Prediction of Phytoplankton Pigment Concentrations from Absorption Spectra

AMY ROSTRON

MSc by Research in Pattern Analysis and Neural Networks

ASTON UNIVERSITY

September 2005

Acknowledgements

I would like to thank everyone at Aston NCRG for all the help and encouragement offered over the last twelve months. The approachability and willingness to help of all within the department has made this year and this thesis possible for me.

In particular thanks to Professor David Saad and Dr. Davide D'Alimonte for their patience and interest. Also to Professor David Lowe for his enthusiasm and teaching me how to present my work and to and Dr. David Evans for his insight and advice.

Special thanks to my supervisor Dr. Dan Cornford for all his time and support. His patience, enthusiasm and guidance have been unending throughout.

Finally, thanks to all the research students for their help and for bringing much fun to a challenging year!

ASTON UNIVERSITY

Prediction of Phytoplankton Pigment Concentrations from Absorption Spectra

Amy Rostron

MSc by Research in Pattern Analysis and Neural Networks, 2005

Thesis Summary

This thesis studies the relationship between light absorption spectra and pigment concentrations in oceanic waters. Neural networks including Multi-layer Perceptrons and Radial Basis Functions will be used in order to model this relationship. The data will first be investigated by a thorough visualisation before attempting to reconstruct the spectra using forward models. Bayesian learning techniques are then discussed and applied to the retrieval of pigment concentrations. A range of data driven models will be implemented and finally a generative model produced, using Hybrid Monte Carlo sampling techniques.

Keywords: absorption spectra, chlorophyll, phytoplankton, pigment concentration retrieval, Principal Components Analysis, Multi-layer Perceptron, Radial Basis Function, Generalised Linear Model, Bayesian methods, Automatic Relevance Determination, Hybrid Monte Carlo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aim of the Thesis

The ultimate aim of the project is to create a model that reliably predicts pigment concentrations from remotely sensed light absorption spectra.

1.2 Background of the issue

The pigments studied are naturally occurring chemicals contained within phytoplankton. The importance of creating such a model then stems from the vital role of phytoplankton in both photosynthesis and the marine food web.

The contribution of phytoplankton to photosynthesis means it is an essential element of the carbon cycle. Models of phytoplankton growth together with satellite data have been used 'to convert maps of pigment concentration into maps of the carbon fixation rate' (Bricaud et al, 1995). A better scientific understanding of the carbon cycle may aid research into global climate change and in particular global warming.

Phytoplankton are also a vital element of various ecosystems, as they form the basis of the marine food chain. They are relied on by many species and therefore set 'a kind of upper limit on the productivity of the entire food chain' (Thomas, 1997).

Phytoplankton contain many pigments including chlorophyll. Chlorophyll concentrations are 'the major determinant of transmissibility of visible light through the ocean' (Chlorophyll Concentration, 2002) and therefore again affect entire ecosystems. A reliable model has many potential applications by facilitating basic monitoring of changes in chlorophyll and/or phytoplankton levels. This may assist investigation into the effects of major events, such as El Niño, on the global marine ecosystem (Thomas, 1997).

Differentiating between pigments is important because it allows identification of phytoplankton groups, each of which may have different properties and varying ecological roles. The size of the oceans means that reliable retrieval of pigment concentrations via remote satellite sensing would be much cheaper and more practical than current alternatives.

1.3 Overview of the Context of the Research

A substantial amount of relevant work surrounding the problem has already been done. Important contributions have come from Morel and Bricaud (1981), Hoepffner and Sathyendranath (1993) and more recently from Wozniak *et al* (2000). Regression frameworks have dominated with models achieving varying degrees of success in modelling the relationship between pigment concentrations and absorption spectra.

Previous models have focused on two central ideas. The first is the simpler approach and is based on the use of derivative analysis. High order derivatives of spectra are examined to identify change points and spectral absorption peaks for each pigment, as the basis for reconstructive models. The peaks may be used as the centres for fitting absorption bands, as by Aguirre-Gomez, Weeks & Boxall (1998), such that derivative analysis is a precursor to other modelling approaches.

The second and recently more popular approach focuses on absorption bands. Different pigments absorb in different areas of the spectrum, so each group of pigments studied has characteristic absorption bands. These are partially known at least for *in vitro* pigments (as dispersed in solution under laboratory conditions). These bands are usually assumed to be Gaussian with fixed centres, but as yet there are no universally agreed parameters. These may

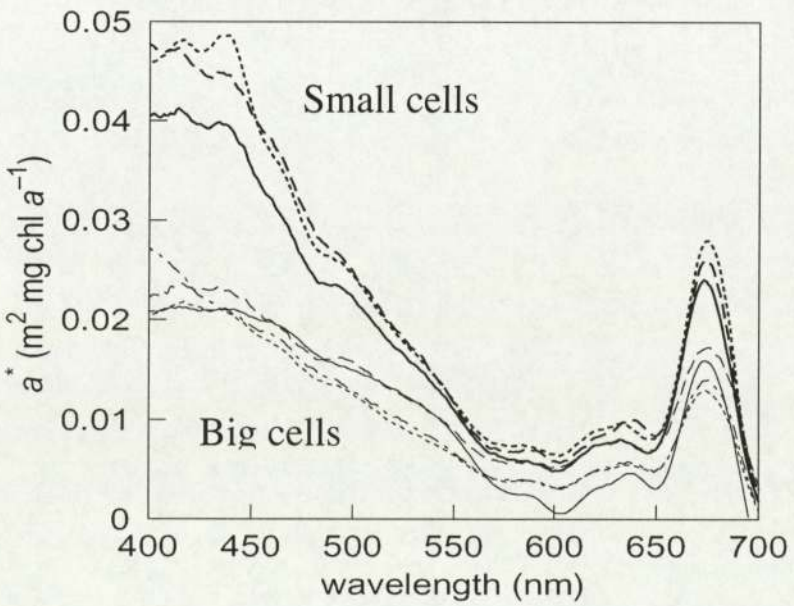
be determined by various methods and subsequently mixture models fitted to the data, as for example by Hoeppffner and Sathyendranath (1993).

The relationship between spectral absorption coefficients and concentrations of chlorophyll a (the dominant pigment) has been shown to vary with pigment composition and pigment packaging effect (Bricaud *et al*, 1995). These are in turn each governed by phytoplankton population levels and photoacclimation, the natural adaptation of the phytoplankton due to varying amounts of light reaching their cells. Population levels and photoacclimation reflect changes in other underlying factors, such as temperature, depth and nutrient concentration (for further detail see Wozniak *et al* (2000) and references therein).

The package effect arises because pigments are contained within cells and not uniformly distributed within phytoplankton. It is a result of variations in cell size and pigment distribution among phytoplankton populations and refers to the corresponding effect upon absorption spectra. Bricaud *et al* (2004) conclude that it is the package effect, rather than the pigment composition, that is often the 'dominant source of biological noise'.

The package effect is quantified as the ratio of pigment absorption *in vivo* versus that *in vitro* (Finkel *et al*, 2002). That is, the actual absorption of the pigment, as naturally 'packaged' within cells, compared to the corresponding absorption when dispersed into solution. In general, greater 'packaging' of pigments (or a larger cell size distribution) effectively dampens the whole of the corresponding absorption spectrum, as illustrated below (figure 1.3.1).

Figure 1.3.1: Effects of cell size on absorption spectra (Finkel 2001).



Morel and Bricaud (1981) discuss this flattening effect further and also propose a parameterisation of the package effect. The package effect is known to depend on cell size and intracellular pigment concentration (Morel and Bricaud, 1981), yet the associated non-linearity and uncertainty remain difficult to model. Properties of the phytoplankton are therefore an important variable in the pigment concentration versus absorption relationship.

Photoacclimation operates in a similar way, yet has the reverse effect on absorption levels. The amount of light reaching the cells, largely affected by varying depth and location of the phytoplankton and by seasonal factors, determines pigment content per cell. The greater the pigment content the larger the corresponding absorption, thus countering the package effect illustrated in figure 1.3.1.

Wozniak *et al* (2000) produced some of the best results to date by modelling both package and acclimation effects. Error bars in estimating mean absorption coefficients for chlorophyll a are stated as 36% - significantly better results than earlier models have produced. This project aims to improve upon previous works and to produce a solid modelling foundation for the reliable prediction of pigment concentrations from remotely sensed spectra.

1.4 The Data

The data were obtained from eight separate cruises each in different regions of oceanic water. Each sample includes a spectrum, which measures the absorption of light by the phytoplankton at 151 wavelengths. For each spectrum there are corresponding measures of concentration of five primary pigment groups:

- Chlorophyll a (Chl-a);
- Chlorophyll b (Chl-b);
- Chlorophyll c (Chl-c);
- Photosynthetic Carotenoids (PSCs); and
- Non-photosynthetic Carotenoids (NPSCs)

In addition there are several parameters relating to the size distribution of the phytoplankton cells. These consist of the estimated percentages of small, medium and large cells and also an overall size index for each sample (see below for further detail).

Also known is the cruise from which each sample was obtained. In total there are 1525 samples. These have been pre-divided into training (1220) and test (305) sets in order to evaluate the performance of any models produced. This predetermined division ensures any results will be comparable with those of related research taking place in France (Bricaud et al, 2003).

Absorption coefficients were measured using either the Quantitative Glass-fibre Filter Technique (QFT) or the "glass slide technique" (cruise 3 only). Contributions of phytoplankton and non-algal particles to total particulate absorption were then scientifically determined.

Pigment concentrations were measured simultaneously using High-pressure Liquid Chromatography (HPLC). The pigments were then grouped into the following five categories based upon absorption characteristics and the total corresponding concentration calculated. Further details of these methods are presented by Bricaud et al (1998).

Size parameters are also estimated using the HPCL concentration measurements together with taxonomic data regarding the sample-specific phytoplankton properties. The relative biomass proportions of picophytoplankton (<2 μm), nanophytoplankton (2-20 μm) and microphytoplankton (20-200 μm) are estimated the using the formulae presented in Bricaud et al (2004). This produces three size parameters specific to each sample from which a size index is then derived. A central value is assumed for each of the three taxonomic subsets and is then weighted by the corresponding biomass. The sum of these produces the size index – an indication of the dominant size structure for each sample.

The samples considered are all collected from the first optical depth to impose some limitation on photoacclimation variability and subsequent influence on the package effect. Samples have been processed to remove both the absorption due to detritus in the water and

that due to the seawater itself. Where necessary correction factors have been applied to measurements potentially introducing systematic errors to the affected samples.

The nature of the data is such that there may be many additional variables impacting upon the concentration and absorption measurements. These include cruises in different depths and temperature waters, cruises in varying locations at different times of year and differences in measuring equipment. The experimental methods then may be subject to some uncertainties and particularly instrumental limitations. Small inconsistencies within the dataset are therefore unavoidable given the substantial time and costs involved in collection and collation of the data.

A summary of cruise locations, times and number of samples collected can be found in Appendix A.1.

1.5 The Approach

The investigation will begin with a visualisation of the data in Chapter 2. This is an opportunity to understand the data and any immediate structure within it and to find any outliers. Dimension reduction techniques will be applied to assist visualisation and also to potentially produce a more relevant and manageable data set for the following modelling process.

A discussion of Bayesian methods follows in Chapter 3 providing a framework for the subsequent models. The remainder of Chapter 3 concerns 'forward modelling' experiments, which attempt to reconstruct the absorption spectra given the concentration measurements. These are tackled first, because they are believed to be simpler and may provide useful insights prior to tackling the reverse problem.

Finally, pigment concentration retrieval will be attempted - firstly using data driven models in Chapter 4 and finally by producing a generative model in Chapter 5. The conclusions reached will be summarised in chapter 6 and suggestions for further work proposed in chapter 7.

Chapter 2

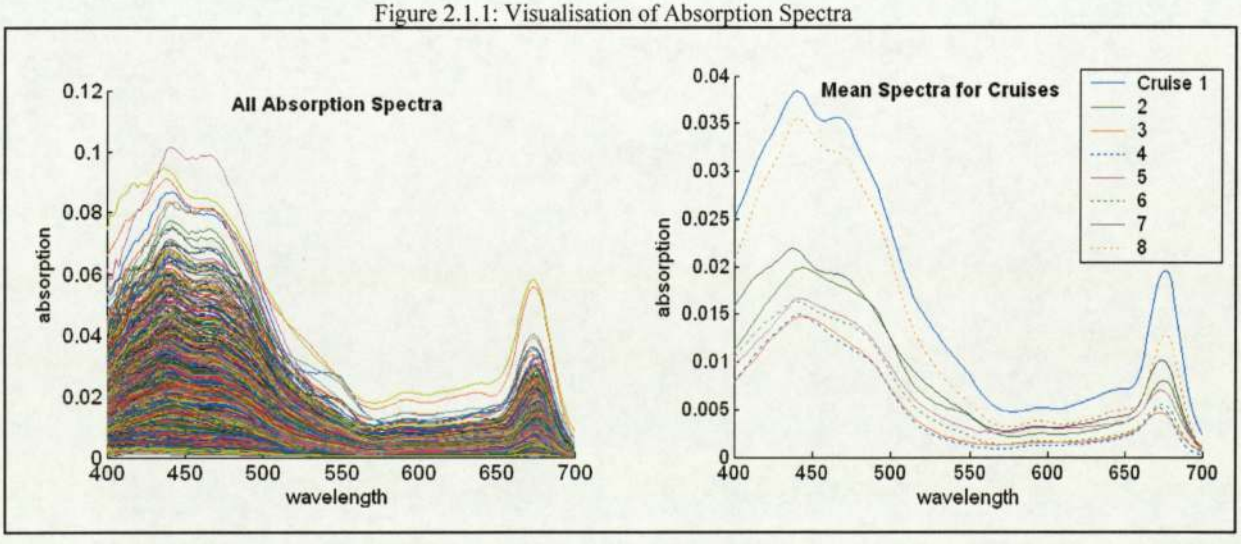
Visualisation & Dimensionality Reduction

The first stage of the research process is to visualise the data. Any immediate correlations, patterns or structure in the data and any outliers may then be identified.

2.1 Basic Spectra Visualisation

Basic plots of the raw spectra were produced, including plots for separate cruises and selected size index groupings. This basic representation showed similar shaped spectra across cruises and size divisions, though differences were evident, particularly at shorter wavelengths. A plot of the mean spectra per cruise shows some significant feature variation (see figure 2.1.1), such as the heights of the major peaks. This may be attributed to differences in concentrations between cruises, but could also indicate other differences, such as variation in the package effect. Additional cruise dependent differences suggest that the modelling framework may need to incorporate additional variables and/ or model cruises separately.

Figure 2.1.1: Visualisation of Absorption Spectra



A plot of the standard deviation of absorption by wavelength showed very similar structure to the mean spectrum suggesting that the variance may be proportional to absorption. Normalisation to unit variance could then enable clearer representation of the data and possibly better separation of certain subsets. Multiplicative variability also indicates that a log transform of the data may be useful.

2.2 Visualisation Issues

The high dimensionality of the data meant informative visualisations were difficult. Also, use of the raw data as an input to a neural network was likely to result in lengthy training and sub-optimal parameters.

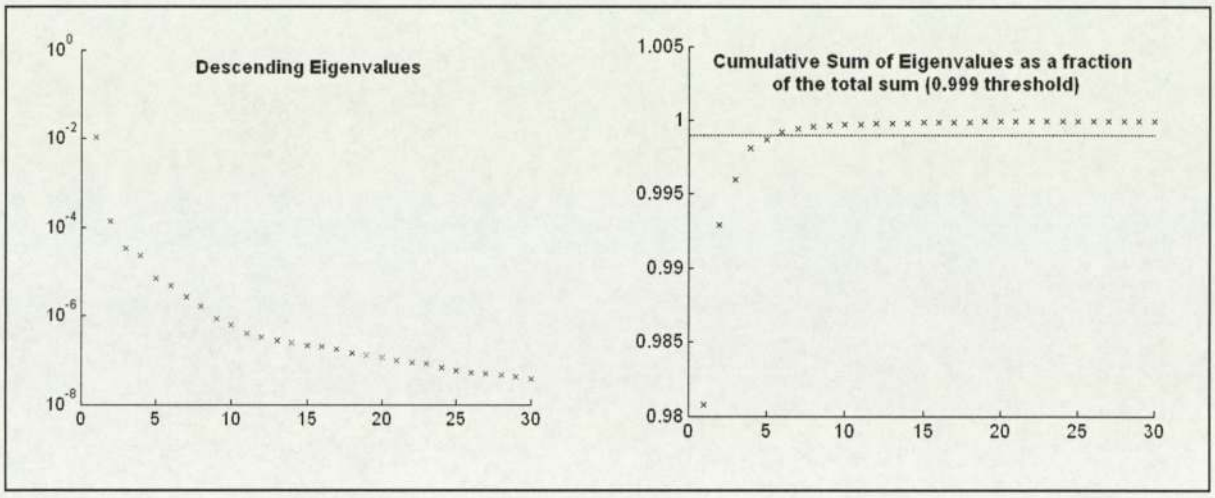
To avoid these problems and enable better graphical representation, dimensionality reduction techniques were used. Both linear and non-linear approaches were considered to try to determine the underlying structure of the data. The particular methods used were Principal Components Analysis (PCA) and Neuroscale, each of which may emphasise different aspects of the data. A lower dimension representation may also help avoid over-fitting of networks and therefore improve following model performance.

2.3 Principal Components Analysis (PCA)

The first approach was to try a simple, linear PCA of the data. PCA was carried out on the raw data set, on the log transformed data set and using data normalised to zero mean and optionally unit variance.

The log-plot in figure 2.3.1 shows the eigen-spectrum relating to the Principal Components (PCs). The relative size of the first and second eigenvalues indicates that a huge amount of information is contained in the first two PCs. The curvature of the descending data points is relatively smooth, so it is difficult to determine at what point noise becomes dominant. This lack of an obvious break in the curvature could indicate that the noise is quite high on even the larger PCs. It appears that around ten to twelve PCs are necessary to capture most of the information.

Figure 2.3.1: Eigenvalue Analysis



A cumulative analysis followed to see how much of the variance was retained by using successive numbers of PCs (see fig 2.3.1). The trend is examined more closely by omitting the largest eigenvalues, normalising and producing log plots. Up to 99.3 percent of the variance in the data (depending on the normalisation method) is retained by the first two PCs. Using just six PCs this increases to over 99.9% (see fig 2.3.1) and using 10 PCs the variance attributable to additional PCs is of order 10^{-4}.

Plotting the PCs (eigenvectors) themselves revealed a similar pattern (see Appendix A.2 for illustration). The first eight to twelve vectors (depending on normalisation and space) have a fairly smooth appearance. As further vectors are plotted there appears to be less structure and erratic behaviour is observed. This is assumed to be the result of noise in the data.

Although the graphical analysis does not give a definitive answer as to the optimal number of principal components, around eight to twelve PCs appear to retain information without excessive noise. These conclusions are supported by previous research (Evans & Cornford, 2003), which used probabilistic PCA and proposed an optimal representation of twelve PCs.

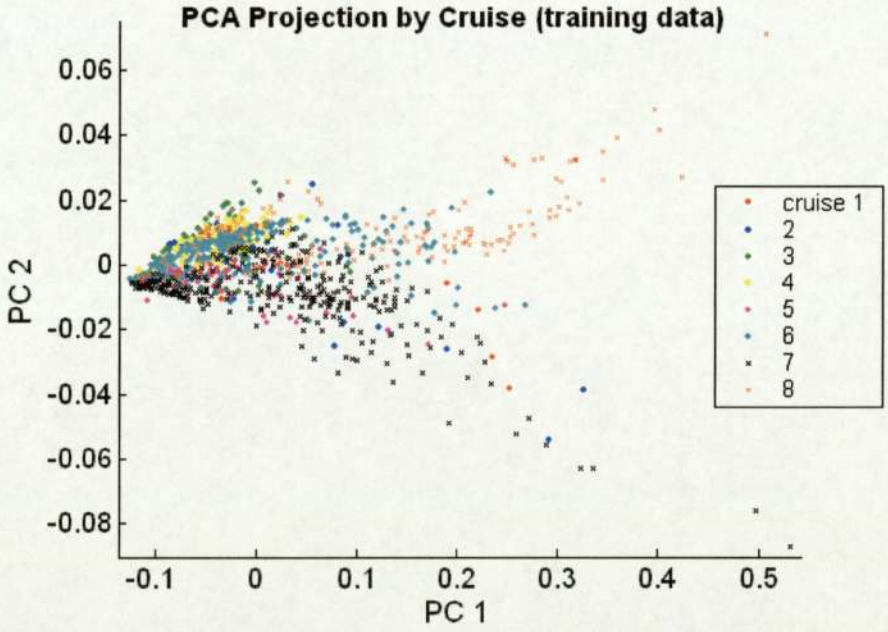
2.3.1 PCA visualisation

To visualise the data in two dimensions it was initially projected onto the first two PCs. The first approach was to project the spectral data, second the concentrations and finally to project a combined dataset comprising both. Labelled plots of training and test data projections confirm that the test set spans a similar range to the training data.

The PCA visualisation was repeated with labelling by cruise and by size index. As with the spectral plots there is evidence of differences between cruises indicated by clustering of the samples by cruise (see figure 2.3.1.1). The size distribution data also appears significant, as despite the approximate nature of the size index, clustering is again evident.

Projecting the data onto the second and third PCs resulted in an alternative representation, but still displayed the same apparent clustering by cruise and size. The same was true when incorporating concentrations and using concentration data alone, so a proportion of this clustering is likely to be the result of differences in concentration between cruises.

Figure 2.3.1.1: PCA Projection of Absorption Spectra



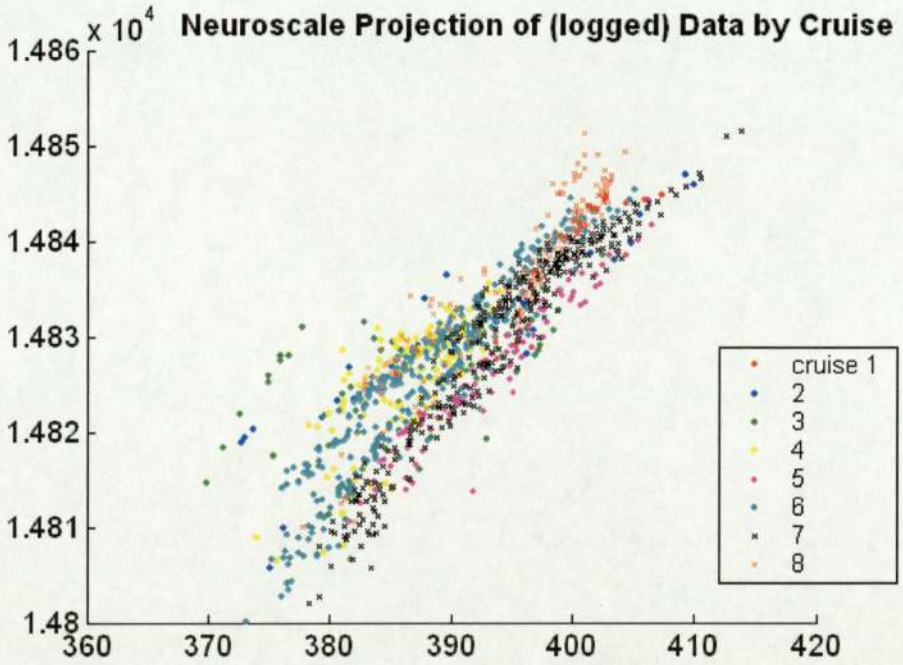
For further illustrative plots and to see the various outliers highlighted by each approach, see Appendix A.3.

2.4 Neuroscale

To investigate the possibility of non-linear structure in the data Neuroscale, a non-linear projection technique (Nabney, 2002), was used. Again both spectra and concentration data were used. Initially, Neuroscale was applied directly to the data without any dimensionality reduction but had limited success. The main approach involved pre-processing the data using PCA with various numbers of PCs. Normalised data was projected onto these PCs and then used as an input to Neuroscale. The log transform is again considered.

The Neuroscale results supported previous PCA findings of clustering by cruise and size with several possible outliers (see figure 2.4.1). Separations were not obviously improved by using Neuroscale, suggesting the data was largely of linear structure. Each approach highlighted outliers, though this was no more conclusive than by use of PCA.

Figure 2.4.1: Neuroscale Projection of Absorption Spectra



2.5 Outlier Analysis

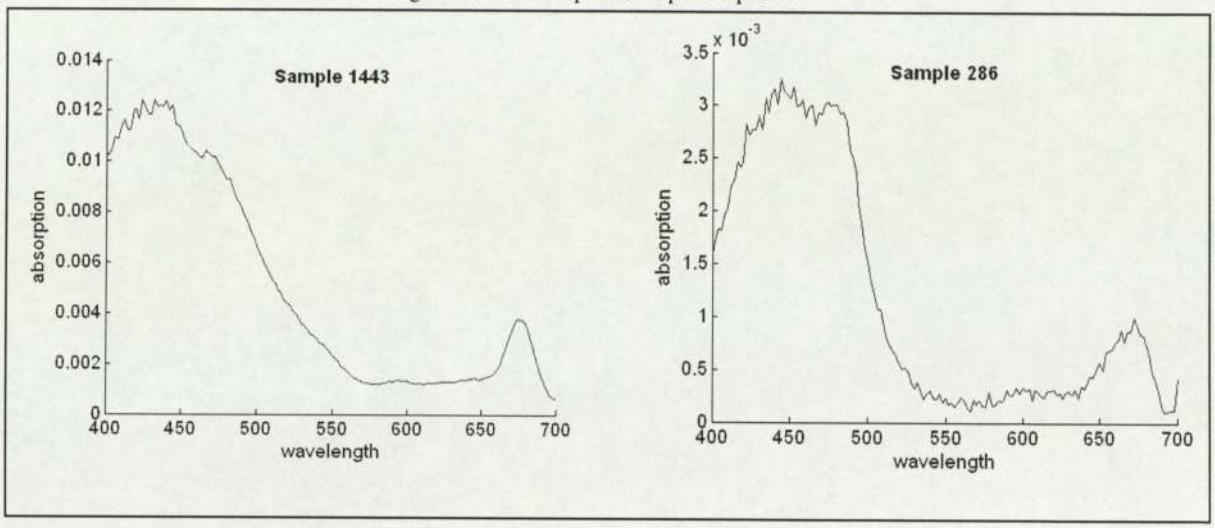
Given the level of uncertainty related to the data and experimental methods it was expected that there be some outlying samples. Identifying and removing these from the dataset should enable a more accurate model.

The various visualisations of the data reveal a number of possible outliers. However, the outliers identified are not consistent for all visualisations. There is dependence on the particular visualisation method (PCA or Neuroscale), input data (spectra and/or concentrations), normalisation, and whether or not the log transform is used. Even using a single plot the analysis is subjective. Density modelling has been avoided here, as although this would attach an objective measure the densities would be based on the same data and therefore not provide an independent threshold.

Having identified potential outliers from each visualisation the individual spectra of these samples were examined for unusual features. Comparison with the total mean spectrum, mean spectrum for the appropriate cruise and the spectra of surrounding points followed. Interestingly, a number of outliers identified by concentrations corresponded to those found using the spectra. This consistency suggests these may be extreme cases rather than 'true' outliers resulting from errors. To address this issue for certain samples the concentrations were examined alongside the spectra and also compared to mean concentrations and several typical concentration profiles.

Several spectra displayed exponential characteristics indicating the influence of detritus in the water. It appears that the detritus effect has not been correctly removed from these samples and so they will be discarded. In examining individual spectra some unusual structures were evident. Certain spectra, for example sample 1443 (see fig 2.6.1), showed noisy measurements at lower wavelengths, while others such as sample 80 were noisier at high wavelengths. Others, including sample 286 (see fig 2.6.1) display noise in all areas of the spectrum. This may indicate differences in noise characteristics and package effect across cruises, suggesting a need for separate noise models. Some spectra, despite obvious influence of noise were retained, as they are still characteristic of the basic underlying structure.

Figure 2.6.1: Example Absorption Spectra



The final analysis classified the majority of samples as extreme yet valid. Forty-one training samples (3.36%) and 8 test samples (2.62%) were identified as outliers. The majority of these came from cruise seven suggesting either problems with the measuring equipment or errors in processing the measurements (see Appendix A.4 for breakdown). For some samples the analyses were inconclusive and so these were retained with caution.

2.6 Visualisation & Dimensionality Reduction Conclusions

Several main findings resulted from visualisation offering some insight into the structure of the data:

Clustering by cruise and clustering by size: The evidence of clustering by cruise and size suggest a definitive need to incorporate these characteristics into the modelling framework. Both are likely to be largely a result of packaging and acclimation effects.

Evidence of largely linear underlying structure: The degree of clustering does not seem to vary greatly using the non-linear projection methods and so the relationship appears largely linear. Outliers were also more successfully identified by PCA than Neuroscale methods.

Dimension reduction: PCA suggests that the data can be reduced to ten dimensions without significant loss of information.

Usefulness of logarithmic space: Log space certainly spread the data more and assisted identification of outlier. It also proved useful in examining the structure of PCs.

Outliers: Irregular outliers have been removed from the dataset, such that subsequent models will disregard these samples.

Differential noise structures: Early evidence suggests a need for cruise specific noise models.

Chapter 3

Spectra Reconstruction and Regression Framework

This stage of research will focus on reconstructing the spectra from the concentration data. The results of this 'forward modelling' will then be used together with previous findings to determine an appropriate regression framework for the following pigment concentration retrieval problem. If a forward model accurately reconstructs the spectra then there is potentially a direct inversion of this model that can predict concentrations from the spectra.

3.1 Bayesian Methods

As there is uncertainty regarding the model parameters a Bayesian approach is preferred. This models uncertainty by using probability densities and incorporates prior knowledge by placing prior distributions over the parameters (Nabney, 2002). Weight priors are comparable to regularisation coefficients, such that many parameters may be used without problems of overfitting. A specific form of prior also facilitates use of Automatic Relevance Determination (ARD).

In the Bayesian framework before any data is observed the parameters are modelled by a prior probability distribution, $p(\mathbf{w})$ that reflects knowledge of the parameters before any training. When the data is observed (as training takes place), the prior is updated using the new information to produce the corresponding posterior distribution for the parameters. This is done using Bayes' theorem (3.1), which combines the prior with the likelihood, $p(D \mid \mathbf{w})$ – the probability of the data given the weights.

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})\, p(\mathbf{w})}{p(D)} = \frac{p(D \mid \mathbf{w})\, p(\mathbf{w})}{\int p(D \mid \mathbf{w})\, p(\mathbf{w})\, d\mathbf{w}} \qquad (3.1)$$

The denominator p(D) is just a normalising constant, so the posterior for the weights $p(\mathbf{w} \mid D)$ is proportional to the product of the likelihood and the prior. The posterior distribution narrows as training incorporates new knowledge as to which weight values are most consistent with the data observed.

Network training then results in a posterior probability distribution for the parameters based on prior beliefs and all the observed data. Forward propagation of new inputs through the trained network then produces corresponding outputs. Bayesian predictions are then found by integrating these outputs with respect to the posterior distribution for the parameters (Neal, 1996).

More specifically, the predictive (or posterior) distribution of a new data point **y,** given input vector **x** and the dataset D, is obtained by integrating over the posterior distribution of network weights, **w** (see (3.2) below).

$$p(y \mid \mathbf{x}, D) = \int p(y \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{w} \mid D)\, d\mathbf{w} \qquad (3.2)$$

where $p(\mathbf{w} \mid D)$ is again the posterior distribution for the weights, **w** given the dataset, D and $p(\mathbf{y} \mid \mathbf{x}, \mathbf{w})$ is the output for the given model with parameters, **w** and input, **x**. Predictions are not then based on a single estimate for the parameters, but effectively integrate over all possible models so that validation is theoretically unnecessary.

3.1.1 Bayesian Methods in Practice – Priors

The current knowledge of parameters in the problem is limited, so the initial prior probability distribution used will be broad. Small weights are preferred initially to allow sufficient flexibility of mapping, so a Gaussian distribution with zero mean of the general form (3.3) will provide a suitable first approximation and may simplify later analysis.

$$p(\mathbf{w}) = (1/ Z_W(\alpha)) \exp(-\alpha E_W(\mathbf{w})) \qquad (3.3)$$

where α is a hyperparameter controlling the prior distribution, E_W is the error function and $Z_W(\alpha)$ is the normalisation factor $Z_W(\alpha) = \int \exp(-\alpha E_W(\mathbf{w}))\ d\mathbf{w}$, such that $\int p(\mathbf{w})\ d\mathbf{w} = 1$.

For a quadratic weight penalty the error function E_W takes the form (3.4):

$$E_W(\mathbf{w}) = (1/2) \sum_i \mathbf{w}_i^2 \qquad (3.4)$$

The prior then becomes (3.5):

$$\begin{aligned} p(\mathbf{w}) &= (1/ Z_W(\alpha)) \exp(-\alpha/2) \|\mathbf{w}\|^2 \qquad (3.5) \\ &= (\alpha/2\pi)^{W/2} \exp(-\alpha/2) \|\mathbf{w}\|^2 \end{aligned}$$

Multiple priors may be used to correspond to groups of weights, each comparable to a weight error term regularising the associated weights (Nabney, 2002) and therefore tackling overfitting.

The prior is governed by the additional hyperparameter, α, which represents the inverse variance of the distribution and controls the distribution of the other model parameters. Such 'hyperpriors' effectively allow a network to choose its own complexity (Saad, 2004) by determining the magnitude of weights allowed and are central to the use of ARD (see section 3.6).

3.1.2 Bayesian Methods in Practice – Likelihood

The likelihood may be expressed similarly using the general form (3.6) with hyperparameter, β.

$$p(D \mid \mathbf{w}) = (1/ Z_D(\beta)) \exp(-\beta E_D(\mathbf{w})) \qquad (3.6)$$

where E_D is the error function and $Z_D(\beta)$ the normalisation factor $Z_D(\beta) = \int \exp(-\beta E_D(\mathbf{w}))\ dD$, such that $\int p(D \mid \mathbf{w})\ dD = 1$. β scales the error function and so controls the variance of the noise.

For a sum-of-squares error function E_D is expressed as (3.7):

$$E_D(\mathbf{w}) = (1/2) \sum_n (\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{y}_{obs\ n})^2 \qquad (3.7)$$

where $\mathbf{y}(\mathbf{x}_n, \mathbf{w})$ is the model prediction for input $\mathbf{x}_n$ and $\mathbf{y}_{obs\ n}$ is the corresponding observed value.

