figure 5.15 are not single-valued functions of the misclassification rate and so T must be calculated as

$$T = \max_{\forall i}(\min(ROC1_i - ROC2_i)) \tag{5.41}$$

with $\min(ROC1_i - ROC2_i)$ denoting the minimium difference between the two ROC curves

at the ith interpolation location. Retaining the minimum difference between the two ROC curves at an interpolation point will produce a smaller test statistic. Consequently, we are less likely to reject the null hypothesis when it is true (a type I error).

At a 95% confidence level, the ROC for the PNN is not significantly different from that of linear discriminant analysis (P-value=0.3735). In contrast, the ROC for quadratic discriminant analysis is significantly different from that of linear discriminant analysis at a 95 % confidence level. In fact, the ROC curves for LDA and QDA are significantly different at any confidence level since P-value=0 for this test.

Analysis of ROC curves with hypothesis testing provides a rigorous basis for evaluating classifiers. In particular, we can definitively say that a particular classifier provided worse performance than the best classifier for that data set. However, we must be cautious when the null hypothesis is retained at a given confidence level. Retaining the null hypothesis does not mean we have proven there is no difference between the performance of two classifiers, it means there is insufficient evidence to reject the null hypothesis. Consider again the two hypothetical ROC curves shown in figure 5.13. According to the KS-test these curves are not significantly different at the 95% confidence level until the number of UXOs in the test data exceeds 160. Does this mean we will attain the same performance from these two classifiers? Certainly not, since one ROC curve requires us to dig all clutter items to find all UXOs whereas the other only requires us to dig 30% of clutter items. This can represent a significant expenditure in a real context. Hence if the KS test indicates that two classifiers are not significantly different, then we must consider their actual performance (the number of clutter items dug in order to find all UXOs) when validating classifiers.

5.8 Discussion and conclusions

In this chapter I have reviewed statistical pattern recognition. I presented generative classifiers, which apply Bayes rule to minimize the probability of misclassifying a feature vector. Discriminative classifiers also attempt to minimize the probability of misclassification, but circumvent the problem of estimating parameters of the likelihood function by directly defining the decision boundary. In the next chapter, I demonstrate the application of statistical

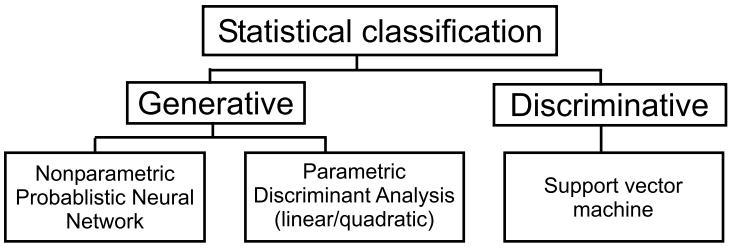


Figure 5.16: Classification algorithms presented in this chapter

classification to several real data sets. Practical applications of statistical classification algorithms have shown that there is no single algorithm that will work best in all situations. While the support vector machine is appealing in its formulation as an optimization problem, generative classifiers can outperform SVMs. It is prudent to use all available classifiers and evaluate their performance as labelled data become available. I will investigate this approach to UXO classification in the next chapter.

Chapter 6

UXO classification

6.1 Introduction

In this chapter, I review research on the application of classification algorithms to the problem of discriminating between UXO and clutter items. I also apply classification algorithms to real data sets from a number of field sites.

The UXO classification problem presents a number of challenges which are not always encountered in pattern recognition. These include:

- Useful features are site-specific.
- There is uncertainty in the feature vectors.
- The UXO class is often far less numerous than the clutter class.
- Training data is limited and is expensive to acquire.

I address the problem of feature selection for magnetic and electromagnetic data using canonical analysis. I also examine the propagation of errors through the classification process with Monte Carlo simulations.

The final two issues in the above list are central to the success of statistical classification. The rarity of UXOs at a field site makes random sampling an inefficient way to generate training data: we are unlikely to obtain enough UXO ground truth to model the UXO class with confidence. In this chapter, I demonstrate how retraining a statistical classifier as the training data set grows can alleviate this problem.

The statistical classifiers described in the previous chapter are not the only available approach for discriminating between UXO and clutter items. A commonly-used discrimination method is to dig every anomaly with an amplitude above a specified threshold. While simple thresholding circumvents the sometimes laborious process of inversion, anomaly amplitude is not a particularly robust feature for discriminating between UXO and clutter. For example, UXO items at depth may produce a comparable anomaly amplitude to shallower clutter items.

Feature extraction via inversion can provide more useful diagnostics for discrimination. In cases where multiple data types are available, we can threshold on some linear combination of the available features. At a higher level of sophistication, we may take nonlinear combinations of features. The remanence classifier developed by Billings is an example of this approach: we project the components of the estimated moment onto a nonlinear function to obtain the estimated remanence [8].

An advantage of these thresholding or "rule-based" classifiers is that they do not require any information regarding the distribution of clutter at a field site. In contrast, statistical classifiers model the decision boundary between UXO and clutter and so we need information about both classes before we can make predictions for the test data. However, I will

demonstrate in this chapter that statistical classifiers can learn from the training data as digging proceeds and thereby improve their performance.

Figure 6.1 shows a hierarchy of classifiers which can be used for UXO classification. It should be emphasized at the outset that increasing sophistication does not necessarily imply better performance. As will be demonstrated in this chapter, the performance of classifiers is dependent upon the particular data set and there is no classifier which performs best in general. Given this result, I propose an algorithm for selecting a classifier during digging. The algorithm selects a classifier from a suite of classifiers (statistical or otherwise) using a performance metric estimated from the training data.

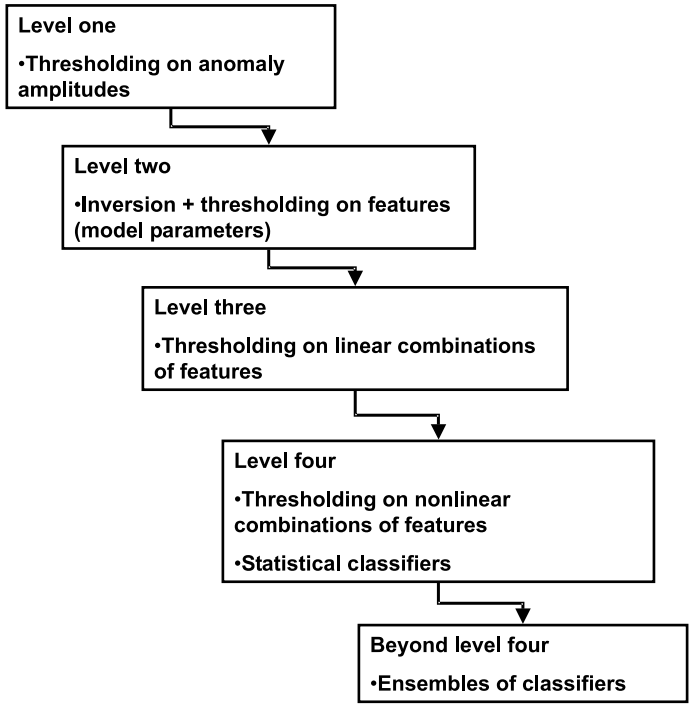


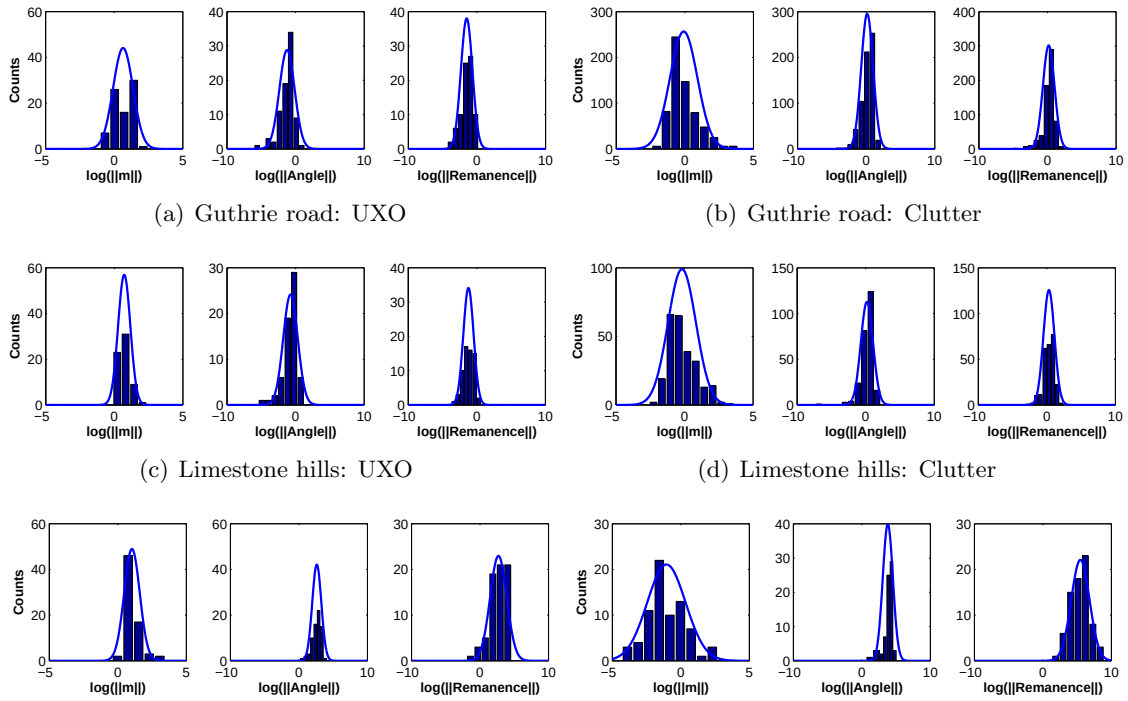
Figure 6.1: Flowchart showing increasing sophistication in UXO classification.

6.2 Feature selection

The existing literature on UXO classification uses a wide variety of features estimated from electromagnetic and magnetic data. The features are most often parameters of the estimated dipole for magnetics and shape and size parameters (eg $k1$,$k2$ in the Pasion-Oldenburg model) for electromagnetics. Some authors have also used goodness-of-fit as a feature, reasoning that UXO items have anomalies which can be better fit by parametric models. Estimated depth can also provide some discrimination between UXO and clutter, especially at sites with large ordnance items that are relatively deep below the surface. In the following sections, I prioritize raw features using canonical analysis of real data.

6.2.1 Canonical analysis of magnetics data

An appropriate feature set for discrimination of magnetics data can be selected with canonical analysis. I consider three data sets acquired at field sites: Guthrie road, Montana, Lime-



(a) Guthrie road: UXO (b) Guthrie road: Clutter

(c) Limestone hills: UXO (d) Limestone hills: Clutter

(e) Badlands range: UXO (f) Badlands range: Clutter

Figure 6.2: Marginal class distributions for log-transformed features. Solid lines are normal distributions with the estimated class means and variances.

stone hills, Montana and Badlands bombing range, South Dakota. Table 6.2.1 summarizes the UXO and clutter items found at these sites.

Data set	**UXO types**	**No. of UXOs**	**No. of Clutter**
Guthrie road	81 mm, 76 mm	80	644
Limestone hills	76 mm, 81 mm, 75 mm, 155 mm, 90 mm, 105 mm, 4.2" illumination round, 2.75" rocket	64	253
Badlands range	M38 practice bombs	71	73

Table 6.1: Summary of ordnance and clutter items at field sites.

The aim of canonical analysis is to prioritize features at each field site based upon their ability to separate UXO and clutter. For a raw feature set, I select the norm of the estimated moment and the angle of the moment with the earth's magnetic field, as well as the minimum estimated remanence.

To meet the assumptions of canonical analysis I log-transform and standardize the features (ie subtract the mean of each feature). Figure 6.2 shows the resulting marginal distributions for UXO and clutter classes at the field sites.

Only the marginal class distribution of $\log(\|\mathbf{m}\|)$ for UXOs at Limestone Hills is normal at the 95% confidence level (Kolmogorov-Smirnov one-sample test). However, these distributions are sufficiently well-behaved (ie approximately unimodal) for the purposes of

canonical analysis. Taking the logarithm of the angle is not strictly necessary for these data sets: an approximately normal marginal can also be obtained by subtracting the mean of this feature. However, I found that the canonical analysis was unchanged by the particular transformation used on the angle. This is because the (natural) logarithm is an approximately linear function for most estimated angles in the data. Consequently, the number of standard deviations between class means is largely unaffected when we take the log of the angle. Figures 6.3 and 6.4 compare the resulting projections onto the first canonical eigenvector with and without the log transformation of the angle.

For all data sets, canonical analysis of the transformed features produces a single eigenvector which describes 100% of the separation between the classes. The contributions from the raw features to these eigenvectors are given in table 6.2.1. At Guthrie road, the most

Data set	$\log(\|\mathbf{m}\|)$	log(Angle)	log(Remanence)
Guthrie road	-0.0287	0.4936	0.8648
Limestone hills	0.3081	0.0737	0.9485
Badlands range	-0.5036	0.7510	0.4270

Table 6.2: Contributions of features to the first canonical axis of each data set.

significant contribution to the separation between UXO and clutter classes is provided by remanence. The angle of the estimated moment with the earth's field also provides some discrimination between the classes. This is consistent with a physical understanding of magnetization: because of demagnetization effects the induced moment of a UXO is expected to be restricted to a range of angles about the earth's field. Figure 6.3 shows the projection direction identified by canonical analysis to provide the maximum separation between UXO and clutter classes for the Guthrie road data.

At Limestone hills the largest contribution to the first canonical variate is again from remanence, but in this case the size of the moment provides more separation between classes than does the angle. This is because the marginal distribution of $\log(\|\mathbf{m}\|)$ for UXO items is shifted to larger values compared to the same distribution at Guthrie road. The marginal distribution of $\log(\|\mathbf{m}\|)$ for clutter items is also very subtly shifted to smaller values compared to the same distribution at Guthrie road. While it is tempting to suggest that there are larger UXOs and smaller clutter items at Limestone hills than at Guthrie road, we cannot make any inferences about target size using magnetics data. The analysis only indicates that there is a difference in the distributions of estimated moments at Guthrie road and Limestone hills.

Canonical analysis indicates that remanence produces the smallest separation between UXO and clutter classes at Badlands bombing range. In contrast with the other sites, a significant number of UXO items at Badlands have a relatively large estimated remanence ($> 50\%$). This result is somewhat surprising, given that estimated moments for UXOs tend to have small angles with the earth's magnetic field. This suggests that their magnetization is primarily induced. However, the dig list shows that large number of ordnance items are "pancaked", crushed or otherwise distorted. With the exception of one outlier, distorted UXOs tend to have a smaller estimated remanence than intact UXOs (figure 6.5) at the Badlands site. A possible explanation for this is that many intact bombs did not strike the ground with sufficient force to fully erase their remanent magnetization. Alternatively, the nominal dimensions of 250 lb. bombs (37 cm diameter, 162 cm length) used to generate the feasibility curve may require some tuning.

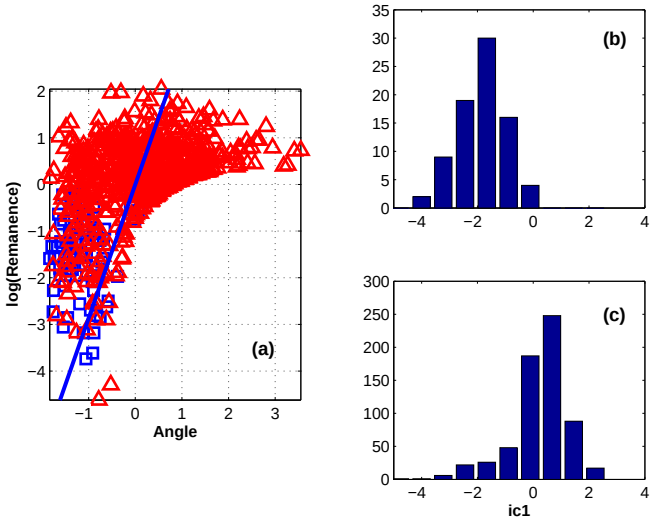


Figure 6.3: (a) Standardized feature space spanned by angle and log(Remanence) showing UXO (squares) and clutter (triangles) classes for Guthrie road data. The solid line is the eigenvector from canonical analysis providing the largest separation between classes. (b) Distribution of UXOs after projection onto the first canonical eigenvector. (c) Distribution of clutter after projection onto the first canonical eigenvector.