Thus the likelihood becomes (3.8):

$$p(D \mid \mathbf{w}) = (1/ Z_T(\beta)) \prod_n \exp((-\beta/2)(\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{y}_{obs\ n})^2) \qquad (3.8)$$
$$\equiv (1/ Z_T(\beta)) \exp(-\beta/2\ \|\mathbf{y} - \mathbf{y}_{obs}\|^2) \qquad \text{in matrix form,}$$

where $Z_T(\beta)$ is the combined normalising constant.

3.1.3 Bayesian Methods in Practice – Posterior

Using the definitions of the prior (3.3) and likelihood (3.6) together with Bayes' rule (3.1) the posterior distribution can now be expressed as (3.9):

$$p(\mathbf{w} \mid D) = (1/ Z_S) \exp (-\mathbf{S}(\mathbf{w})) \qquad (3.9)$$

where $\mathbf{S}(\mathbf{w}) = \beta E_D + \alpha E_W$ and $Z_S = \int \exp(-\beta E_D - \alpha E_W)\ d\mathbf{w}$.

Expanding $\mathbf{S}(\mathbf{w})$ about the minimum using Taylor series gives the following approximation (3.10):

$$\mathbf{S}(\mathbf{w}) \approx \mathbf{S}(\mathbf{w}_{MP}) + (1/2)(\mathbf{w} - \mathbf{w}_{MP})^T \mathbf{A} (\mathbf{w} - \mathbf{w}_{MP}) \qquad (3.10)$$

where $\mathbf{w}_{MP}$ is the weight vector at the minimum of $\mathbf{S}$ and $\mathbf{A}$ is the Hessian of $\mathbf{S}$ ($\mathbf{A} = \nabla\nabla\mathbf{S}$).

The posterior can then be expressed as a Gaussian approximation (3.11), which becomes exact for a linear model.

$$p(\mathbf{w} \mid D) \approx (1/ Z_{S*}) \exp (-\mathbf{S}(\mathbf{w}_{MP}) - (1/2)(\mathbf{w} - \mathbf{w}_{MP})^T \mathbf{A} (\mathbf{w} - \mathbf{w}_{MP})) \quad (3.11)$$

where Z_{S*} is a normalising constant for the approximating Gaussian.

This is known as the Laplace approximation and effectively finds the most likely parameters by identifying the modes of the posterior and modelling each by a Gaussian distribution. The above analysis however, assumes the hyperparameters to be known and fixed. This is not the case here and so will be dealt with shortly in section 3.1.4.

3.1.4 Bayesian Methods in Practice – Hyperparameters & Evidence

The normalisation factor, p(D) is known as the evidence. As seen in (3.1) it is obtained by integration over the weights, which may only be analytically tractable for particular forms of prior and likelihood (Nabney, 2002). Also, true Bayesian prediction requires integration over the posterior (see (3.2)), which can cause similar problems. Consequently, the data driven networks will use 'evidence' approximation in calculation of the posterior.

The fully Bayesian approach integrates over all unknown weights including hyperparameters, so the posterior may be represented as in (3.12).

$$\begin{aligned} p(\mathbf{w} \mid D) &= \int\int p(\mathbf{w}, \alpha, \beta \mid D) \; d\alpha \; d\beta \qquad (3.12) \\ &= \int\int p(\mathbf{w} \mid \alpha, \beta, D) \; p(\alpha, \beta \mid D) \quad d\alpha \; d\beta \end{aligned}$$

Evidence approximation avoids this multi-dimensional integral by instead seeking optimal hyperparameters based on information from the training data. By assuming the posterior of the hyperparameters $p(\alpha, \beta \mid D)$ to be sharply peaked around their most probable values α_{MP} and β_{MP}, the weight posterior in (3.12) reduces to (3.13).

$$p(\mathbf{w} \mid D) \approx p(\mathbf{w} \mid \alpha_{MP}, \beta_{MP}, D) \iint p(\alpha, \beta \mid D)\ d\alpha\ d\beta \approx p(\mathbf{w} \mid \alpha_{MP}, \beta_{MP}, D) \qquad (3.13)$$

To make the approximation the hyperparameters must be fixed to these 'optimal' values α_{MP} and β_{MP} that maximise the posterior $p(\mathbf{w} \mid D)$. The posterior $p(\mathbf{w} \mid \alpha, \beta, D)$ however, may contain multiple modes due to non-linear mapping and/or network symmetries (Nabney, 2001). Therefore α_{MP} and β_{MP} are found from the modes of the posterior distribution of the hyperparameters (see (3.14)).

$$p(\alpha, \beta \mid D) = \frac{p(D \mid \alpha, \beta)\, p(\alpha, \beta)}{p(D)} \qquad (3.14)$$

p(D) is the integral of the numerator in (3.14) and therefore not relevant in determining the modal values. Also, the prior $p(\alpha, \beta)$ can be assumed uniform, such that to find the peaks it is necessary only to maximise $p(D \mid \alpha, \beta)$ – the probability of the data for the given hyperparameters. This probability can be expressed as:

$$p(D \mid \alpha, \beta) = \int p(D \mid \mathbf{w}, \alpha, \beta)\, p(\mathbf{w} \mid \alpha, \beta)\ d\mathbf{w} = \int p(D \mid \mathbf{w}, \beta)\, p(\mathbf{w} \mid \alpha)\ d\mathbf{w} \qquad (3.15)$$

α and β are effectively scale parameters determining the magnitude of $\|\mathbf{w}\|^2$ and the noise respectively. The weights then have no dependence on the noise parameter, β, and the likelihood is independent of α. Using (3.3) and (3.6) this is equivalent to (3.16):

$$p(D \mid \alpha, \beta) = (1/Z_W(\alpha)) * (1/Z_D(\beta)) \int \exp\{-\alpha E_W(\mathbf{w}) - \beta E_D(\mathbf{w})\}\ d\mathbf{w} \qquad (3.16)$$

To find hyperparameters the logarithm of this likelihood, $\ln\{p(D \mid \alpha, \beta)\}$, is optimised separately with respect to α and β (as outlined by Bishop (1995)). This results in the following equations ((3.17) and (3.18)) for α and β in terms of the eigenvalues ξ_i of the data Hessian ($\mathbf{H} = \beta \nabla\nabla E_D$):

$$2\,\alpha\,E_W^{MP} = \sum_i \{\,\xi_i / \xi_i + \alpha\} = \upsilon \qquad (3.17)$$

$$2\,\beta\,E_D^{MP} = N - \sum_i \{\,\xi_i / \xi_i + \alpha\} = N - \upsilon \qquad (3.18)$$

which reduce to the following re-estimation formulae for α and β ((3.19) and (3.20) respectively):

$$\alpha^{new} = (\upsilon / 2E_W) \qquad (3.19)$$

$$\beta^{new} = (N - \upsilon / 2E_D) \qquad (3.20)$$

Using iterative optimisation of weights and re-estimation of hyperparameters it is then possible to find the optimal values for α, β and **w,** which should eventually converge. The hyperparameters are then fixed to these values and the posterior recalculated using equation (3.13). The output of this trained network corresponds to the mean of the predictive distribution, which as before is found using equation (3.2) following substitution of the newly approximated posterior.

3.2 Networks – The Multi-Layer Perceptron (MLP)

The basic neural network used will be a two layer Multi-Layer Perceptron (MLP). This is compatible with use of Bayesian methods, regularisation and Automatic Relevance Determination (ARD). Pre-processing methods will include PCA, as the full dimensionality of data may again cause problems through computational complexity and sub-optimal training. The PC reconstruction of the spectra is very good reducing dimension without significant loss of information: using ten PCs the correlation coefficient exceeds 0.999 and the mean absolute error is 0.00008.

Inputs to the network itself will be the five pigment concentrations and (optionally) cell size distribution data. Outputs will be either normalised estimates of total absorption or alternatively projected spectra, as reduced to ten dimensions by PCA. The networks will be trained using scaled conjugate gradient optimisation (Nabney, 2002) with a sum-of-squares error function. Performance will be evaluated using the designated test set and reconstruction

errors calculated by comparing observed and predicted spectra. Error measures will include the root mean squared error (RMSE), mean absolute error (MAE) and bias (see Appendix A.5 for full error formulae). The analysis will compare final reconstructed spectra rather than raw network outputs, so as to optimise with respect to any data processing as well as the network mapping.

Experiments will include training models for the whole spectrum, at individual wavelengths and by separate cruises. The basic normalisations and a log transform of the data will also be considered. The log transform has the advantage of constraining outputs to be positive as well as potentially fitting underlying structure in the data.

3.3 Validation

The strict application of Bayesian methods implies that any 'large' network can produce reliable results and that validation for network selection is unnecessary. However, excessive numbers of hidden units result in the modelling of unnecessary noise and highly complex networks may result. There is also a required knowledge of what constitutes a large enough network. To avoid these issues a more flexible approach will investigate the performance of various network structures using a validation set.

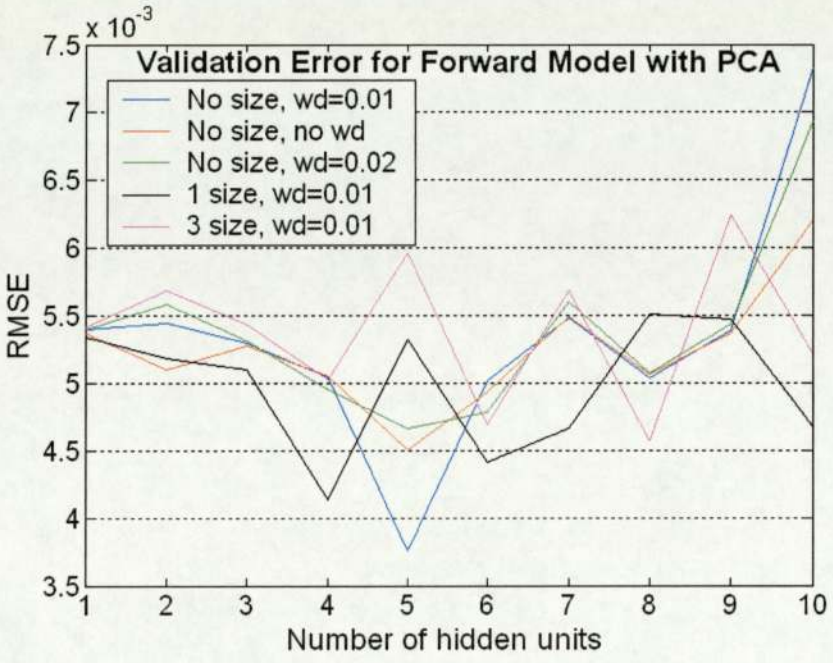
The validation set consists of a random sample of 20% of the training data. The remainder of the training data will be used to train the validation networks with various numbers of hidden units between one and sixty. Validation experiments will be carried out with no regularisation and using several different weight decay priors (Nabney, 2002).

3.3.1 Validation Outcomes

Training and testing validation networks results in several error minimums for each structure. Errors are quite erratic and also vary with optimisation parameters, but approximate global error minimums are identified in the following analysis. The basic regression model for all cruises taking concentration inputs and using the ten PC representation of spectra gives a global minimum at five to six hidden units (see figure 3.3.1.1). Incorporating size distribution

inputs into the model has some effect on errors, but this is neither conclusively good nor bad. Use of small weight decay coefficients appears beneficial but the effects are again small.

Figure 3.3.1.1: Validation Error for the basic forward model with PCA pre-processed spectra.

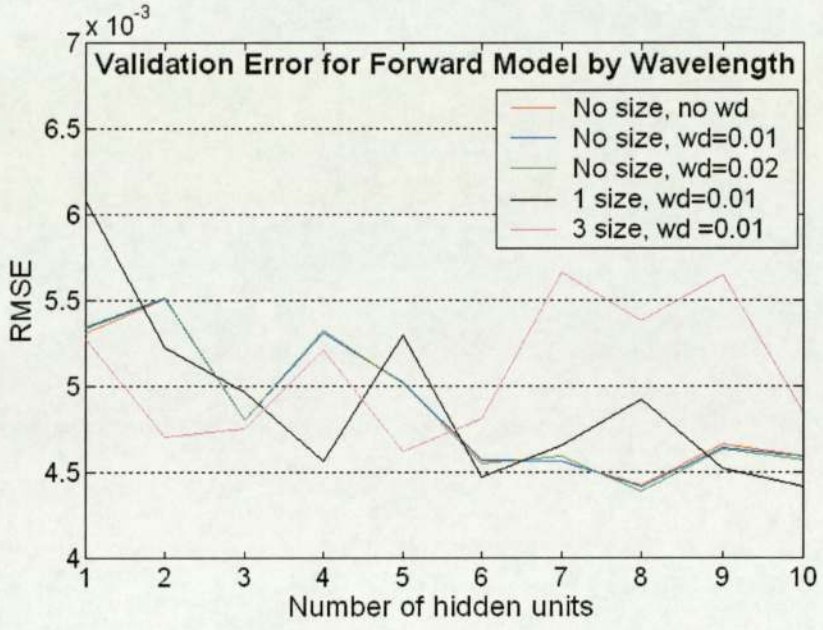


Note: wd denotes weight decay coefficient

The alternative model not using PCA and predicting absorption at individual wavelengths supports a network with seven to eight hidden units (dependent on optimiser parameters, see figure 3.3.1.2). While further minimums occur around 16 to 18 units the improvement is minimal. Training is already lengthy and so further units would appear to unjustifiably overcomplicate networks.

Weight decay appears to marginally improve model performance, while size inputs have varying effects and cause some erratic error behaviour. Despite these differences in the magnitude of errors there is limited effect on positioning of global minimums and the inferred optimal networks. The optimum number of hidden units remains between five and eight for each variant of model.

Figure 3.3.1.2: Validation Error for the forward model by wavelength.



Note: wd denotes weight decay coefficient

A given structure is assumed to be at least as effective in fitting the data when trained on individual cruises as when modelling the whole data set. While there may be more hidden units than necessary for individual cruise models hyperparameters will adjust accordingly. Using a standardised structure determined using all the data will better facilitate performance comparisons. Limiting the hidden layer to eight units constrains potential structural noise, but can still be considered a 'large' network relative to the estimated optima. Further forward models will use this pre-determined structure and revert to use of the full training set (including the validation data). The original test set will be used to evaluate model performance.

3.4 Reconstruction

Initially, basic non-Bayesian reconstructions of the data were produced to understand the general mapping. Attempts to reconstruct the whole spectra from the MLP directly with 151 outputs result in non-smooth predicted spectra. It is concluded as expected that such a model is too complex and fits the noise in the data. Models were more successful where the target spectral data was reduced in dimension (by PCA) prior to training.

3.4.1 Bayesian Modelling 1: Full Spectrum using PCA

Bayesian models are first implemented using ten PCs and eight hidden units, as determined by validation. The basic model takes the five concentrations as inputs with variations also incorporating the three size data variables or size index. A Generalised Linear Model (GLM) with the same inputs and outputs will also be tested alongside the MLP to provide a benchmark. The results of training several preliminary models are seen in table 3.4.1.1.

Table 3.4.1.1: Bayesian forward model errors. MLP has 8 hidden units; all models use 10 PC representation of spectra.

	MLP No size Mu norm*	**MLP** No size SD norm*	**MLP** Size index SD norm*	**MLP** 3 sizes SD norm*	**GLM** No size SD norm*	**GLM** Size index SD norm*	**GLM** 3 sizes SD norm*
CORR	0.965	0.967	0.966	0.966	0.954	0.955	0.956
Mean % error	15.44	15.10	14.91	16.01	18.04	17.47	17.56
Bias	-6.70e-05	9.49e-06	-2.10e-05	-1.17e-05	9.71e-05	9.32e-05	9.32e-05
MAE	0.0014	0.0014	0.0014	0.0014	0.0016	0.0016	0.0016
RMSE	0.0025	0.0024	0.0025	0.0025	0.0029	0.0028	0.0028

CORR=Correlation coefficient, RMSE=Root mean squared error, MAE=Mean absolute error.

*Data normalisation method (spectra, concentrations and size each normalised separately): Mu refers to removal of mean, SD removes mean and sets to unit variance.

Normalisation of inputs to zero mean and unit variance tends to improve results. The overall correlation between true and predicted values is in excess of 0.95 for all models, including the GLM. Incorporating the size data there are only small changes, which appear more positive when using the GLM. The true effect however is difficult to assess at this stage, as the additional data sometimes improves performance and other times worsens it. It may be that small improvements are disguised by an increase in noise resulting from the more complex network created.

Modelling by cruise significantly improves the overall predictive performance for all of the models. Table 3.4.1.2 compares two of the models from table 3.4.1.1 with the average errors found using the by cruise model. Calculating average reconstruction errors (excluding any problem cruises) there is consistently greater accuracy in the mapping. Average correlations improve for all networks compared to the corresponding overall model, as do error measures RMSE and MAE. Size effects remain unclear but again seem more positive using the GLM.

Table 3.4.1.2: Bayesian forward model errors – comparison with by cruise averages. MLP has 8 hidden units; all models use 10 PC representation of spectra, SD norm.

	MLP with size index		**GLM with 3 size inputs**	
	Overall	**By cruise average***	**Overall**	**By cruise average**
CORR	0.966	0.968	0.956	0.982
Mean % error	14.91	13.43	17.56	11.00
Bias	-2.10e-05	0.000132	9.32e-05	6.48e-05
MAE	0.0014	0.0011	0.0016	0.00098
RMSE	0.0025	0.0020	0.0028	0.0017

*MLP by cruise average excludes cruise 1 due to extreme error.

Cruises with few samples however, such as cruise 1, are prone to extreme or infinite errors when using particular networks. This is illustrated by the example breakdown of by cruise results in table 3.4.1.3, which correspond to the MLP from table 3.4.1.2.

Table 3.4.1.3: Bayesian forward model errors - breakdown by cruise for the MLP, modelled in linear space with size index input, 8 hidden units, 10 PCs, SD norm. **A** refers to the overall model and **B** to the model trained on individual cruises.

Cruise	**CORR**		**Mean % error**		**Bias**		**RMSE**	
Model	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
1	0.99	0.32	11.58	9.9e+152	0.0020	-1.5e+149	0.0025	2.3e+149
2	0.91	0.94	25.41	19.30	0.0012	-3.90e-04	0.0037	0.0029
3	0.92	0.86	24.02	24.42	0.0015	5.92e-04	0.0033	0.0030
4	0.96	0.97	16.11	15.04	4.07e-04	4.62e-04	0.0015	0.0014
5	0.98	0.98	23.52	14.53	-0.0022	-3.83e-04	0.0042	0.0025
6	0.95	0.96	17.67	15.79	-5.17e-04	2.14e-04	0.0023	0.0021
7	0.98	0.99	11.44	9.07	5.52e-04	5.70e-05	0.0020	0.0016
8	0.98	0.99	10.95	9.85	-3.10e-04	4.17e-05	0.0023	0.0020
Average *	0.97	0.97	15.74	13.43	-2.25e-05	0.00013	0.0024	0.0020

* Averages for by cruise model exclude Cruise 1.

This comparative breakdown also offers an insight as to how the two types of model perform on individual cruises. Aside from the unusual cruise 1 each cruise is better predicted where the model is trained specifically on that cruise. This may be attributed to implicit learning of cruise specific package effects. The difference in performance is most evident for cruise 5

where the change in percentage error is almost 9%. The errors from the by cruise model are quite low suggesting that the difference in performance might be caused by a stronger presence of cruise specific attributes relative to the other cruises. This could be an external non-modelled variable that has greater influence on the fifth cruise.

An alternative approach is to carry out the modelling in log space. The variables are transformed by taking their natural logarithm prior to any other processing. The model is applied as normal and the transform later reversed with an exponential transform. The correlations are calculated directly in log space, but to allow some degree of comparability the predictions in log space are exponentially transformed before calculation of the errors. Although this reversal of the original transform may introduce a small bias (Smith & Barnett, 2004) it will give a close approximation to the errors to help establish whether the transform is useful.

Results for four overall models in log space and the corresponding by cruise model are shown in table 3.4.1.4. Results produced by the overall models are very similar to those using the standard, untransformed space and modelling by cruise similarly improves predictions. Relative to the standard space the by cruise averages tend to be slightly better correlated and produce smaller errors. Also the incorporation of the extra size inputs seems more successful when working within the log space. The differences overall however are not significant enough to conclude that either space is superior particularly as transformation bias has not been removed. At this point the GLM by cruise in linear space with three size inputs (table 3.4.1.2) is the best performing model.

Table 3.4.1.4: Bayesian forward model errors - log space models. All use 10 PC representation of spectra. MLP has 8 hidden units.

	MLP No size		**MLP Size index**		**GLM No size**		**GLM 3 size inputs**	
	Overall	**By cruise***	**Overall**	**By cruise***	**Overall**	**By cruise***	**Overall**	**By cruise***
CORR	0.967	0.973	0.968	0.970	0.959	0.977	0.963	0.978
Mean % error	16.06	12.52	15.18	11.95	18.39	12.28	17.03	11.70
Bias	-1.57e-04	-3.05e-05	-1.31e-04	1.40e-05	-1.43e-04	-5.06e-05	-1.41e-04	-3.99e-05
MAE	0.0014	0.0011	0.0013	0.0011	0.0016	0.0011	0.0015	0.0010
RMSE	0.0025	0.0018	0.0024	0.0019	0.0029	0.00178	0.0027	0.0017

*MLP by cruise averages exclude cruise 1 due to extreme error.

3.4.2 Bayesian Modelling 2: By wavelength

An alternative network model predicts absorption separately at each wavelength. The inputs are the concentrations and optional size data as before, but the output is a single prediction of total absorption at a given wavelength. Iteration produces a prediction at each wavelength to give a complete reconstruction of each spectrum.

The resulting predictions are some of the best thus far in terms of both correlations and errors and the best overall models in both linear and log space are produced. The by cruise averages also improved further, though predictions for cruises with few samples remain erratic. In particular for cruise 1 and sometimes also cruise 2 the algorithm completely collapses as before. Using the log transform in this model improves predictions and avoids collapse for cruise 2, though cruise 1 remains problematic containing only two test samples. There does not appear to be any benefit to modelling by wavelength for the GLM relative to the previous models.

Results for the most successful experiments are seen in table 3.4.2.1 alongside the previous best model for comparison purposes. The log transform on the by wavelength model produces the best model so far, though it only marginally outperforms the previous best GLM.

Table 3.4.2.1: Bayesian forward model errors – by wavelength models.

	MLP by wavelength Size index linear space **(best overall model)**	**MLP** by wavelength 3 size inputs log space **(best overall log space model)**	**MLP** by wavelength by cruise* No size inputs log space **(best model so far)**	**GLM** by wavelength by cruise* No size inputs linear space	**GLM** 10 PCs by cruise* 3 size inputs linear space **(best GLM-previous best model)**
CORR	0.971	0.971	0.980	0.982	0.982
Mean % error	14.23	14.44	10.74	11.40	11.00
Bias	-4.04e-005	-1.48e-004	-5.38e-005	3.16e-005	6.48e-005
MAE	0.0013	0.0013	0.00091	0.0010	0.00098
RMSE	0.0023	0.0023	0.0016	0.0017	0.0017

*MLP by cruise errors are averages per cruise excluding cruise 1. GLM by cruise errors are averaged across all cruises.

3.4.3 GLM comparison

The GLM performs quite consistently, giving similar results whether trained by wavelength or using the full projected spectra (see tables 3.4.1.1 – 3.4.2.1). For the basic versions of both models the correlations are slightly smaller and errors significantly bigger than those found using the non-linear models. However, training the GLM by cruise is much more successful and even gives better average results than the corresponding MLP in many cases. The GLM also does not encounter difficulties in modelling any individual cruise and is fast to train. Table 3.4.2.1 illustrates how close the GLM (particularly in linear space) comes to matching the optimal MLP at this stage.

Compared to the MLP-based networks inclusion of size inputs improves results more widely when using the GLM. Log transformation improves results for the overall models whether using PCA or modelled by wavelength, but appears to have detrimental effect for the 'by cruise' variants. The improvement from including size seems greater when working in log space, but the best predictors are derived in linear space.

The good performance of the GLM relative to the non-linear models suggests that a large part of the basic relationship is linear. It may only be additional factors such as noise and acclimation effects that introduce non-linearity. Overall the MLP log space model is marginally more effective (table 3.4.2.1) indicating that some non-linear element may exist.

Having determined the best forward models using the standard Bayesian evidence approach, several of these will be adapted to incorporate Automatic Relevance Determination (ARD).

3.5 Automatic Relevance Determination (ARD)

ARD is a method for identifying the most significant inputs and automatically adjusting weights to reflect this. Thus, it may directly limit the number of input variables, such that only relevant ones are retained. Alternatively, calculation and analysis of ARD coefficients provides a relative measure of input importance. This may guide the complete removal of irrelevant inputs from the dataset, thereby eliminating all noise contributed by unnecessary inputs.

ARD coefficient analysis is also potentially useful when inverting and producing models for concentration retrieval. Analysis by wavelength may determine significant inputs for prediction in each part of the spectrum. For example, if a pigment is identified as irrelevant in predicting absorption between 600 and 700nm then it may be assumed that these absorption measurements are irrelevant when it comes to the reverse problem of predicting concentrations.

The high dimensionality of the data means ARD has the potential to significantly improve model performance by limiting the noise in both the forward and inverse models. It may also determine the usefulness of specific pigments and size data for predicting absorption and thus has implications for any further modelling.

3.5.1 Implementing ARD

ARD is implemented by adapting the prior so that each input variable has a separate regularisation coefficient. The weights associated with each input then have independent Gaussian prior distributions, each governed by a hyperparameter representing the inverse variance of the corresponding weights. The hyperparameters or ARD coefficients then indirectly control the magnitude of the weights. A small hyperparameter specifies large variability in the weights, which means they are likely to be big and the associated input important.

ARD works iteratively within the evidence framework (described in section 3.1.4) gradually re-estimating hyperparameters to reflect the degree of relevance of inputs. Normalisation of inputs to zero mean and unit variance before training ensures comparability of the coefficients.

3.5.2 ARD Priors for Best Forward Models

Re-training the best models identified previously with ARD priors has almost no effect on the errors. For each model implemented MAE and RMSE do not change (to the 4 d.p. accuracy

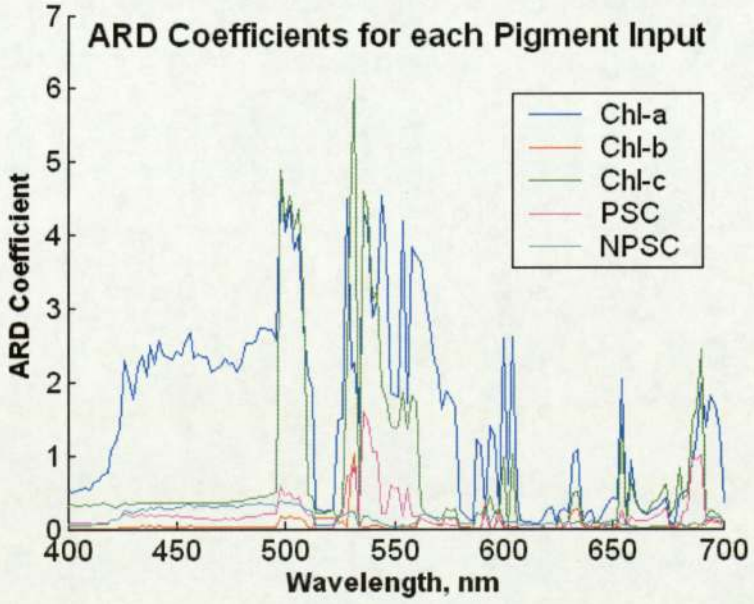
recorded). The only changes are small improvements to percentage errors (less than 0.03%) and to the bias (less than 2 x 10^{-6}). Individual cruise models show similarly marginal improvements for some cruises yet experience further problems of instability and extreme errors.

3.5.3 ARD Coefficients – Wavelength Analysis

The ARD coefficients are the hyperparameters and so their inverse gives a measure of importance of the corresponding input. Using Bayesian training to predict absorption from the concentration inputs at each individual wavelength generates a sequence of ARD coefficients for each pigment. Variations include incorporating size data, using log transformation of variables and using different network structures and optimisation parameters.

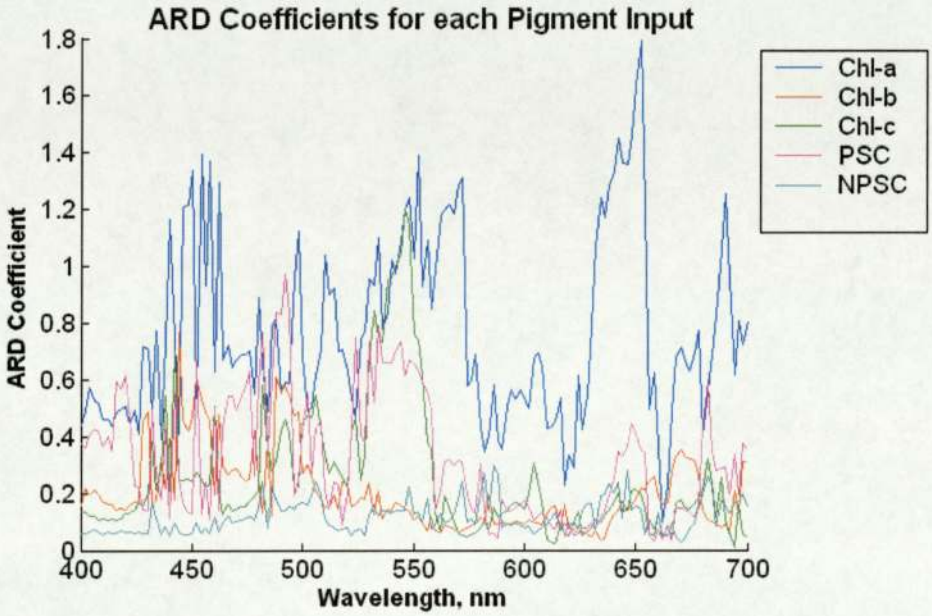
Results vary greatly depending on the particular model and parameters chosen, but most seem to agree that certain pigments have little importance for absorption prediction in certain areas of the spectrum. In the standard space, as expected ARD finds chlorophyll-a to be dominant throughout the whole of the visible spectrum (see figure 3.5.3.1). In contrast, the NPSCs and chlorophyll b have a much smaller contribution to overall absorption. PSCs also have much less significance but display two peaks at around 540nm and 690 nm.

Figure 3.5.3.1: ARD Coefficients for the MLP with 3 hidden units in linear space and 1000 optimiser iterations.



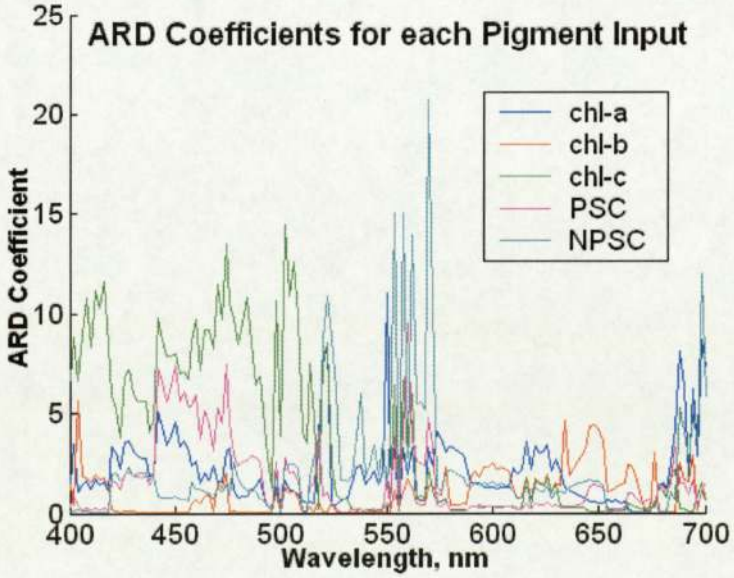
Using eight hidden units however, chlorophyll b no longer appears insignificant in most areas of the spectrum. The more complex network seems able to capture the more subtle predictive information provided by the chlorophyll b input. Chlorophyll a displays a similar dominance but the effect of other pigments, particularly chlorophyll b and the PSCs is magnified.

Figure 3.5.3.2: ARD Coefficients for the MLP with 8 hidden units in linear space and 100 optimiser iterations.



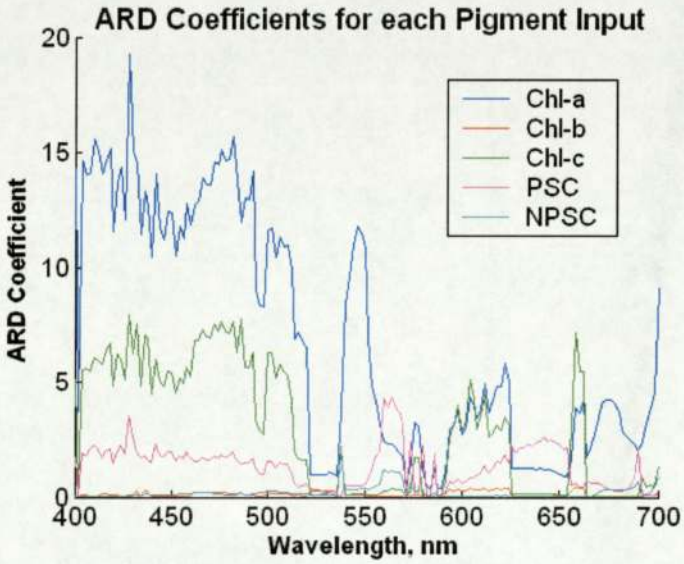
Log space experiments however give very different results suggesting much less influence of chlorophyll a. Chlorophyll c dominates in the first half of the measured spectra and each of the other pigments appears to dominate a small part of the spectrum at higher wavelengths.

Figure 3.5.3.3: ARD Coefficients for the MLP with 3 hidden units in log space and 1000 optimiser iterations.



Using a log transform on the spectra, thus constraining the absorptions to be positive, whilst retaining linear space for the concentrations gives different results again. This experiment is again consistent with the dominance of chlorophyll a, yet indicates greater influence of the PSCs and chlorophyll c (see figure 3.5.3.4). This is especially true for the PSCs at higher wavelengths and both pigment groups become the dominant input at certain wavelengths.

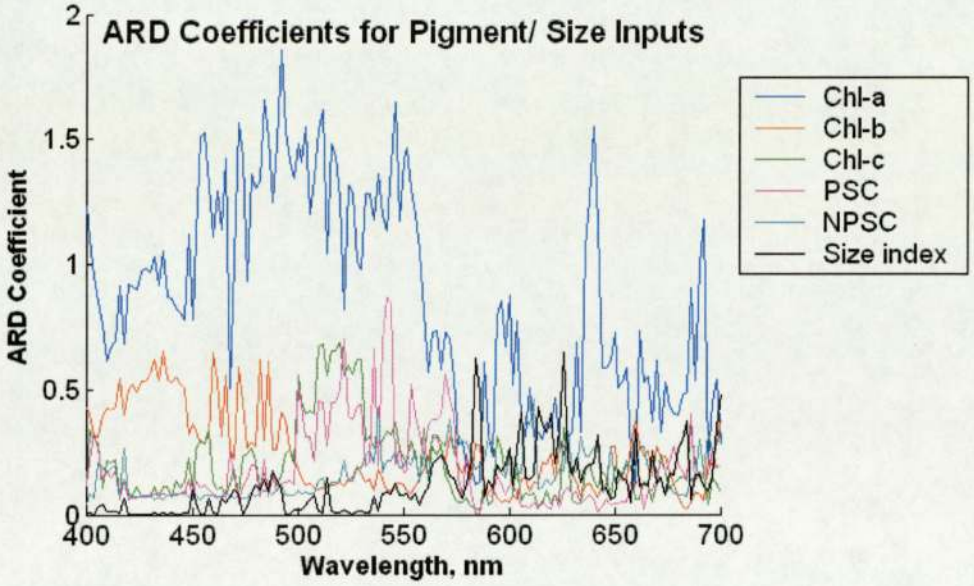
Figure 3.5.3.4: ARD Coefficients for the MLP with 3 hidden units, concentration in linear space, log transformed spectra and 1000 optimiser iterations.



Running ARD on both concentration and size inputs mostly show the concentrations to be much more influential (see figure 3.5.3.5). The size index has its greatest impact at higher wavelengths between 575 and 650nm where it becomes more important than several pigments for certain network structures. Plotting ARD coefficients for the pigments and all three size inputs it becomes difficult to assess the effect of individual size inputs.

ARD coefficients vary vastly with both network structure and the parameters of the optimiser. It is difficult then to draw many firm conclusions and this remains a possible avenue for further exploration. Considering alternative initialisations of each network and/or averaging across models may retrieve more a more stable result or decide that noise is dominant. Another possibility is to iteratively apply ARD and then remove the implied least important inputs. This would gradually reduce network sizes and may more accurately remove noise sources. Despite this instability however, recurring themes include the dominance of chlorophyll a across the spectrum as expected and the small but present influence of size inputs particularly at higher wavelengths.

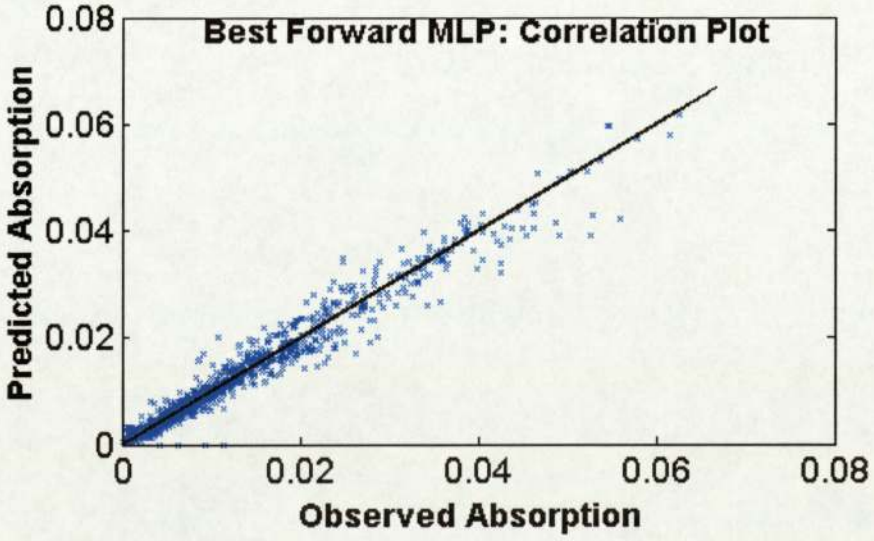
Figure 3.5.3.5: ARD Coefficients for the MLP with 8 hidden units in linear space and 100 optimiser iterations.



3.6 Optimal Reconstruction Models

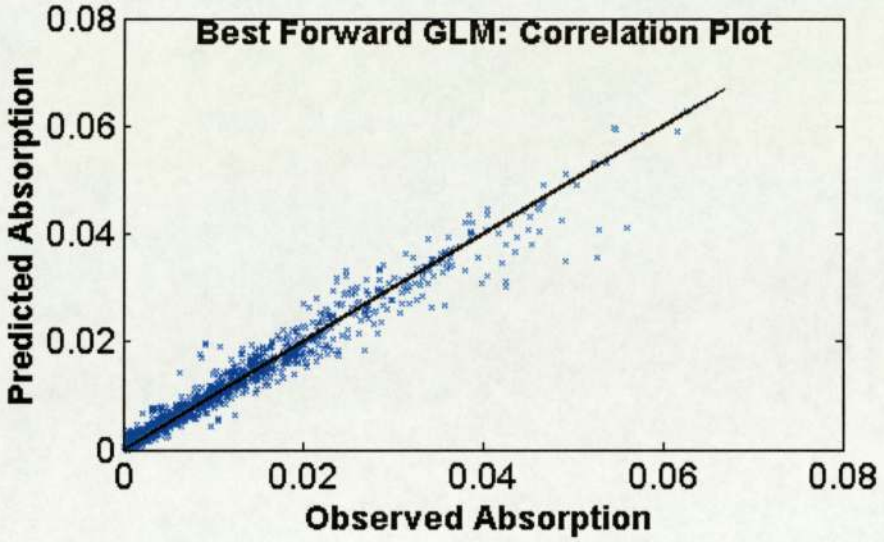
The small gains from using ARD priors appear insignificant relative to the instability introduced into the models. The best performing models across the monitored statistics then are those detailed previously in table 3.4.2.1. The reconstruction performance of two of these is illustrated graphically in figures 3.6.1 and 3.6.2. Predicted versus observed values of absorption are plotted for six feature wavelengths selected using the spectra visualisation.

Figure 3.6.1: Predicted versus observed absorptions for the best MLP - by wavelength, by cruise, in log space, no size inputs.



Note: Absorptions are plotted for all test samples at six selected feature wavelengths (430, 470, 540, 590, 640 and 674nm)

Figure 3.6.2: Predicted versus observed absorptions for the best GLM - by cruise, in linear space, three size inputs.



Note: Absorptions are plotted for all test samples at six selected feature wavelengths (430, 470, 540, 590, 640 and 674nm).

The two plots actually appear very similar to the extent that the errors seem to occur on the same samples. This may suggest excessive noise or that factors such as the package effect are more prevalent in these samples, as both models predict absorption badly. Visual similarity of the plots, although subjective, is further evidence that the MLP does not add much predictive ability relative to the GLM.

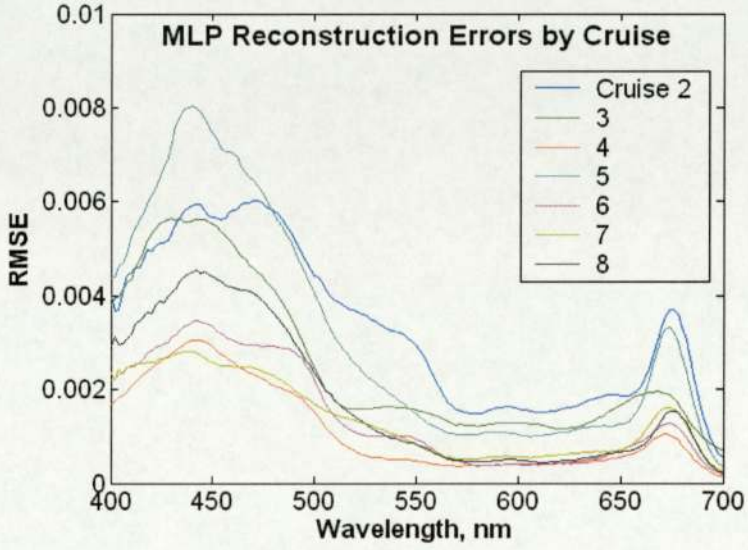
Error magnitudes appear only slightly larger where observed absorption is greater suggesting that there may be a proportional element to errors, but that they are largely additive. Both figures also indicate some negative bias not obvious (particularly for the GLM) from the statistics alone. This bias is most evident for larger absorptions though does not appear extreme. The structure of errors is investigated further in the following section (3.7).

3.7 Error Structure

Plotting the errors found across the spectrum offers further insight. Error magnitudes for each model, whether trained at individual wavelengths or using PCA reduced spectra, reveal a structure reflecting the mean spectrum, as in the example in figure 3.7.1. The errors then do appear proportional to absorption. This apparently multiplicative error may favour the use of log transformed spectra in further models, such that errors would become additive. This

structure also means that the largest errors occur at lower wavelengths between 400 and 500nm (nanometers). However, whilst error magnitudes are smaller in certain areas of the spectrum particularly around 550 to 650 nm, absorption varies correspondingly such that relative errors actually tend to be higher.

Figure 3.7.1: Error Structures by cruise (on the test set) for the MLP by cruise, linear space, three size inputs, 10 PCs.



Note: Cruise 1 is omitted due to unstable and non-typical behaviour

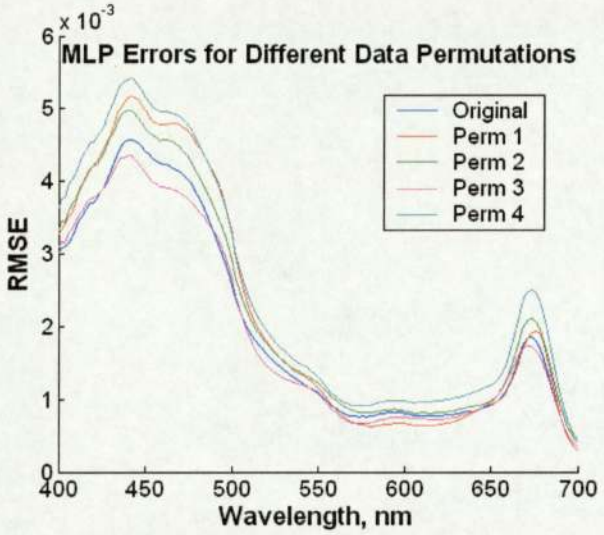
Comparing the errors for each cruise as in figure 3.7.1 reveals some structural differences. Some display individual characteristics, such as slight differences in shape and positioning of the major peaks and possible secondary peaks. This may be related to differences in concentration, but might also indicate certain differences in combinations of pigments or additional factors, such as the package effect as modelled by Morel and Bricaud (1981). This is further evidence to support modelling by cruise or at least inclusion of a cruise related component for the inverse model.

3.8 Permutations

The previous forward modelling experiments have indicated that problems occur in some cases where there are few samples available. This suggests possible dependence on the split of the data. To investigate this several of the models will be applied to four new divisions of data set where the samples are designated to test and training sets by random methods (for a breakdown see Appendix A.6).

Experiments for the overall models on each permutation of the data show the error magnitudes to display a structure similar to the mean as before (see figure 3.8.1). The errors for each permutation are very closely correlated with those of the original set. There are however small differences in magnitude throughout the spectrum, which are most prominent at the lower wavelengths between 400 and 500nm. The same relation is true for the variance for each set of predictions. These conclusions remain unchanged for models including size inputs.

Figure 3.8.1: Error Structures (on the test set) for the MLP in linear space, no size inputs, 10 PCs using the original designated data set and four random permutations.



The effect of the permuting the data set is much more apparent however in the models by cruise. The performance on several individual cruises is clearly much more sensitive to the split of training and test set, as is the case for cruise 2 (see fig 3.8.2). Error magnitudes vary significantly here with each permutation.

Error structures for a given cruise though are mostly of very similar shape and resemble the mean. All permutations for Cruise 6 (figure 3.8.3) for example display the same secondary error peak around 550nm. Between cruises however, there are obvious differences in structure as well as magnitude. This is seen by comparing just two cruises using figures 3.8.2 and 3.8.3 and is further evidence of additional cruise related differences, as well as varying pigment concentrations.

Figure 3.8.2: Error Structures (on the test set) for the MLP by cruise, in linear space, no size inputs, 10 PCs using the original designated data set and four random permutations.

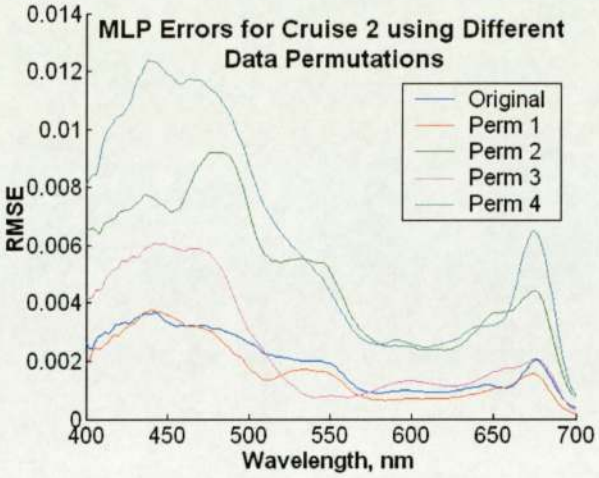
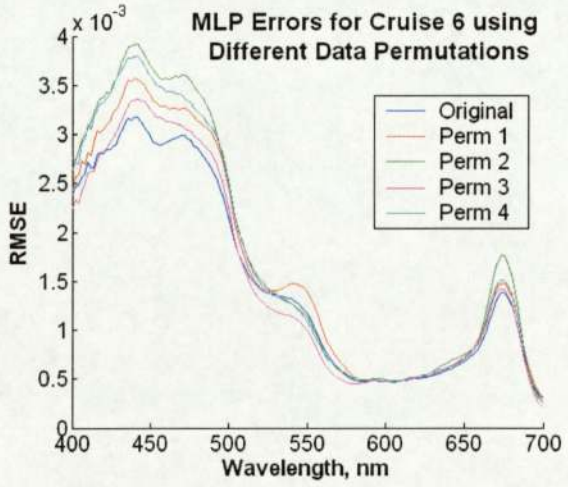


Figure 3.8.3: Error Structures (on the test set) for the MLP by cruise, in linear space, no size inputs, 10 PCs using the original designated data set and four random permutations.



Several permutations however do result in some unusual features specific to that permutation and cruise. The original data set for example produces for cruise 3 a secondary error peak around 525 nm, which is either much more subdued or absent for the other permutations (see figure 3.8.4). Interestingly this is the split with least test data samples for cruise 3, suggesting the error may be disproportionately skewed by an unusual sample. As expected such differences tend to be most prominent where there are few samples in the given cruise and where the biggest changes in sample size occur. This is confirmed by comparing the average RMSE for each cruise and permutation of data.

Figure 3.8.4: Error Structures (on the test set) for the MLP by cruise, in linear space, no size inputs, 10 PCs using the original designated data set and four random permutations.

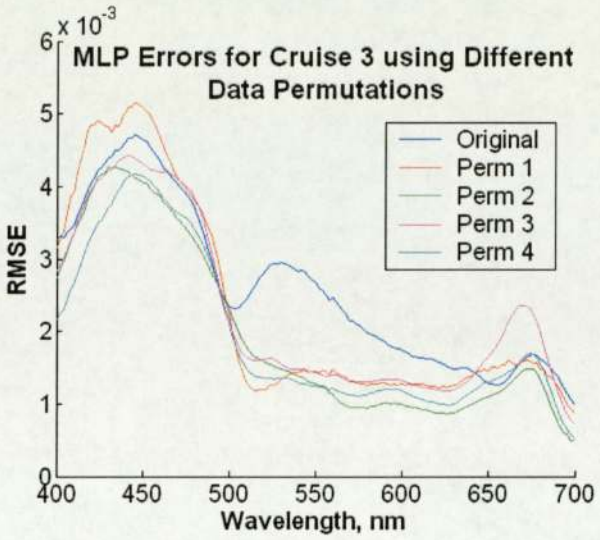
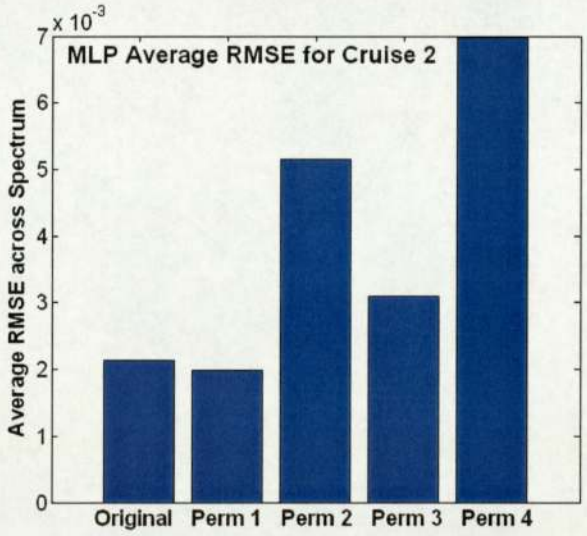


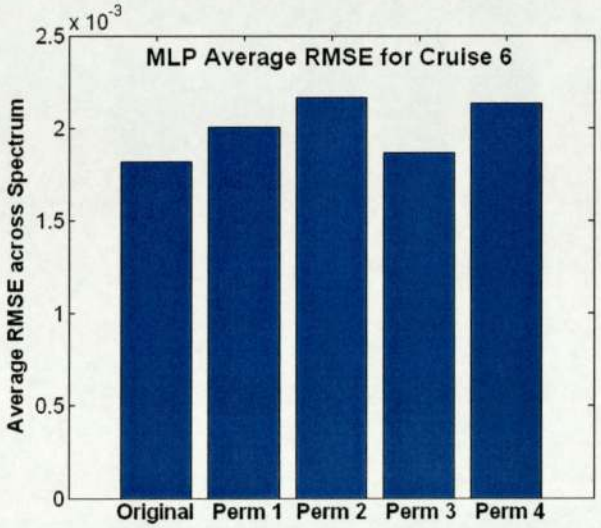
Figure 3.8.5 shows the average errors for cruise 2 – the smallest cruise except for cruise 1. The errors display more inconsistency than any of the other cruises and vary by up to 0.005 between data sets. The error is greatest on the fourth permutation where the number of data samples designated for training is least.

Figure 3.8.5: Average Reconstruction errors (on the test set) for the MLP by cruise, in linear space, no size inputs, 10 PCs using the original designated data set and four random permutations. Average error across all wavelengths for cruise 2.



In contrast, errors for the largest cruise (cruise 6) are plotted in figure 3.8.6. These have the least variability of errors – less than 0.0005 across the permutations tested.

Figure 3.8.6: Average Reconstruction errors (on the test set) for the MLP by cruise, in linear space, no size inputs, 10 PCs using the original designated data set and four random permutations. Average error across all wavelengths for cruise 6.



The extent of these effects of permuting the data on several cruises mean it may be necessary to model this as a further source of uncertainty resulting in extended confidence intervals.

3.9 Forward Modelling (Spectra Reconstruction) Conclusions

Good reconstruction: Models reconstruct spectra from pigment concentrations with correlations in excess of 0.98.

Network structure: Eight hidden units appear to be an appropriate forward MLP framework so this may be a starting point for the inverse model.

GLM performance: Impressive results using the GLM compared to the MLP suggest some potential for a linear element to models. The small differences though do suggest that the MLP implicitly captures at least some of the natural variability due to package and acclimation effects.

Log transform: The transform appears to improve model performance though the effect is small for the forward model. Proportional error structures also indicate that transform of the spectra may be useful.

Size distribution data: Experiments are inconclusive as to the benefits of incorporating the additional size data. ARD suggests limited relevance yet there are some positive effects on model performance, so its inclusion will be considered for the inverse models.

Important pigments: ARD confirms that the pigments have varying importance with regard to predicting total absorption and identifies chlorophyll a as the dominant pigment. This may suggest a need to model individual pigments separately. There is potential to further explore ARD output, particularly an iterative application.

ARD priors: Models incorporating ARD tend to improve predictions but the effect is marginal.

Dataset dependence: The sensitivity of models to alternative selections for training and test data indicate further uncertainty in the data possibly to be incorporated in the modelling framework.

Chapter 4

Concentration Retrieval – Data Driven Methods

The project will now investigate the retrieval of concentrations. Concentration retrieval will begin with data driven networks based on direct inversion of the forward models and then later focus on a generative absorption model. In both cases a Bayesian approach will again be adopted, as the same uncertainties exist.

The data driven retrieval models are basically reversals of the 'forward' networks. These models will take spectral data and various other inputs with the ultimate aim of accurately retrieving the concentrations of one or more of the five pigment groups. As with the forward models they can be trained on individual cruises and incorporate PCA, log transformations and ARD priors. The natural variability due to package effect and photoacclimation is modelled implicitly by the data driven approach.

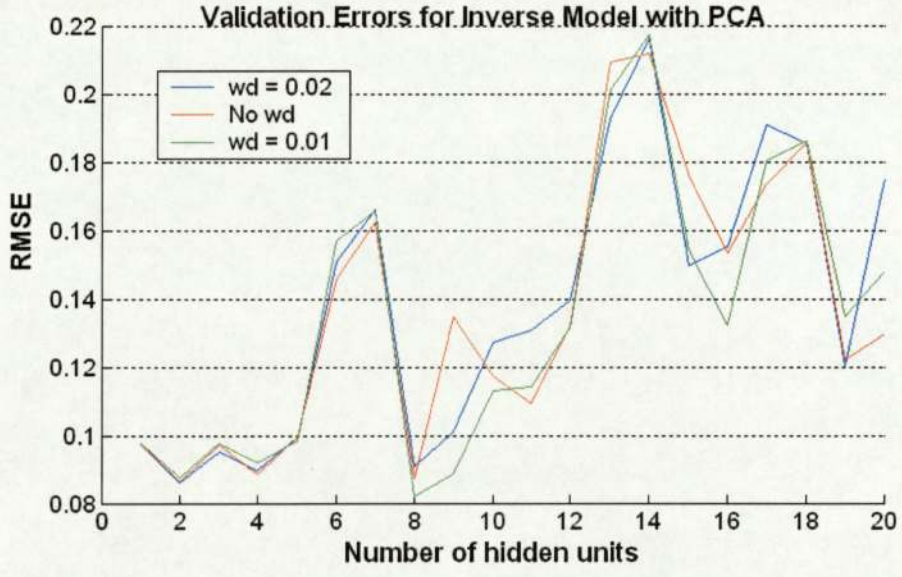
4.1 Validation

As with the forward models the first step will be a brief validation experiment. This may give some insight as to whether it is useful to include size distribution data, what prior might be appropriate and primarily the number of hidden units to be used for an MLP-based model. As before several models will be used with various combinations of input, normalisation and weight decay coefficient. Outputs will be concentration predictions for a single pigment or all five pigments.

4.1.1 Validation Outcomes

The first model takes as inputs the ten dimensional PC projections of the spectra and (optionally) size data and predicts the five pigment concentrations for each sample. For the basic model taking only absorption inputs the main minimum occurs at eight hidden units with and without weight decay (see figure 4.1.1.1). Weight decay has some effect on errors and correlations and generally improves predictions. This indicates that even a simple weight decay prior will prove useful. Incorporating some form of projected size data reduces errors for certain network architectures shifting the minimum to five to six units (see figure 4.1.1.2). Using in excess of ten units the errors start to magnify, so eight units appear to be the best structure for this particular model.

Figure 4.1.1.1: Validation errors for direct inverse model with PCA with and without weight decay. No size inputs, 1000 optimiser iterations.



Networks trained to retrieve individual pigment concentrations favour slightly different networks. For Chl-b and the NPSCs RMSE is minimised at six units while for Chl-c four units are optimal and for the PSCs thirteen units produce the minimum (see figure 4.1.1.3). Chl-a retrieval errors have three similar minima at nine, thirteen and sixteen units. The separate predictions reveal that although similar network structures may be optimal for selected pigments the magnitude of errors varies significantly depending on the pigment being retrieved. The Chl-b, Chl-c and NPSC pigment groups score mean absolute errors less than

0.05 for each structure implemented (with two or more hidden units). In contrast the Chl-a and PSC mean errors are more than twice as large and are largely in the range 0.08 to 0.12.

Figure 4.1.1.2: Validation errors for direct inverse model using various size inputs, 0.01 weight decay coefficient, 1000 optimiser iterations.

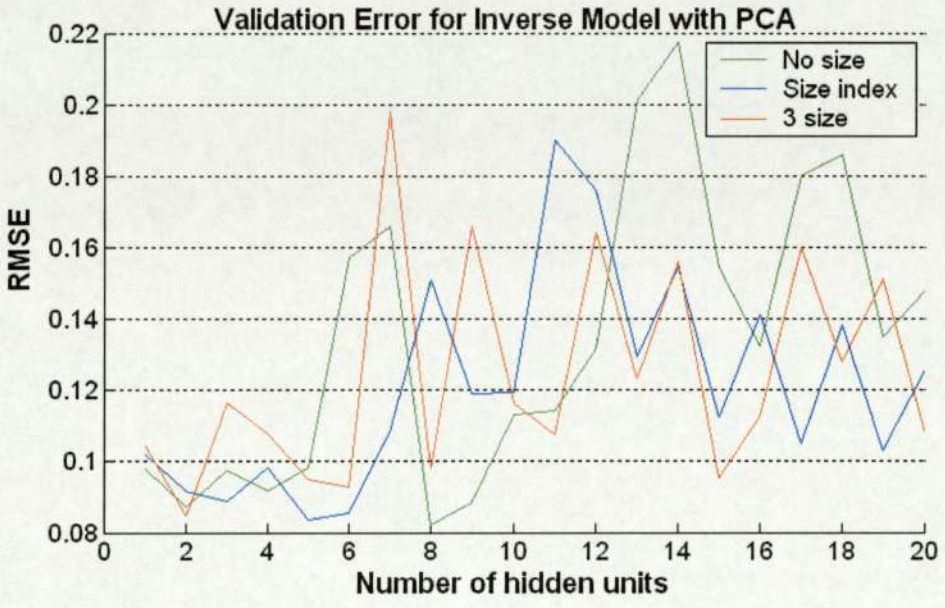
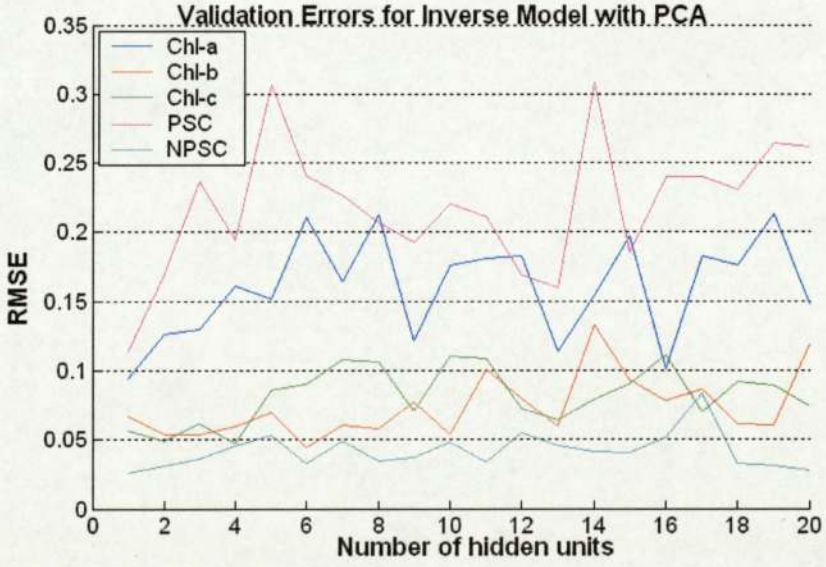
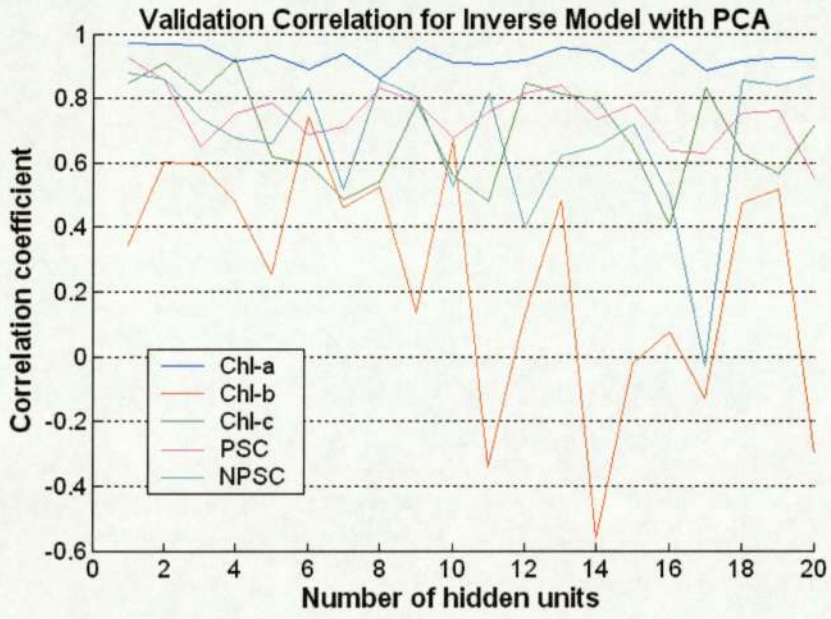


Figure 4.1.1.3: Validation errors for direct inverse model with PCA retrieving concentrations for individual pigments. No size inputs, and 0.01 weight decay coefficient. 1000 iterations.



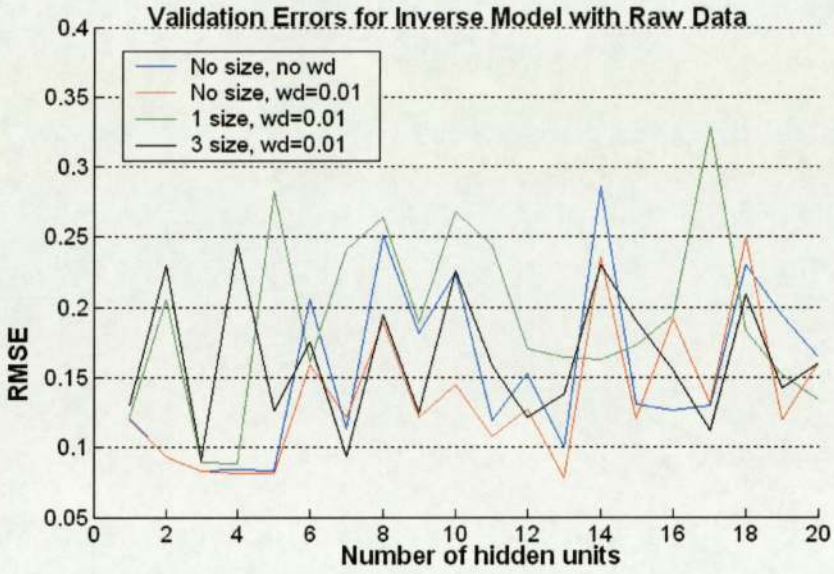
In terms of correlations however, chlorophyll a retrieval is far superior and significantly poorer results are achieved for other pigments, particularly chl-b and NPSCs (see figure 4.1.1.4). The results are highly erratic for chlorophyll b suggesting there may be large errors in the data or that the relationship between absorption and this pigment is rather weak. Nevertheless, maximum correlation for all but chl-c is achieved using between six and nine units. Average reconstruction errors and correlations across all pigments remains close to those found using a single model, so the individual retrieval does not appear to greatly benefit predictive capability.

Figure 4.1.1.4: Validation correlations for direct inverse model with PCA retrieving concentrations for individual pigments. No size inputs, 0.01 weight decay coefficient, 1000 optimiser iterations.



A second MLP is constructed taking all 151 absorption measurements as direct inputs. In the absence of size inputs the minimums coincide at five and thirteen hidden units. Incorporating the size index reduces correlations and increases errors both with and without weight decay. Using the three size measures has less impact but also seems to increase errors. For the models including size the minimums are even less clear at around three to seven units.

Figure 4.1.1.5: Validation errors for direct inverse model taking raw inputs, with different size inputs and weight decay coefficient combinations, 2000 optimiser iterations.



Training these models to retrieve pigments individually results in large error variances again (see figure 4.1.1.6). This difference in magnitude of errors means that certain pigments dominate with regard to where overall minimums are situated. Errors for Chl-a and the PSCs are again significantly larger and thus their minimums coincide with those of the overall model at five and thirteen units respectively. Chl-a predictions are again the most closely correlated whilst results for chl-b are much worse relative to the other pigments (see figure 4.1.1.7).

Figure 4.1.1.6: Validation errors for direct inverse model taking raw inputs and retrieving pigments individually. No size input and 0.01 weight decay coefficient. 500 iterations.