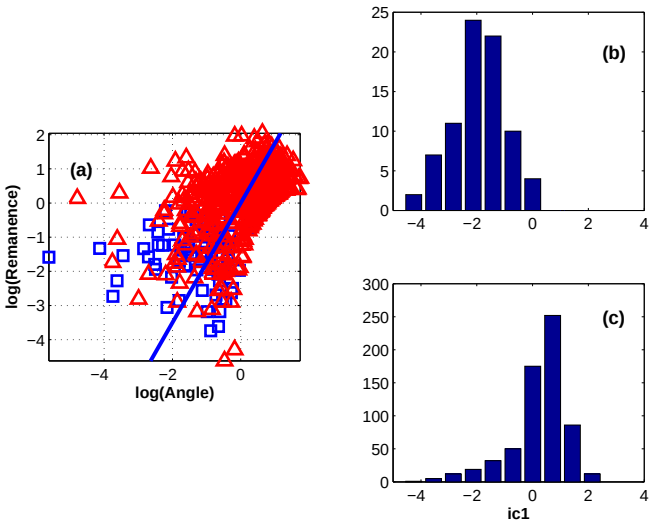


Figure 6.4: (a) Standardized feature space spanned by log(Angle) and log(Remanence) showing UXO (squares) and clutter (triangles) classes for Guthrie road data. The solid line is the eigenvector from canonical analysis providing the largest separation between classes. (b) Distribution of UXOs after projection onto the first canonical eigenvector. (c) Distribution of clutter after projection onto the first canonical eigenvector.

Figure 6.6 compares the receiver operating characteristic curves (ROC) generated by thresholding on the first canonical variate with that generated by thresholding on remanence alone. Although the canonical variate provides an initial decrease in the false alarm rate at Guthrie road and Limestone hills, remanence ultimately requires us to dig fewer clutter items in order to find all UXOs. However, at these sites a two-sample Kolmogorov-Smirnov

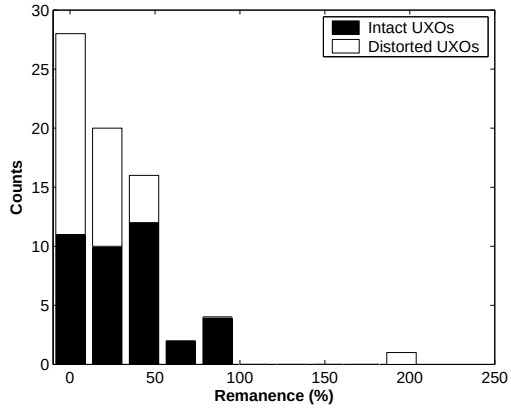


Figure 6.5: Estimated remanence for distorted and intact ordnance items at Badlands bombing range.

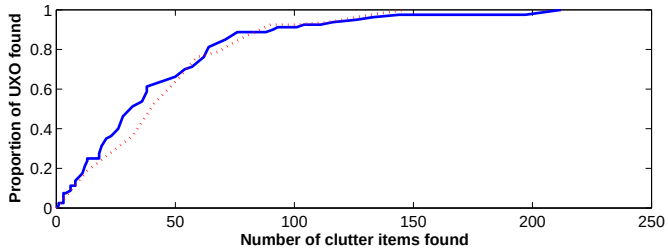
test indicates with 95% confidence that there is no significant difference between the two ROC curves. Therefore, we cannot conclude that thresholding on either remanence or the first canonical variate provides better classification performance. However, canonical analysis does indicate that, in comparison with the size of the moment and the angle with the earth's magnetic field, remanence provides the largest separation between UXO and clutter classes for these sites.

Thresholding on remanence produces a respectable false alarm rate at Badlands bombing range: in the process of digging all 71 UXOs we must dig only 20 clutter items. The canonical variate significantly improves this false alarm rate, requiring only 7 false alarms. However, the preceding analysis identified useful features using all available ground truth. In general, we will use a representative training set of labelled data to prioritize features with canonical analysis. A smaller training set will likely result in a higher false alarm rate when thresholding on canonical variates. This is an advantage of the rule-based remanence classifier: the performance of the classifier depends only upon prior knowledge of the ordnance items at a site.

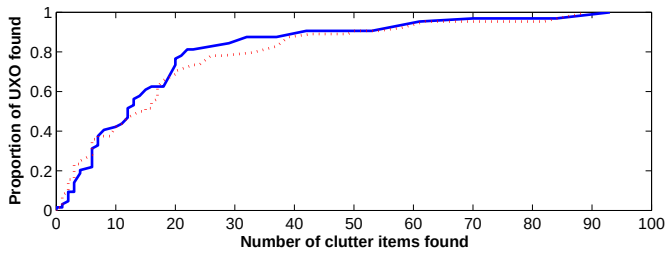
6.2.2 Canonical analysis of electromagnetic data

As a second example of canonical analysis applied to UXO discrimination, I consider features estimated from the Yuma test site. Table 6.2.2 summarizes the canonical analysis of these data. Unlike the analysis of magnetics data in the previous section, in this case there are two canonical variates which together describe 98.6% of the separation between classes. The parameters $k1$ and $k2$ contribute the most to the corresponding eigenvectors, with the β parameters providing negligible separation. This is consistent with the analysis of Pasion-Oldenburg parameters presented in chapter 4: the k parameters are more sensitive to target shape and size than are the β parameters.

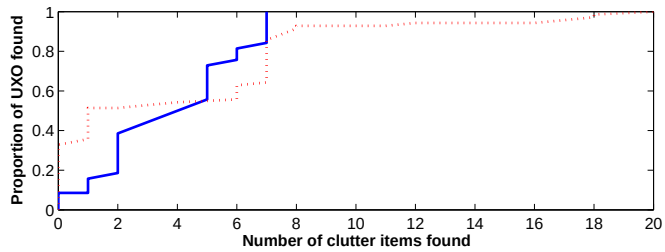
Projection onto the first two eigenvectors $(ic1, ic2)$ produces a feature space which appears very similar to a projection onto the original k parameters (figure 6.7). However, projection onto the second eigenvector ($ic2$ in figure 6.7) increases the separation between the M75 and 40 mm ordnance types at the expense of an increased overlap between 40 mm and 60 mm ordnance classes. In problems with multiple classes (> 2), the between-class scatter matrix will be dominated by classes whose means deviate the most from the



(a) Guthrie road



(b) Limestone hills



(c) Badlands bombing range

Figure 6.6: ROC curves generated by thresholding on the first canonical variate (solid line) and remanence (dashed line).

total mean of the feature data. The first few eigenvectors of $\mathbf{D}$ will try to separate these classes, sometimes to the detriment of the separation between other classes. Figure 6.7 clearly shows that the dimensionality of the feature space can be reduced to two features: a projection onto the last two eigenvectors gives no separation between the classes.

Index	1	2	3	4
Eigenvalue	0.8285	0.1577	0.0127	0.0011
$\log_{10}(k1)$	0.3256	0.5733	0.7800	-0.5364
$\log_{10}(k2)$	0.9384	-0.7423	-0.5442	0.7296
β_1	-0.1114	-0.3468	0.2800	-0.1366
β_2	-0.0310	-0.0061	0.1309	0.4017

Table 6.3: Eigenvalues and eigenvectors of $\mathbf{D}$ for Yuma data. Eigenvalues are normalized to sum to one.

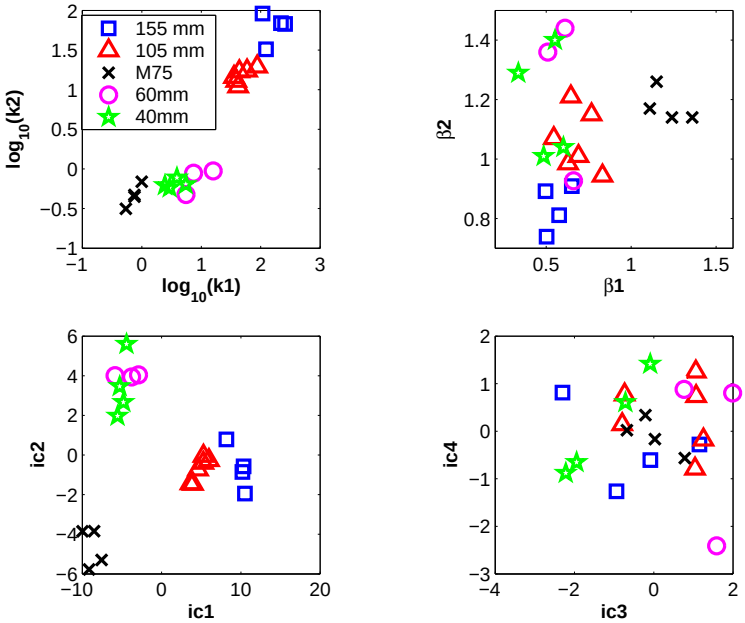


Figure 6.7: Canonical analysis of Yuma test data. Top row shows projections in the original feature space and bottom row shows projections onto the canonical eigenvectors.

6.3 Classification of magnetics data

The canonical analysis of feature data presented in the previous sections provide canonical variates which are linear combinations of the raw features. In general, statistical classifiers such as a PNN or nonlinear SVM try to separate classes using nonlinear combinations of the feature data. In this section, I compare the performance a suite of statistical classifiers with the rule-based remanence classifier.

Hart et al. trained a probabilistic neural network to discriminate between UXO and scrap using features extracted from magnetics data [35]. They found that a PNN trained using size (a proxy for strength of the dipole moment), depth, and inclination of best-fitting dipoles provided a higher rate of correctly classified UXOs and fewer false alarms than linear discriminant analysis.

6.3.1 Training for UXO classification

Estimating priors

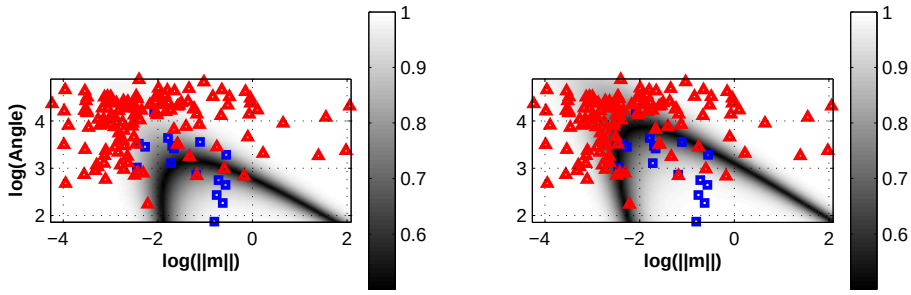
At first consideration, the estimate of the prior probability appears to be crucial to obtaining optimal performance from a generative classifier. For example, if we estimate the priors from the training data, then the prior probability of the UXO class will typically be quite small and will produce undesirably small posterior probabilities. This does not reflect our prior expectation that it is far more costly to misclassify a UXO item than it is to misclassify a clutter item.

However, the prior probability turns out to be something of a red herring in the context of UXO discrimination. Recall that for linear discriminant analysis the discriminant

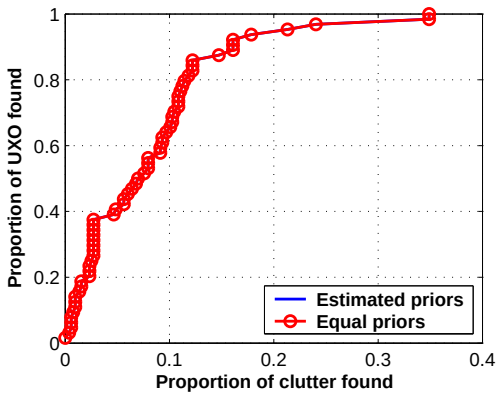
function is expressed as

$$f_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma_p^{-1} \mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_j^T \Sigma_p^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_i^T \Sigma_p^{-1} \boldsymbol{\mu}_i) + \frac{1}{2} \log \left(\frac{P(\omega_i)}{P(\omega_j)} \right) \tag{6.1}$$

with $f_{ij} = 0$ the decision boundary between classes ω_i and ω_j. In this expression the prior probabilities only appear in the constant term $\frac{1}{2} \log \left(\frac{P(\omega_i)}{P(\omega_j)} \right)$. Therefore the prior probabilities only affect the location of the decision boundary. This effect is demonstrated for a random realization of the Guthrie road data (figure 6.8(a)). Using equal or estimated priors with quadratic discriminant analysis merely shifts the decision boundary. The aim of UXO classification is not to dig all items on one side of the decision boundary but rather to generate a prioritized dig list by thresholding on classifier output. Consequently, while the values of the posterior probabilities will depend on the priors, the order in which items are dug is independent of the prior. This effect is shown in figure 6.8(b): using equal or estimated priors produces exactly the same ROC curve.



(a) Quadratic discriminant analysis applied to a random realization of Guthrie road training data. Squares are UXO items and triangles are clutter items. Left: estimated priors. Right: equal priors



(b) ROC curves generated by quadratic discriminant analysis with equal priors and estimated priors (the two curves are identical in this plot).

Figure 6.8: Effect of priors for UXO classification.

Regularization

Training a PNN or SVM for UXO classification requires estimation of a regularization parameter via cross-validation. In the previous chapter, the regularization parameter was found by minimizing the estimated probability of misclassification.

$$\Phi = \sum_{j=1}^{K} \sum_{i=1}^{N_{\omega_j}} (1 - \hat{P}(\hat{\omega} = \omega_j | \mathbf{x}_i)) \tag{6.2}$$

Hart et al. chose to square the leave-one-out error [35]

$$\Phi^* = \sum_{j=1}^{K} \sum_{i=1}^{N_{\omega_j}} (1 - \hat{P}(\hat{\omega} = \omega_j | \mathbf{x}_i))^2 \tag{6.3}$$

Although no justification was provided for this error function, squaring the probability of misclassification increases the importance of highly uncertain feature vectors in the cross-validation error function. Though Φ and Φ^* are nonlinear functions of the smoothing parameter σ, both functions tend to be smooth with well-defined minima (figure 6.9). The leave-one-out estimate also provides a reasonable approximation to the computed value for the test data. However, minimizing either error function does not provide particularly good

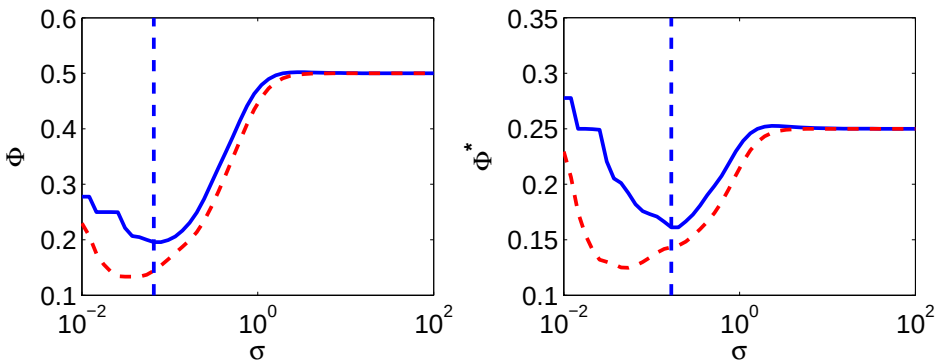


Figure 6.9: Cross validation error functions Φ (left) and $\Phi*$ (right) evaluated as functions of the kernel smoothing parameter σ. Solid line is the error function estimated from the training data and dashed line is the error function computed from the test data.

performance for a PNN (figure 6.11). The decision boundary is overly complex and produces a high false alarm rate. These error functions provide poor performance because most of the terms in the summation are from feature vectors which are correctly classified. Consequently, the error functions are insensitive to misclassified feature vectors. While squaring the probability of misclassification emphasizes the contribution from misclassified feature vectors, the smoothing parameter estimated from Φ^* still produces a high misclassification rate.

I have found that an effective heuristic for the kernel smoothing is the average within-class standard deviation, where the average is taken over all classes (see figure 6.10). This heuristic is motivated by analogy with the averaging of the class covariances used in linear discriminant analysis. As shown in figure 6.11, this estimate tends to produce a smooth decision boundary and better average performance than cross-validation methods.

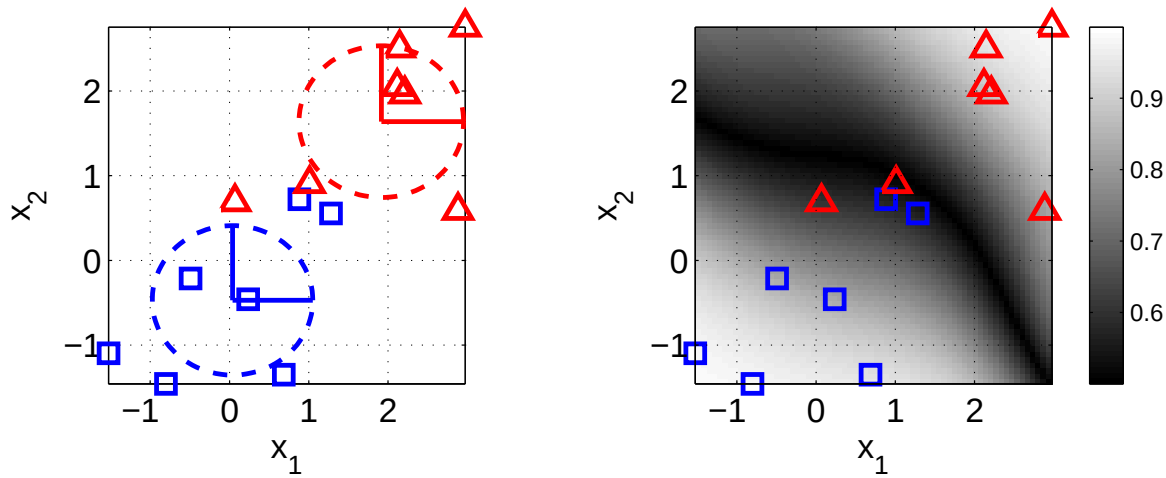


Figure 6.10: A heuristic for estimating PNN kernel smoothing. Left: synthetic feature data belonging to two classes. For each class I compute the standard deviation of that class along each feature (solid lines enclosed by dashed ellipses). The kernel smoothing is then the average standard deviation over all classes. Right: PNN decision boundary using a smoothing estimated with this heuristic.

6.3.2 Results

In this example, the training data are generated as random subsets of the entire data set. The performance of a statistical classifier will depend upon the particular realization of the training data set. I therefore compute an average ROC curve and its standard deviation over twenty realizations of the training and test data. To average ROC curves over multiple realizations, I interpolate the ROC curve for each realization at fixed values of the false alarm rate. I then compute a mean and performance bounds of the ensemble of ROC curves at each interpolation location. In addition to representing the average performance of a classifier, the interpolated curve also provides an indication of the worst case performance. This is because the curve is interpolated out to the maximum false alarm rate seen for the twenty realizations of the training data.

I also consider the effect of training data size on classifier performance. As the size of the training data set increases, we might expect classifier performance to improve. All classifiers are trained in a two-dimensional feature space spanned by the logarithm of the dipole moment magnitude and the logarithm of the angle with the earth's field.

Figure 6.12 shows the decision boundaries for all classifiers for a realization of the Guthrie road training data set. In this figure the training data are a random sample of 10 percent of all feature data. Quadratic discriminant analysis and the PNN produce similar decision boundaries in the feature space. Interestingly, the SVM output does not attain a value of +1 in the feature space. This means that there are no support vectors for the UXO class and the decision boundary is comprised of a nonlinear combination of support vectors from the clutter class. The shape of the decision boundary is nonetheless quite similar to that of the PNN.

Figure 6.13 shows the performance of linear and quadratic discriminant analysis for the Guthrie road data. Mean ROC curves generated by thresholding on remanence and angle are also shown for comparison. Linear discriminant analysis (LDA) produces an ROC curve which is quite similar to that generated by thresholding on angle. The performance of this

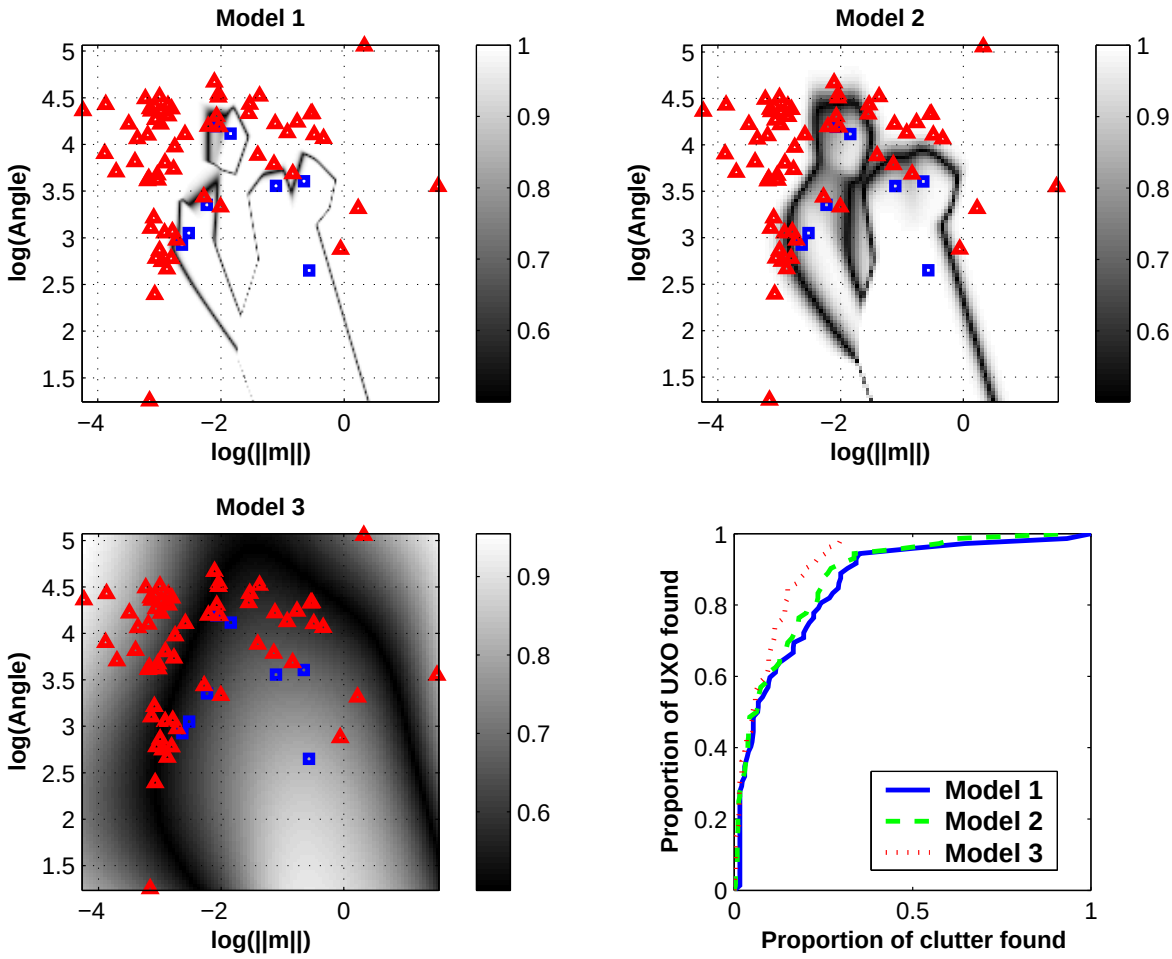


Figure 6.11: Models of the decision boundary for a realization of the Guthrie road training data. Model 1: PNN with σ estimated by minimizing Φ. Model 2: PNN with σ estimated by minimizing Φ^*. Model 3: PNN with σ estimated as the mean standard deviation within the training data. Bottom right: ROC curves generated from the test data using the PNN models

classifier does not significantly improve as the size of the training data set increases. This is because the pooled covariance matrix used in LDA is estimated from all the training data, regardless of class. Even for relatively small training data sets, there is sufficient information to obtain an accurate estimate of the pooled covariance. Consequently, LDA provides constant performance as the training data grows. In contrast, quadratic discriminant analysis (QDA) estimates a separate covariance for each class, and so it requires information about both classes. Small training data sets generated by random sampling contain few UXO items, resulting in poor performance. As more information is obtained, the estimate of the UXO class covariance significantly improves.

The mean ROC curves generated by thresholding on remanence and angle do not depend upon the size of the training data set. This is because the outputs of these rule-based classifiers are independent of the training data. While remanence provides the optimal classification performance at this site, the rule-based approach does not learn from the available information.

The performance of nonlinear classifiers on the Guthrie road data is shown in figure 6.14. The PNN with Gaussian kernels provides similar average performance to QDA. This classifier is even more sensitive to the realization of training data because the likelihood function depends directly upon the training data. Therefore for small training data sets there is relatively large variance in the performance of the classifier. As the training data set grows, the PNN estimates an improved approximation of the class likelihood functions and there is a corresponding decrease in the false alarm rate. The SVM shows a similar improvement in classification performance as the training set grows.

At Badlands bombing range, a linear decision boundary does a reasonable job at separating the two classes (figure 6.15). Regularization of the PNN and SVM yields a boundary which is approximately linear for this realization, while QDA models a curved boundary.

In contrast to the Guthrie road data, QDA does quite poorly on the Badlands data (figure 6.16). The additional structure in the QDA decision boundary hinders classification in this case. The SVM yields excellent performance on these data followed closely by the PNN (figure 6.17). For this data set, there is no significant improvement in the performance of classifiers as the size of the training data set is increased. This is because there is a large separation between UXO and clutter classes so that an approximately linear decision boundary can be estimated from small training sets.

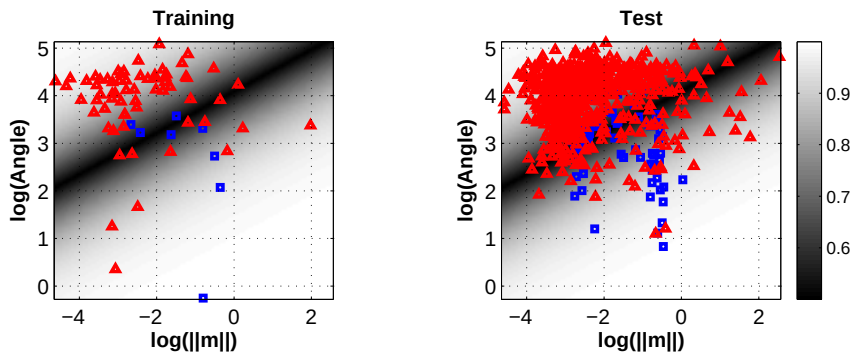
Statistical classifiers show more variability in their performance on the Badlands data than on the Guthrie road data. This is because the training data sets for the Badlands data are small relative to those used for Guthrie road. At Badlands there is more variability in the realizations of the training data and so there is more variability in the outputs of statistical classifiers. Interestingly, the variability of rule-based classifiers *grows* as the size of the training data set grows (see figure 6.17). This is an effect of the random sampling used in this study: as the training data set grows there is a commensurate decrease in the size of the test data. The realizations of the test data are increasingly variable and so the performance of rule-based classifiers for a given realization is more likely to deviate from the mean.

6.3.3 Retraining

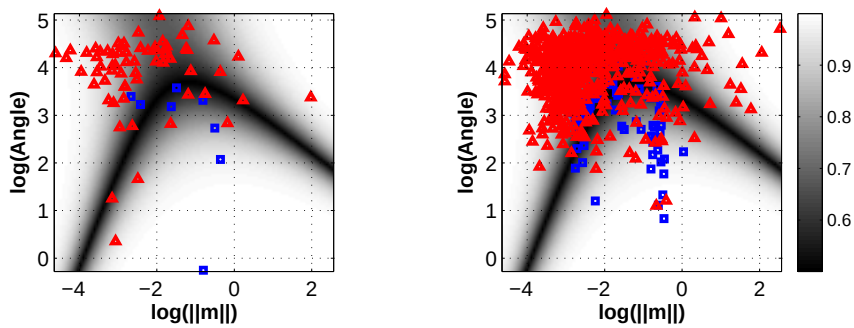
Relying on random training data is a risky proposition. If we are unlucky in our realization of training data, then our statistical classifier will perform quite poorly. Increasing the size of the training data set can alleviate this problem, though this will increase costs and is not guaranteed to lower the false alarm rate.

As items are excavated during field operations, our training data set grows. This additional information can be used to retrain the classifier and revise the dig list. I simulate this procedure by starting with a small random training data set (F=0.1). I train a PNN on these data and then dig 10 items most likely to be UXOs. The PNN is retrained with the new training data and the next 10 items are identified from the updated probabilities. Figure 6.18 compares the performance of the retrained PNN with a PNN that is trained using only the initial realization of training data. In this figure I show an ensemble of ROC curves generated from 20 realizations of the initial training data.

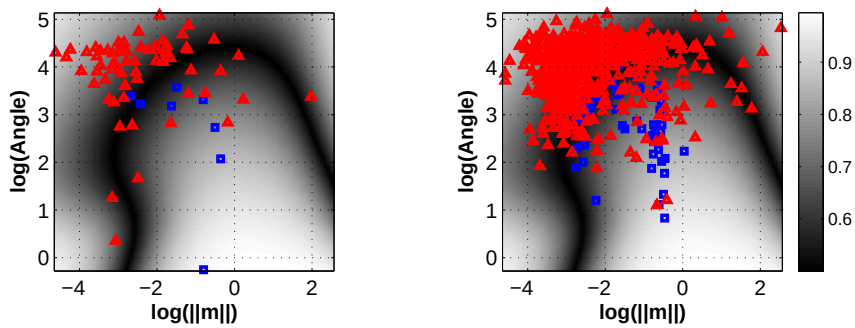
In the worst case scenario at Guthrie road, a PNN without retraining would require us to dig all clutter items. Retraining reduces this worst case false alarm rate to about 0.35. By learning from the training data as items are excavated, the retrained PNN is able to guard against "bad" realizations of the training data. At Badlands bombing range, the separation between UXO and clutter is so large that retraining does not improve upon the