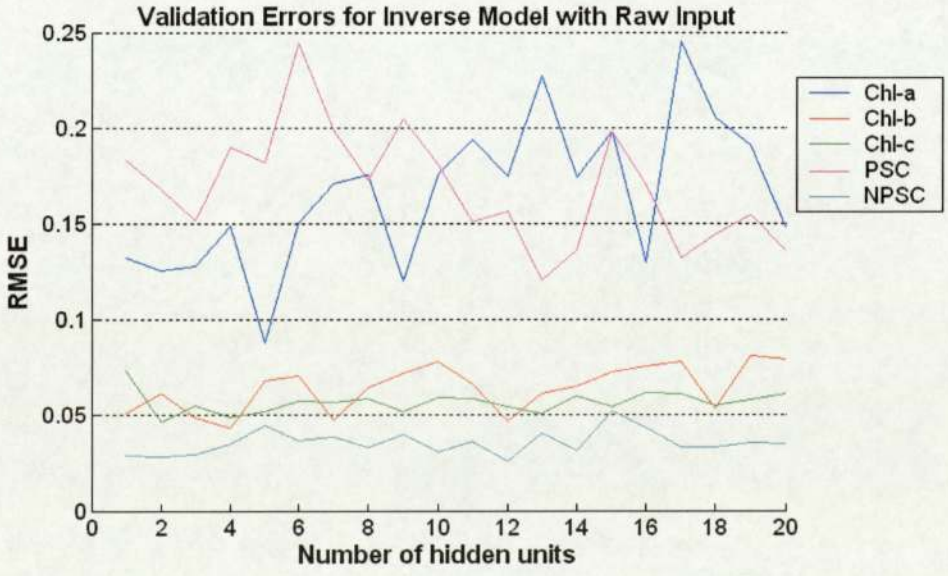
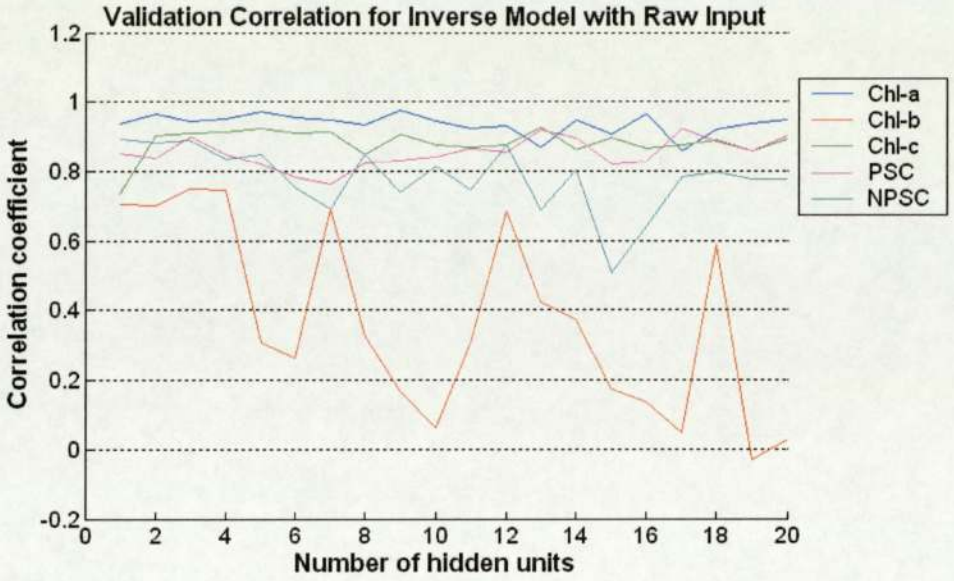


Figure 4.1.1.7: Validation correlations for direct inverse model taking raw inputs and retrieving pigments individually. No size input and 0.01 weight decay coefficient. 500 iterations.



Comparing the two main types of model (with raw data input and projected inputs) reveals interesting differences. The error magnitudes using the raw inputs generally seem bigger for the majority of structures tested (between one and twenty units) - possibly a result of the more complex network necessary. Results from the raw input model however, appear more stable with respect to the number of hidden units and display less sharp error peaks.

As both models have potential yet display quite different minimums two different starting structures for the direct inverse networks will be preferred: nine hidden units for the projected data model and thirteen units for the raw data model. Size data experiments will continue as validation and previous experiments remain inconclusive as to whether the size data provides a useful input variable. The variation in errors in the retrieval of individual pigments suggests that individual models ought to be considered further.

Errors as a whole are larger and generally more erratic than for the forward models. The similar size of network indicated by validating the first MLP, suggests some parallels with the forward problem and is indicative of linearity in the mapping. As with validation of the forward models there is potentially a more stable result to be achieved by averaging across many network architectures and initialisations. The analysis however appears to provide a reasonable first estimate.

4.2 Bayesian Pigment Concentration Retrieval

As a result of validation the following three model structures are implemented:

1 MLP – 9 hidden units
Input dimension reduced to ten using PCA

2 GLM – (no hidden units)
151 raw (normalised) spectral inputs

3 GLM – (no hidden units)
Input dimension reduced to ten using PCA

Initially the second model was an MLP with 13 hidden units taking raw inputs, but this proves to be too complex and collapses. Given the success of the GLM in forward modelling a second GLM taking raw inputs is to be used. The basic experiment will take only spectral inputs and retrieve all five pigments simultaneously. The three models produce the following results (table 4.2.1):

Table 4.2.1: Average Concentration Retrieval Errors for all Pigments

Model	1 – **MLP with PCA**		2 – **GLM, raw input**		3 – **GLM with PCA**	
Normalisation Method	MU	SD	MU	SD	MU	SD
CORR	0.973	0.974	0.975	0.976	0.970	0.970
Mean % error	31.07	30.56	29.91	29.90	32.52	31.50
Bias	-5.52e-04	0.0015	0.0012	5.01e-04	5.91e-04	6.46e-04
MAE	0.0250	0.0240	0.0238	0.0237	0.0255	0.0252
RMSE	0.0408	0.0399	0.0391	0.0379	0.0428	0.0425

Note: MU refers to normalisation removing mean, while SD normalisation also divides by the standard deviation

As in the majority of previous experiments the SD normalisation improves results for both the GLM and non-linear models. All three produce correlations close to 0.97 and perform similarly on the error measures. The GLM with raw inputs is consistently slightly better for each measure followed by the MLP with PCA and then the GLM with PCA. This suggests that the latter are not complex enough at this stage to capture all the available information in the mapping.

4.3 Individual Pigment Concentration Retrieval

Adapting the above models to retrieve each pigment individually has varying effects (see table 4.3.1). The GLM with PCA appears not to be affected at all by the change. For the GLM with raw input, which performed best in the previous experiment, RMSE decreases yet MAE becomes worse. In contrast the MLP results improve such that MAE is better than for all the other models. The GLM with raw input though remains the best predictor regarding correlation and RMSE. Individual retrieval has some positive effect then in the non-linear case, but has little if any beneficial effect on the GLM.

Table 4.3.1: Average Concentration Retrieval Errors for models trained by pigment (All SD normalised)

Model	1 – **MLP with PCA**	2 – **GLM, raw input**	3 – **GLM with PCA**
CORR	0.975	0.976	0.970
Mean % error	26.13	30.49	31.48
Bias	0.0018	4.95e-04	6.46e-04
MAE	0.0221	0.0240	0.0252
RMSE	0.0397	0.0382	0.0425

The following table (4.3.2) breaks down the results of the two PCA based models by pigment. Each individual pigment is better predicted by the MLP except for the PSCs where RMSE is actually less using the GLM. This may indicate that this particular pigment has a more linear relation to the absorption spectra relative to the other pigments and might support the use of different models for the different pigment groups.

Table 4.3.2: Concentration Retrieval Errors for models trained by pigment - breakdown by pigment (PCA-based models)

Pigment	**CORR**		**Mean % error**		**Bias**		**RMSE**	
Model	**MLP**	**GLM**	**MLP**	**GLM**	**MLP**	**GLM**	**MLP**	**GLM**
Chl-a	0.98	0.97	19.55	25.57	0.0052	7.41e-04	0.056	0.067
Chl-b	0.91	0.89	30.77	35.43	-8.65e-05	0.0011	0.022	0.025
Chl-c	0.94	0.92	31.91	40.06	0.0014	-0.0012	0.022	0.025
PSC	0.94	0.95	27.90	31.87	0.0028	0.0020	0.059	0.053
NPSC	0.91	0.84	22.84	27.29	-2.34e-04	5.96e-04	0.018	0.023

Finally, the MLP is compared to the corresponding model retrieving the pigments simultaneously. Each pigment tends to be equally well or better predicted by individual retrieval, though for the PSCs RMSE actually increases. This may suggest that there is simply more noise in the PSC measurements. Both breakdowns show chl-a to be most accurately retrieved in terms of the percentage errors.

Table 4.3.3: Concentration Retrieval Errors - comparative breakdown by pigment (MLP with PCA)
A refers to the model retrieving all pigments simultaneously and **B** to individual pigment retrieval.

Pigment	**CORR**		**Mean % error**		**Bias**		**RMSE**	
Model	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Chl-a	0.97	0.98	22.41	19.55	0.0030	0.0052	0.064	0.056
Chl-b	0.92	0.91	31.97	30.77	-5.70e-04	-8.65e-05	0.021	0.022
Chl-c	0.94	0.94	33.68	31.91	7.46e-04	0.0014	0.022	0.022
PSC	0.96	0.94	39.69	27.90	0.0041	0.0028	0.051	0.059
NPSC	0.88	0.91	25.70	22.84	3.91e-04	-2.34e-04	0.020	0.018

4.4 Separate Cruise Models

Training networks on individual cruise data sets for both the MLP and the raw input GLM, correlations and RMSE become worse, while MAE and percentage errors improve (see below). The MLP also encounters problems with the first cruise, lacking in samples.

Table 4.4.1: Average Concentration Retrieval Errors by cruise (All SD normalised, not individual pigment retrieval)

Model	1 – **MLP with PCA**	2 – **GLM, raw input**	3 – **GLM with PCA**
CORR	0.971	0.966	0.9813
Mean % error	23.45	28.71	23.02
Bias	0.0031	0.0029	0.0012
MAE	0.0218	0.0235	0.0195
RMSE	0.0435	0.0472	0.0337

Note: MLP results do not include cruise 1

For the GLM with PCA however, both correlation and errors significantly improve. It becomes the best model based on correlation, MAE, RMSE and also mean percentage error which falls to its lowest at 23%.

In addition, models were trained both by cruise and by pigment. Retrieving individual pigment concentrations for each cruise the performance of the raw input model actually deteriorates on each error measure relative to the by cruise only model. Minimal negative changes occur for each cruise and similarly for each pigment group with a small adverse effect overall. For the GLM with PCA by cruise, retrieving pigments individually changes results only marginally- some becoming better and others worse.

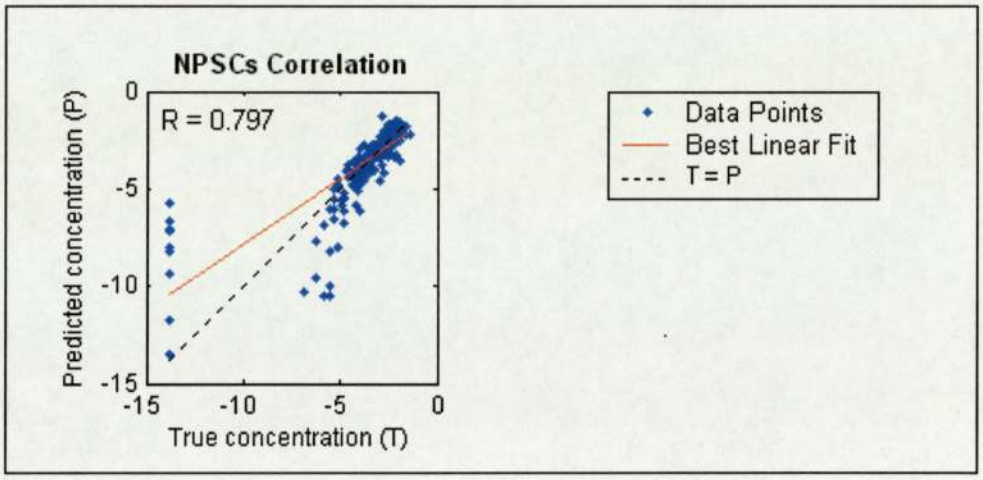
Adapting the MLP to model by cruise and by pigment results in collapse for all of the first three cruises. On the remaining cruises predictive performance varies. Effects again are small with predictions improving for cruises 7 and 8 yet worse for 5 and 6. Each individual pigment group is predicted slightly better aside from the PSCs. Overall results are slightly better than modelling by cruise alone, but exclude those cruises with fewest samples and extreme results and so improvements cannot be considered significant.

Individually retrieving pigments from the 'by cruise' models is not conclusively beneficial and appears to overcomplicate particularly the non-linear models. It may still be considered later when attempting to optimise the better models found.

4.5 Logarithmic Transformation

Applying the log transform to concentration inputs encountered problems, specifically the significant number of zero measurements relating to chl-b, chl-c and the NPSCs. Despite having amended these to take on small values the effect on the model is demonstrated clearly by the following regression plot (figure 4.5.1). On the left of the plot the vertically aligned data points are those for which the true (observed) concentration has been amended. Their positioning appears inconsistent relative to those in the main cluster, yet may be an accurate representation reflecting the absence of certain pigments in these samples.

Figure 4.5.1: Log space correlation plot. (R is the correlation coefficient between true and predicted NPSC concentrations)



As a possible solution all samples containing one or more zero concentrations were removed. This creates an alternative dataset with the number of training samples reduced by 212 to 967 and the number of test samples reduced by 62 to 235.

The new dataset was first tested on several models in linear space to provide a benchmark and assess the effect of removing the samples. Little difference was found in correlations though percentage errors improved, particularly for the MLP. MAE and RMSE however, actually became worse with the new dataset for every model. Analysis by pigment indicates that this may be because the loss of more accurate data relating to other pigments outweighs the benefit of removing the noisy data. Results for chl-a for example, which previously contained no zeros, become worse with the amended data.

In log space the trend is similar (see table 4.5.1). Correlations are all slightly worse, but percentage errors improve. MAE and RMSE both increase compared to the results in linear space (from both the previous dataset and the new one with zeros removed). In some cases there are small positive changes for individual pigments but clearly not the majority. The exception to the above trend is the MLP by cruise for which all the errors improve in log space with the new dataset. Again though the improvement is not enough to match the GLM.

Table 4.5.1: Concentration Retrieval Errors– log space comparison, (all SD normalised)

Model	1 – **MLP with PCA** (by cruise)		2 – **GLM, raw input**		3 – **GLM with PCA** (by cruise & pigment)	
	linear	**LOG**	linear	LOG	linear	LOG
CORR	0.9710	0.9703	0.9761	0.9591	0.9812	0.9727
Mean % error	23.4517	19.8739	29.8979	23.6231	23.1113	17.6473
Bias	0.0031	0.0018	5.01e-004	-0.0015	0.0011	-0.0019
MAE	0.0218	0.0217	0.0237	0.0248	0.0196	0.0200
RMSE	0.0435	0.0396	0.0379	0.0416	0.0338	0.0342

Note: MLP results do not include cruise 1, LOG results use amended dataset, linear results use previous data.

The GLM with PCA by cruise and by pigment in log space now produces the lowest percentage error for any of the models. However, given that improvement is also seen using linear space with the new dataset and that other errors increase, at least part of this improvement is attributed to the removal of small and zero concentrations, which have a disproportionate effect on percentage error. Bearing this in mind and that transform bias has not been removed use of the transform is still not obviously beneficial to any form of GLM.

A final variant of model takes a log transform of the spectra only. While individual pigment results may be better for each model whether or not trained by pigment and/or cruise the overall correlations reduce and errors increase.

4.6 Size Inputs

Inclusion of size data potentially assists the implicit modelling of the package effect and is considered in two forms. Firstly, for those models using PCA the two types of size input (size index and the three size measurements) will be included in the projection. This method increases the total input data without affecting network structure. Secondly, for all models the two types of size data will be used as direct inputs. These are therefore additional to the spectral input units, increase the dimension and alter the structure of the network.

4.6.1 Projected Size Inputs

Incorporating the single size index into the projection data affects each PC based model slightly differently. The general trend however is a reduction in correlation and increase in the majority of errors, as is the case for the MLP in table 4.6.1.1, such that overall it appears detrimental to model performance. Similarly including the three projected size measures has negative effects for most models. It does however improve predictive output for the MLP when trained by cruise (table 4.6.1.1).

In log space the effects differ and are slightly clearer. The overall MLP predictions are worsened by including the size index, but improve where 3 sizes are used. For the MLP by cruise both size options and particularly using three sizes improves predictions, which become better than those produced in linear space.

For the GLM with PCA both forms of projected size input are not helpful for the overall model, though when modelled by cruise using the three size inputs becomes useful.

Table 4.6.1.1: Average Concentration Retrieval Errors using various size inputs for the MLP by cruise (all SD normalised, all pigments)

Model	No size	Projected size index	Projected 3 size inputs	Non- projected size index	Non-projected 3 size inputs
CORR	0.9710	0.9676	0.9720	0.9756	0.9736
Mean % error	23.4517	22.8898	21.4439	20.8874	22.5810
Bias	0.0031	0.0033	-5.719e-004	0.0024	0.0029
MAE	0.0218	0.0217	0.0209	0.0204	0.0210
RMSE	0.0435	0.0460	0.0416	0.0397	0.0407

Note: Results do not include cruise 1.

4.6.2 Additional (non-projected) Size Inputs

Using additional network inputs has a different impact. For the size index results are very mixed with some models, such as the MLP by cruise (table 4.6.1.1), improving and others

becoming worse predictors. Using three inputs effects remain small, but results are generally more positive. Aside from the overall MLP, correlations and errors improve or are at least maintained.

In log space including the size index worsens overall MLP predictions but improves them where the three size measurements are used. The improvement is better than when using the three projected inputs and significantly improves on all other variants thus far. Modelling by cruise each size input improves prediction surpassing those in linear space.

Including the size index in log space for the raw input GLM is unhelpful. Using three sizes RMSE and percentage error improve but MAE gets worse. By cruise, size inputs improve outputs but linear results remain superior. For the GLM with PCA size mostly has a negative effect. By cruise however the reverse is true and (non-projected) size inputs improve the model.

The size index alone does not appear to capture the information well enough to be useful to the models. Results are often worse presumably because the network is made unnecessarily complex without adding useful information. Regarding the three size inputs results are more positive, particularly where taken as additional (non-projected) inputs. Working with log-transformed spectra and concentrations, the extra size inputs appear to have more impact and often improve results, with selected models exceeding performance in linear space.

The effect of size inputs is not consistent, though in some forms and for certain networks it appears a useful input. The available measures may not contain enough information to well represent the cell size distributions and/or package effect and may also be subject to noise.

4.7 Optimal Models 1

The optimal models in each space thus far and their corresponding performance measures are detailed below (table 4.7.1). The direct inverse model and overall GLM with PCA are excluded here, as the alternative models are significantly better predictors at this stage. The GLM with PCA modelled by cruise remains most successful, though now the log transform (remembering that the transform bias must be removed for an exact measure) appears useful.

Table 4.7.1: Concentration Retrieval Errors – optimal models
(all SD normalised, all pigments retrieved simultaneously unless specified)

Model	1 – **MLP with PCA** (all cruises)		1 – **MLP with PCA** (by cruise)		3 – **GLM with PCA** (by cruise)	
Space Size Inputs	**Linear** no size by pigment	**LOG** 3 sizes (nonproj)	Linear size index (nonproj)	**LOG** 3 sizes (projected)	**Linear** no size	LOG 3 sizes (nonproj)
CORR	0.9746	0.9700	0.9756	0.9752	0.9813	0.9790
Mean % error	26.1292	19.8660	20.8874	17.6812	23.0236	16.0752
Bias	0.0018	-0.0018	0.0024	-0.0012	0.0012	-9.58e-004
MAE	0.0221	0.0217	0.0204	0.0203	0.0195	0.0189
RMSE	0.0397	0.0365	0.0397	0.0362	0.0337	0.0332

Note: MLP by cruise results do not include cruise 1, LOG results use amended dataset with zero samples removed, linear results use previous data.

These and several other models showing potential will now be considered with further adaptations to inputs, outputs, network structure and priors.

4.8 Alternative Network Structures - Extra PCs /hidden units

Variants of several superior models will now be explored for further predictive potential firstly using changes to the number of inputs (by amending the PC dimension) and secondly to the number of hidden units. Although validation has been carried out previously this was with regard to broadly defined models. Also, for the simpler GLM (with no hidden units) a greater number of inputs ought to be viable without creating an overly complex network and introducing unnecessary noise.

4.8.1 Increasing the PC dimension

Regarding the MLP, for both the overall model and the model trained by cruise the trend is of decreasing errors and increasing correlation as the number of PCs increases. Taking the overall case to the extreme of 151 PCs maximises correlation in linear space, though does not minimise all error measures, presumably because there is more potential for noise to interfere.

In log space additional PCs also reduce errors but again there is a point, thought to be close to 40 PCs, at which errors begin to rise again. The optimal variants of this overall MLP are produced using log space, but still do not surpass previous results for individual cruise models. Several higher dimensional PC projections of inputs also cause large shifts in error values and cause erratic behaviour of individual pigment retrieval models.

Regarding the MLP by cruise there seems to be an optimal dimension again close to 40 PCs at which minimum errors are achieved in both linear and log space. The optimal MLP found has a PC dimension of 40, takes three projected size inputs and is in log space. The following table (table 4.8.1.1) compares this with the corresponding model in linear space and with use of alternative size inputs.

The comparison shows that selected size inputs have a positive effect and again appear most influential in log space. This adapted MLP now succeeds the GLM as the best model, though is now quite sensitive to changes to inputs and is liable to collapse on the small cruises using this larger numbers of PCs.

Table 4.8.1.1: Concentration Retrieval Errors for the MLP by cruise with increased PC dimension and varying size inputs (non projected unless otherwise stated) in both linear and log space. (All SD normalised)

Model Dynamics	**Linear space, 40 PCs**			**LOG space, 40 PCs**		
Size Inputs	No size	1 size	3 size	No size	3 size	3 size proj
CORR	0.9759	0.9681	0.9759	0.9646	0.9774	0.9794
Mean % error	22.5954	23.7296	22.6236	20.4683	16.7191	16.0817
Bias	0.0015	0.0031	0.0020	8.00e-004	0.0034	-3.76e-05
MAE	0.0206	0.0214	0.0199	0.0222	0.0195	0.0182
RMSE	0.0392	0.0455	0.0392	0.0413	0.0369	0.0323

Note: These averages do not include cruise 1

The GLM however may also use additional PCs and ought to support a much greater number of inputs before noise interferes compared to the MLP, which is much more complex due to its hidden layer. Increasing the PC dimension for the GLM with PCA improves performance though it still cannot match the by cruise models (seen in table 4.7.1).

Adapting the GLM by cruise however improves results further and re-establishes the GLM as the optimal model. In linear space the maximum number of PCs (151) with no size input produces the best correlation and minimises RMSE at 0.0323, while including size minimises MAE at 0.0185. Log space improves on these further and is optimised using only 48 to 50 PCs plus the three size inputs. Minimums are produced for MAE and RMSE at 0.0181 and 0.0323 respectively with either projected or non-projected size inputs. Retrieving pigments individually does not appear to improve the adapted PC models.

Table 4.8.1.2: Concentration Retrieval Errors for the GLM by cruise with increased PC dimension and varying size inputs (non projected unless otherwise stated) in both linear and log space. (All SD normalised)

Model Dynamics	**Linear space**				**LOG space**		
Size Inputs	No size 50 PCs	No size 100 PCs	No size 151 PCs	3 size proj 151 PCs	No size 50 PCs	3 size proj 50 PCs	3 size proj 100 PCs
CORR	0.982	0.983	0.983	0.982	0.977	0.982	0.982
Mean % error	21.97	21.63	21.54	21.36	16.73	15.15	15.07
Bias	0.0016	0.0016	0.0014	0.0013	-1.88e-04	-4.56e-04	-2.24e-004
MAE	0.0190	0.0188	0.0187	0.0185	0.0192	0.0181	0.0182
RMSE	0.0332	0.0325	0.0323	0.0334	0.0332	0.0323	0.0324

Additional PCs certainly seem to gradually add more useful information, but at the cost of a more complex model. This is clearly more of an issue regarding the MLP than the GLM. The by cruise models are again confirmed as superior and clearly surpass the raw input GLM and overall models. Log space results are also better and including three size inputs again appears most useful.

4.8.2 Increasing the number of hidden units

Given that the optimal models so far are by cruise, use some form of size data and take additional inputs in the form of a greater PC dimension it may be that more or less hidden units provide a more appropriate structure for the MLP. Several by cruise models will be adapted to investigate this.

Firstly the hidden layer was reduced to eight units. The effect as expected is an increase in MAE and RMSE relative to the corresponding model with nine units. Subsequently increasing the PC dimension reduces errors but not to the extent of rivalling the nine unit model. The same was true using both seven and four units for the hidden layer.

Increasing the hidden layer by one unit at first appears helpful, as the error on the standard 10 PC model reduces. However, when trying to use in excess of 30 PCs problems occur for cruises where there are few samples. Further hidden units (in both linear and log space) generally appear to introduce more noise and overcomplicate the network, such that results tend to get worse or the model collapses. No significant improvements on the original nine unit hidden layer structure are found.

Adjusting network architecture causes some erratic behaviour, such that certain combinations of inputs and hidden layer size give particularly good or bad results. These do not seem systematic or to follow any obvious pattern and so may be the result of noise. The current structure performs relatively well and is generally stable and will therefore be retained.

4.9 ARD

An ARD prior potentially improves a given model by adjusting weights to reflect importance of inputs. The optimal non-linear models at this stage will be adapted to incorporate ARD. The GLM is excluded, as it does not have a hidden unit layer or therefore the weights that ARD controls. However, using the forward model ARD coefficients it may be speculated as to which pigments are most important in predicting absorption at each wavelength. This data can potentially be analysed and converted to weightings for spectral inputs for the reverse models, including the GLM.

4.9.1 ARD Prior for the direct inverse model

Adapting the prior for several of the better MLP models to incorporate ARD has mixed effects. Some models, such as the first example in table 4.9.1.1 become worse, whilst others

like the second example show small improvements. Overall the effect is more often positive for those models it is applied to. Certain by cruise models however become very sensitive to collapse for the smaller cruises, depending on the combination of PC dimension and size inputs.

Applying ARD to the MLP by cruise, it succeeds the GLM as the best model. The margin though is small as shown by table 4.9.1.1. Minimum errors are achieved in log space using 40 PCs and 3 projected size inputs. MAE falls to 0.0180 and RMSE to 0.0318.

Table 4.9.1.1: Concentration Retrieval Errors for MLP by cruise with ARD compared to the previous best model (all SD normalised)

Model	**MLP by cruise**				**GLM by cruise**
Space	Linear		LOG		LOG
Size Inputs	No size		3 size proj		3 size proj
PC dimension	40 PCs		40 PCs		50 PCs
Prior	No ARD	ARD	No ARD	ARD	(No ARD)
CORR	0.9759	0.9756	0.9794	0.9800	0.982
Mean % error	22.60	23.21	16.08	15.87	15.15
Bias	0.0015	0.0021	-3.76e-05	5.98e-06	-4.56e-04
MAE	0.0206	0.0210	0.0182	0.0180	0.0181
RMSE	0.0392	0.0394	0.0323	0.0318	0.0323

4.9.2 Applying ARD Information from forward modelling

Although some improvements were found using ARD in the inverse model the effects are small. Therefore given the significant uncertainty related to the ARD data collected from forward models the removal and/or additional weighting of inputs will not be pursued here.

4.10 Constrained Concentrations

Given that the concentration predictions ought to be positive they will be constrained within the models. Following forward propagation and post-processing each negative prediction will simply be set to zero.

Applied to the GLM by cruise with no additional inputs and the standard 10 PCs the results for each cruise are identical where there are no negative concentration predictions and improve slightly where there are. In these cases retrieval errors for each pigment are also slightly better such that the overall result is a small improvement. The log transform implicitly applies this constraint so any log space models ought not to be affected by the change. As log space models are optimal at this stage the constraint may be unnecessary.

4.11 Additional Slope & Curvature Information

The final adaptation of direct inverse model to be considered incorporates gradient and/or curvature data related to the spectra. This potentially adds a further dimension to the information that can be captured by the model. The slope is calculated numerically using the MATLAB 'gradient' function, which takes forward differences for the end absorptions and central differences for interior points. The curvature is similarly calculated by applying the 'gradient' function twice. The data produced is then normalised, PCs calculated and various combinations of the new data then incorporated into the GLM and MLP.

4.11.1 Concentration Retrieval using Slope & Curvature Data alone

Firstly a new model will be considered for retrieving the concentrations without using the standard absorption inputs. The GLM by cruise will be used to assess the viability of retrieval by this method.

In linear space retrieval of concentrations using gradient data alone is quite successful. Additional PCs also improve results but the new model still fails to match the standard spectral PC model. The best result found using gradient data alone is produced using the first

35 PCs. This maximises correlation at 0.9805 and reduces RMSE to 0.0344. In log space correlations are still close to 0.95 but the errors much worse than for the standard model.

Using the curvature data alone it is again possible to retrieve the concentrations to some degree. Around twenty PCs appear optimal producing correlations as high as 0.973. Errors though in both linear and log space are significantly worse than when using either absorption data or the absorption gradient data and RMSE exceeds 0.040.

Using equal numbers of PCs of slope and curvature data does not appear to improve on results from using gradient PCs alone. The information provided by each evidently overlaps, so that where gradient data is used adding in curvature data complicates the network without bringing new information.

Slope and/or curvature data does not in itself contain enough information for accurate concentration retrieval, but there is certainly evidence that the variables are related to the pigment concentrations. It may be that each could provide useful additional information to that contained in the absorption spectra alone.

4.11.2 Incorporating Additional PC Inputs

The gradient/curvature inputs will now be incorporated into previous models with only absorption data inputs. The effect of additional inputs will be assessed by comparison with corresponding models in terms of the number of standard spectral inputs, size inputs and network structure. Gradient data is initially included in equal proportion to the absorption data, such that the basic model takes the first ten PCs of the spectra and the first ten PCs of the gradient of the spectra. Similarly a second variant includes equal numbers of spectral and curvature PCs and a third model takes equal numbers of PCs for the raw spectra, the gradient and the curvature.

The MLP and GLM will both be investigated in linear and log space and with varying network structures. Both overall and by cruise models are considered, as is modelling by pigment and the use of ARD priors.

4.11.3 Inclusion of Gradient PCs

The first experiments test the overall GLM but show no radical improvement from previous models. Given the already superior performance of the by cruise models the focus from here on will shift to these variants of the GLM and MLP.

In linear space the effect of including gradient data (with or without size inputs) on the GLM is mixed. Using log space however, the model performs consistently better with the extra data. Small numbers of additional PCs improve this further as does modelling by pigment to give the following best model (table 4.11.3.1):

Table 4.11.3.1: Concentration Retrieval Errors for the GLM by cruise with additional gradient inputs. Modelled in log space by pigment with 3 non-projected size inputs, (SD normalised). Standard model takes 11 absorption spectra PC inputs.

	Standard model	**Plus 11 gradient PCs**
CORR	0.9790	0.9803
Mean % error	15.9763	15.1732
Bias	-9.1877e-004	-0.0013
MAE	0.0189	0.0180
RMSE	0.0331	0.0309

For the MLP the extra inputs do not appear useful and results are worse than both the GLM and the standard MLP.

4.11.4 Inclusion of Curvature PCs

Adding in the curvature rather than gradient data improves predictions for all of the models tested in linear space, but changes are minimal relative to the standard absorption input models. In log space the performance is also better than for the standard GLM, but not as good as the corresponding model using gradient inputs. The curvature model does however improve on the gradient inclusive model when using larger numbers of inputs, but does not significantly or consistently outperform previous models. Again adaptation is unsuccessful for the MLP, which becomes unstable quite quickly with additional PCs.

4.11.5 Inclusion of both Gradient and Curvature PCs

For the model including both gradient and curvature data the linear space predictions generally become worse. Utilising the log transform as well predictions are mostly better or equivalent to both the standard model and those inclusive of gradient/curvature data alone. On exceeding around 20 PCs the errors do however start to increase at a faster rate than for the corresponding models due to the more complex structure.

The combined inputs produce a new best model taking 11 PCs each from the spectra, gradient and curvature data for which the results can be seen in table 4.11.6.1. The model is implemented in log space, takes three non-projected size inputs and retrieves pigments individually. Both MAE and RMSE are minimised by the model at 0.179 and 0.0309 respectively.

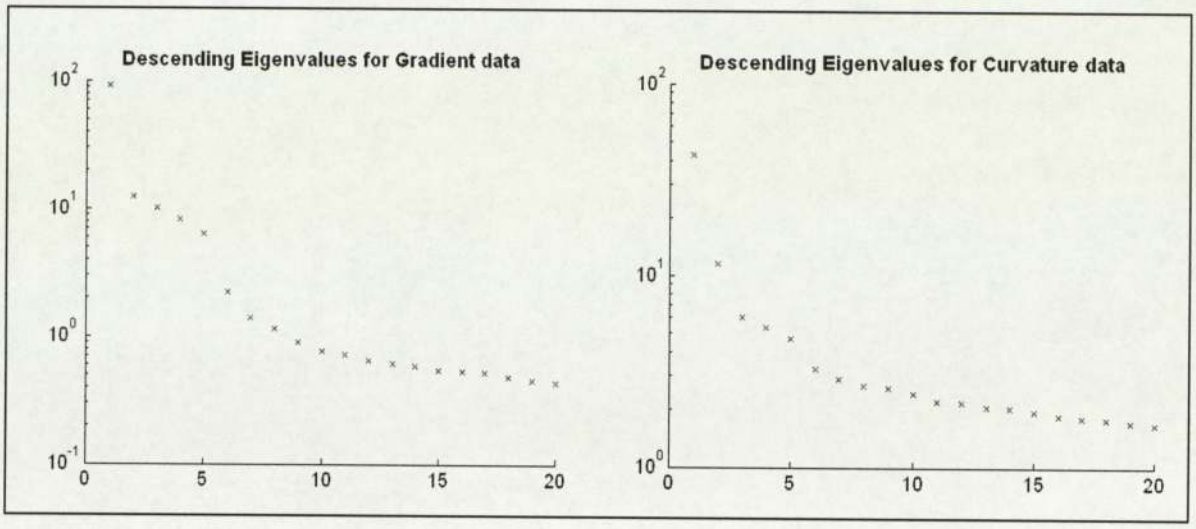
Having tried several combinations of PC inputs there is clearly some useful information contained in the gradient and/or curvature PCs. It may be that alternative proportions of each input are optimal. A more systematic approach to selecting optimal gradient and curvature inputs will use a PC-based analysis of the data.

4.11.6 Gradient and Curvature data – PCA and Concentration Retrieval

Plotting eigenvalue spectra for the gradient and curvature data there are big differences (see table 4.11.6.1), such that each may be best represented by different numbers of PCs. For the gradient data the first eleven PCs are needed to capture 90% of the variance. A noticeable kink in the trend does occur between the fifth and sixth eigenvalues suggesting that noise is being introduced.

The variance explained by the curvature data PCs is much more evenly spread between the PCs. The first PC only captures 28% of the variance and in excess of fifty PCs are necessary to retain 90% of the variance. The kink again appears at five to six eigenvalues (see figure 4.11.6.1). Examination of the corresponding eigenvectors (PCs themselves) for each dataset reveals very noisy structure even in the early PCs. Six to eight of the gradient PCs display most structure yet only two to five of the curvature PCs.