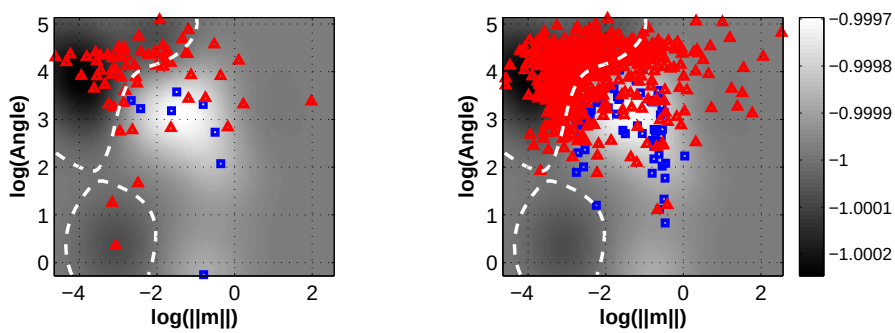
(a) Linear discriminant analysis



(b) Quadratic discriminant analysis

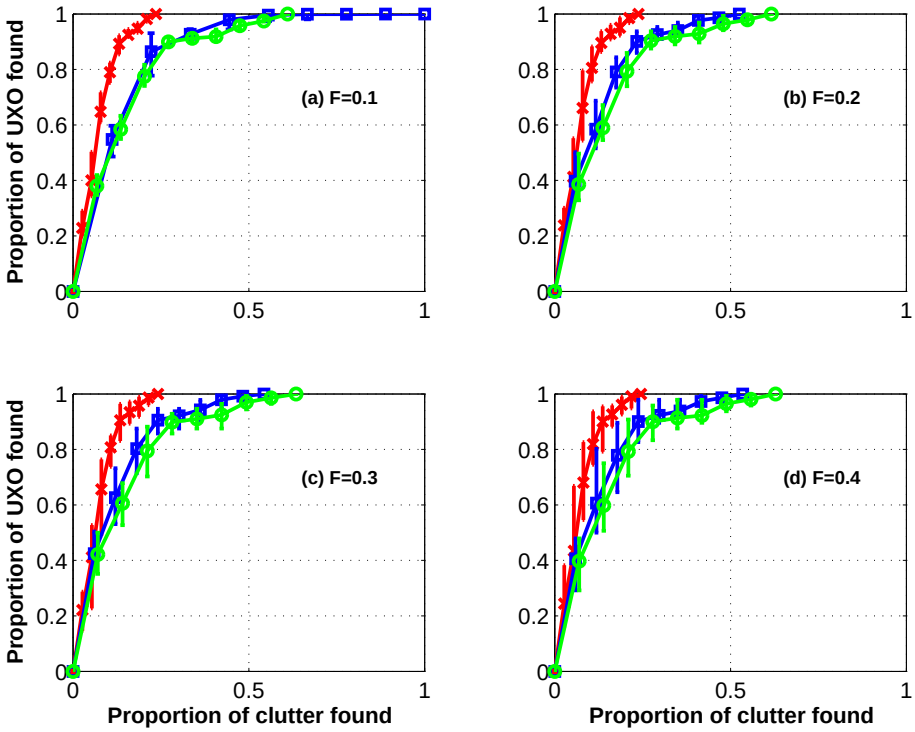


(c) Probabilistic neural network

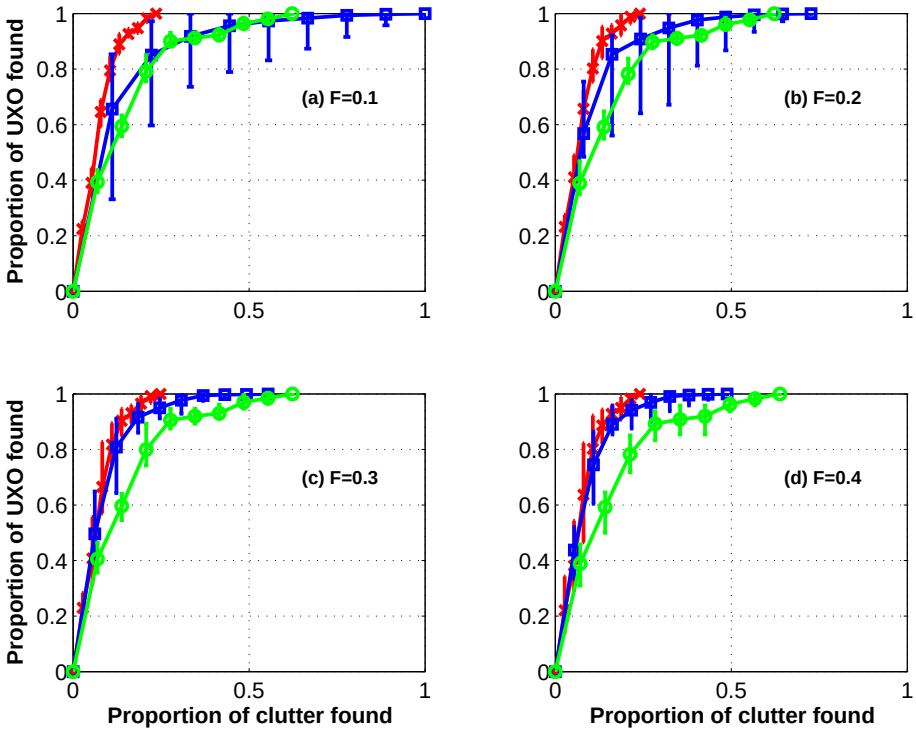


(d) Support vector machine: solid line shows the decision boundary and dashed lines show the support planes (only the support plane for the clutter class is visible in this example).

Figure 6.12: Classifier solutions for a single realization of the Guthrie road data. Squares are UXO items and triangles are clutter items. Training data are a random sample of 10% of all feature data.

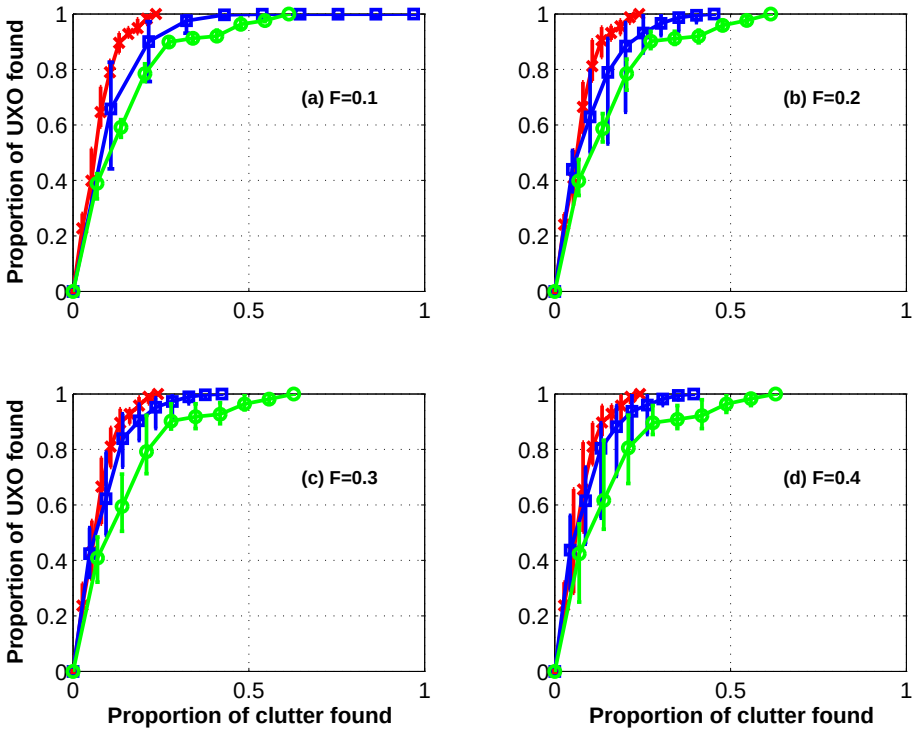


(a) Linear discriminant analysis

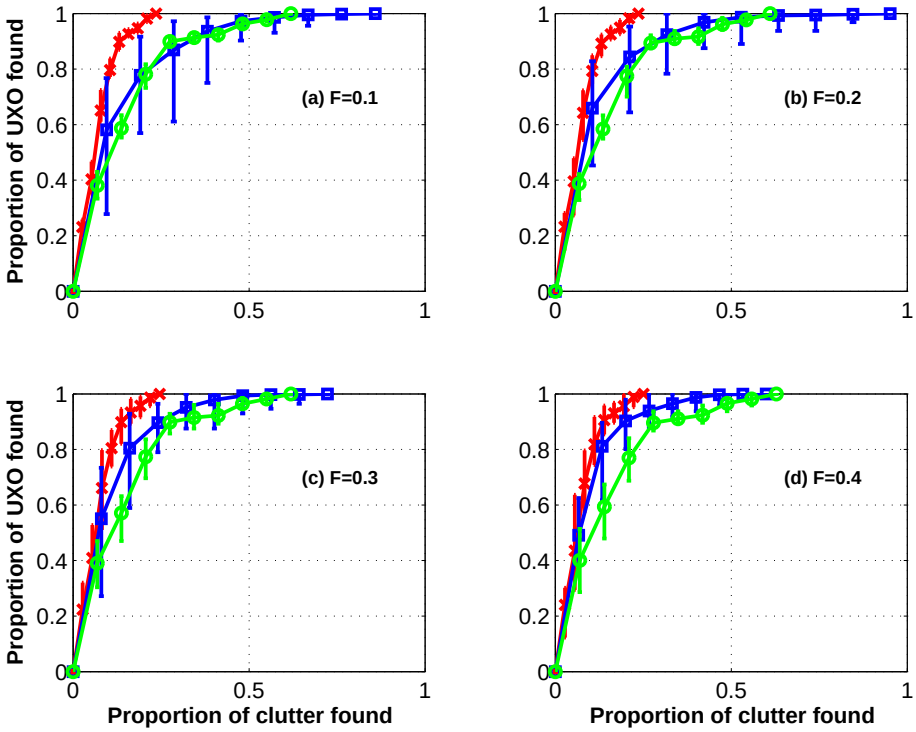


(b) Quadratic discriminant analysis

Figure 6.13: ROC curves of Guthrie road magnetics data for discriminant analysis (squares). Also shown are ROC curves generated by thresholding on remanence (crosses) and angle (circles). Subplots (a) through (d) show the average performance of classifiers when the size of the training data set is a fraction F of all feature data. Error bars indicate minimum and maximum performance bounds seen over all realizations for that classifier.

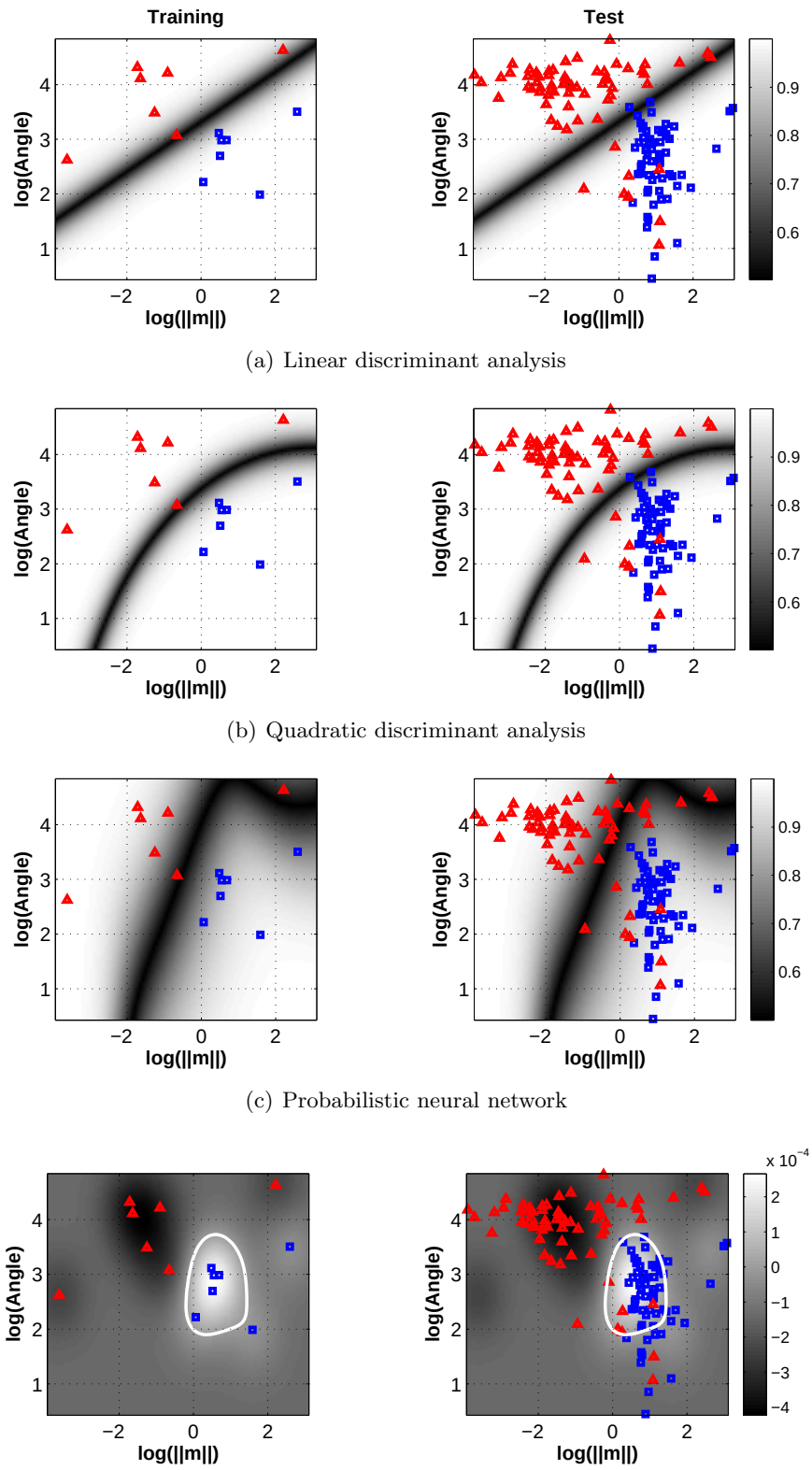


(a) Probabilistic neural network



(b) Support vector machine

Figure 6.14: ROC curves of Guthrie road magnetics data for nonlinear classifiers with Gaussian kernels (squares). Also shown are ROC curves generated by thresholding on remanence (crosses) and angle (circles). Subplots (a) through (d) show the average performance of classifiers when the size of the training data set is a fraction F of all feature data. Error bars indicate minimum and maximum performance bounds seen over all realizations for that classifier.



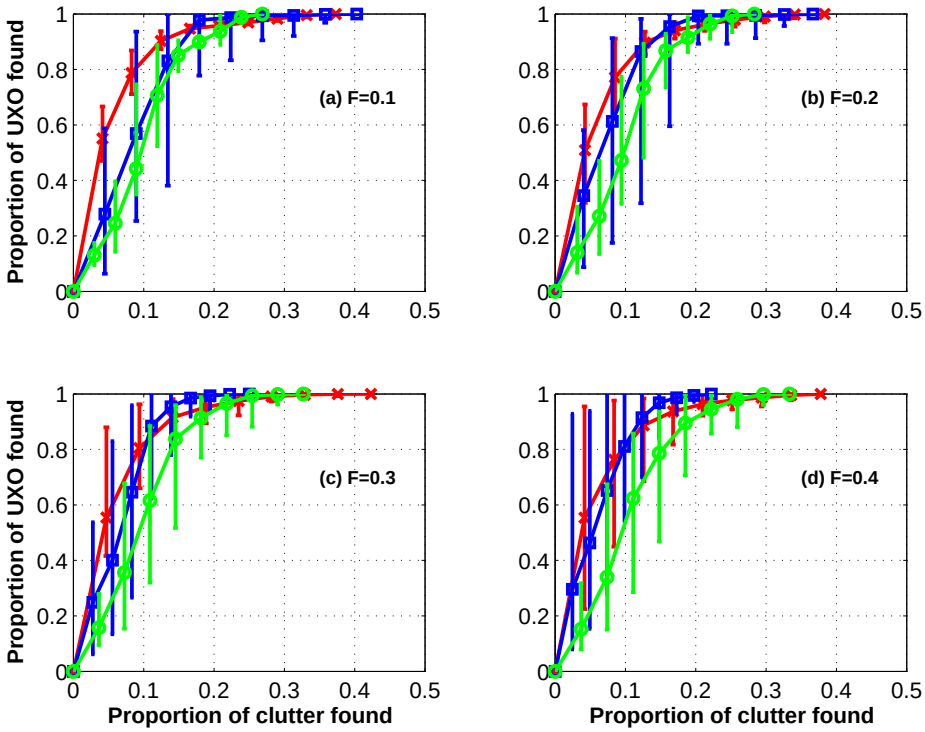
(a) Linear discriminant analysis

(b) Quadratic discriminant analysis

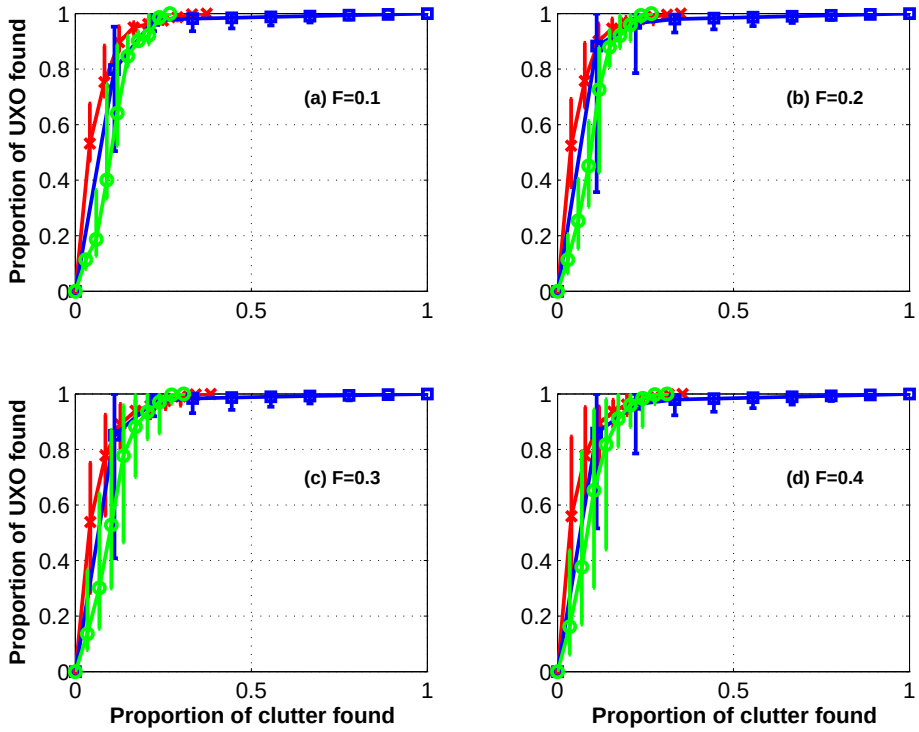
(c) Probabilistic neural network

(d) Support vector machine: solid line shows the decision boundary.