Figure 4.11.6.1: Eigenvalue spectra for gradient and curvature of the absorption data.



Using this information as a starting point several further combinations of PC inputs are tried. The optimal models discovered are detailed below (table 4.11.6.1). Individual pigment retrieval is again found to have small but positive effects.

Table 4.11.6.1: Concentration Retrieval Errors for GLM with various PC inputs (All SD normalised)

Model	GLM with PCA, by cruise			
Space	**Linear**	**LOG**	**LOG**	**LOG**
Size Inputs	3 sizes (nonproj)	3 sizes (nonproj)	3 sizes (nonproj)	3 sizes (nonproj)
PC Inputs	20 spec PCs	11spec PCs	11 spec, 11 grad,	100spec PCs
	20 grad PCs	11 grad PCs	11 curv PCs	10 grad PCs
		by pigment	by pigment	by pigment
Correlation	0.9821	0.9803	0.9806	0.9835
Mean % error	21.8337	15.1732	15.1593	14.2000
Bias	0.0014	-0.0013	-0.0013	-0.0010
MAE	0.0195	0.0180	0.0179	0.0170
RMSE	0.0329	0.0309	0.0309	0.0299

Note: LOG results use amended dataset, linear results use previous data.

One of the better linear models is shown as a benchmark to illustrate the benefit of applying the log transform. Each selected log model is clearly more successful. The change from the second to the third model is minimal confirming previous beliefs of limited impact of the

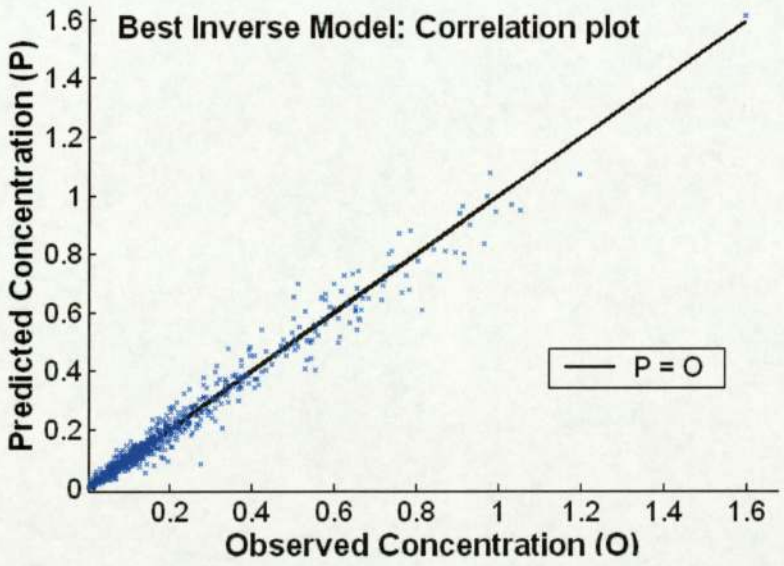
curvature data when added to the direct absorption data and gradient data. Improvements are maximised by using one hundred spectral PCs with an additional ten gradient PCs and this model is discussed further in the following section.

The derivative data is clearly relevant to some extent and improves model performance. For the MLP however, it appears that the increase in inputs and therefore complexity outweigh the potential benefit of the additional information.

4.12 Optimal Models 2

A correlation plot for the optimal direct inverse model found is shown in figure 4.12.1. The errors do appear to increase slightly with the magnitude of the observed concentration, though there is no particular range where predictions appear worse or any obvious bias.

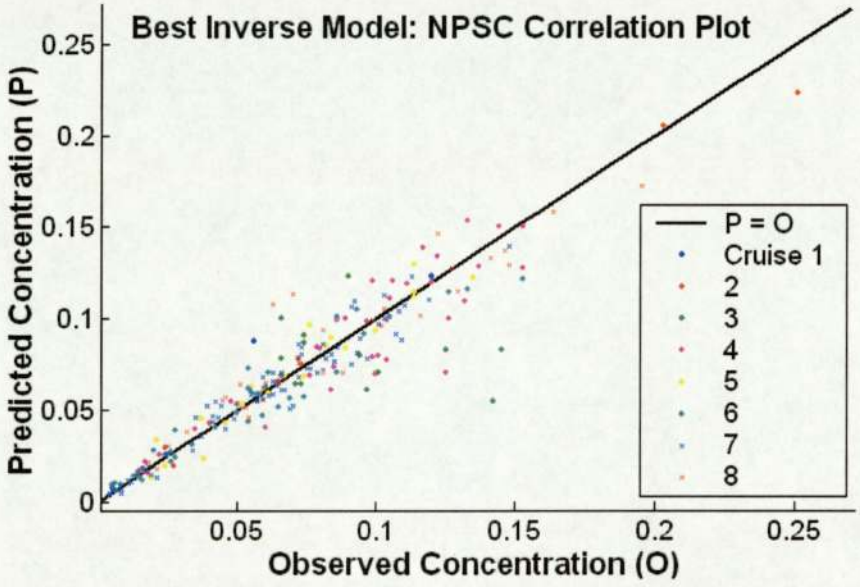
Figure 4.12.1: Observed versus true concentrations (all pigments) for the GLM by cruise modelled in log space, 100 PC spectral inputs plus 10 PC gradient inputs, 3 non-projected size inputs, modelled by pigment (SD normalised)



Analysing the correlation by pigment and by cruise some further conclusions may be drawn. Outlying predictions do not appear particular to select pigment groups, though certain cruises do appear to be more prone to errors. The NPSC concentrations, as plotted in figure 4.12.2, highlight cruises three and eight as the greater sources of error. Plots for each of the other

pigment groups can be found in Appendix A.7. The Chl-a and Chl-b analyses show the greatest errors on cruise two and again three. Chl-c on the other hand highlights several outliers from cruises seven and five and the PSCs again highlight cruise seven. Those from cruise seven and possibly several others seem likely to be noise affected samples that ought to have been removed earlier. This analysis confirms that particular cruises are modelled better than others, whether due to non-modelled natural cruise differences, such as photoacclimation effects, or different noise levels between cruises.

Figure 4.12.2: Observed versus true concentrations of NPSCs for the GLM by cruise modelled in log space, 100 PC spectral inputs plus 10 PC gradient inputs, 3 non-projected size inputs, modelled by pigment (SD normalised)



The breakdown of results by cruise in table 4.12.1 supports these findings and especially highlights cruise 3. This may be largely because the cruise contains few samples, but may also be indicative of an un-modelled variable, which is particularly relevant to cruise 3.

Table 4.12.1: Concentration Retrieval Errors for the optimal direct inverse model (from table 4.11.6.1) broken down by cruise

Cruise	CORR	Mean % err	Bias	MAE	RMSE
1	0.9775	29.3980	0.0102	0.0199	0.0266
2	0.9855	14.6069	0.0059	0.0141	0.0223
3	0.9089	29.3514	-0.0160	0.0304	0.0408
4	0.9754	15.7543	0.0013	0.0121	0.0180
5	0.9867	15.4696	0.0034	0.0169	0.0321
6	0.9801	15.3591	-0.0033	0.0144	0.0266
7	0.9902	11.0848	-9.3877e-04	0.0195	0.0350
8	0.9899	10.8413	-3.5383e-04	0.0153	0.0264
Average	0.9835	14.2000	-0.0010	0.0170	0.0299

4.12.1 Overall versus By Cruise Models

By cruise models are clearly superior in terms of predictive performance, however their practical application is quite limited. It is currently unknown what causes different characteristics between cruises and thus the models could only reliably be applied to concentration retrievals of a very specific nature, for example regarding location. It would be useful then to understand how predictive capabilities of the two model types differ with the aim of finding an optimal model in terms of both prediction and applicability.

Scatter plots of predictions for each pigment produced by each type of model are compared in Appendix A.8. Errors are not identical and analysis is quite subjective, but there is no immediate structure evident. The overall model does not appear to do anything consistently 'wrong' and plotting the corresponding predictions against one another shows no particular pattern of errors. This may suggest that the two models learn equally well the basic underlying mapping, but that by cruise models are able to learn cruise specific detail. This could include effects of unknown variables, such as temperature or cruise specific noise, such as systematic measurement error.

Overall models are still very relevant and could benefit from inclusion of further variables, such as temperature, which might influence the relationship between pigment concentrations and absorption spectra.

4.13 Direct Inverse Modelling Conclusions

Model Performance: The errors for the optimal direct inverse model found are listed in table 4.13.1. It is a variant of the GLM modelled by cruise, which uses the logged dataset, three non-projected size inputs, one hundred absorption spectra PCs and ten spectra gradient PCs and retrieves by pigment.

Table 4.13.1: Concentration Retrieval Errors for the optimal direct inverse model (from table 4.11.6.1)

Pigment	CORR	Mean % error	Bias	MAE	RMSE
Chl-a	0.9845	11.5342	3.4076e-004	0.0331	0.0483
Chl-b	0.9634	18.5127	-0.0018	0.0104	0.0176
Chl-c	0.9694	15.753	-0.0010	0.0094	0.0188
PSC	0.9849	11.8639	-8.7146e-004	0.0229	0.0358
NPSC	0.9726	13.9437	-0.0017	0.0093	0.0144
OVERALL	**0.9835**	**14.2000**	**-0.0010**	**0.0170**	**0.0299**

Modelling by cruise: Individual cruise models are again superior in terms of error performance, though overall models appear quite similar regarding where errors occur.

Network structure: Nine hidden units and forty PC inputs appears the optimal structure for the MLP modelled by cruise. For the GLM each additional input PC appears to improve results in linear space, while in log space the error minimum seems to occur around fifty PCs.

GLM performance: For direct inverse modelling the GLM generally outperforms the MLP suggesting the relation is largely linear.

Log transform: Transforms of the spectra and concentrations appear useful though require removal of zeros, which reduce the dataset further. Earlier findings regarding multiplicative properties of errors suggest possible usage as part of a noise model.

Size distribution data: Size inputs have a positive effect on predictive ability, particularly in log space, and are included in the optimal model thus far. Log space improvements may suggest a link to the noise model.

Individual pigments: Retrieving each pigment separately seems to improve or at least maintain performance, though effects are small for the GLM and by cruise models.

ARD priors: ARD priors have a small positive effect when incorporated into the MLP.

Dataset dependence and small cruises: Small cruises continue to be problematic particularly when using the MLP.

Slope and curvature data: Inclusion of slope data produces better results though the further addition of curvature data does not appear to add any useful information.

Noise and/or external variables: Each cruise appears to be subject to individual noise sources and/or additional non-modelled variables. There is scope to improve models by identifying and incorporating these features. Inclusion of further variables relating to package and acclimation effects, such as depth and temperature, potentially improve the implicit model.

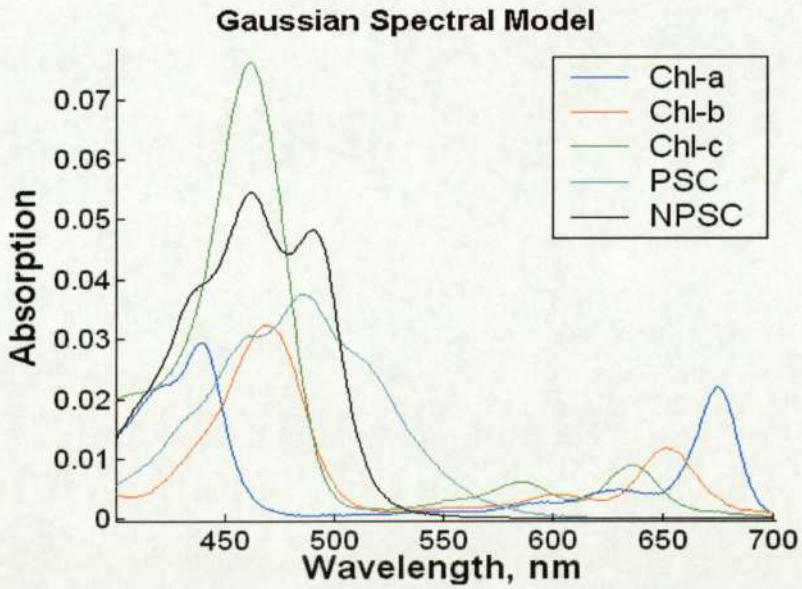
The data driven models can be adapted quite easily to incorporate additional variables and generally perform well. Errors though are still at 14% suggesting that there is something missing from the model, most likely an accurate representation of the package effect, and/or that the noise levels are simply too high for more accurate retrieval.

Chapter 5

A Generative Model for Concentration Retrieval

Some of the most successful research undertaken to date has focused on the Gaussian spectral model (see figure 5.1). Underlying the theory of the model is an assumption that the absorption due to each pigment can be represented by a number of Gaussian distributions. A specific absorption spectrum can then be calculated for each pigment, which estimates the absorption at each wavelength for unit concentration of the given pigment. Multiplying this specific absorption by a measured concentration of the corresponding pigment then ought to give the absorption due to that pigment.

Figure 5.1: Gaussian Spectral Model: the specific absorption spectra for each pigment as estimated by Annick Bricaud.



Note: These spectra are derived from *in vitro* measurements for each individual pigment (in solvent). The band maxima have subsequently been shifted to match the known *in vivo* maxima. This means that the package effect is not included here.

The model estimates that total absorption is equal to the sum of products of concentrations and specific absorption (5.1), subject to the package and photoacclimation effects.

$$a(\lambda) = \sum_i C_i \sum_j a^*_{ij} \phi_{ij}(\lambda, \mu_{ij}, \sigma^2_{ij}) \qquad (5.1)$$

where λ is the wavelength

$a(\lambda)$ is the total absorption (a function of wavelength)

i is the pigment index (from one to five)

j is the index of the Gaussian band used to represent the i^{th} pigment

a^*_{ij} are the heights of the Gaussians

ϕ is a function of the wavelength, λ and also μ_{ij} and σ^2_{ij}, the centres and widths of the Gaussian bands respectively. Therefore, $\phi_{ij}(\lambda, \mu_{ij}, \sigma^2_{ij})$ is the j^{th} Gaussian band for the i^{th} pigment.

$\sum_j a^*_{ij} \phi_{ij}(\lambda, \mu_{ij}, \sigma^2_{ij})$ is then the specific absorption for the i^{th} pigment. Given this model there are four unknowns: j, a^*_{ij}, μ_{ij} and σ^2_{ij}, whilst the known variables are the observed data D = $\{a(\lambda), C_i\}$- the set of absorption spectra and corresponding concentrations.

5.1 Estimation of Model Parameters

The first stage then is to estimate the unknown parameters j, a^*_{ij}, μ_{ij} and σ^2_{ij}. The Gaussian spectral model is similar to a Radial Basis Function (RBF), which has the general form seen in (5.2). The RBF then will be the starting point for creating a generative Gaussian based model.

$$y_i(\mathbf{x}) = \sum_j w_{ij} \phi_j(\mathbf{x}) \qquad (5.2)$$

The findings from both visualisation and the direct inverse model suggest that that the relation between concentrations and spectra is largely linear. To incorporate this information and simplify the model the parameters j, μ_{ij} and σ^2_{ij} will be fixed. This means that the number of Gaussian bands, their positions and widths will be fixed, such that the only unknown in

specific absorption is a^*_{ij}, the heights of the bands. Having fixed the other three unknowns it will be possible to learn distributions for a^*_{ij} in the Bayesian framework.

To fix these three parameters a simple non-Bayesian RBF is created. The first aim of the model will be to fit the Gaussian absorption bands proposed by Annick Bricaud (figure 5.1). The RBF is trained using a single wavelength input and target data consisting of Annick's absorption bands. The model output and errors are then analysed numerically and graphically to assess how many basis functions and what parameters are most appropriate to accurately model Annick's estimated spectra.

As a preliminary investigation the model was implemented without fixing the basis function parameters. The centres and width parameters (μ_{ij} and σ^2_{ij}) are initialised randomly and allowed to change to reflect a Gaussian mixture model fitted to the data. Using this flexible structure each of Annick's pigment spectra are best modelled using eight hidden units (to 4 d.p). Additional units offer no improvement on this fit and tend to marginally increase the error, such that eight 'bands' per pigment appears the optimal network architecture with the following errors (table 5.1.1).

Table 5.1.1: RMSE for RBF fit to Annick's specific absorption bands with non-fixed basis functions.

Pigment Group	**Number of Bands** (RBF centres, j)	**RMSE** (4 d.p)
Chl-a	8	0.0029
Chl-b	8	0.0031
Chl-c	8	0.0064
PSC	8	0.0008
NPSC	8	0.0033

The RBF was next adapted to use fixed, equally spaced basis functions of equal width to see if this affects the number of basis functions needed or the performance of the model. The basic model was run for each pigment using up to sixty basis functions. The centres (μ_{ij}) are fixed with one centred on each end of the spectrum (400nm and 700nm) and the others positioned at equally spaced intervals determined by the value of j. Experiments are run using several different widths of Gaussian with a common width for all bands modelling a given pigment.

Training is much faster using fixed basis functions and it is possible to get an excellent fit to Annick's specific absorption spectra using the networks detailed in table 5.1.2. The fixed RBF models use more basis functions but fit the spectra much more closely than those with non-fixed centres. A fixed width of 1000 appears to give good model flexibility to fit the spectra. Increasing the scale tends to reduce the number of basis functions needed but results in a bigger minimum.

Three alternative models are identified each using a different threshold for identifying optimal structures. The first finds the minimum to 4d.p, the second uses an RMSE threshold of 0.001 and the third an RMSE threshold equal to 1% of maximum absorption for the given pigment. The corresponding correlations for each pigment spectrum are in excess of 0.99 and RMSE less than 9e-004 for each pigment.

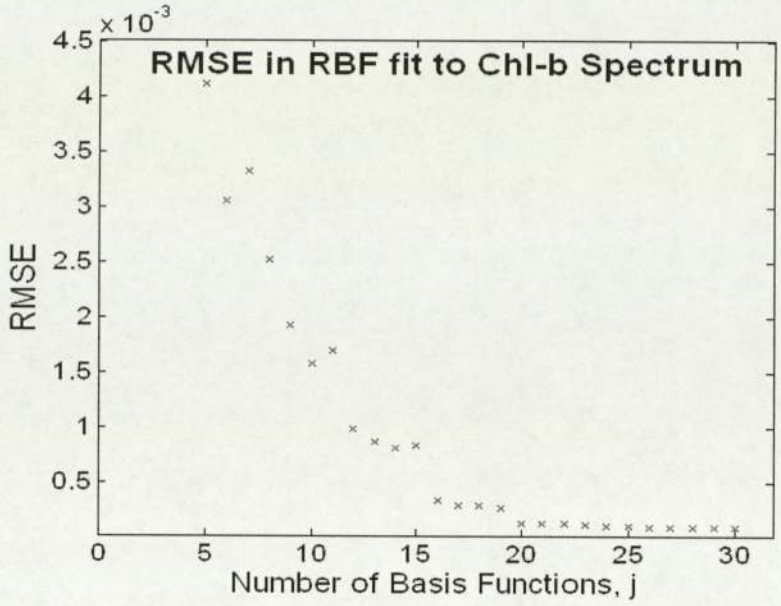
Table 5.1.2: RMSE for RBF fit to Annick's specific absorption bands with fixed basis functions. All widths (σ^2_{ij}) set to 1000. All RMSE figures given to 4 d.p.

Pigment Group	**Number of Gaussian Bands (RBF centres, j) & RMSE**					
	RMSE 4 d.p. minima		RMSE < 0.001		RMSE < 1% max(a(λ))	
	j	RMSE	j	RMSE	j	RMSE
Chl-a	28	0.0002	18	0.0009	28	0.0002
Chl-b	20	0.0001	13	0.0009	17	0.0003
Chl-c	22	0.0001	15	0.0006	15	0.0006
PSC	26	0.0001	7	0.0008	22	0.0004
NPSC	28	0.0002	22	0.0009	24	0.0005

The Chl-a group clearly requires the most complex network as expected, given the more complex absorption spectrum estimated by Annick. Graphical analysis shows the errors for each pigment follow a similar trend and converge to a minimum. The RMSE plot for chl-b (figure 5.1.1) levels off around 20 basis functions, which matches the 4d.p minimum identified previously (table 5.1.2).

In log space the error curves tend to be less smooth with several minima. Using the 1% maximum absorption RMSE threshold the number of bands used to fit chl-a, chl-b, chl-c, PSC and NPSC spectra are estimated at 22, 16, 18, 21 and 23 respectively. The number of bands then is less clear but fewer than the corresponding linear space estimates for all except chl-c.

Figure 5.1.1: RMSE in fitting Annick's estimated chl-b spectrum with various RBF networks



As a further benchmark the number of bands required to model the forward model estimate of specific total absorption from the raw data is explored. A model is produced for each pigment mapping it's individual concentration to the total observed absorption spectrum. A unit concentration of that pigment is then forward propagated to produce a corresponding estimate of absorption. It is then noted how many basis functions best model the relation between wavelength and this specific total absorption spectrum estimated using the given pigment.

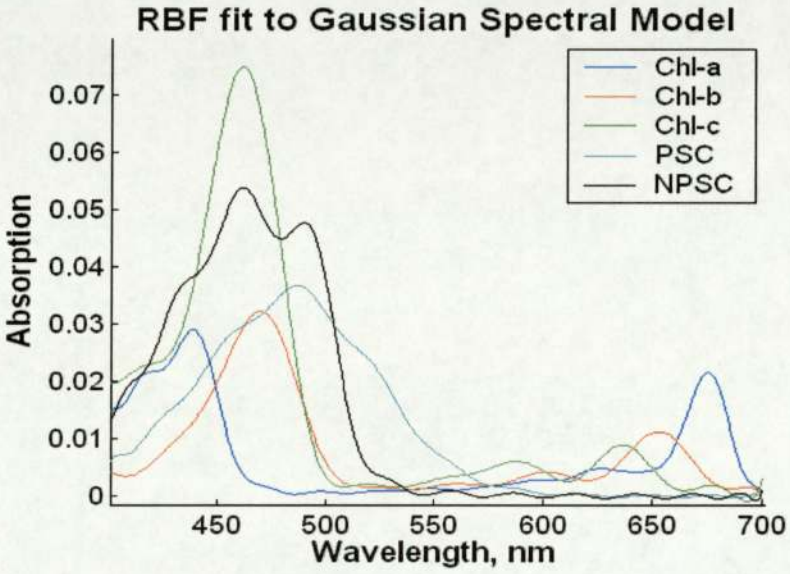
Using non-fixed basis functions to fit these forward model absorption estimates eight units for each pigment is appropriate again, as in modelling Annick's individual spectra. Fixing the basis functions and using the 1% maximum absorption threshold supports the following network structure (table 5.1.3).

Table 5.1.3: RMSE for RBF fit to forward estimates with fixed basis functions. All widths (σ^2_{ij}) set to 1000. Structure selected using 1% maximum absorption threshold for RMSE, RMSE given to 4 d.p.

Pigment Group	**Number of Bands** (RBF centres, j)	**RMSE**
Chl-a	24	0.0001
Chl-b	19	0.0003
Chl-c	14	0.0007
PSC	20	0.0003
NPSC	14	0.0005

By fixing the basis functions the computational complexity is significantly reduced. Therefore, the choice of structure will be based on the most reasoned approach despite not being the simplest network. By selecting the network with RMSE threshold of 1% maximum absorption (table 5.1.2) the threshold is appropriate to the specific pigment. The corresponding reconstructed specific absorption spectra are shown in figure 5.1.2.

Figure 5.1.2: RBF fit to the specific absorption spectra proposed by Annick Bricaud (using the structure detailed in table 5.1.2 for RMSE < 1% max(a(λ)))



Comparing this to the model itself in figure 5.1 the network seems to offer a good approximation. Using different numbers of bands per pigment in this way may slightly complicate software, but allows the model to be more flexible where most needed.

5.2 The Generative Model & Bayesian Methods

Having now fixed the values of the parameters j, μ_{ij} and σ^2_{ij} a distribution for the weights can be learnt by training the network using the observed dataset. These weights are the heights, a^*_{ij} and the value of the true concentration given the observed value. They are represented in the following discussion by the probabilities $p(\mathbf{w})$ and $p(\mathbf{x}|\mathbf{x}_{obs})$ respectively.

Using Bayes' theorem (as described in section 3.1) the posterior distribution of weights (**w**) given the observed data (D) is given by (5.3), which may then be expressed as in (5.4).

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})\, p(\mathbf{w})}{p(D)} = \frac{p(D \mid \mathbf{w})\, p(\mathbf{w})}{\int p(D \mid \mathbf{w})\, p(\mathbf{w})\, d\mathbf{w}} \qquad (5.3)$$

$$p(\mathbf{w} \mid D) = p(\mathbf{w} \mid \mathbf{y}_{obs}, \mathbf{x}_{obs}) = \frac{p(\mathbf{y}_{obs} \mid \mathbf{x}_{obs}, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{x}_{obs})}{\int p(\mathbf{y}_{obs} \mid \mathbf{x}_{obs}, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{x}_{obs})\, d\mathbf{w}} \qquad (5.4)$$

The known data are the observed values of **y** given the true concentrations **x**, that is $(\mathbf{y}_{obs} \mid \mathbf{x})$. The posterior (5.4) though contains the term $(\mathbf{y}_{obs} \mid \mathbf{x}_{obs})$, which is unknown. However, this posterior may be re-expressed to incorporate the known data by rewriting $p(\mathbf{y}_{obs} \mid \mathbf{x}_{obs}, \mathbf{w})$ in terms of the true **x** as in (5.5).

$$p(\mathbf{y}_{obs} \mid \mathbf{x}_{obs}, \mathbf{w}) = \int p(y_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, d\mathbf{x} \qquad (5.5)$$

Substituting this into the expression for the weight posterior (5.4) results in equation (5.6). This in turn reduces to (5.7) by recognising that $p(\mathbf{w})$ represents the prior belief about the weights and so has no dependence on $\mathbf{x}_{obs}$.

$$p(\mathbf{w} \mid D) = \frac{\int p(y_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{w} \mid \mathbf{x}_{obs})\, d\mathbf{x}}{\int\int p(y_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{w} \mid \mathbf{x}_{obs})\, d\mathbf{x}\, d\mathbf{w}} \qquad (5.6)$$

$$p(\mathbf{w} \mid D) = \frac{\int p(y_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{w})\, d\mathbf{x}}{\int\int p(y_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{w})\, d\mathbf{x}\, d\mathbf{w}} \qquad (5.7)$$

Again using Bayes rule ((3.1)):

$$p(\mathbf{x}_{obs} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{x}_{obs})}{p(\mathbf{x})} \qquad (5.8)$$

With the assumption of a generative noise model: $\mathbf{x}_{obs} = \mathbf{x} + \varepsilon$, where ε is a noise term the unconditional distributions $p(\mathbf{x})$, $p(\mathbf{x}_{obs})$ are assumed identical, so that $p(\mathbf{x}) = p(\mathbf{x}_{obs})$. This then

implies $p(\mathbf{x}_{obs} \mid \mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{x}_{obs})$ are symmetrical distributions, such that $p(\mathbf{x}_{obs} \mid \mathbf{x}) = p(\mathbf{x} \mid \mathbf{x}_{obs})$. The posterior then can be expressed as (5.9).

$$p(\mathbf{w} \mid D) = \frac{\int p(\mathbf{y}_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x}_{obs} \mid \mathbf{x})\, p(\mathbf{w})\ d\mathbf{x}}{\int\int p(\mathbf{y}_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x}_{obs} \mid \mathbf{x})\, p(\mathbf{w})\ d\mathbf{x}\ d\mathbf{w}} \qquad (5.9)$$

To progress each of the probabilities in (5.9) must be replaced by an appropriate probability distribution. These will be chosen consistent with prior beliefs, so that the generative model incorporates as much knowledge as possible and also implicitly imposes known constraints.

The spectral data $\mathbf{y}_{obs}$ is assumed to be generated by the model (5.1) plus Gaussian noise, so that:

$$p(\mathbf{y}_{obs} \mid \mathbf{x}, \mathbf{w}) \sim N(\mathbf{y}_{obs}, C_{y\,obs}) \qquad (5.10)$$

where the mean $\mathbf{y}_{obs}$ is the observed training data and the standard deviation $C_{y\,obs}$ is a fixed value equal to 10% of the mean absorption across all wavelengths and all samples in the training data.

A gamma prior of the form (5.11) will be used for both the concentrations and the heights, so as to constrain each to be positive. Given the model (5.1) this then also constrains the absorption to be positive, as is the prior belief for all pigment concentrations.

$$\mathbf{x} \sim Ga(\alpha, \beta) \Rightarrow p(\mathbf{x}) = (\beta^{\alpha} / \Gamma(\alpha))\, \mathbf{x}^{(\alpha-1)} \exp(-\beta\mathbf{x}) \qquad (5.11)$$

where $E(\mathbf{x}) = \alpha/\beta$, $Var(\mathbf{x}) = \alpha/\beta^2$

The conditional distribution of concentrations then is given by (5.12). The values of α and β are determined using the observed training data concentrations, which are assumed to be subject to noise of 5%. The mean is set to the observed values of the concentrations plus a small additive component to ensure a reasonable flexibility on even the smaller weights. The variance will be set to the square of the assumed error plus a similar additive component. Using the definition of the gamma distribution (5.11) this implies the following values for α and β.

$$p(\mathbf{x}|\mathbf{x}_{obs}) \sim Ga(\alpha_C, \beta_C) \qquad (5.12)$$

where $\alpha_C = (\mathbf{x}_{obs} + \kappa)^2 / \{(\mathbf{x}_{obs} * 0.05)^2 + \kappa^2\}$

$\beta_C = (\mathbf{x}_{obs} + \kappa) / \{(\mathbf{x}_{obs} * 0.05)^2 + \kappa^2\}$

for $\kappa = 0.05 * \tilde{\mathbf{x}}_{obs,}$ (5% of the mean concentration across all samples for each pigment).

The prior distribution for the heights (5.13) is set similarly with parameters estimated using Annick's specific absorption spectra. Taking a larger variance on the heights of 20% will reflect the uncertainty in Annick's estimates and the derived initialisation. Again an additive component will be incorporated into both the mean and variance.

$$p(\mathbf{w}) \sim Ga(\alpha_w, \beta_w) \qquad (5.13)$$

Justified approximations have now been found for each distribution in (5.9). However, there is not an obvious analytic solution and so sampling will be required.

5.3 Sampling

The basic principle of sampling is to replace a complex integral with a finite sum. To evaluate an integral of the form:

$$I = \int F(\mathbf{w})\, p(\mathbf{w} \mid D)\, d\mathbf{w} \qquad (5.14)$$

weight vectors $\mathbf{w}_i$ are sampled from the posterior $p(\mathbf{w} \mid D)$, so that the approximation becomes:

$$I \approx (1/L) \sum_i F(\mathbf{w}_i) \qquad \text{where } i=1,2\ldots,L \qquad (5.15)$$

The difficulty is in ensuring that the finite set of sample vectors $\mathbf{w}_i$ are appropriately distributed so as to be representative of the true posterior $p(\mathbf{w} \mid D)$. There are many possible sampling methods with varying degrees of complexity, though each has certain limitations.

Gibbs sampling for example requires conditional distributions, which in this case are unknown, while other methods are simply poor at finding representative samples or very time consuming.

Markov Chain Monte Carlo (MCMC) methods attempt to generate representative samples by using random walks in parameter space with the aim of finding areas of weight space where $p(\mathbf{w} \mid D)$ is reasonably large. A sequence of vectors is generated where each has a dependence on the previous vector and also a random component, so that:

$$\mathbf{w}_{new} = \mathbf{w}_{old} + \boldsymbol{\epsilon} \qquad (5.16)$$

where $\boldsymbol{\epsilon}$ is a small random vector.

One of the most widely used MCMC methods is the Metropolis-Hastings algorithm. This approach attempts to preferentially sample regions of higher posterior probability. Steps resulting in a reduction in $p(\mathbf{w} \mid D)$ are proportionally rejected with the intention of generating a truly representative sample set. This method produces some good results though may encounter problems due to strong correlations in the posterior (Bishop, 1995). It is also only feasible for smaller networks, as the non-systematic approach to exploring the distribution can be lengthy with many steps taken in the 'wrong' direction away from areas of high probability density.

Hybrid Monte Carlo (HMC) methods are a progression from the Metropolis-Hastings algorithm and attempt to increase efficiency by introducing a systematic element to the sampler. The HMC algorithm incorporates gradient information from the sampled distribution to bias the direction of movement of the otherwise random sampler. This reduces the number of steps necessary to explore the distribution, thus making a more efficient sampler. This is therefore the only practical method applicable to this particular problem.

5.3.1 Implementing Hybrid Monte Carlo (HMC) Sampling

The posterior for the weights $p(\mathbf{w}| D)$ will be sampled using its corresponding cost and gradient functions. The posterior to be sampled is given by (5.9), though for the purpose of

generating samples the normalisation constant can be omitted as in (5.17). The exponential of the weights will be stored in the network to ensure the positivity of absorption estimates.

$$p(\mathbf{w} \mid D) \propto p(\mathbf{y}_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{w}) \qquad (5.17)$$

The related cost function **E** is expressed in (5.19) and consists of three error components.

$$\mathbf{E} = -\ln(p(\mathbf{w} \mid D)) \qquad (5.18)$$

$$\mathbf{E} \propto -\ln \{p(\mathbf{y}_{obs} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{x} \mid \mathbf{x}_{obs})\, p(\mathbf{w})\}$$

These components are the data error (5.19), the weight error (5.20) and the concentration error (5.21) and are each calculated by taking the negative logarithm of the distributions defined previously in (5.10), (5.12) and (5.13).

$$-\ln\{p(\mathbf{y}_{obs} \mid \mathbf{x}, \mathbf{w})\} = \sum_i ((\beta_D/2).* \sum_n (\mathbf{y}_n - \mathbf{y}_{obs\,n})^2 \qquad (5.19)$$

$$-\ln \{p(\mathbf{w})\} = \sum_i (\beta_w * \mathbf{w}) + \ln(\Gamma(\alpha_w)) - ((\alpha_w - 1).*\ln(\mathbf{w})) - (\alpha_w * \ln(\beta_w)) \qquad (5.20)$$

$$-\ln\{p(\mathbf{x} \mid \mathbf{x}_{obs})\} = (\boldsymbol{\beta}_C * \mathbf{x}) - ((\boldsymbol{\alpha}_C - 1).*\ln(\mathbf{x})) + \ln(\Gamma(\alpha_C)) - \alpha_C * \ln(\beta_C) \qquad (5.21)$$

where i = 1,2,3,4,5 and is the pigment index.

The corresponding gradient function (5.22) is simply given by the derivative of (5.18) with respect to **w**. Together these provide the necessary input functions to the HMC sampler.

$$(d\mathbf{E}/d\mathbf{w}) = (d/d\mathbf{w})[-\ln\{p(\mathbf{w} \mid D)\}] \qquad (5.22)$$

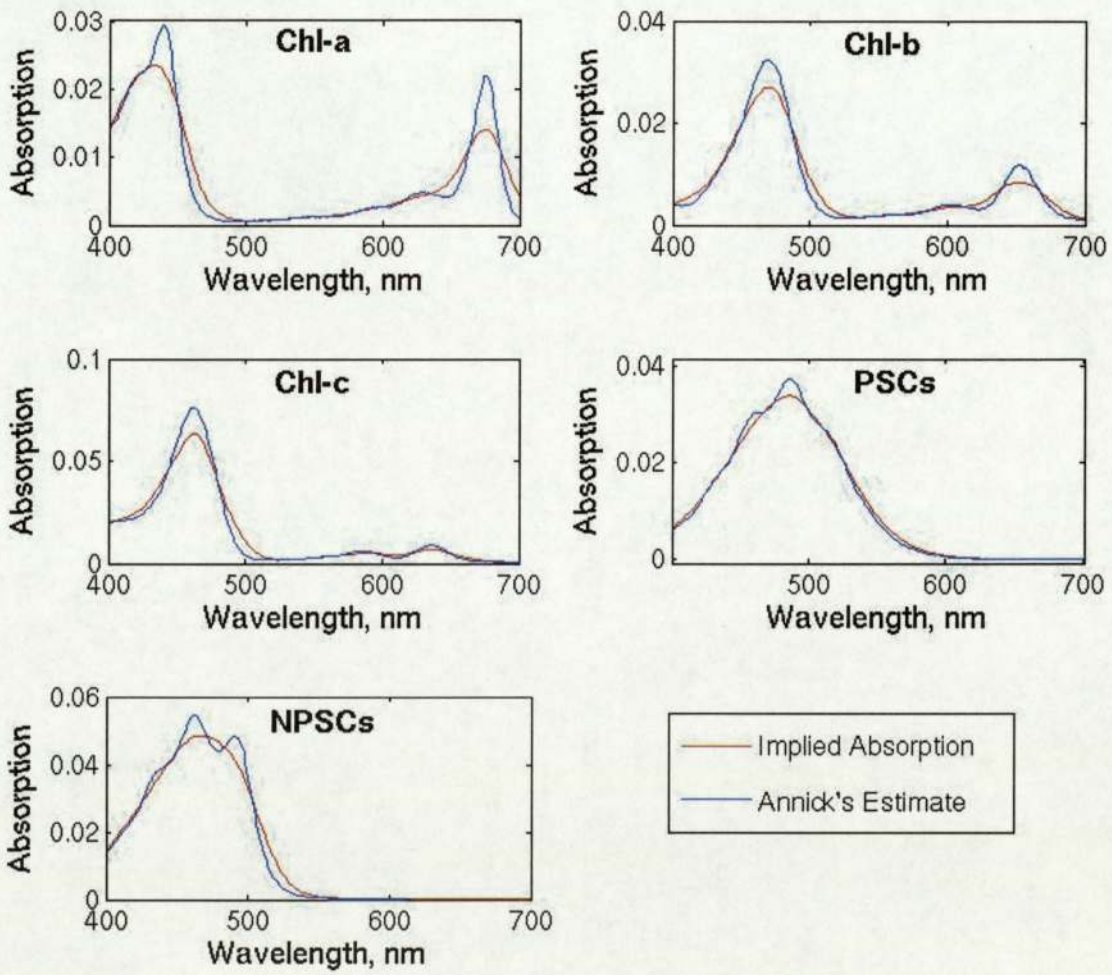
The posterior is sampled (using only the training data) generating samples $\mathbf{x}_n$ and $\mathbf{w}_n$ for the concentrations and the heights, given the observed training data $\{\mathbf{y}_{obs}, \mathbf{x}_{obs}\}$. In this first stage of sampling only the weight samples for the heights will be retained giving the set of vectors:

$\{\mathbf{w}_n ;\ n = 1, 2, 3, ...N\}$, where N is the total number of samples taken.

As it may be difficult for the sampler to find the region(s) of high posterior probability a reasonable initialisation may be critical. To this purpose the heights will first be initialised using Annick's estimated specific absorption bands and the concentrations using the observed data. Scaled-conjugate-gradient optimisation will follow to find an approximate numerical solution for these weights as a first estimate for the sampler. Over-optimisation however will be avoided, as this can result in a high level of rejection and difficulty in exploring the whole posterior.

To ensure that initialisation of the heights is satisfactory the implied specific absorption spectra for each pigment will be compared with Annick's estimates of specific absorption. It was clear from early plots that the mapping was initially too smooth and producing a poor fit to Annick's estimates. Thus the Gaussian widths were reduced and as the comparison in the following plot (figure 5.3.1.1) shows, a reasonable fit is now achieved.

Figure 5.3.1.1: Comparison of specific absorption spectra as implied by initialised weights and Annick's estimates

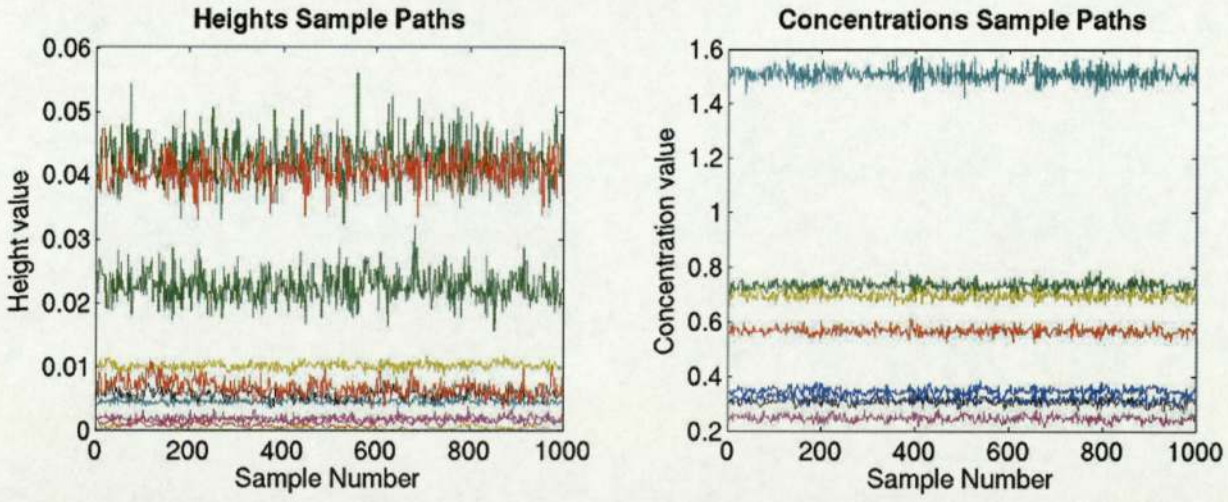


Initially 1000 samples will be generated following a burn in period of 1000 samples. Discarding the first 1000 samples in this way ought to allow the sampler to converge, such that samples are taken from the stationary distribution. While this is a relatively small sample size it appears appropriate for this linear-based model with fixed basis functions. Step sizes will be set by experimentation ensuring that acceptance rates fall in the range 0.6 to 0.9. A systematic sub-sample of 200 will be taken from the 1000 to produce a more manageable dataset to carry forward to the second stage of sampling. This sampling process is repeated per cruise and using the complete set of training data.

5.3.2 Sampled Heights

The sample paths for both the height and concentration samples are plotted to ensure that convergence has occurred and that the samples are being taken from a relatively stable distribution. The sampling paths for a number of randomly chosen heights and concentrations using the cruise 2 data are shown in figure 5.3.2.1. While the paths appear a little erratic there is no major swapping of weights. It seems that the sampler is sufficiently burned in, but that there is simply not enough information in the data to more accurately determine the weights.

Figure 5.3.2.1: Sample paths for random selections of weights from cruise 2 (untransformed from log space)

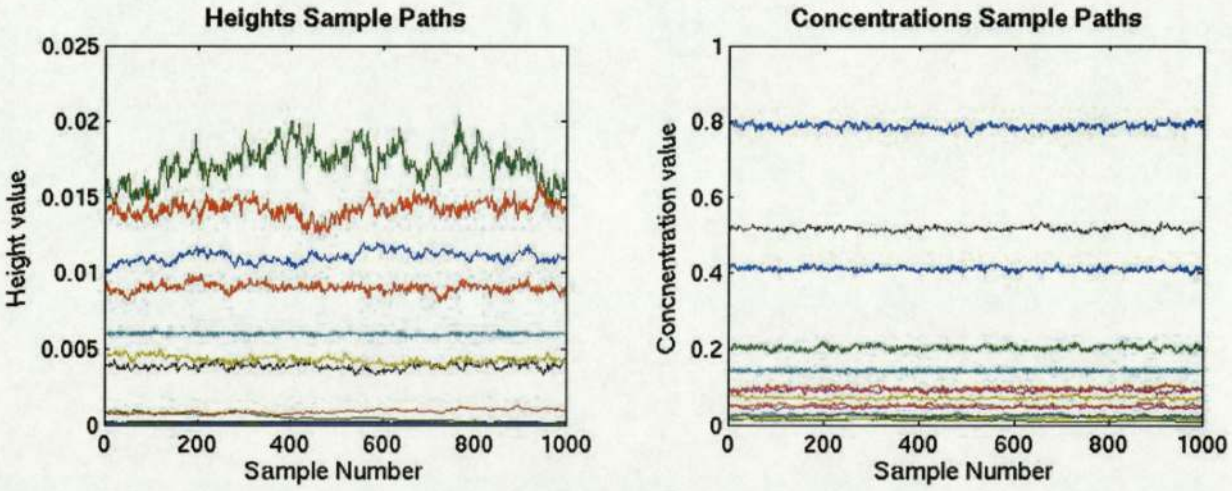


Similarly, for cruise 1 the sample paths are quite erratic yet appear to have converged as far as possible given the small number of samples in the data set. The variance in the sample paths is therefore greater, though there is no particular trend to suggest non-convergence. The other cruises seem similarly well mixed, such that they are as close to convergence to the stationary distribution as the data will allow.

The only obvious swapping of weights is displayed by the heights from cruise 6, which as the largest cruise, was thought to offer more potential for a 'by cruise' model. Within cruises however, samples come from various water types and locations. As the largest cruise it may contain more information, thus necessitating a more complex mapping and taking longer to converge. Experimenting with a longer burn in however does not produce more stable results. It may be that there is more noise present than previously thought or that an un-modelled variable, such as the package effect, is particularly significant for this cruise.

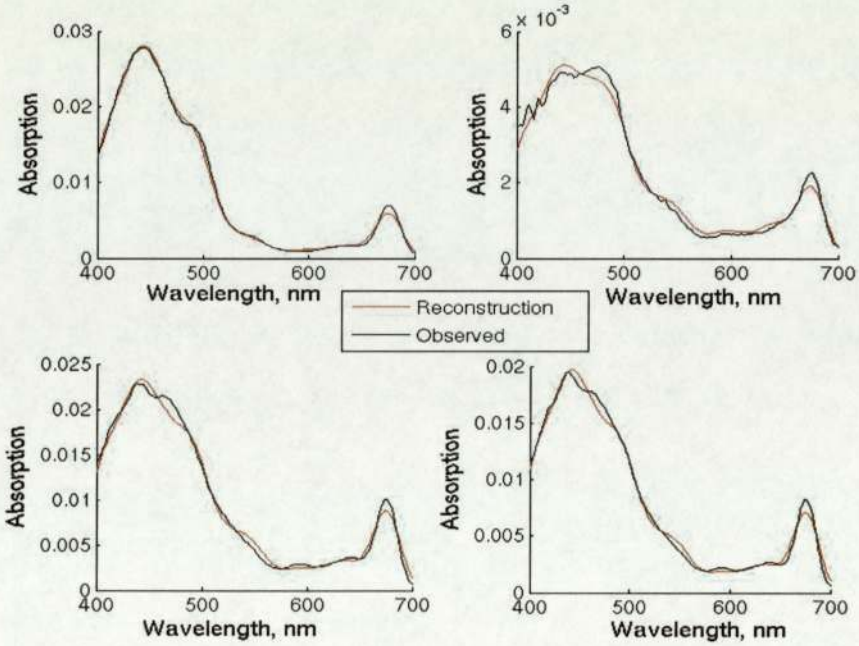
Sample paths generated using all of the training data are more promising (see figure 5.3.2.2). The paths for both heights and concentration are relatively stable with no swapping of the most significant weights.

Figure 5.3.2.2: Sample paths for a random selection of weights from all cruises (untransformed from log space)



To assess graphically the accuracy of the height estimations several spectra will be reconstructed using the optimised network. Figure 5.3.2.3 shows the observed and reconstructed spectra for four examples from the training data of cruise 2.

Figure 5.3.2.3: Spectra Reconstructions using the optimised weight vector compared to true observed spectra. The four spectra are randomly selected samples from the training set for cruise 2



The implied reconstructions show a close approximation to the true spectra and similar results are achieved for other individual cruises and also when sampling weights for the whole data set. However, as figure 5.3.2.4 shows cruise 6 is again an exception. Equivalent plots generated from the cruise 6 data show a much worse fit with significant errors for each of the four samples across the whole spectrum. This is likely to be reflected in poorer concentration retrievals later on.

Reconstruction performance on the test data is also quite different. Using the optimised network and test concentrations to estimate spectra has hugely variable results, such that some reconstructions are quite accurate whilst others are poor. Figure 5.3.2.5 shows the observed spectra and corresponding reconstruction for every sample in cruise 2. Despite the good performance on the training set seen in figure 5.3.2.3 the approximations range from an RMSE value of 0.00095 for the third sample up to 0.029 on the seventh. This variability suggests that the network is lacking in information relevant to the mapping. The performance on unseen data is consequently poor and likely to limit the accuracy of later concentration retrieval.

Figure 5.3.2.4: Spectra Reconstructions using the optimised weight vector compared to true observed spectra. The four spectra are randomly selected samples from the training set for cruise 6.

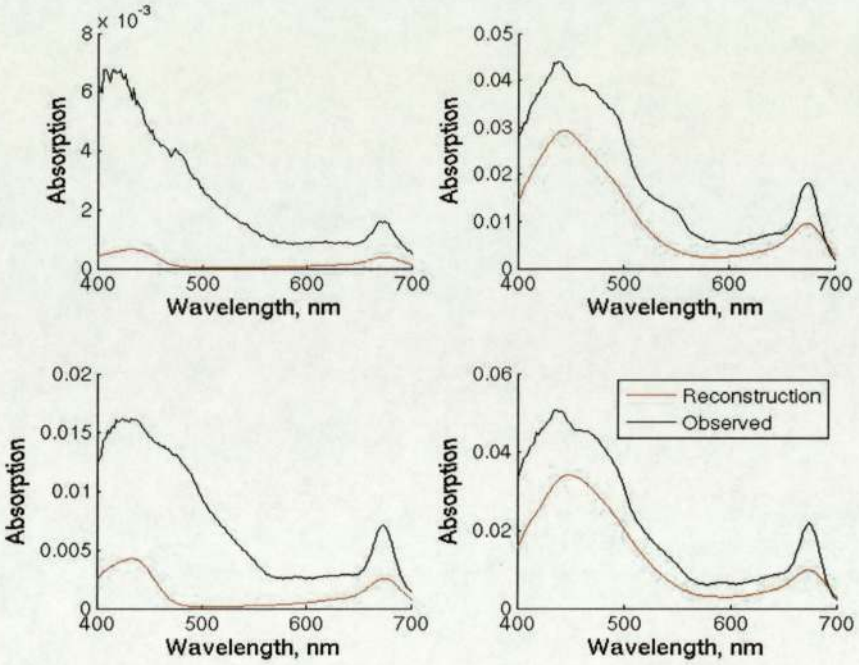
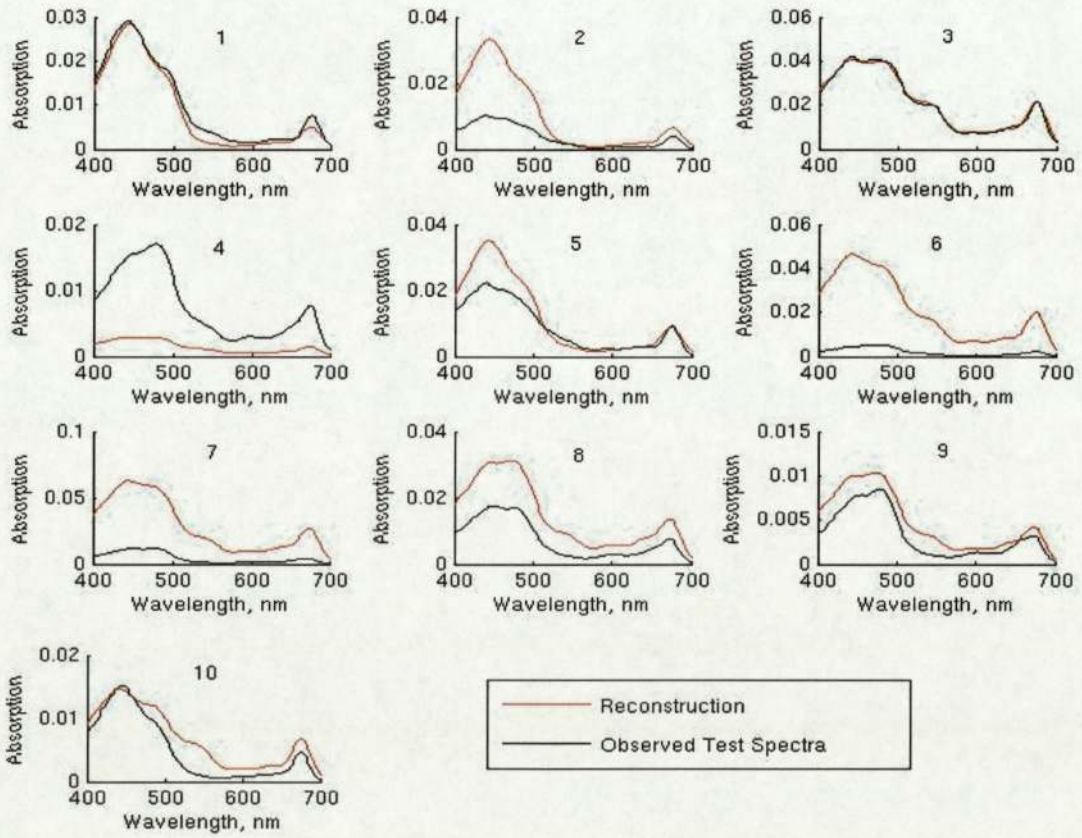


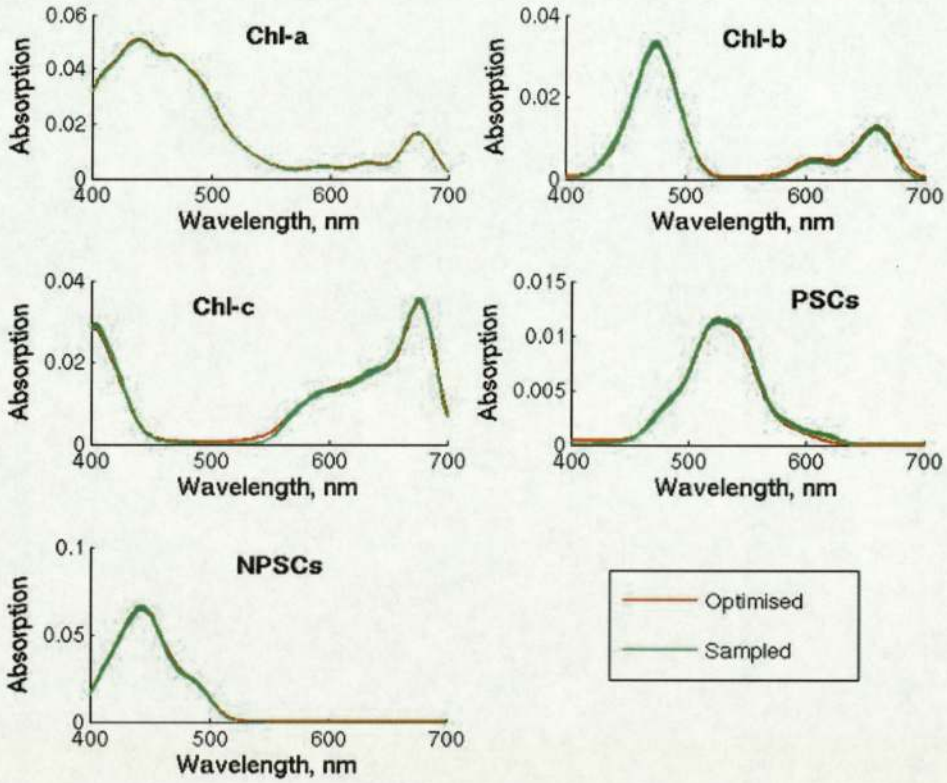
Figure 5.3.2.5: Spectra Reconstructions using the optimised weight vector compared to true observed spectra. The ten spectra are the full set of samples from the test set for cruise 2.



Finally, the implied reconstructions of specific absorption will be examined. The following plots show the specific absorption implied by the initial optimisation of the weight vector and the specific absorptions implied by fifty sample vectors from the posterior weight distribution. This enables pigment specific analysis of the variability of estimates across the sampled vectors for each cruise data set.

Looking first at the results using the full training set in figure 5.3.2.6, the variances are small. All of the sampled vectors produce implied absorptions with a very small range across the spectrum and are very close to the reconstruction produced by the optimised weight vector. These implied spectra however, are also very similar to the initialisation (figure 5.3.1.1). This may suggest that the model has not actually learnt from the data, which is quite possible given the large number of inputs and likely dependencies between them. It would also explain why the sample paths were more stable than for the separate cruises.

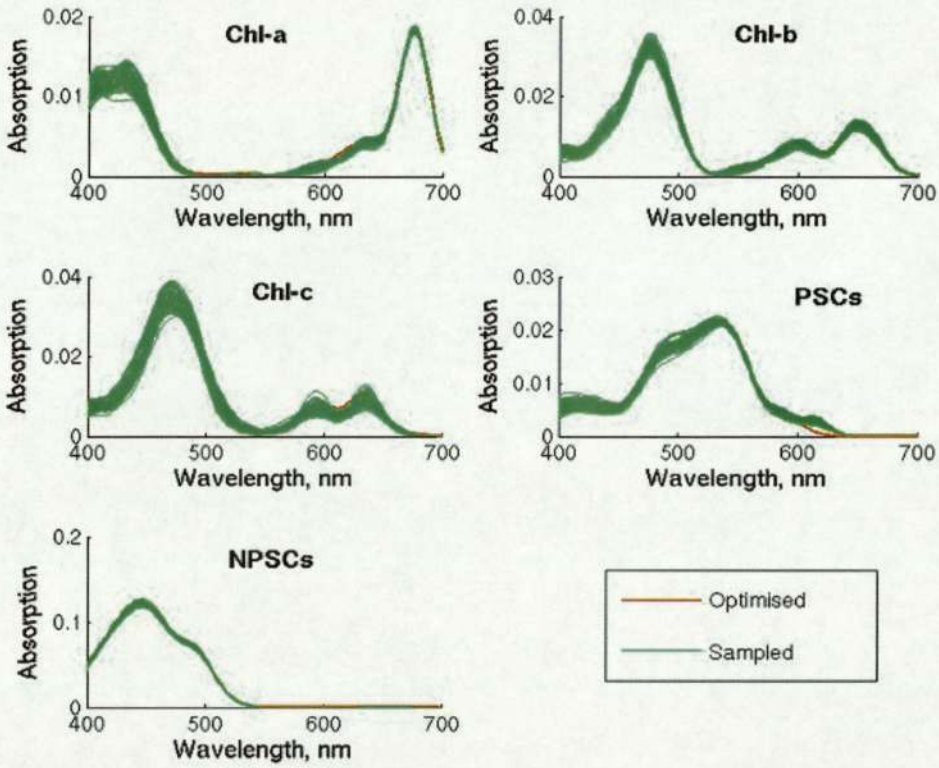
Figure 5.3.2.6: Implied specific absorption spectra using the optimised weight vector and fifty samples using all data



The results by cruise are quite different to those from the overall sampler and also vary significantly between cruises. The ranges are significantly bigger suggesting less certainty in the weights.

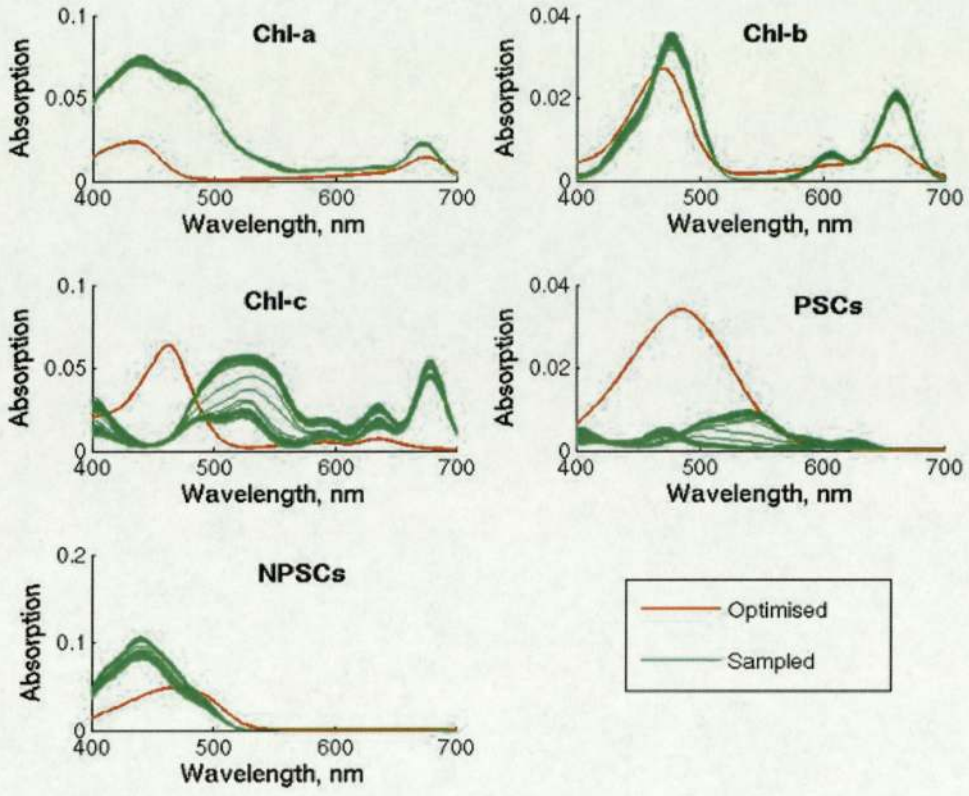
Results for cruise 2 (figure 5.3.2.7) and most other individual cruises also have a very similar structure to that estimated by the optimised weight vector, but have larger range of estimates. This variability appears greater at the lower wavelengths suggesting less certainty in these weights, as was indicated by previous error analysis (see section 3.7). The shapes of the spectra however have changed, indicating that the model has been able to learn characteristics specific to the cruises.

Figure 5.3.2.7: Implied specific absorption spectra using the optimised weight vector and fifty samples using cruise 2 data



Cruise 6 is slightly different and the weight samples produce a particularly large range of estimates especially for chlorophyll c (figure 5.3.2.8). These differ significantly from the optimised estimates suggesting that the optimisation may not have converged, that the data is simply more variable or that the data does not contain enough information to translate into an accurate mapping.

Figure 5.3.2.8: Implied specific absorption spectra using the optimised weight vector and fifty samples using cruise 6 data



The chl-c and NPSC plots show quite different implied spectra from the optimised values and from other cruises. Other cruises such as cruise 4 show some such discrepancies though not to this same extent. Previous analysis has not indicated any major characteristic differences between cruise 6 and the other cruises. These individual pigment absorption structures may again then be indicative of external influences, in particular the package effect, which has not been modelled here.

Having found a high level of uncertainty in some of the heights distributions the effect on the parameters of altering the variance will be briefly examined. Using the cruise 2 data the sampler is implemented with the noise on the heights amended to 10% and 30%. The greater variance tends to slightly increase the sampler rejection rates and reconstruction variability, but this simple experiment does not immediately highlight any major effects. The 20% noise estimate will therefore be considered satisfactory.

While there is clearly uncertainty in the samples there are some successes in reconstructing the spectra. The sampled weight values then will be accepted as reasonable estimates for the following concentration retrieval.

5.4 Concentration Retrieval

The predictive distribution is given by (5.23)

$$p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, D) = \int p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, \mathbf{w}) \; dp(\mathbf{w} \mid \mathbf{y}_{obs}, \mathbf{x}_{obs})$$
$$= \int p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{y}_{obs}, \mathbf{x}_{obs}) \; d\mathbf{w} \qquad (5.23)$$

Using Bayes' theorem (3.1) the posterior for a new unknown concentration $\mathbf{x}_{new}$ can be expressed as:

$$p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, \mathbf{w}) = \frac{p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w})\, p(\mathbf{x}_{new})}{p(\mathbf{y}_{new})} \qquad (5.24)$$

This can then be substituted into the predictive distribution (5.23) so that:

$$p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, D) \propto \int p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w})\, p(\mathbf{x}_{new})\, p(\mathbf{w} \mid \mathbf{y}_{obs}, \mathbf{x}_{obs}) \; d\mathbf{w} \qquad (5.25)$$

The weights can now be fixed to the sampled values so that for each sample $\mathbf{w_n}$, there is a corresponding posterior proportionality (5.26).

$$p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, D) \propto \int p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w_n})\, p(\mathbf{x}_{new}) \; d\mathbf{w} \qquad (5.26)$$

where n = 1,2,... 200.

The values $(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w_n})$ are assumed to be generated by the given model and again subject to Gaussian noise, so

$$p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w_n}) \sim N(\mathbf{y}_{new}, C_{y\,new}) \qquad (5.27)$$

where the mean $\mathbf{y}_{new}$ is the observed test absorption spectra and the standard deviation $C_{y\,new}$ is a fixed value equal to 10% of the mean absorption across all wavelengths and all samples for the test data.

The concentration prior $p(\mathbf{x}_{new})$ will be assumed gamma distributed, thus again ensuring positivity of the concentration measures. It parameters will be set based on only the training data to avoid any biasing of test experiments. The mean is set equal to $\tilde{\mathbf{x}}_{obs}$ - the mean concentration across all training examples per pigment and the variance to $C_{x\ obs}$ – the corresponding variance in concentration by pigment.

$$p(\mathbf{x}_{new}) \sim Ga(\alpha_P, \beta_P) \qquad (5.28)$$

where $\alpha_P = (\tilde{\mathbf{x}}_{obs})^2 / C_{x\ obs}$ and $\beta_P = \tilde{\mathbf{x}}_{obs} / C_{x\ obs}$

It is now possible to sample from $p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, D)$ to generate samples of predicted concentration for each of the sampled height vectors. The cost function (**F**) is derived from (5.26) and given by (5.29). By taking the fixed sample values $\mathbf{w_n}$ for the heights the integral over **w** becomes redundant.

$$\mathbf{F} = -\ln\{p(\mathbf{x}_{new} \mid \mathbf{y}_{new}, D)\} \qquad (5.29)$$
$$\mathbf{F} \propto -\ln\{\int p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w_n})\, p(\mathbf{x}_{new})\ d\mathbf{w}\}$$
$$\mathbf{F} \propto -\ln\{p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w_n})\, p(\mathbf{x}_{new})\}$$

The retrieval cost function has two components – the data error (5.30) and concentration error (5.31).

$$-\ln\{p(\mathbf{y}_{new} \mid \mathbf{x}_{new}, \mathbf{w_n})\} = \sum_i ((\beta_D/2).* \sum_n (\mathbf{y}_n - \mathbf{y}_{obs\ n})^2 \qquad (5.30)$$

$$-\ln\{p(\mathbf{x}_{new})\} = (\boldsymbol{\beta}_P * \mathbf{x}) - ((\boldsymbol{\alpha}_P - 1).*\ln(\mathbf{x})) + \ln(\Gamma(\alpha_P)) - \alpha_P * \ln(\beta_P) \qquad (5.31)$$

Together with the corresponding gradient function these provide the appropriate inputs to the sampler for retrieval of unknown concentrations. Sampling predicted concentrations, $\mathbf{x}_{new}$ from the posterior generates a Markov chain of samples for each $\mathbf{w_n}$:

$$[\{\mathbf{x}_{new\ m}\} \mid \mathbf{w}_1, \{\mathbf{x}_{new\ m}\} \mid \mathbf{w}_2, \ldots \{\mathbf{x}_{new\ m}\} \mid \mathbf{w}_{200}] \qquad (5.32)$$

where m=1,2,…M (the number of samples).

Optimisation will again be carried out prior to sampling to provide a reasonable initialisation. 50 burn in samples will be discarded and 50 samples then collected for each sampled height vector **w**.

5.5 Sampled Concentrations

Firstly, to check the underlying model performance the concentrations retrieved by optimisation alone will be analysed and compared to the corresponding true (observed) values. The heights will first be optimised and fixed and the concentrations also then optimised, thus providing a rough estimate without use of sampling. This ought to indicate whether the model has the potential to produce good results. The height optimiser uses 400 iterations followed by 10000 for the concentrations.

The initial results are seen in table 5.5.1 and are significantly poorer than those achieved by data driven methods. Modelling cruises separately is again the most successful method, but comparison with the error breakdown by cruise in table 4.12.1 highlights the inferior performance. Comparison with early versions of the data driven model however, particularly for cruises 2 and 4, the results achieved are much closer.