Figure 6.15: Classifier solutions for a single realization of the Badlands bombing range data. Squares are UXO items and triangles are clutter items. Training data are a random sample of 10% of all feature data.

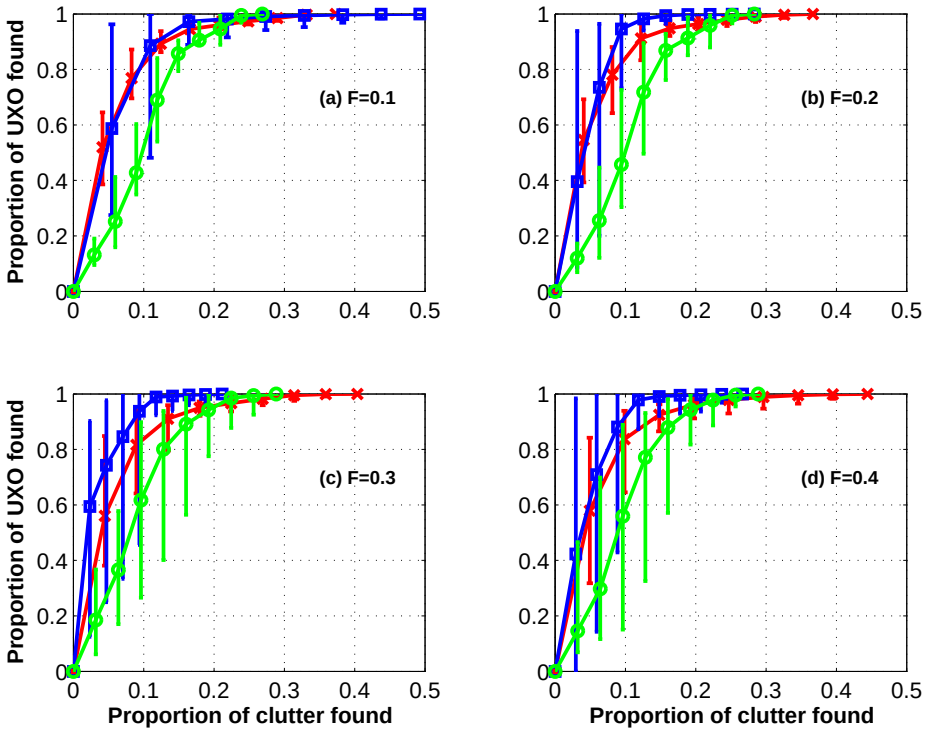


(a) Linear discriminant analysis

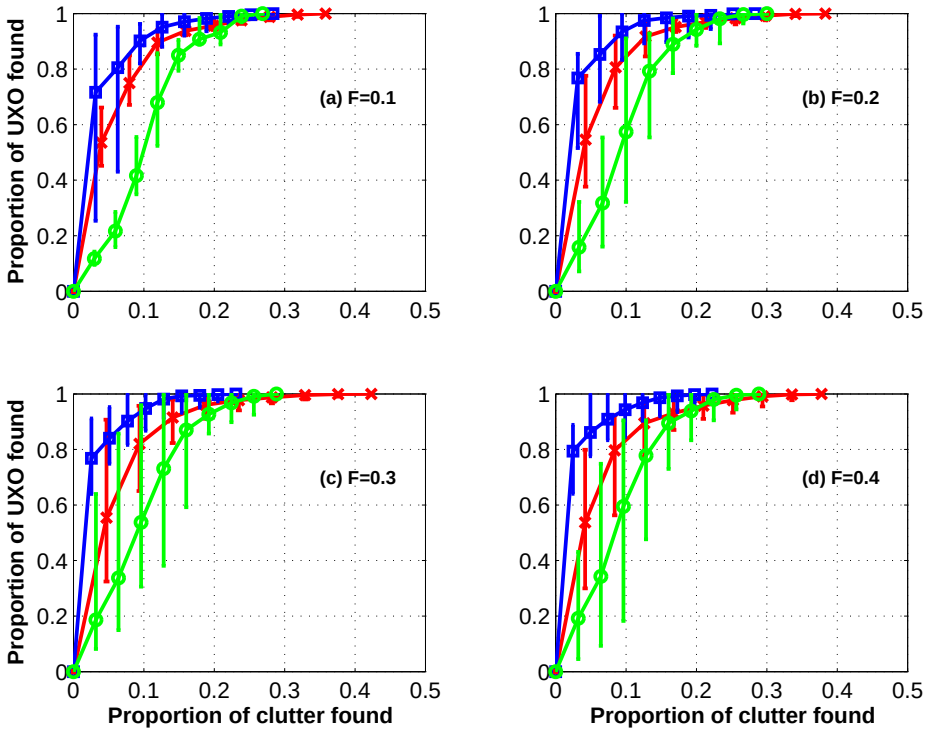


(b) Quadratic discriminant analysis

Figure 6.16: ROC curves of Badlands bombing range magnetics data for discriminant analysis (squares). Also shown are ROC curves generated by thresholding on remanence (crosses) and angle (circles). Subplots (a) through (d) show the average performance of classifiers when the size of the training data set is a fraction F of all feature data. Error bars indicate minimum and maximum performance bounds seen over all realizations for that classifier.

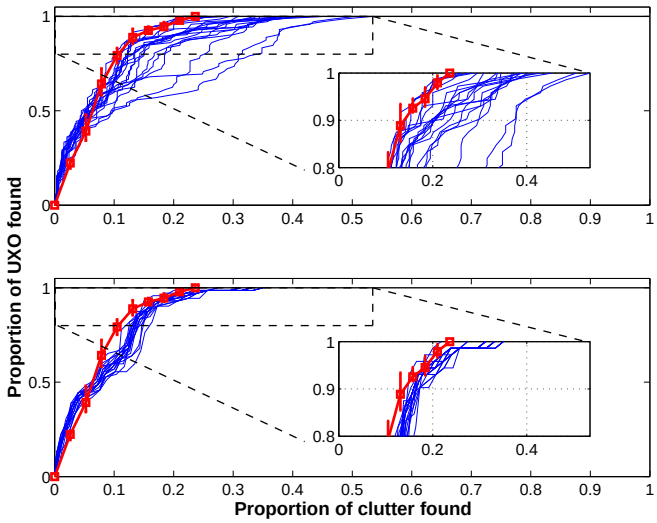


(a) Probabilistic neural network

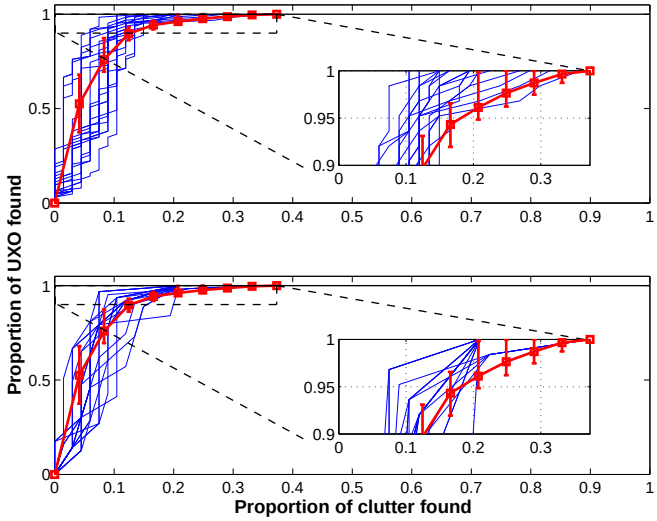


(b) Support vector machine

Figure 6.17: ROC curves of Badlands bombing range magnetics data for nonlinear classifiers with Gaussian kernels (squares). Also shown are ROC curves generated by thresholding on remanence (crosses) and angle (circles). Subplots (a) through (d) show the average performance of classifiers when the size of the training data set is a fraction F of all feature data. Error bars indicate minimum and maximum performance bounds seen over all realizations for that classifier.



(a) Guthrie road



(b) Badlands bombing range

Figure 6.18: Ensemble of ROC curves generated by a PNN. Heavy solid line is the mean ROC curve for remanence with error bars indicating performance bounds seen over all realizations. Top plot is for 20 random realizations of the training data ($F = 0.1$). Bottom plot uses the initial random realizations with retraining in increments of 10 items. Dashed lines indicate range of inset plots.

worst case false alarm rate (figure 6.18).

Figure 6.19 shows the evolution of training and test data sets for a retrained PNN applied to the Guthrie road data. The initial realization is quite poor in this example; a PNN without retraining requires us to dig nearly all clutter items in order to find all UXOs. Retraining quickly corrects the initial decision boundary: after only a few iterations of the algorithm the decision boundary is quite similar to its final form. The retrained PNN requires us to dig 30 % of all clutter items to find all UXOs.

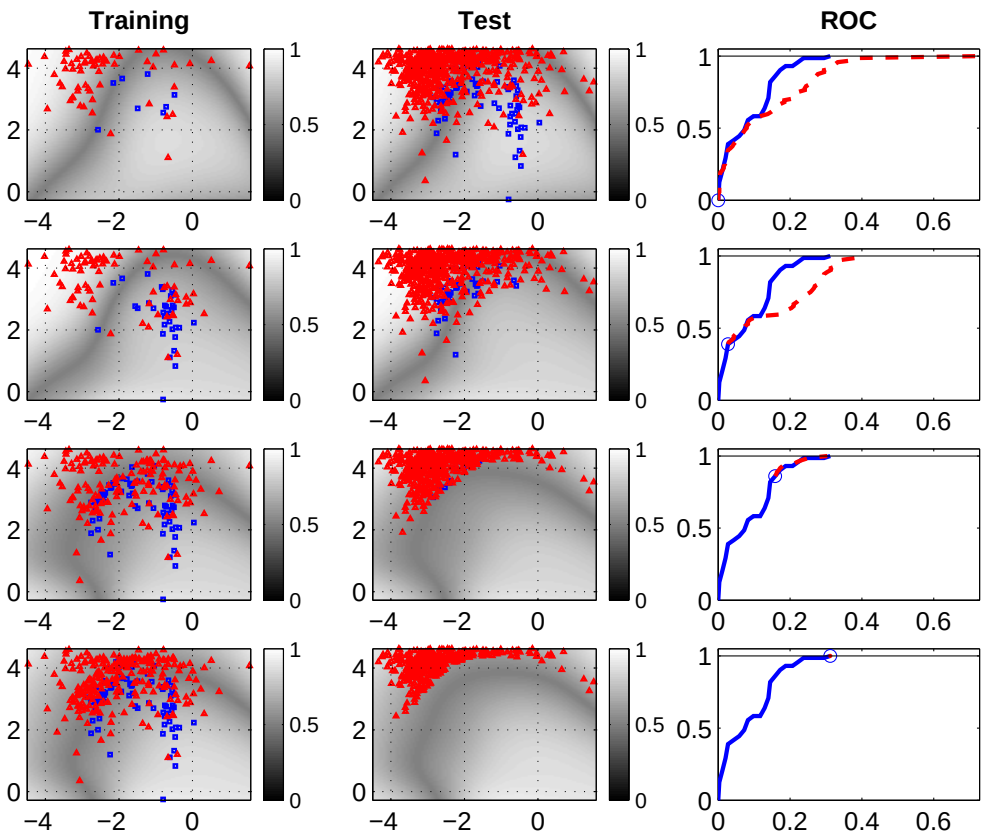


Figure 6.19: Evolution of the training and test data for a retrained PNN. Each row shows the training and test sets and retrained PNN decision boundary. The rightmost column shows ROC curves for the retrained PNN (Solid) and a PNN without retraining for the remainder of the digging process (Dashed). Open circle indicates the point on the ROC curve for the training and test sets in the corresponding row.

The PNN is particularly suited to the retraining process. By modelling the likelihood function as a superposition of Gaussian kernels, it is flexible enough to incorporate new information into the class distributions as items are excavated. Linear discriminant analysis, for example, is unlikely to show much improvement with retraining. As was shown in the previous section, the linear decision boundary will not change very much as the training set grows.

Care must be taken, however, not to overfit the training data at each iteration of retraining. Information acquired during the digging process only comes from items which are likely to be UXOs. Hence our training data becomes increasingly biased at each iteration of retraining. Previous experiments using retraining with a cross-validation error function

degraded classifier performance. This was because the PNN overfit the training data at each iteration.

The interval at which a classifier is retrained is governed by practical field considerations. At one extreme we might imagine retraining the classifier after every item is excavated. This would require some logistical effort coordinating field crews. Updating the dig list after every item would also prevent a field crew from efficiently managing its movements on a site. It is more reasonable to generate a dig list for each crew at the start of the field day and then update the classifier with whatever information is available at the end of the day.

6.4 Selecting a classifier

While statistical classifiers can outperform rule-based classifiers, there is no way to know *a priori* which classifier is optimal for a given data set. In this section, I address this problem with an algorithm for selecting a classifier as the learning process proceeds.

Classifier selection or combination appears in various forms in the statistical literature. "Bagging" is a democratic approach: a test vector is classified based upon the majority vote of an ensemble of classifiers. A single algorithm is used to generate the ensemble, with each classifier trained using a random subset of the training data. "Boosting" applies classifiers sequentially, with each classifier in the sequence placing increased importance on training vectors which were misclassified by the previous classifier. This scheme also uses a single classifier to generate predictions at each stage. Finally, "bumping" selects a single best classifier from a set of classifiers. The best classifier is chosen based upon some optimality criteria estimated from the training data [25].

For UXO classification, I propose an algorithm which is a variant of bumping. The classifier selection algorithm is summarized in pseudocode in figure 6.20.

I start with an initial training data set and an "active" classifier selected from the set of available classifiers C. The selection algorithm needs an initial training data set because statistical classifiers require information about both UXO and clutter classes to make classification decisions. If we were to generate the training data set by only digging items which are likely to be UXO, then we would not have a representative distribution of clutter at the beginning of the algorithm.

At the first iteration, I dig (i.e. label) the batch of N_{dig} items identified by the active classifier as most likely to be UXOs. The numbers of UXO and clutter found in the current batch are then added to running totals $NUXO$ and $NScrap$. The newly-labelled items are appended to the training data.

Next, I evaluate the predicted performance of the classifiers in the ensemble C using a version of cross-validation known as "k-fold" cross-validation. A criticism of leave-one-out cross-validation is that the training data sets at each stage of the process are not significantly different and so the resulting decision boundary may fit the training data too closely. This can be addressed by leaving out a proportion of the training data for cross-validation. In k-fold cross-validation, a random hold-out subset is used for validation. The method provides an asymptotic estimate of a classifier's actual performance on the test data and has been successfully used in model selection problems [25].

To implement k-fold cross-validation, I randomly divide the training data set into temporary training and test sets ($Training\_temp$ and $Test\_temp$ in figure 6.20). The size of the temporary training set is a specified proportion F of the total training data set, so that

the temporary test set is $(1 - F)$ of the training data.

All statistical classifiers are trained using the temporary training set. The expected performance of all classifiers (rule-based and statistical) is then evaluated with the temporary test set. I compute the estimated false alarm rate FAR_{est} for all classifiers for the current realization of temporary test data. Here the false alarm rate is the maximum of the ROC curve, i.e. it is the proportion of clutter items which must be dug in order to find all UXOs.

The estimated false alarm rate for each classifier will depend upon the particular realization of the temporary training and test sets. Consequently, the false alarm rates for all classifiers are estimated using a large number ($N_{realizations}$) of realizations. However, it is important that all classifiers are evaluated on the same set of realizations of the temporary training and test sets. If different realizations were used for each classifier, then it is possible that one classifier might see lucky realizations of the training data. This could produce overly optimistic estimates of that classifier's performance. For a fair comparison, all classifiers must be evaluated using the same temporary training and test sets.

Once I have computed FAR_{est} using cross-validation, I compare the mean false alarm rates of all classifiers. I set the active classifier as the classifier with the minimum mean FAR_{est}. However, if the actual performance of the active classifier is optimal (i.e. no clutter items were found in the current batch of labelled items), then there is no reason to change classifiers. I only switch classifiers if the active classifier has uncovered clutter items and cross-validation has identified another classifier which is expected to do better.

When the active classifier is chosen, the process begins anew and the next batch of test items is labelled with the active classifier. Implicit in this algorithm is retraining of statistical classifiers as each batch of targets is labelled. The algorithm proceeds until some stopping criterion is achieved (e.g: no UXOs are found in the current batch of labelled items).