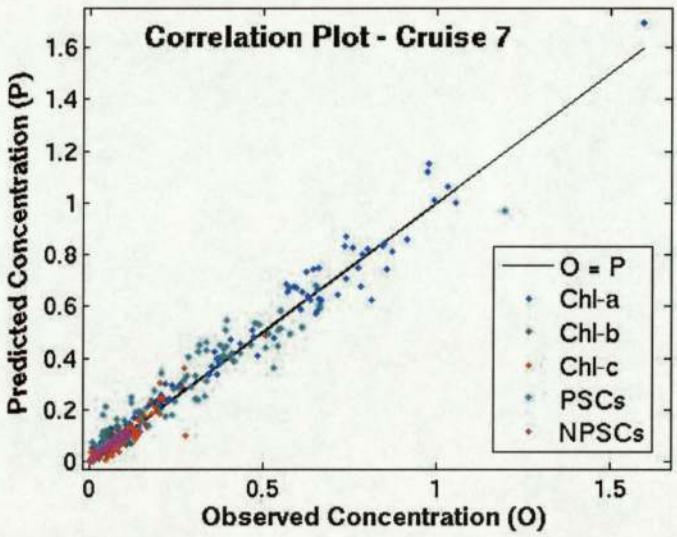
Table 5.5.1: Concentration retrieval errors from the optimised concentrations versus the true measures (test set)

Cruise	CORR	Mean % err	Bias	MAE	RMSE
1	0.949	129.57	0.093	0.0996	0.1373
2	0.949	56.99	0.013	0.0448	0.0703
3	0.424	88.27	0.0099	0.0738	0.1396
4	0.927	25.31	-0.0031	0.0188	0.0265
5	0.502	90.04	-0.036	0.0886	0.1522
6	0.506	101.46	0.0067	0.0503	0.0784
7	0.982	28.68	0.0095	0.0312	0.0471
8	0.960	36.19	0.016	0.0351	0.0518
ALL	0.867	61.17	-0.0015	0.0539	0.0896

There is a lot of variability between the cruises - more so than for the direct model and bad results are especially noticeable for cruises 1,3 5 and 6. A second experiment increases the height optimiser iterations to a maximum of 30000 to see if this affects the retrievals. The amendment produces some small improvements for individual cruises, but actually makes the overall model results slightly worse. On the problem cruises 5 and 6, the results were identical (to the recorded accuracy) and further investigation suggests that the concentration optimiser is becoming stuck in local minima.

Plotting these retrieved values against the observed concentrations confirms these differences between cruises. Cruises 2, 4, 7 and 8 produce quite structured correlation plots as in figure 5.5.1 with no particular pigments consistently badly predicted, while cruise 3 shows a much wider scatter.

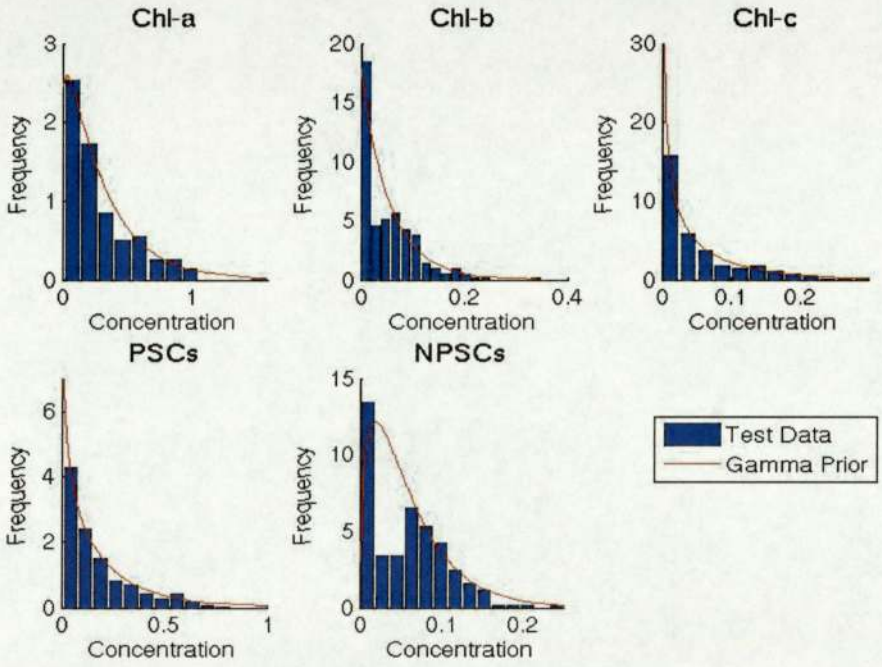
Figure 5.5.1: Retrieved versus true concentrations using the optimised heights and concentrations for cruise 7



Given that these errors are currently still quite large the sampling process is unlikely to retrieve concentrations accurately. The results though are quite promising for this early stage of modelling and there is no collapse of the model on any individual cruise. A small number of experiments implementing sampling will therefore be carried out to further assess the potential of the model. As described previously it will be a two stage sampling process where the heights will first be optimised and sampled and subsequently the concentrations.

Having carried out the retrieval the final gamma distributions implied for the concentrations will be compared to the true distribution of data. The gamma prior distributions appear to offer a reasonable fit whether modelling each cruise separately or altogether, though they do vary from model to model and between pigments. An example of the fit using the model for all cruise data is seen in figure 5.5.2. The fit appears to be best for chl-a, chl-c and the PSCs, so it is expected that these will be modelled slightly more successfully.

Figure 5.5.2: Final gamma prior distributions compared to the true test data distributions (all cruises)



An estimation of expected concentration for each pigment in each data sample can now be produced by averaging the sampled concentrations $\mathbf{x_m}$ over all sampled heights $\mathbf{w_n}$. This is then used to calculate a measure of retrieval error comparable with the data driven models, as seen in table 5.5.2.

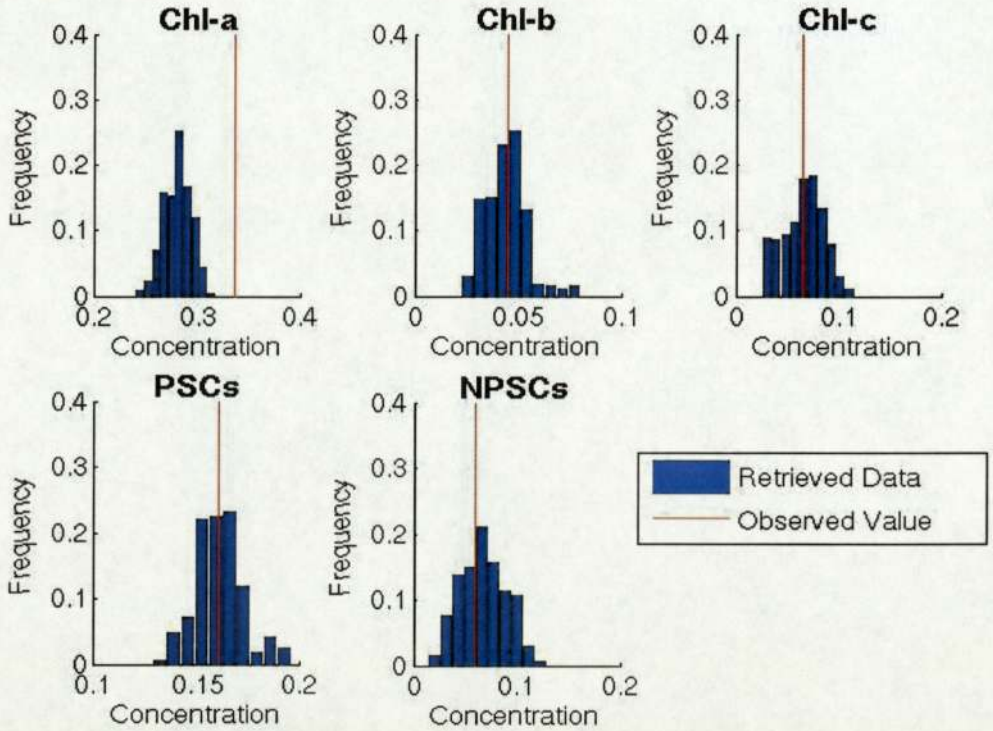
The results as expected following the optimisation based retrieval analysis, are far from rivalling those of the direct models at this stage. The fact that several cruises generate percentage errors of around 25 - 30% though, is encouraging and together with relatively high correlations shows the model to at least be functional. Cruises 1, 3, 5 and 6 are still problematic, as was indicated by the prior analysis, though could be improved using alternative initialisation and/or extended optimisations. Results are not particularly good at this stage but provide a foundation for further work.

Table 5.5.2: Concentration retrieval errors for the mean of the sampled concentrations versus the true measures (test set)
Height optimisation – 400 iterations, concentration optimisations – 10000 iterations

Cruise	CORR	Mean % err	Bias	MAE	RMSE
1	0.9451	119.62	0.0775	0.0879	0.1283
2	0.9532	41.595	0.0128	0.0403	0.0673
3	0.4521	81.59	0.0047	0.0715	0.1333
4	0.9256	25.0575	-0.0038	0.0191	0.0272
5	-0.0166	273.1394	0.1341	0.2816	0.6061
6	0.7938	63.7326	-0.0108	0.0308	0.0575
7	0.9805	31.5544	0.0090	0.0322	0.0488
8	0.9498	36.5908	0.0170	0.0387	0.0575
ALL	0.8685	62.59	-0.0021	0.0547	0.0897

Finally to illustrate the actual output of the model the distribution of retrieved samples for a single test set example is plotted as a histogram in figure 5.5.3. The true value of the concentration for each pigment is superimposed.

Figure 5.5.3: Distributions of retrieved concentration samples for the first test set example taken from cruise 8



This particular example shows relatively accurate retrieval of all pigments except chl-a with true values approximately coinciding with modal retrieved values. As is the case here for chl-

a however, at least one of the true pigment concentration values frequently falls completely outside the retrieved distribution. This illustrates that the retrieval is currently subject to significant errors. The high level of confidence reflected by these distributions may be due to the small number of samples taken and is a likely area for future improvement.

5.6 Generative Modelling Conclusions

The basic model appears to function well and provides a reasoned framework for the mapping from absorption spectra to pigment concentrations. Despite the limited number of experiments thus far, results are encouraging and several useful conclusions may be drawn.

Gaussian bands: The RBF offers a good fit to the estimated specific absorption spectra.

Gamma priors: The gamma approximation for the prior distribution of concentrations also looks to be appropriate and a good fit.

Individual cruise models: Results by cruise are better and there is evidence that these models are capable of learning cruise specific weights. Highly variable results between cruises are indicative of cruise specific noise sources and/or non-modelled relevant variables, such as the package effect and photoacclimation.

Model parameters: The model is likely to benefit from improved initialisation, better optimisation and increased sample sizes. An optimal 'burn in' period may also be identified using convergence diagnostics (Nabney, 2002).

The data: Uncertainty regarding the data remains a source of difficulty in modelling and understanding the results obtained.

While results at this stage are much worse than those from data driven retrieval there remains much scope for further experimentation and improvement. This includes the opportunity to incorporate cell size and spectra gradient data and to integrate more fully the log transform, which has appeared useful throughout. Parameterisation of the package effect appears essential and may largely explain performance differentials relative to the data driven model.

Chapter 6

Conclusions

Modelling the relation between absorption spectra and pigment concentrations is a difficult problem exacerbated by the lack of a larger, more reliable data set. It has however still been possible to retrieve concentrations to some degree of accuracy and also to make some valid inferences.

6.1 A Summary of Outcomes

Visualisation highlighted several concepts, which have since been confirmed as relevant aspects of the modelling process. Evidence of linear characteristics within the data were first noted during visualisation and later verified by the success of both the GLM and the linear based, generative model. Cruise differentials were similarly evident at an early stage and cruise specific models since have been proved to be far superior. The same is true of log transformation of variables and also the size data. Visualisation also facilitated removal of outliers and produced a useful lower dimension representation of the data.

Forward modelling successfully reconstructed the spectra with mean errors of less than 11%. Several important factors were highlighted including the recurring themes of linearity, cruise differentials and log transform benefits, whilst the effects of size inputs were inconsistent. ARD priors improved models marginally though analyses by wavelength produced inconsistent results across the various models. Significant dependence on the division of the data was also demonstrated, thus emphasising the problems of working with a limited data set.

Data driven inverse models have produced useful estimates of pigment concentration. The most success is achieved by working with log transforms of the data and by modelling cruises and pigments individually. Further improvements although small are achieved using size distribution data and absorption spectra gradient data alongside the standard absorption inputs. Pigment concentrations were retrieved with errors of approximately 14% and several useful conclusions reached, which provided a strong basis for a generative absorption model.

Generative modelling has so far shown that it is possible to map to some degree a relation between spectra and concentrations. Concentration retrievals however, are currently much worse than by direct methods. Uncertainty in the data was a problem and it was difficult to distinguish between the effects of noise and the influence of non-modelled variables. This model though attaches more physical meaning to the data and there remains much scope for adaptation and further improvement.

6.2 Limitations and Constraints

The project has encountered several limitations, which have made it difficult to draw many firm conclusions.

The data set itself has been a constraint and is subject to many potential sources of noise and uncertainty. The data has been collated over 12 years and so changes in measurers and instruments were unavoidable. Therefore, despite strict and precise measurement protocols some degree of error is inevitable. There are also still possible unknowns including missing (not measured) pigments and light absorbing compounds within the water and phytoplankton.

Several potentially relevant variables could not be considered directly due to the limited data set. These include temperature, nutrient concentration and depth of waters. The data set is also relatively small given the variety of potential noise sources, which is particularly relevant when modelling cruises individually.

The final key limitation has been time. Even with limited data availability the amount of information to be considered and the many permutations thereof is vast. There remain many unexplored avenues some of which will be suggested for further work in chapter 7.

Chapter 7

Further Work

As discussed previously availability of data has imposed some limitations on the work to date. The current data set appears almost exhausted in terms of what further information can be gained from it. Further work would almost certainly benefit from a larger and more comprehensive dataset, which might enable more solid inferences. Ideally measurers, methods and equipment used would be identical for all data collection and subsequent pre-processing. Additional data collection however, would require a substantial investment of both time and money.

Visualisation was indicative of size related structure and it may be that the phytoplankton cell size would prove more useful given an alternative measure, or if modelled explicitly as part of the package effect. There certainly appears to be something missing from both data driven and generative models and so experimentation with additional variables, such as water depth, temperature and time of year could be useful. Feature selection methods such as Independent Component Analysis (ICA) could be applied to a data set containing such extra variables to determine which are useful inputs. A larger data set would in itself engender greater confidence, particularly on a by cruise basis.

ARD is an area, which given more time could be explored more fully. In particular there is scope to build upon the wavelength analysis produced during forward modelling. Possibilities include using averaging of ARD output and iterative applications. If a more stable result is achieved then application to the inverse problem may be more practicable. Experiments thus far indicate that the relevance of pigments varies greatly across the spectrum, so there is certainly potential to significantly reduce the dimension of the problem given a reliable result.

A significant amount of linearity is clearly present in this problem and is captured successfully by the models. Advancement of the models however, is limited by the absence of components accurately representing packaging and acclimation effects and therefore the non-linear aspect of the mapping. Inclusion of size parameters produced some improvement, further confirming the relevance of the missing components and the need for a more informative associated parameterisation. Further work then would almost certainly benefit from focusing on these non-linear elements.

There is certainly scope to improve both the data driven model and the generative model. The generative model in particular may benefit from simple adaptations, such as more accurate initialisation and more thorough optimisation, burn in and sampling. The model is yet to fully capitalise on the findings of the data driven model and may benefit from introducing elements such as gradient and size data. Subsequent improvements would rely on more significant changes, in particular inclusion of variables better reflecting package and acclimation effects. The data driven model on the other hand, implicitly models both the package effect and photoacclimation. Additional inputs representative of these could however improve the model further, for example a more detailed size breakdown or a depth parameter for each sample.

References

Cited References

Aguirre-Gomez, R., Weeks, A. R. & Boxall, S. R. 1998. The Identification of Phytoplankton Pigments from Absorption Spectra. *International Colloquium on Photosynthesis and Remote Sensing,* pp.191-205

Bishop, C. M. 2003. *Neural Networks for Pattern Recognition.* Oxford University Press, New York.

Bricaud, A., Renard, F., Claustre, H., Thiria, S., Crepon, M., Mejia, C., Cornford, D., Evans, D., 2003. WP3 "Performing different NN-based techniques to retrieve ocean constituent concentrations from sea-level reflectance spectra" Part 1. Determination of MLPs linking phytoplanktonic absorption to pigment concentrations in Case 1 waters. *Second year report February 2002- February 2003.*

Bricaud, A., Claustre, H., Ras, J., and Oubelkheir, K., 2004. Natural variability of phytoplanktonic absorption in oceanic waters: Influence of the size structure of algal populations. *Journal of Geophysical Research*, **109**, C11010.

Evans, D., & Cornford, D., 2003. Modelling absorption spectra: pigment concentration retrieval. *Technical Report Draft 1.0.,* pp. 1-27.

Finkel, Z. V. 2001. Light absorption and size scaling of light-limited metabolism in marine diatoms. *Limnology and Oceanography.* **46(1)** pp.86-94.

Finkel, Z.V., Irwin, A.J., and Schofield, O. 2002. (Institute of Marine and Coastal Sciences, Rutgers University). *Using Cell Size to Predict Intracellular Pigment Concentrations Under Sub-saturating Growth Irradiance.* [Online] [Accessed 2005] Available from World Wide Web: <http://www.imcs.rutgers.edu/cool/coolresults/agu2002/posters/0211_oceanopticsXVI_zfinkel.ppt>

Hoepffner, N., & Sathyendranath, S., 1993. Determination of the Major Groups of Phytoplankton Pigments From the Absorption Spectra of Total Particulate Matter. *Journal of Geophysical Research* **98,** pp. 22789-22803.

Morel, A., & Bricaud, A., 1981. Theoretical results concerning light absorption in a discrete medium, and application to specific absorption of phytoplankton. *Deep-Sea Research* **28A,** pp.1375-1393.

Nabney, I. T. 2002. *Netlab: Algorithms for Pattern Recognition.* Springer, Gateshead.

Neal, R., M. 1996. *Bayesian Learning for Neural Networks.* Springer, New York.

Ocean Science Bedford Institute of Oceanography, 2002. *Chlorophyll Concentration* [Online]. [Accessed 2004]. Available from World Wide Web: <http://www.mar.dfo-mpo.gc.ca/science/ocean/ias/seawifs/seawifs_2.html >

Saad, D., 2004. Bayesian Methods – *Aston NCRG Lectures.*

Smith, M., and Barnett., P., 2004. *What is retransformation bias, and how can it be corrected?* [Online][Accessed 2005] Available from World Wide Web: <www.herc.research.med.va.gov>

Thomas, D.A, (Network Montana Project (NMP)) 1997. *Phytoplankton Background - Exploring Phytoplankton Pigment Concentrations* [Online][Accessed 2004] Available from World Wide Web: <http://www.math.montana.edu/~nmp/materials/ess/hydrosphere/expert/phyto/phytobackg.html>

Wozniak, B., Dera, J., Ficek, D., Majchrowski, R., Kaczmarek, S., Ostrowska, M., Koblentz-Mishke, O. I., 2000. Model of the in vivo spectral absorption of algal pigments. Part 1. Mathematical apparatus. *Oceanologia* **42(2),** pp.177-190.

Wozniak, B., Dera, J., Ficek, D., Majchrowski, R., Kaczmarek, S., Ostrowska, M., Koblentz-Mishke, O. I., 2000. Model of the in vivo spectral absorption of algal pigments. Part 2. Practical applications of the model. *Oceanologia* **42(2)**, pp.191-202.

Additional References

Bricaud, A., Babin, M., Morel, A., & Claustre, H., 1995. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: Analysis and parameterization. *Journal of Geophysical Research* **100,** pp13321-13332.

Stuart, V., Sathyendranath, S., Platt, T., Maass, H., & Irwin, B.D., 1998. Pigments and species composition of natural phytoplankton populations: effect on the absorption spectra. *Journal of Plankton Research* **20,** pp.187-217.

Sathyendranath, S., Hoge, F. E., Platt, T., & Swift, R. N., 1994. Detection of phytoplankton pigments from ocean color: improved algorithms. *Applied Optics* **33,** pp.1081-1089.

Appendices

Appendix A.1 - Details of data collection

The known details for each cruise are as follows:

Cruise	Location	Date	No. of samples
1	NW Mediterranean	April 1990	9
2	N tropical Atlantic	October 1991	48
3	Pacific	September 1994	77
4	Pacific	November 1994	180
5	E & W Mediterranean	May 1996	106
6	Mediterranean	September 1999	446
7	Alboran sea	January 1998	446
8	N Atlantic	2001	213
			1525

Appendix A.2 - Eigenvectors

Figure A.2.1: Eigenvectors from linear space, normalised by removing mean

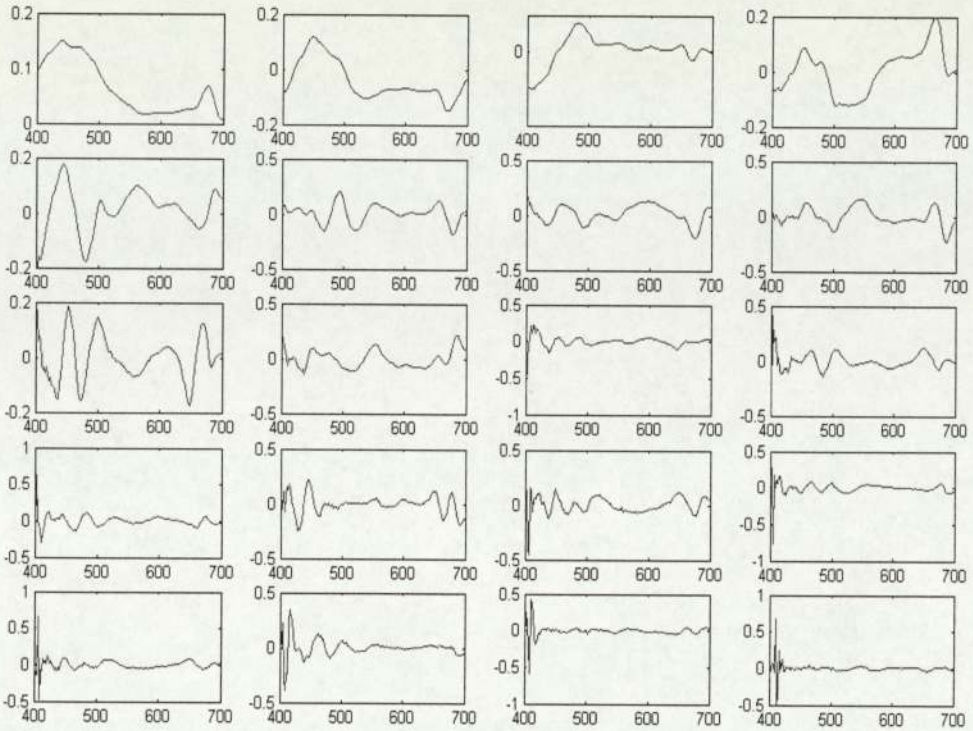
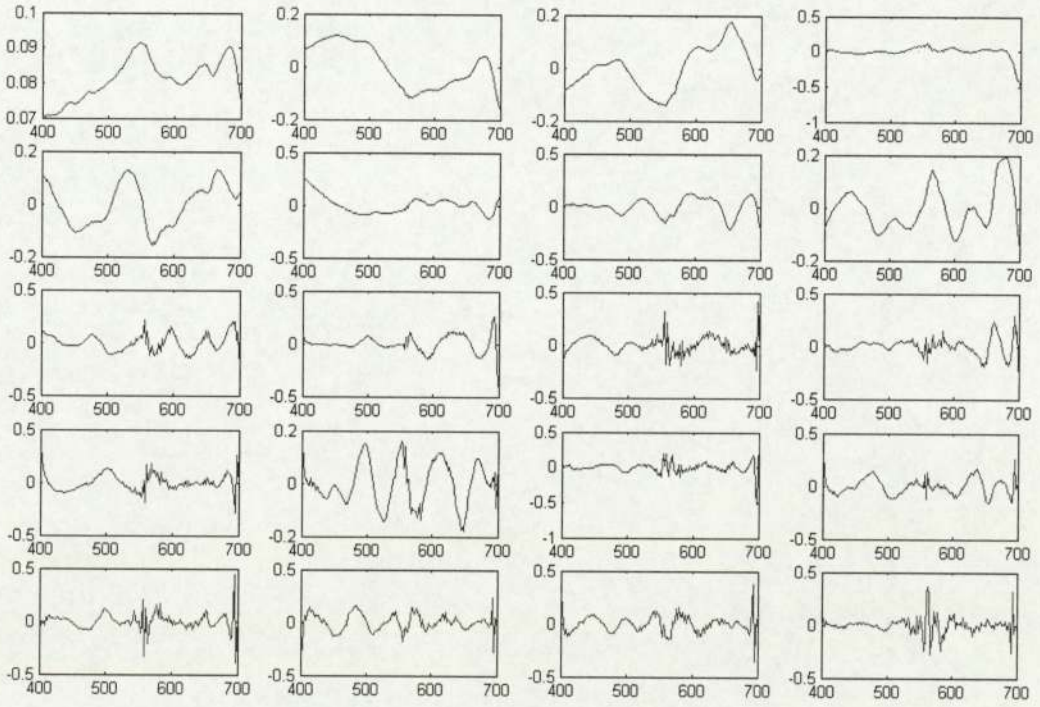


Figure A.2.2: Eigenvectors from log space, normalised by removing mean



Appendix A.3 – PCA visualisation plots

Figure A.3.1:

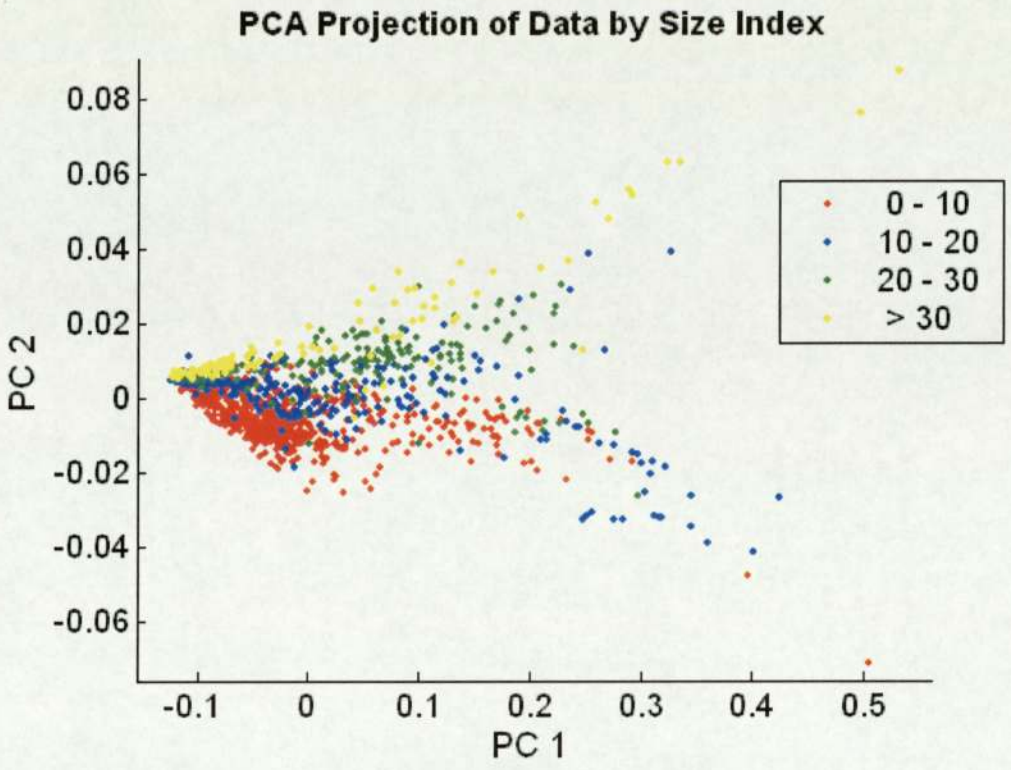
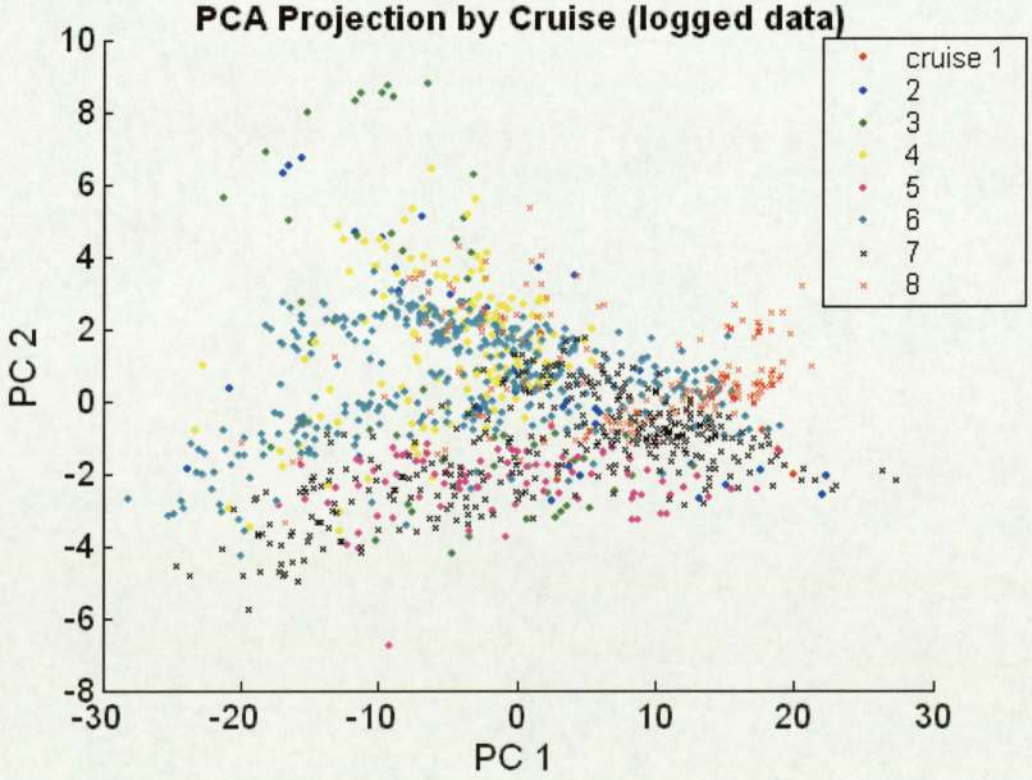


Figure A.3.2:



Appendix A.3 – PCA visualisation plots

Figure A.3.1:

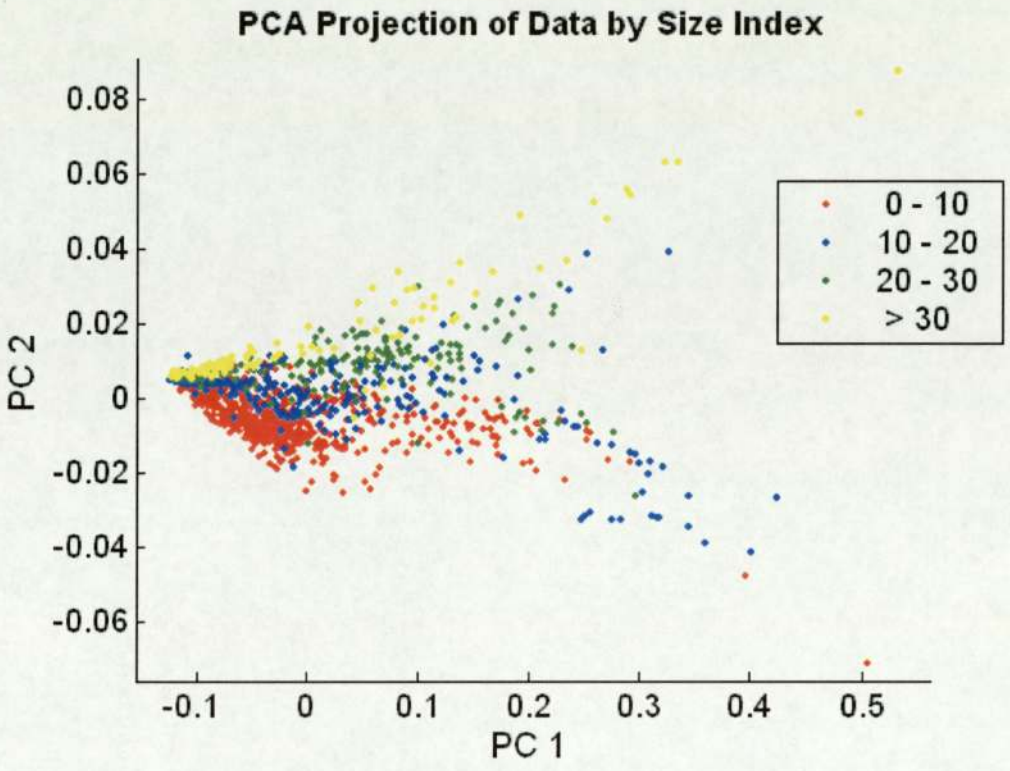
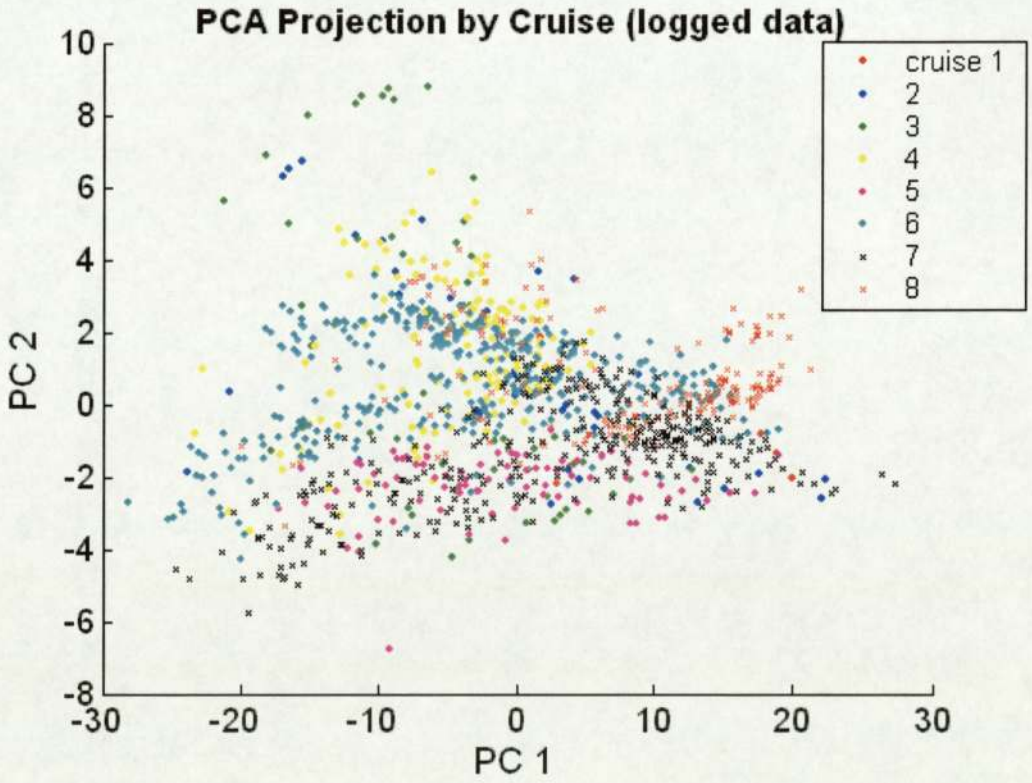


Figure A.3.2:



Appendix A.4 – Removed Outlier Breakdown

TRAINING DATA			TEST DATA		
TRAINING INDEX	SAMPLE	CRUISE	TEST INDEX	SAMPLE	CRUISE
4	5	1	154	762	6
6	7	1	203	983	7
7	9	1	204	984	7
12	14	2	211	1024	7
91	114	3	212	1026	7
256	320	5	222	1084	7
432	542	6	228	1106	7
601	753	6	233	1142	7
652	813	6			
694	867	7			
702	878	7			
716	900	7			
753	951	7			
773	972	7			
774	973	7			
780	982	7			
788	993	7			
789	994	7			
790	995	7			
796	1004	7			
797	1005	7			
806	1014	7			
814	1025	7			
815	1027	7			
823	1035	7			
826	1038	7			
836	1050	7			
843	1058	7			
854	1072	7			
861	1082	7			
862	1083	7			
872	1095	7			
920	1154	7			
940	1178	7			
957	1202	7			
958	1203	7			
997	1250	7			
1032	1291	7			
1041	1302	7			
1048	1309	7			
1153	1443	8			

Appendix A.5 – Error formulae

All errors are calculated in MATLAB and make use of MATLABs built-in functions.
'resid' is the prediction error vector or matrix.