Line	Pseudocode	Line	Comments
1	Require	1	
2	*Training*	2	initial training set
3	*Test*	3	initial test set
4	*C={T1, T2, T3, … TL}*	4	set of *L* classifiers
5	*Ndig*	5	number of items to dig at each iteration
6	*Tactive* ε *C*	6	active classifier from the set *C*
7	*F*	7	size of random samples of training set as a proportion of set size
8	*Nrealizations*	8	number of random samples of training set
9	*iteration=1*	9	iteration counter
10	Repeat	10	
11	P(*Test*)=Predict(*Training,Test,Tactive*)	11	Train the active classifier and generate predictions *P* of membership in the UXO class
12	*wdig*=Dig(*P(Test), Ndig*)	12	Generate set of labels *wdig* for the *Ndig* items identified by *Tactive* as most likely to be UXOs
13	*Training*=Update(*Training,wdig*)	13	Append newly labelled feature vectors to training set
14	*NUXO*(*iteration*)=length(*wdig* ε *wUXO*)	14	number of UXO found
15	*NScrap*(*iteration*)=length(*wdig* ε *wScrap*)	15	number of Scrap found
16	For(*i*=1 to *Nrealizations*)	16	loop on realizations
17	For(*j*=1 to *L*)	17	loop on classifiers
18	*Training_temp*=RandTrain(*Training, F*)	18	Generate random training set which is a proportion *F* of the existing training set
19	*Test_temp=Training U Training_temp*	19	Generate a random test set with the remaining items in Training
20	*P(Test_temp)*=Predict(*Training_temp,Test_temp,Tj*)	20	Train classifier and generate predictions *P(Test_temp)*
21	*threshold* = min(*P(Test_temp(w e wUXO))*)	21	Set a threshold as the minimum prediction for items belonging to the UXO class
22	*FARtemp*(*i*)=length(*P(Test_temp(w e wScrap))>=threshold*)	22	Estimated false alarm rate for this realization is the number of Scrap items which would be dug in order to find all UXOs
23	EndFor	23	End loop on classifiers
24	*FARest*(j)=mean(*FARtemp*)	24	Estimated false alarm rate for classifier is the mean FAR over all realizations
25	EndFor	25	End loop on realizations
26	If((*FARest(Tactive)>FARest(Tj)*) and (*sum(NScrap)>0)*)	26	Change active classifiers if the
27	*Tactive=Tj*	27	active classifier does not have the minimum
28	EndIf	28	*FARest* and scrap items have been found.
29	*iteration=iteration+1*	29	
30	Until(StopDig(*NUXO, Nscrap*))	30	Stop digging when a convergence criterion StopDig is achieved

Figure 6.20: An algorithm for selecting classifiers during learning

6.4.1 Synthetic example

As a first example, I simulate the classifier selection algorithm for the Yuma EM feature space. I assume that the generative model used in linear discriminant analysis is the true model (i.e. I use a pooled covariance matrix for all classes). I then randomly generate 200 test items from normal distributions with the mean and covariance of each class. The relative frequencies of each ordnance type in the synthetic test data are equal to the frequencies in the training data. Figure 6.21 shows the observed training data and synthetic test data for this simulation.

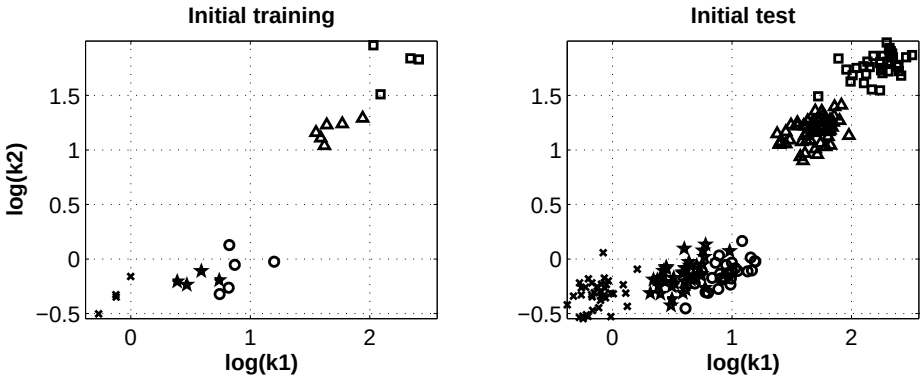


Figure 6.21: Yuma EM training set and synthetic test set for simulation of classifier selection algorithm.

Figure 6.22 shows the performance of the classifier selection algorithm on the synthetic test data. The ensemble of classification algorithms includes linear discriminant analysis, quadratic discriminant analysis and a PNN. The active classifier is selected after batches of $N_{dig} = 20$ items are labelled. Cross-validation is carried out with $F = 0.9$. In this example, linear discriminant analysis represents the "true" model of the decision boundary, and so its performance should, on average, be optimal for these test data. Figure 6.22 shows that LDA does provide a minimal misclassification rate for these data. Training QDA using only the initial realization of training data produces very poor performance; the initial training data is too small to provide accurate estimates of the class covariances.

To test the selection algorithm, I set the initial active classifier to be quadratic discriminant analysis. Although this classifier does quite poorly using only the initial training data, QDA does reasonably well with retraining for the first couple of iterations of the selection algorithm. Once a misclassification occurs, the algorithm selects linear discriminant analysis as the active classifier for the remainder of the digging process. The resulting ROC curve is comparable to that generated by the optimal LDA classifier. By evaluating the performance of classifiers as digging proceeds, the algorithm was able to detect that QDA was an inappropriate classifier for these data.

As the algorithm proceeds, the choice of the active classifier becomes less clear. At early iterations, LDA has the minimal false alarm rate by a comfortable margin. However, for the last three iterations of the algorithm there is no classifier which is clearly optimal (rightmost plot in figure 6.22). This effect can be understood by considering histograms of the estimated false alarm rates at different iterations (figure 6.23). Because the training data set is relatively small at the second iteration, there is enough variability in the realizations of the temporary training data to resolve differences in the expected performance of classifiers.

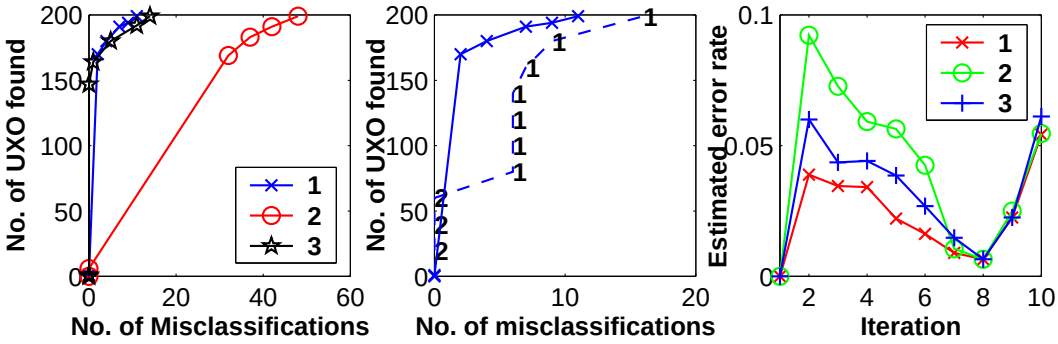


Figure 6.22: Left: performance of statistical classifiers on synthetic Yuma test data. Classifiers are: (1) Linear DA, (2) Quadratic DA, (3) PNN. Middle: performance of classifier selection algorithm (dashed line) on Yuma test data. Numbers indicate the active classifier for that point on the ROC. Solid line with crosses shows the performance of the optimal LDA classifier. Right: estimated false alarm rate for classifiers as a function of selection algorithm iteration.

By the tenth iteration, the realizations of training data are quite similar, and so cross-validation yields very similar distributions of FAR_{est}.

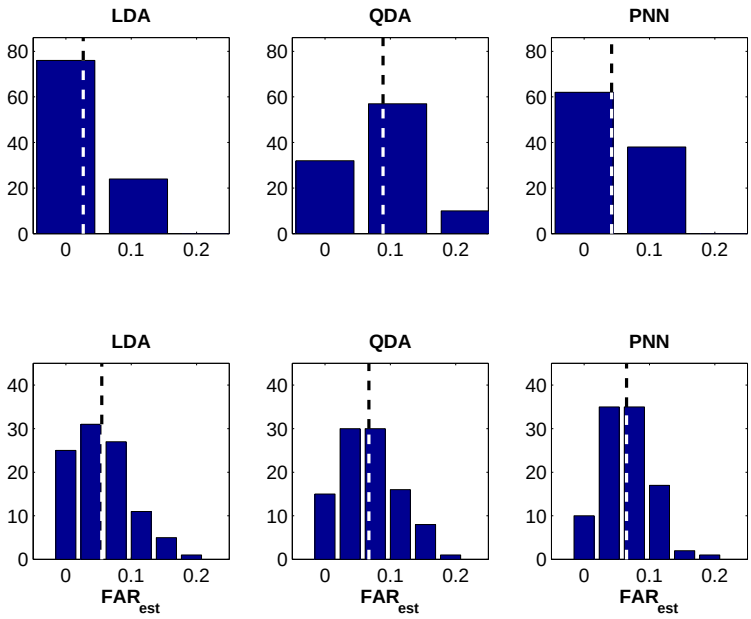


Figure 6.23: Estimated false alarm rates (FAR_{est}) for classifiers at two iterations of the selection algorithm applied to synthetic Yuma test data. Vertical dashed lines indicate the mean FAR_{est} for each classifier. Histograms are generated by 100 realizations of temporary training and test data. Top row is for the second iteration of the algorithm, bottom row is for the tenth iteration.

One way to address this problem is to adjust the proportion F as the training data set grows. To obtain more variability in the realizations of the temporary training data, I adopt a cooling schedule for F. When the training data is sufficiently large (e.g. more than

100 feature vectors), I decrease F at each iteration by a factor $\alpha < 1$

$$F \leftarrow \alpha F. \tag{6.4}$$

I delay the cooling schedule for the first few iterations because the training data set is so small that the temporary training data must initially be a sizeable (i.e. $F = 0.9$) proportion. Figure 6.24 shows the resulting performance of the algorithm with $\alpha = 0.75$. The performance is the same as before, but there are larger differences between the estimated false alarm rates at later iterations. Decreasing F with iteration strongly affects the distributions of FAR_{est} (figure 6.25). There is more variability in the estimated performance of the classifiers and so it is easier to identify the optimal classifier at each iteration.

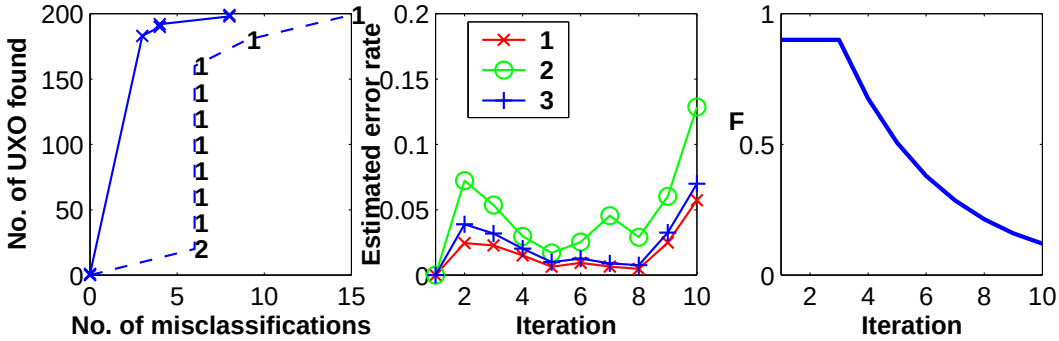


Figure 6.24: Classifier selection algorithm with variable F. Left: performance of classifier selection algorithm (dashed line) on Yuma test data. Numbers indicate the active classifier for that point on the ROC. Classifiers are: (1) Linear DA, (2) Quadratic DA, (3) PNN. Solid line with crosses shows the performance of the optimal LDA classifier. Middle: estimated false alarm rate for classifiers as a function of selection algorithm iteration. Right: F as a function of iteration.

6.4.2 Application to magnetics data

For magnetics classification, the initial training data set is generated as a small random subset of the feature data. I choose the set of classifiers to be remanence, a PNN and an SVM. The statistical classifiers are trained in a feature space spanned by log(Moment) and log(Angle). I select remanence as the initial active classifier.

Figure 6.26 shows the performance of the algorithm for a single realization of the initial training data at Badlands bombing range. Although the selection algorithm depends upon the initial realization of training data, figure 6.26 shows its typical performance. The algorithm keeps remanence as the active classifier for the first few iterations. At this early stage of the digging process, no false alarms have occurred and so there is no reason to change classifiers. By the third iteration, the statistical classifiers have gained enough information about the UXO class that cross-validation predicts that they will outperform remanence. The selection algorithm chooses nonlinear statistical classifiers for the remainder of the digging process, resulting in a smaller false alarm rate than would be obtained by thresholding on remanence alone. The resulting performance is comparable to that obtained with retrained PNNs for this data set.

The estimated false alarm rate increases for all classifiers as the algorithm proceeds. This is because in the early stages of digging we are labelling "easy" targets which all

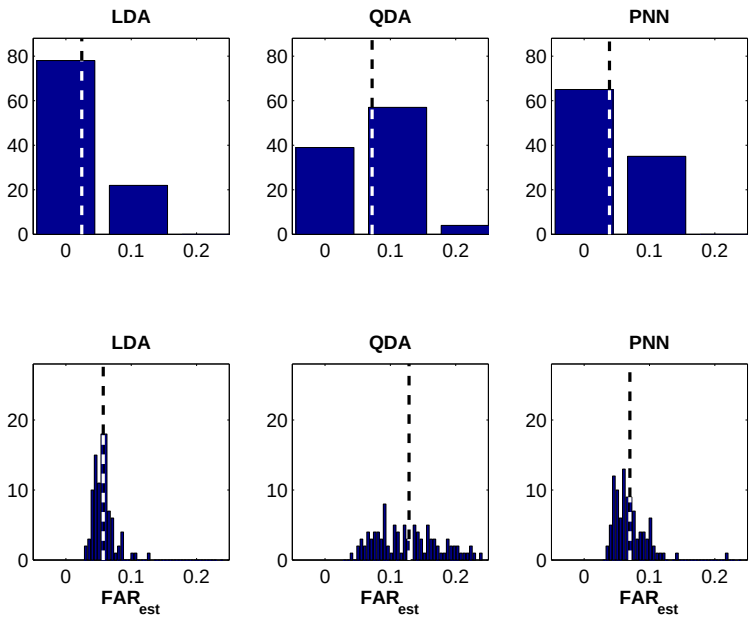


Figure 6.25: Estimated false alarm rates (FAR_{est}) for classifiers at two iterations of the selection algorithm with variable F. Vertical dashed lines indicate the mean FAR_{est} for each classifier. Histograms are generated by 100 realizations of temporary training and test data. Top row is for the second iteration of the algorithm, bottom row is for the tenth iteration.

classifiers can identify as UXO. As digging proceeds, we encounter more clutter and the classification task becomes more difficult. Consequently, the estimated false alarm rate increases for all classifiers.

The estimated false alarm rate is quite optimistic: in figure 6.26 the estimated FAR indicates that we will dig no more than 8% of all clutter items. However, the remanence classifier actually requires us to dig 25% of all clutter items. This discrepancy arises from the fact that the actual false alarm rate is governed by a small number of UXO items which are difficult to find and are discovered in the final few iterations of digging. The estimated FAR provides a measure of the relative merits of classifiers using the available training data, and does not necessarily represent the actual performance. While this is an obvious weakness of the selection algorithm, I have found that the algorithm provides near optimal performance for the data sets and classifiers considered.

The performance of the selection algorithm is quite sensitive to the interval N_{dig} at which statistical classifiers are retrained and the active classifier is selected. Figure 6.27 compares the performance of the algorithm for the same initial realization of training data but with intervals $N_{dig} = 10$ and $N_{dig} = 20$. When the algorithm selects the active classifier at the larger interval $N_{dig} = 20$, the training data becomes increasingly biased in favor of remanence. By the time the algorithm switches from remanence to statistical classifiers at the third iteration, there is no difference in the dig lists provided by the ensemble of classifiers. The resulting ROC curve for the selection algorithm is identical to that produced by simply thresholding on remanence (bottom row of figure 6.27).

Figure 6.28 shows the application of the selection algorithm to the Guthrie road data, with a retraining interval of $N_{dig} = 10$. The algorithm chooses remanence as the best avail-

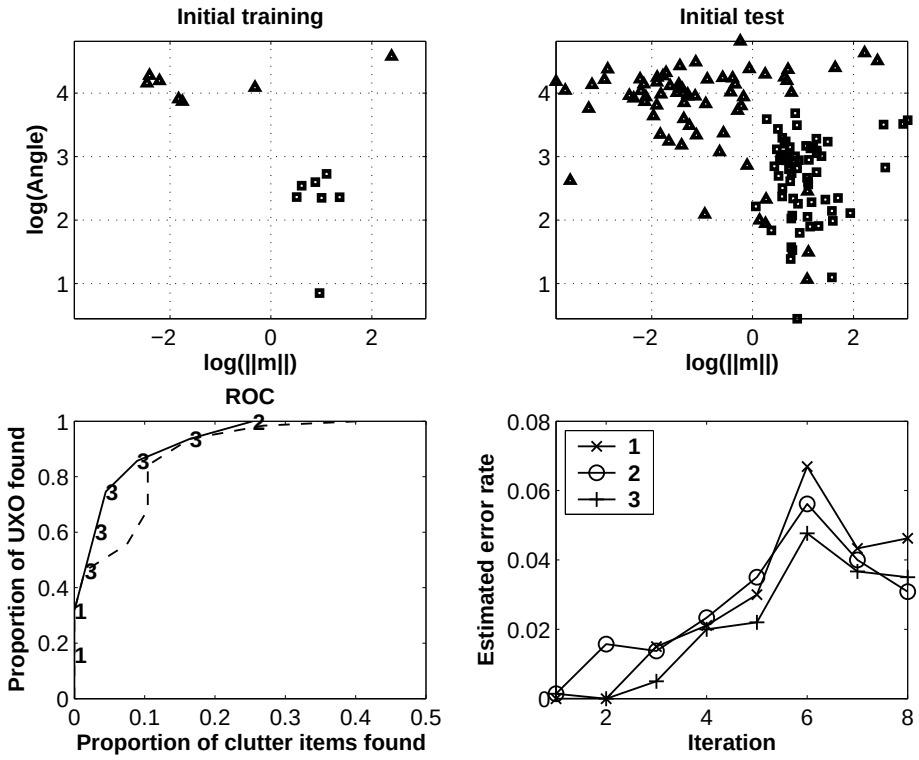


Figure 6.26: Performance of the classifier selection algorithm for Badlands bombing range. Top row shows the initial realization of test and training data. Squares are UXO and triangles are clutter. Bottom left: ROC curves generated by thresholding on remanence (dashed), and by the classifier selection algorithm (solid). Numbers indicate the active classifier for the point on the ROC curve: (1) Remanence, (2) PNN, (3) SVM. Bottom right: estimated false alarm rates for classifiers at each iteration of the classifier selection algorithm.

able classifier throughout the digging process. This is consistent with the experiments in the previous sections: remanence was the optimal classifier amongst all available classifiers for the Guthrie road data.

In this section, I have proposed an algorithm for selecting a classifier during the digging process. My approach is to evaluate the performance of classifiers using the available information. While there is no guarantee that this algorithm will provide optimal performance for a given data set, the continual evaluation of classifiers can at least guard against a poor choice of classifier. Some care must be taken with the setting parameters (F, α,N_{dig}, $N_{realizations}$) and further experiments may suggest more sophisticated means of tuning parameters. The algorithm is easily expanded to include a wide variety of classifiers and is not computationally intensive. The simulations presented here each required less than a minute to execute on a standard Pentium 4 laptop. Ultimately, the utility of this algorithm at field sites will depend upon close coordination between field crews and an operations centre which generates updated dig lists.

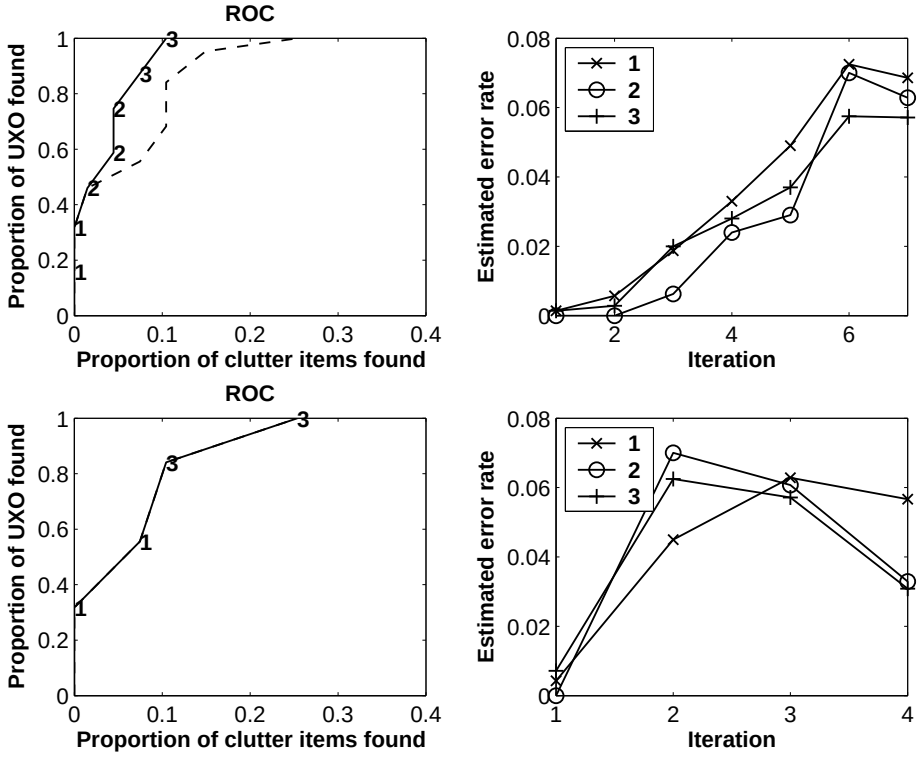


Figure 6.27: Dependence of the classifier selection algorithm upon retraining interval. Top row shows the ROC curves and estimated classifier error rates (as in figure 6.26) for the selection algorithm with $N_{dig} = 10$. Bottom row is for $N_{dig} = 20$, ROC curves for remanence and the selection algorithm are identical in this example.

6.5 Classification of electromagnetic data

Electromagnetic data is better suited to statistical classification techniques. The sensitivity of this data type to target shape and size means that multiple features can be used to define the feature space. However, the large number of features which have been developed for this data type means that we must carefully choose our features to avoid the curse of dimensionality.

Previous studies with statistical classifiers and electromagnetic data have yielded promising results. Pasion proposed a simple rule-based algorithm based on thresholding ratios of decay parameters from his model [10].

Norton and Won proposed a "fingerprinting" approach to discrimination using frequency-domain EM data [36]. They extracted the elements of the polarization tensor from the observed data and then computed a metric which compares the estimated values of the polarization tensor with a library of polarizations for ordnance items. They showed using laboratory measurements that this metric can be used to discriminate between different ordnance types.

A weakness in this approach is that their metric assumes that two estimated polarizations which are the same "distance" from a library item are equally likely to be that item. However, as was shown in chapter 4, there is a strongly nonlinear relationship between target shape and parametric model parameters. Therefore a Euclidean metric may not have

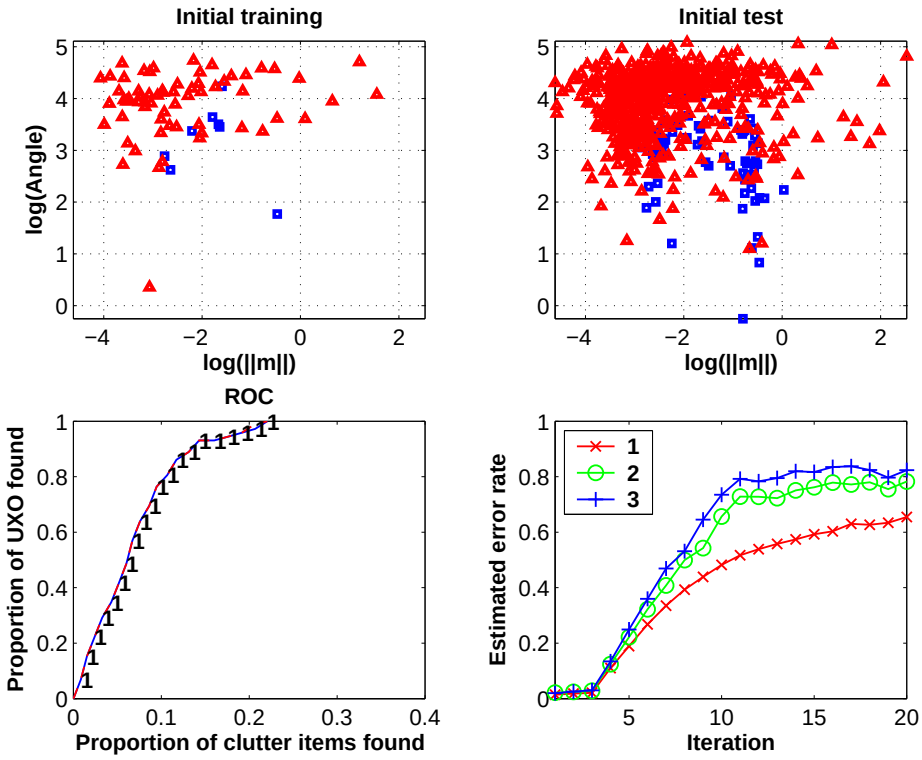


Figure 6.28: Performance of the classifier selection algorithm for Guthrie road. Top row shows the initial realization of test and training data. Squares are UXO and triangles are clutter. Bottom left: ROC curves generated by thresholding on remanence (dashed), and by the classifier selection algorithm (solid). Numbers indicate the active classifier for the point on the ROC curve: (1) Remanence, (2) PNN, (3) SVM. Bottom right: estimated false alarm rates for classifiers at each iteration of the classifier selection algorithm.

the same meaning for two points which are equidistant from a library item. Barrow and Nelson accounted for this nonlinearity by showing experimentally that the distributions of polarization tensor elements for UXO items were log-normally distributed [28]. They devised a rule-based classifier by estimating the mean and covariance of log-transformed model parameters for 81 mm mortars in a calibration grid. To generate a dig list at a field site, they then computed the number of standard deviations between an observed test vector and the class mean. An advantage of this method is that it requires no prior information regarding clutter.

Statistical classifiers have also been applied to discrimination with EM data. Zhang et al. compared the performance of a support vector machine and Gaussian likelihood ratio [1] for discrimination using both magnetic and frequency-domain EM data [37]. EM features included the estimated peak frequency of the axial excitation and the ratio of peak frequencies discussed in chapter 4. They found that the best performance was obtained using the SVM applied to EM features estimated by cooperative inversion. However, given the small number of UXO items in the test data (16), a KS test of the resulting ROC curves would likely not reject the null hypothesis that these classifiers are significantly different. A feature space which combined features from EM and magnetics data actually degraded

[1] equivalent to quadratic discriminant analysis

the performance of both classifiers, suggesting that careful assessment of the feature space, perhaps using canonical analysis, may be necessary before attempting joint classification.

A study by Collins et al. using a larger suite of classifiers produced similar results: while statistical classifiers typically outperform simple thresholding algorithms, there is no classifier whose performance is optimal in all cases [38]. These results highlight the need for an algorithm to choose a classifier as digging proceeds, as discussed in the previous section.

Unfortunately, no EM data suitable for classification was available for this thesis. However, future work will use EM features for statistical classification. In the next section, I examine the effect of noise upon classification using the Yuma training data set.

6.6 Uncertainty in classification

As discussed in chapter 3, model parameters estimated with inversion have an associated uncertainty. This uncertainty can be estimated using a linearized or nonlinear uncertainty analysis. Uncertainties estimated with nonlinear uncertainty analysis are typically larger than those obtained with a linearized uncertainty analysis. Uncertainties in the feature vectors are always neglected in the classification process, primarily because most classification algorithms have been developed for applications where there is negligible uncertainty in the feature vectors.

Propagating uncertainties through a classifier is a difficult task, since we must deal with both with errors in the parameters of the classifier and errors in the test feature vectors. The former must be estimated by propagating errors from the training data to the parameters of the classifier. For example, if we are to propagate errors through discriminant analysis, then we must estimate the uncertainties in the class means and covariances.

This problem has been tackled in the context of regression using variational methods [39]. In this work the independent variable (input) with associated uncertainty $\mathbf{x} + d\mathbf{x}$ is propagated through a neural network to the dependent variable (output) $\mathbf{y}$ and its associated uncertainty $d\mathbf{y}$. This analysis identified three sources of uncertainty which contribute to $d\mathbf{y}$: the uncertainty in the parameters of the neural network (which are themselves estimated from uncertain training data), the uncertainty in the inputs ($d\mathbf{x}$), and a term involving the interaction of the first two sources of uncertainty.

In classification we might expect the same sources of error. Analytic propagation of errors is beyond the scope of this thesis, and so I propagate uncertainties through the Yuma training data with a Monte Carlo simulation. The aim of this simulation is to gain a first-order understanding of how errors could propagate and how the final output could be used.

As in the previous chapter, I assume that the model generated by linear discriminant analysis is the "true" model for the Yuma test data. However, I account for uncertainty with multiple realizations of the training data. In each realization, I generate each training datum as a random sample from a Gaussian distribution with the mean and variance estimated by linearized uncertainty analysis. I then compute the class means and pooled covariance to generate a model of the decision boundary for this realization. At each point in the feature space I can then compute maximum posterior probability. Finally, I compute the mean and variance of the posterior probability over all realizations of the training data.

This simulation was carried out with 1000 random realizations of the Yuma feature data in a two-dimensional feature space spanned by $k1$ and $k2$ (figure 6.29). The effect of uncertainty is to blur the decision boundaries (figure 6.30). In regions of the feature

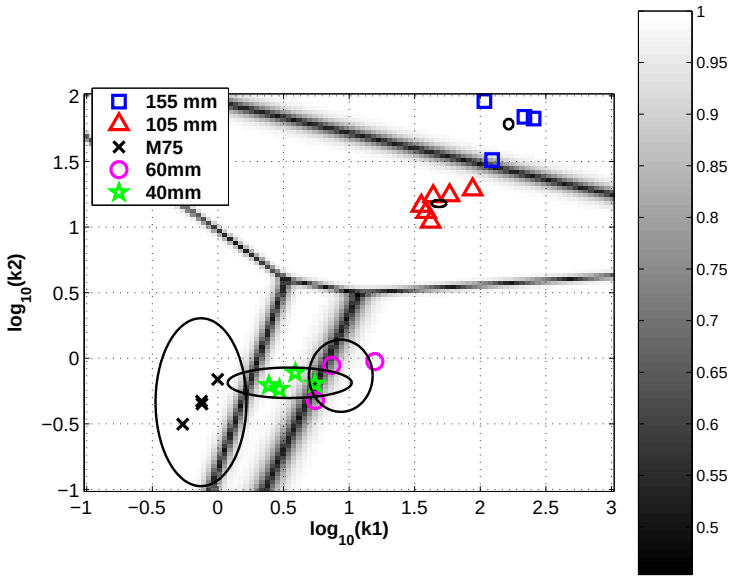


Figure 6.29: Yuma feature data with decision boundaries defined by linear discriminant analysis. Heavy solid lines indicate the uncertainty in the class mean for each class.

space where the training data have a small relative uncertainty (ie for large items), there is a small uncertainty in the posterior probability. For small items, however, the relative uncertainty in the training data produces an uncertainty of up to ± 0.2 in the posterior probability. As was shown in the analysis of magnetics data, linear discriminant analysis

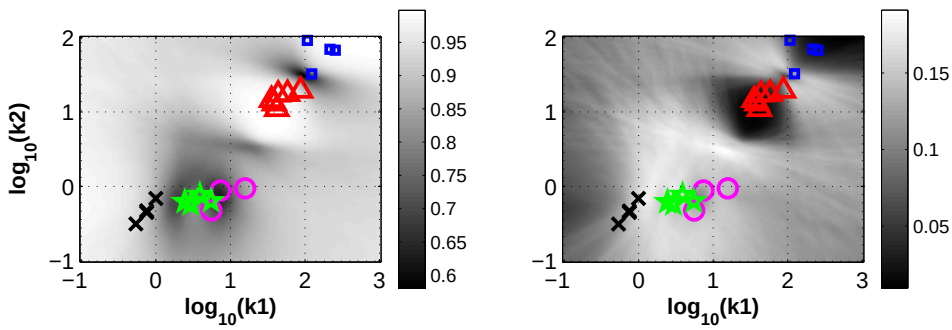


Figure 6.30: Left: Mean posterior probability from Monte Carlo simulation of discriminant analysis with uncertainty in the feature vectors. Right: Standard deviation of the posterior probability, estimated by Monte Carlo simulation.

is relatively insensitive to the realization of the training data. Classifiers which are more sensitive to the particular realization of the training data (eg QDA or PNNs) will therefore produce more uncertainty in the posterior probability.

Estimates of the uncertainty in the posterior probability can be used to revise the dig list. Again, rigorous development is postponed for future research. However, even first order ideas can be insightful. For instance, by subtracting the uncertainty from the posterior we obtain a worst case estimate of the posterior (figure 6.31). This corrected posterior can then be used to generate a prioritized dig list. For EM data, the effect will be to prioritize large items ahead of small items in the dig list.

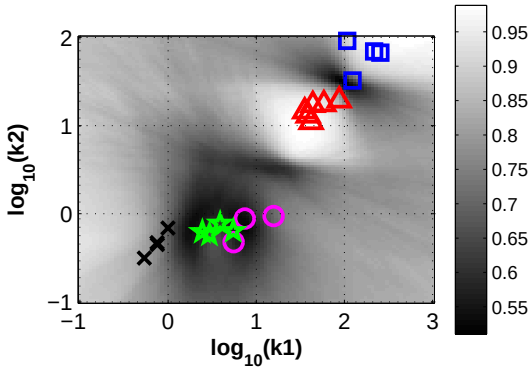


Figure 6.31: Mean posterior minus the standard deviation of the posterior, estimated from Monte Carlo simulation.

6.7 Discussion and conclusions

Useful features are site-specific: a parameter which separates classes at one site may not work at another. Canonical analysis is one way to identify the relevant features at a site. At Guthrie road and Limestone hills, canonical analysis of raw features showed that remanence provided the largest separation between UXO and clutter. For the Badlands bombing range data set, remanence was ranked behind moment and angle.

The results of canonical analyses were supported by applying a suite of classifiers to the feature space spanned by moment and angle. In no case did statistical classifiers outperform remanence at Guthrie road and Limestone Hills. At Badlands bombing range, the best classifier provided a modest reduction in false alarm rate in comparison to remanence.

Classifiers depend upon the input features and so they are also site specific. For example, quadratic discriminant analysis did quite well at Guthrie road but performed poorly on the Badlands data. Though it was outperformed in some cases, the PNN consistently provided reliable classification of the data sets considered in this chapter. I found that a simple heuristic for estimating the kernel smoothing provided better performance than cross-validation techniques. A similar heuristic for estimating kernel smoothing for SVMs produced comparable performance to that of the PNN. This classifier performed best on data which had a relatively large separation between UXO and clutter classes.

In practice, we cannot know the point on the ROC curve where we have found all UXO items. Previous authors have suggested setting a threshold on classifier output and digging everything up to this threshold [35]. The threshold may be estimated from the training data or using prior information from other sites. A more sensible way to choose a threshold might be to monitor the false alarm rate as targets are excavated. If a batch of excavated targets contains no UXOs, then this is probably an indication that we are nearing the maximum of the ROC curve. It is crucial however, that a classifier does not generate an ROC curve which levels off near its maximum. This corresponds to the worst case scenario where we must dig many clutter items to find the last few UXOs. If the digging process is terminated based upon the false alarm rate, then these last few UXO items will never be found. Ultimately, the difficulty of the UXO classification problem comes down to finding those UXOs which "look" like clutter. This is a function of ordnance type: 250 lb bombs are easy to classify correctly but smaller items (e.g. cluster bombs) are often hard to find within the clutter "cloud" in the feature space. Retraining can alleviate this problem to

some extent: by adapting to the changing feature space a classifier is better able to identify UXOs which were not seen in the initial training data. Retraining thereby lessens our dependence on the initial training set.

In this chapter, I have assumed that the training data are generated by random sampling. While this is a weakness of the analysis, random sampling will probably always be a component of UXO remediation, both for initial site assessment and follow-up quality control. For example, during wide-area assessment and vegetation clearance, both clutter and UXO will probably be encountered on the surface. These items could be used as the initial training data set for statistical classifiers once digging begins.

Experiments with a classifier selection algorithm show that it can be a viable way of selecting classifiers for a given data set. The algorithm aims to combine the advantages of all classifiers (rule-based and statistical) to minimize the false alarm rate. In future work I will apply this algorithm to discrimination with EM data.

Initial simulations of uncertainty propagation show that it will be an important consideration for classification of electromagnetic data. Future work will focus upon finding analytic methods of propagating uncertainty through classification.

Chapter 7

Conclusions and future work

This thesis examined the process of discriminating between unexploded ordnance and clutter items. I have discussed issues arising at several stages of this process, beginning with forward modelling (chapters 2 and 4), inversion (chapter 3) and discrimination (chapters 5 and 6).

Chapter 2 presented forward models for magnetic and electromagnetic data. I discussed in detail a numerical forward model, the method of auxiliary sources (MAS), which can be used to model both frequency and time-domain EM data, as well as magnetic data. I implemented MAS for the frequency-domain response of a sphere and examined the distributions of auxiliary charges and the secondary fields produced by these charges. I showed that an equivalent dipole response can be computed from the MAS solution by summing the dipole moments produced by auxiliary source charge bands. This simplification may provide a way to speed-up MAS forward modelling for inversion.

In chapter 3, I implemented the Levenberg-Marquardt algorithm to recover estimates of model parameters from observed data. I applied this iterative algorithm to a synthetic problem using the analytic forward model for a sphere in the frequency-domain. This simple "toy" problem highlighted some of the complications which arise for EM models such as MAS which are parameterized in terms of physical properties. In particular, the strong positive correlation between conductivity and permeability at higher frequencies prevents unique estimation of these parameters. I also showed that a linearized uncertainty analysis tends to underestimate the uncertainty in these parameters. This error is directly related to the correlation between these parameters.

Future inversions using MAS as a forward model will face a similar correlation between conductivity and permeability. One option is to reparameterize in terms of the ratio. Alternatively, we might infer the magnetic properties of the target using other information (eg magnetic data or β parameters of the Pasion-Oldenburg model) and then fix the permeability at a reasonable value.

Initial attempts at inversion with MAS were prohibitively slow for application to real data, and so in chapter 4 I used MAS as a benchmark forward model for verifying the dependence of parametric model parameters upon target shape. I showed that for the Pasion-Oldenburg and power law models the ratio of $k1/k2$ is a diagnostic of target shape. This confirms the previous results Pasion obtained with laboratory measurements [10]. In addition, I showed that the spatial response of two-dipole models becomes more accurate as the distance between target and observation location increases (and the contribution of higher order moments decays).

Chapter 5 outlined statistical classification algorithms for discrimination between UXO and clutter. I derived the canonical analysis algorithm for identifying linear combinations of features which provide the maximum separation between classes. This algorithm can be used to prioritize features for classification and reduce the dimensionality of the feature space. I then described classification algorithms for partitioning the feature space using the available training data. Generative classifiers model the underlying class distribu-

tions, either parametrically (e.g. discriminant analysis) or nonparametrically (e.g. PNN). Discriminative classifiers model the decision boundary directly by maximizing the margin between classes (e.g. SVM).

In chapter 6, I applied statistical classification algorithms to features extracted from three magnetic data sets. Canonical analysis was used to prioritize features, ranking remanence ahead of estimated moment and angle in two of three cases. These results were confirmed by application of statistical classifiers to a two-dimensional space spanned by moment and angle. Statistical classifiers were outperformed by remanence at Guthrie road and Limestone hills, and provided a modest reduction in false alarm rate at Badlands bombing range. I demonstrated how retraining a classifier can improve its performance by guarding against poor initial realizations of the training data.

An important result of this thesis (and of several decades of pattern recognition research) is that no single algorithm performs best on all data sets. To address this problem, I proposed a classifier selection algorithm for choosing a classifier during the digging process. At each iteration, the algorithm uses k-fold cross validation to compare the expected performance of all classifiers. The classifier with the best estimated performance is then chosen as the "active" classifier for the next iteration of digging. I found that cross-validation worked best when the size of the temporary training set (specified as a proportion F) was decreased as the full training set grew.

Experiments with the classifier selection algorithm on synthetic and real data sets provided near-optimal performance. Here optimality is defined by the best performance seen for all classifiers on a given data set. There is no theoretical guarantee that the algorithm will always produce an optimal false alarm rate. However, with continual assessment of classifier performance we can, at the least, guard against a particularly unsuitable choice of classifier. In addition, the algorithm provides an easy way to combine classifiers at all levels of sophistication, from simple thresholding to statistical classifiers.

Finally, I investigated the effects of feature uncertainty on classifier output with a Monte Carlo simulation. I found that propagating uncertainty through a classifier may reprioritize the order in which items are dug.

Future work stemming from this thesis will focus on two areas: inversion of magnetic and EM data using MAS and reliable application of classification algorithms to UXO problems. MAS can model both EM and magnetics data with a single parameterization and so it is a leading candidate for joint inversion of these data types.

A priority of future work will be the application of statistical classification to EM data. Joint classification will also be investigated: the combination of remanence with EM features may provide a powerful means of discriminating between UXO and clutter at some sites.

The reliance of statistical classification algorithms upon a random training data set is an obvious weakness in this thesis. Recent work by Zhang et al. has focussed upon generating the initial training data set from the unlabelled feature space [40]. This is an unsupervised learning problem: we must identify the classes without any information about UXO and clutter in the feature space. Initial results with this approach have significantly improved over random sampling. In future, other unsupervised methods of identifying clusters of feature vectors using only unlabelled feature data will be investigated.

Finally, analytic methods for propagating uncertainty will provide rigorous means to account for uncertainties at the final classification stage.

Bibliography

[1] Technology Office of the Under Secretary of Defense For Acquisition and Logistics. *Report of the Defense Science Board on Task Force on Unexploded Ordnance.* 2003.

[2] Human Rights Watch. *Cluster munitions: measures to prevent explosive remnants of war and to protect civilian populations.* 2003.

[3] S. H. Ward and G. W. Hohmann. Electromagnetic theory for geophysical applications. In M.N. Nabighian, editor, *Electromagnetic Methods in Applied Geophysics*, volume 1, pages 131–311. Society of Exploration Geophysicists, 1991.

[4] S. D. Billings, L. R. Pasion, and D. W. Oldenburg. Discrimination and identification of UXO by geophysical inversion. Phase II: inversion of total-field magnetics. Technical Report ERDC/GSL TR-02-16, U.S. Army Corps of Engineers, Engineer Research and Development Center, 2002.

[5] Y. Li and D. W. Oldenburg. 3-d inversion of magnetic data. *Geophysics*, 61:394–408, 1996.

[6] P. G. Lelievre. Forward modelling and inversion of geophysical magnetic data. M.Sc. thesis, University of British Columbia, Vancouver, B.C., Canada, 2003.

[7] R. F. Butler. *Paleomagnetism: Magnetic Domains to Geologic Terranes.* Blackwell Science, 1992.

[8] S. D. Billings. Discrimination and classification of buried unexploded ordnance using magnetometry. *IEEE Transactions Geoscience & Remote Sensing*, 2004, In press.

[9] S. H. Ward. Unique determination of conductivity, susceptibility, size and depth in multifrequency electromagnetic exploration. *Geophysics*, 24(3):531–546, 1959.

[10] L. R. Pasion. Detecting unexploded ordnance with time-domain electromagnetic induction. M.Sc. thesis, University of British Columbia, Vancouver, B.C., Canada, 1999.

[11] C. E. Baum. *Detection and identification of visually obscured targets.* Taylor & Francis, 1998.

[12] F. Shubitidze, K. ONeill, S. A. Haider, K. Sun, and K. D. Paulsen. Application of the method of auxiliary sources for electromagnetic induction problem. *IEEE Trans. Geoscience & Remote Sensing*, 40:928–942, 2002.

[13] K. O'Neill. Ultra-wideband electromagnetic induction for UXO discrimination. Technical Report SERDP Project CU 1122, U.S. Army Corps of Engineers, Engineer Research and Development Center, 2002.

[14] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* SIAM, 1996.

[15] D. Marquardt. An algorithm for least squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.

[16] J. R. Wait and K. P. Spies. Quasi-static transient response of a conducting permeable sphere. *Geophysics*, 34:789–792, 1969.

[17] W. Menke. *Geophysical Data Analysis: Discrete Inverse Theory.* Academic Press, 1989.

[18] S. E. Dosso. Quantifying uncertainties in geoacoustic inversion I: a fast Gibbs sampler approach. *J. Acoust. Soc. Am.*, 111:143–159, 2002.

[19] K. Sun, K. ONeill, F. Shubitidze, I. Shamatava, and K. D. Paulsen. Theoretical analysis and range of validity of TSA formulation for application to UXO discrimination. *IEEE Transactions Geoscience & Remote Sensing*, 42:1871–1881, 2004.

[20] C. C. Chen and L. Peters. Buried unexploded ordnance identification via complex natural resonances. *IEEE Transactions on Antennas & Propagation*, 42:1645–1654, 1997.

[21] W. L. Anderson. Numerical integration of related Hankel transforms of orders 0 and 1 by adaptive digital filtering. *Geophysics*, 44:1287–1305, 1979.

[22] Y. Das, J.E McFee, J. Toews, and G.C. Stuart. Analysis of an electromagnetic induction detector for real-time location of buried objects. *IEEE Transactions Geoscience & Remote Sensing*, 28:278–287, 1990.

[23] J. E. McFee. Electromagnetic remote sensing: low frequency electromagnetics. Technical Report DRES SSP 124, Defence Research Establishment Suffield, 1989.

[24] S. D. Billings, L. R. Pasion, and D. W. Oldenburg. Inversion of magnetics for UXO discrimination and identification. In *Proc. 2002 UXO Forum.* Orlando, FL, Sep 2002.

[25] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer-Verlag, 2001.

[26] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 22:4–37, 2000.

[27] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In S. Becker T. Dietterich and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS).* 2002.

[28] B. Barrow and H. H. Nelson. Model-based characterization of electromagnetic induction signatures obtained with the MTADS electromagnetic array. *IEEE Transactions Geoscience & Remote Sensing*, 39:1279–1285, 2001.

[29] R. Gittins. *Canonical analysis: a review with applications in ecology.* Springer-Verlag, 1985.

[30] B. D. Ripley. *Pattern recognition and neural networks.* Cambridge University Press, 1996.

[31] P. Legendre and L. Legendre. *Numerical ecology.* Elsevier, 1998.

[32] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:955–974, 1998.

[33] P. Pavlidis, I. Wapinski, and W. S. Noble. Support vector machine classification on the web. *Bioinformatics*, 4:586587, 2004.

[34] W. H. Press et al. *Numerical recipes in C : the art of scientific computing.* Cambridge University Press, 1992.

[35] S. J. Hart, R. E. Shaffer, S. L. Rose-Pehrsson, and J. R. McDonald. Using physics-based modeler outputs to train probabilistic neural networks for unexploded ordnance (UXO) classification in magnetometry surveys. *IEEE Transactions Geoscience & Remote Sensing*, 39:797–804, 2001.

[36] S. J. Norton and I. J. Won. Identification of buried unexploded ordnance from broadband electromagnetic induction data. *IEEE Transactions Geoscience & Remote Sensing*, 39:2253–2261, 2001.

[37] Y. Zhang, L. Collins, H. Yu, C. E. Baum, and L. Carin. Sensing of unexploded ordnance with magnetometer and induction data: theory and signal processing. *IEEE Transactions Geoscience & Remote Sensing*, 41:1005–1015, 2003.

[38] L. M. Collins, Y. Zhang, H. Wang, L. Carin, S. J. Hary, S. L. Pehrsson, H. H. Nelson, and J. R. McDonald. A comparison of the perfomance of statistical and fuzzy algorithms for unexploded ordnance detection. *IEEE Transactions on Fuzzy Sytems*, 9:17–30, 2001.

[39] N. W. Townsend and L. Tarassenko. Estimation of error bounds for neural-network function approximators. *IEEE Transactions on Neural Networks*, 10:217–230, 1999.

[40] Y. Zhang, X. Liao, L. Collins, and L. Carin. Detection of buried UXO via active selection of labeled data. In *Proc. 2004 UXO Forum.* 2004.