Correlation coefficient (CORR) = corr2(A, B)

- calculates the correlation between two variables A and B

Root mean squared error (RMSE) = sqrt(mse(resid));

Mean absolute error (MAE) = mae(resid);
Or equivalently MAE = mean(abs(resid));

- the mean magnitude error over all absorption predictions

Bias = mean(resid); for error vectors
Bias = mean2(resid); for the case of an error matrix

Percentage errors are calculated only after omitting those points which are less than or equal to 0.01. This is to avoid extreme error values where the target values are very small.
idx = find(abs(target)>0.01);
mean percentage error = mean(abs(resid(idx)./target(idx))).*100;
maximum percentage error = max(abs(resid(idx)./target(idx))).*100;

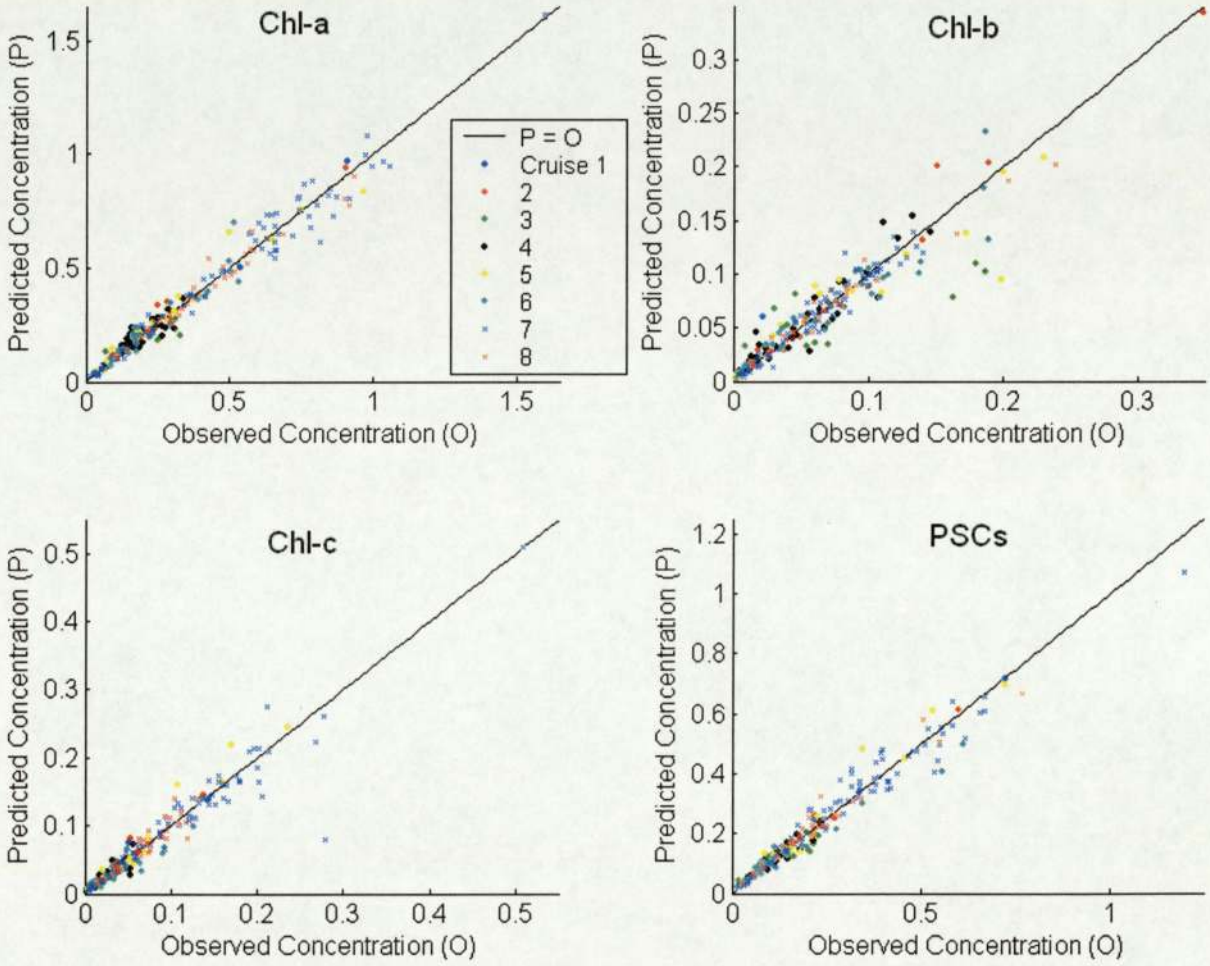
In forward models the variance in the errors at each wavelength is also calculated using:
variance=var(resid)

Appendix A.6 – Permuted Data Set Splits

Cruise	1	2	3	4	5	6	7	8	Total
Total Data	6	47	76	180	105	442	408	212	1476
Original Trn	4	37	61	144	84	354	326	169	1179
Original Test	2	10	15	36	21	88	82	43	297
PERM1	4	7	19	40	18	99	84	26	297
PERM2	2	5	19	31	27	89	91	33	297
PERM3	2	8	17	31	25	89	93	32	297
PERM4	4	13	17	33	27	96	83	24	297

Appendix A.7 – Correlation Plots

Figure A.7.1: Optimal Direct Inverse Model - Correlation Plots by Pigment



Appendix A.8 – Overall/ By Cruise Model Comparison Plots

The model compared is the GLM using the full 151 PC inputs and no size inputs, in linear space and simultaneously retrieving all pigments.

Figure A.8.1: Overall versus By Cruise Model Comparison for Chl-a

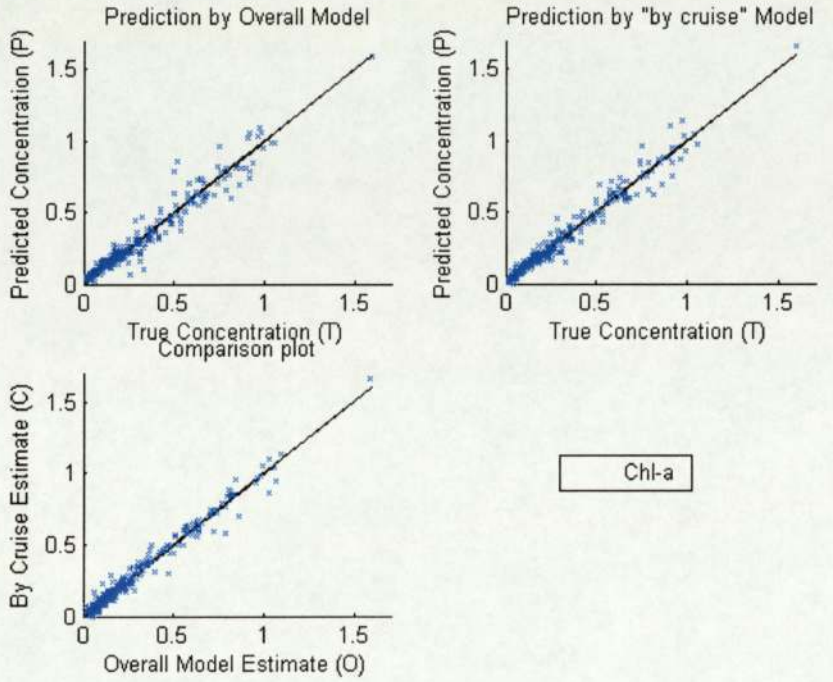


Figure A.8.2: Overall versus By Cruise Model Comparison for Chl-b

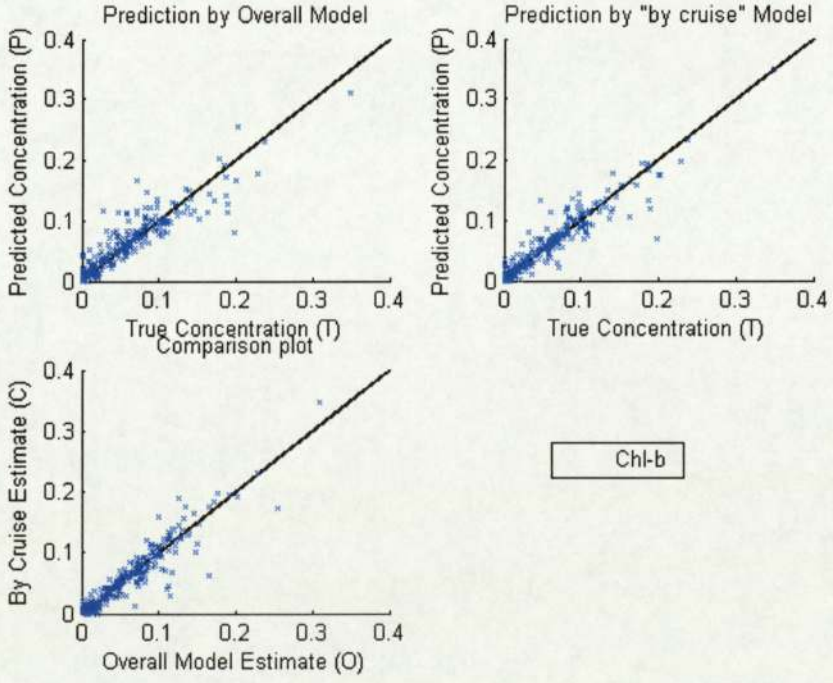


Figure A.8.3: Overall versus By Cruise Model Comparison for Chl-c

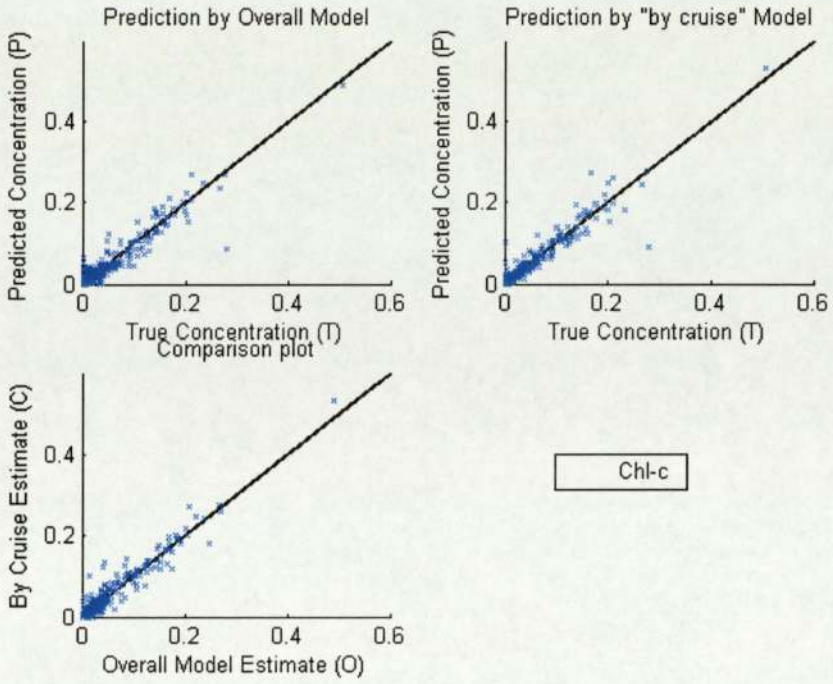


Figure A.8.4: Overall versus By Cruise Model Comparison for PSCs

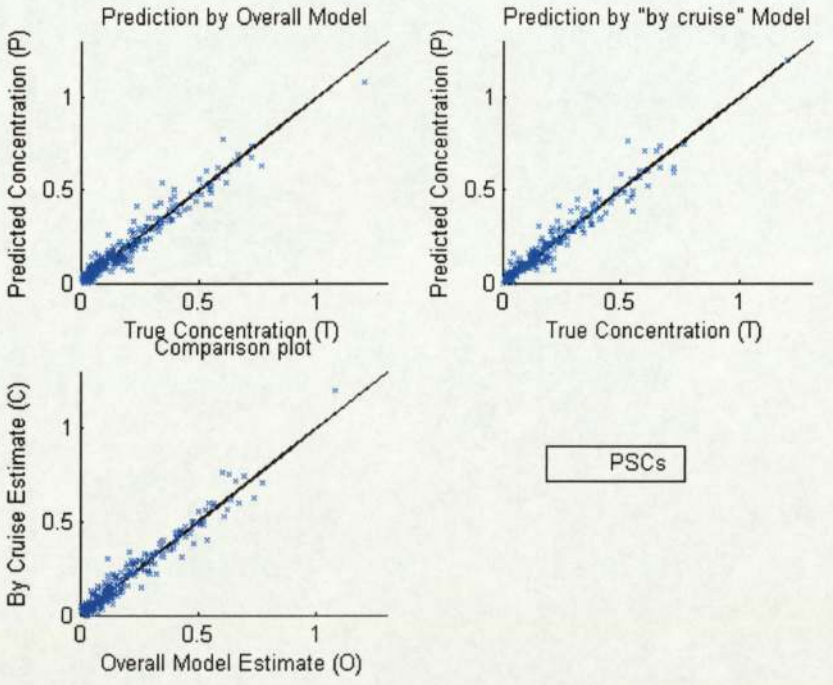
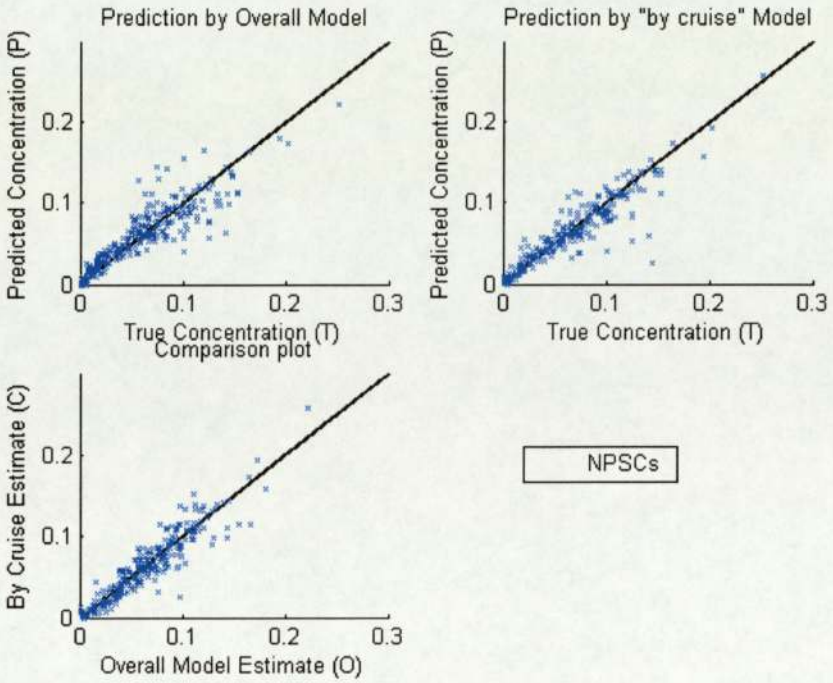


Figure A.8.5: Overall versus By Cruise Model Comparison for NPSCs



Abstract

In this project I will explore the concept of entropy. I will study different measures of entropy and the context in which they are used. I will discuss methods for efficiently approximating probabilities by maximising the entropy of the distribution given a set of constraints.

I will first look at the requirements a measure of entropy should fulfil and then go on to look at measures that have been suggested by mathematicians through history. Following on from this, I will then use the method of optimization, which involves maximisation and minimisation of functions. Despite all the mathematical techniques available in entropy optimization, Lagrange's method of constrained optimization will be the preferred method for entropy optimization.

The next part of the project will be to explore principles of entropy maximisation. This will analyze the available information in order to determine a unique probability distribution, avoiding using any information not given to us. The probability distribution that will be used consistent with the given constraints is the one that has maximum uncertainty. I will then go on to lok at maximum entropy principles suggested by mathematicians over the years and the context in which they are used.

The concept of entropy or uncertainty has played an increasingly significant role in the formulation of probabilistic systems, which are encountered in a variety of disciplines. In this project I will be referring to the information-theoretic entropy rather than the thermodynamic entropy.

What is entropy?

Entropy is a measure of the disorder suggesting a transformation from order to disorder. Entropy can be considered simply as the dictionary definition for uncertainty.

Uncertainty plays a very significant role in our perceptions about the world around us. As our perception of the world becomes more and more complex, the number of phenomena about which we are uncertain increases and the uncertainty about each of the phenomenon also increases. One way of decreasing our uncertainty is to collect information, but it is normally this information that increases our uncertainty, rather than decreasing our uncertainty.

Entropy can be defined in different ways depending on the application in which it is being used, in a probabilistic sense it can be defined a measure of the probability of a particular result. One example of this is if we consider the probabilities for the sum of two dice in backgammon. Below, the left column indicates the sum of the two dice, the next column lists all the possible combinations that give that sum, the third column counts the number of combinations for that sum. As you can see, there are a total of 36 different combinations for the two die, and each are equally probable to occur for "honest" dice. Thus the probability of getting a particular sum, as shown in the last column, is just the number of combinations divided by 36.

Sum	Combinations	Number	Probability
2	1-1	1	1/36=3%
3	1-2, 2-1	2	2/36=6%
4	1-3, 3-1, 2-2	3	3/36=8%
5	2-3, 3-2, 1-4, 4-1	4	4/36=11%
6	2-4, 4-2, 1-5, 5-1, 3-3	5	5/36=14%
7	3-4, 4-3, 2-5, 5-2, 1-6, 6-1	6	6/36=17%
8	3-5, 5-3, 2-6, 6-2, 4-4	5	5/36=14%
9	3-6, 6-3, 4-5, 5-4	4	4/36=11%
10	4-6, 6-4, 5-5	3	3/36=8%
11	5-6, 6-5	2	2/36=6%
12	6-6	1	1/36=3%

The most probable result, occurring one-sixth of the time, is to get *seven*.

Here, without going into too much detail, you can clearly see that a seven is the result with the highest probability, and a 2 or a 12 have the lowest probability.

The probabilities for dice lead us to our first definition of the entropy:

Entropy: A measure of the probability of a particular result.

The equivalent of this in information theory is that the entropy is the measure of disorder in a system.

Entropy: A measure of the disorder or randomness in a closed system.

We can think of entropy as a measure of the disorder in a system. This is reasonable because what we think of as "ordered" systems tend to have very few configurational possibilities, and "disordered" systems have very many. Consider, for example, a set of 10 coins, each of which is either heads up or tails up. The most "ordered" states are 10 heads or 10 tails, in either case; there is exactly one configuration that can produce the result. In contrast, the most "disordered" state consists of 5 heads and 5 tails, and there are 252 ways to produce this result.

Under the statistical definition of entropy, the second law of thermodynamics states that the disorder in an isolated system tends to increase. A broader interpretation of this is that the uncertainty in the world always tends to increase. This can be understood using our coin example. Suppose that we start off with 10 heads, and re-flip one coin at random every minute. If we examine the system after a long time has passed, it is possible that we will still see 10 heads, or even 10 tails, but that is not very likely, it is far more probable that we will see approximately as many heads as tails.

During this project I will be concentrating my efforts on the case of probabilistic uncertainty. For example, as I have shown before there may be uncertainty on the result when flipping a coin or throwing a die. There may be n possible outcomes in each of the situations I have just mentioned, and there probabilities may be:

$p_1, p_2, \ldots, p_n$ where $p_1 \geq 0, p_2 \geq 0, \ldots\ldots\ldots, p_n \geq 0, \sum_{i=1}^{n} p_i = 1.$

We are uncertain as to what the actual outcome will be, we may not even know the values of $p_1, p_1, \ldots\ldots, p_n$.

Different probability distributions have different uncertainties associated with them, for example, it can easily be seen that the uncertainty of a probability distribution (0.5, 0.5) for a head or a tail is much more than the uncertainty of the probability distribution (0.00001, 0.99999) of winning the lottery. The uncertainty with the probability of outcomes is called probabilistic uncertainty, which I will now refer to as entropy.

Measures of entropy

General requirements of a measure of uncertainty of a probability distribution

If the probabilities of n possible outcomes $A_1, A_2,\ldots\ldots,A_n$ of an experiment are $p_1, p_2,\ldots\ldots, p_n$, this gives rise to the probability distribution:

$P = (p_1, p_2,\ldots., p_n); \sum_{i=1}^{n} p_i = 1$ where $p_1 \geq 0, p_2 \geq 0,\ldots\ldots\ldots, pn \geq 0$.

There is an uncertainty as to the outcome when the experiment is performed. Any measure of this uncertainty should satisfy the following requirements:

1. It should be a function of $p_1, p_2,\ldots., p_n$ so that we can write it as:
   $S = S_n(P) = S_n(p_1, p_2,\ldots., p_n)$

2. It should be a continuous function of $p_1, p_2,\ldots., p_n$, i.e. small changes un p_1, $p_2,\ldots., p_n$ should cause small changes in S_n.

3. $S_n(p_1, p_2,\ldots., p_n)$ should be permutationally symmetric, i.e. S should not change when $p_1, p_2,\ldots., p_n$ are permuted among themselves, since uncertainty should not change when outcomes are labelled differently.

4. it should not change if an impossible event is added to the probability scheme, i.e. $S_{n+1}(p_1, p_2,\ldots., p_n, 0) = S_n(p_1, p_2,\ldots., p_n)$.

5. it should be minimum and possibly zero when there is no uncertainty about the outcome. Thus it should vanish when one of the outcomes is certain to happen so that $S_n(p_1, p_2,\ldots., p_n) = 0$ when $p_i = 1$, $p_j = 0$ when $j \neq i$.

6. It should be maximum when there is maximum uncertainty which arises when the outcomes are equally likely so that S_n should be maximum when
   $p_1 = p_2 = \ldots\ldots = p_n = \frac{1}{n}$.

7. The maximum value of S_n should increase as n increases.

8. For two independent probability distributions .

$$P = (p_1, p_2, \ldots, p_n),\ Q = (q_1, q_2, \ldots, q_m);\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{m} q_j = 1.$$

The uncertainty of the joint probability distribution should be the sum of their uncertainties, i.e.

$$S_{n\,m}(PUQ) = S_n(P) + S_m(Q)$$

Shannon's measure of uncertainty

In 1948 Claude E Shannon suggested the following measure:

$$S_n = (p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i.$$

It can easily be seen that this measure is a function of $p_1, p_2, \ldots, p_n$, satisfying [1]. It is also a continuous and symmetric function and if we replace 0 ln 0 by 0, it does not change when an impossible event is added, satisfying [2], [3] and [4]. When one of the probabilities is unity and the others are zero, its value is zero and this is its minimum value, satisfying requirement [5]. To find its maximum value, we can use Lagrange's method to maximise,

$$-\sum_{i=1}^{n} p_i \ln p_i - \lambda \left[\sum_{i=1}^{n} p_i - 1 \right].$$

Since $p_i \ln p_i$ is a convex function, $\sum_{i=1}^{n} p_i \ln p_i$ is a convex function, therefore

$-\sum_{i=1}^{n} p_i \ln p_i$ is a concave function and its local maximum is a global maximum. The maximum value of S_n is:

$$-\sum_{i=1}^{n} \frac{1}{n} \ln \left(\frac{1}{n}\right) = \ln n$$

and this goes on increasing as n increases, hence satisfying requirements [6] and [7] from earlier. For requirement [8] to be satisfied, I have to show that

$$S_{n\,m}(PUQ) = S_n(P) + S_m(Q)$$

Hence

$$S_{n\,m}(PUQ) = -\sum_{j=1}^{m}\sum_{i=1}^{n}(p_i\,q_j)\,\ln\,(p_i\,q_j)$$

$$= -\sum_{j=1}^{m} q_j\,[\sum_{i=1}^{n} p_i\,\ln\,p_i\,] - \sum_{i=1}^{n} p_i\,[\sum_{j=1}^{m} q_j\,\ln\,q_j\,]$$

$$= \sum_{j=1}^{m} q_j\,S_n(P) + \sum_{i=1}^{n} p_i\,S_n(Q)$$

$$= S_n(P) + S_m(Q)$$

This shows that requirement [8] is also satisfied by Shannon's measure of entropy. Shannon showed that any measure that satisfied all these requirements must be in the form:

$$-K\sum_{i=1}^{n} p_i \ln p_i,$$

where K is an arbitrary positive constant. Any other measure of uncertainty can only be obtained by modifying one or more of the requirements.

There are also properties which Shannon's measure of entropy satisfies, but which are not general requirements. They are the following:

- if **p** and **q** are not necessarily independent, let

  $P(X = x_i) = p_i \qquad P(Y = y_j \mid X = x_i) = q_{ij}$,

  So that

  $P(X = x_i, Y = y_j) = p_i\,q_{ij}$

  and

$$S_{mn}(p*q) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p_i\,q_{ij}\,\ln p_i\,q_{ij}$$

$$= -\sum_{i=1}^{n}[\,p_i\,\ln\,p_i\sum_{j=1}^{m} q_{ij}\,] - \sum_{i=1}^{n}[\,p_i\sum_{j=1}^{m} q_{ij}\,\ln\,q_{ij}\,].$$

Now

$$\sum_{j=1}^{m} q_{ij} = \sum_{j=1}^{m} P(Y = y_j \mid X = x_i) = 1.$$

Since it is certain that y must take one of the values $y_1, y_2, \ldots\ldots, y_m$ when $X = x_i$. Thus $(q_{i1}, q_{i2}, \ldots\ldots, q_{im})$ represents the conditional probability distribution of the outcomes of the second experiment when the first experiment has resulted in $X = x_i$. Thus,

$$S_{mn}(p*q) = S_n(p) + \sum_{i=1}^{n} p_i S_m(q_j)$$

Where $S_m(q_j)$ is the entropy of the conditional probability distribution of outcomes of the second experiment when the first experiment has resulted in the *i*th outcome. Thus, the entropy of the joint distribution is equal to the entropy of the first experiment plus the expected value of the conditional entropy of the second experiment.

- Another property Shannon's measure satisfies is,

$$S_{n-1}(p_1 + p_2, p_3, \ldots\ldots, p_n) = -(p_1 + p_2)\ln(p_1 + p_2) - \sum_{i=1}^{n} p_i \ln p_i$$

$$= -(p_1 + p_2)\ln(p_1 + p_2) + p_1 \ln p_1 + p_2 \ln p_2 - \sum_{i=1}^{n} p_i \ln p_i$$

so that

$$S_n(p_1, p_2, p_3, \ldots\ldots, p_n) = S_{n-1}(p_1 + p_2, p_3, \ldots\ldots, p_n) + (p_1 + p_2)\left[-\frac{p_1}{p_1 + p_2} \ln \frac{p_1}{p_1 + p_2} - \frac{p_2}{p_1 + p_2} \ln \frac{p_2}{p_1 + p_2}\right]$$

$$= S_{n-1}(p_1 + p_2, p_3, \ldots\ldots, p_n) + (p_1 + p_2)\, S_2\left[\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right]$$

so that

$$S_{n-1}(p_1 + p_2, p_3, \ldots\ldots, p_n) \leq S_n(p_1, p_2, p_3, \ldots\ldots, p_n)$$

If the two outcomes are combined, the entropy is reduced. This property is desirable because if the two outcomes are combined the uncertainty should not increase.

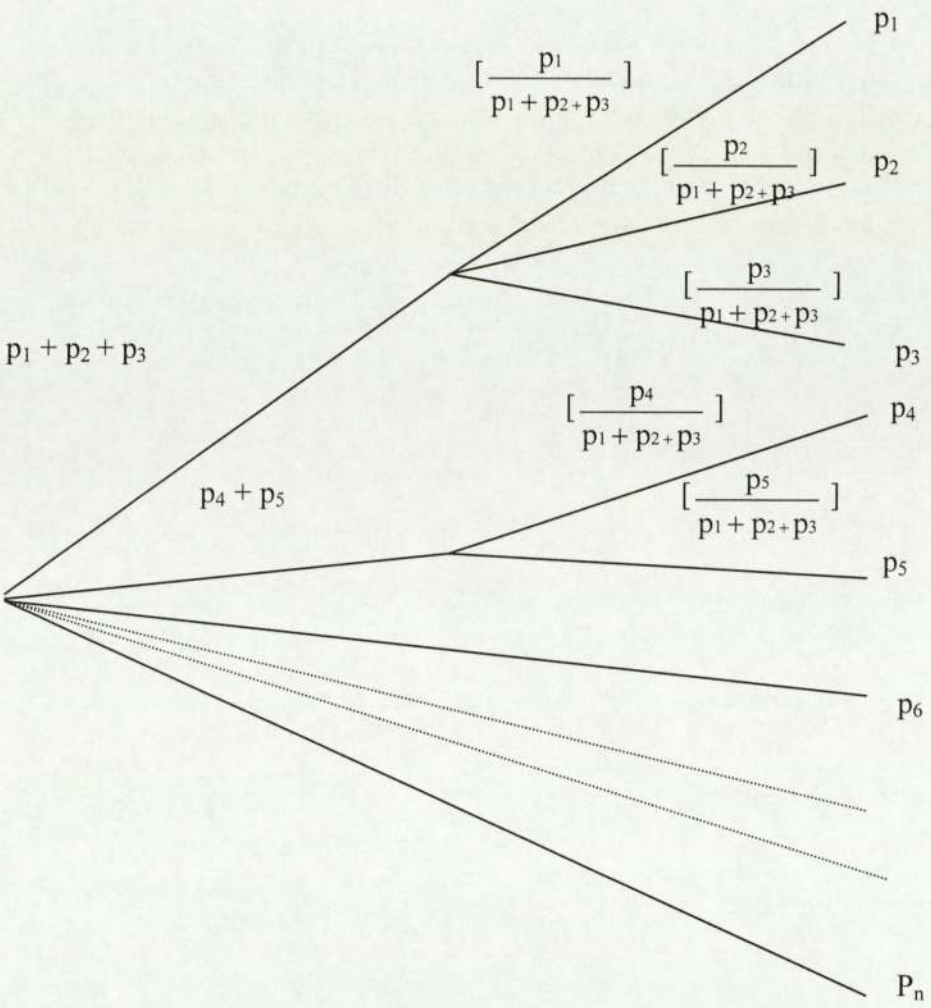


Figure: The Branching Principle

We proceed in two stages to get $S_n (p_1, p_2, p_3,\ldots\ldots,p_n)$. in the first stage we get S_{n-1} $(p_1 + p_2, p_3,\ldots\ldots,p_n)$. In the second stage, we have a probability $p_1 + p_2$ divided into two parts, p_1 and p_2 , giving rise to the probability distribution $[p_1 / (p_1 + p_2), p_1 / (p_1 + p_2)]$ with an entropy $S_2 [p_1 / (p_1 + p_2), p_1 / (p_1 + p_2)]$ and with a weight $p_1 + p_2$. this also explains the earlier formula,

$$S_n (p_1, p_2, p_3,\ldots\ldots,p_n) = S_{n-1} (p_1 + p_2, p_3,\ldots\ldots,p_n) + (p_1 + p_2) S_2 \left[\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right]$$

This recursive property is also called the branching principle, as shown above.

The main properties of Shannon's measure are:

- S should depend on all probabilities $p_1, p_2,\ldots, p_n$. i.e., S should be a function of p.

- $S_n(p_1, p_2, \ldots, p_n)$ should be a continuous function of $p_1, p_2, \ldots, p_n$. i.e small changes in $p_1, p_2, \ldots, p_n$ should lead to small changes in S.
- $S_n(p_1, p_2, \ldots, p_n)$ should be permutationally symmetric i.e., S should not change when $p_1, p_2, \ldots, p_n$ are permuted among themselves since uncertainty should not change when outcomes are labelled differently.
- Shannon's measure is a concave function of $p_1, p_2, \ldots, p_n$.
- Continuity.

  When $p_i = 0$, $p_i \ln p_i$ is not defined, but

  $$\lim_{pi \to 0} p_i \ln p_i = \lim_{pi \to 0} \frac{\ln pi}{\frac{1}{pi}}$$

  $$= \lim_{pi \to 0} -\frac{\frac{1}{pi}}{\frac{1}{(pi)^2}}$$

  $$= \lim_{pi \to 0} (-p_i)$$

  $$= 0.$$

- Entropy does not change with the inclusion of an impossible event.

  $$S_{n+1}(p_1, p_2, \ldots, p_n, 0) = -\sum_{i=1}^{n} p_i \ln p_i - 0 \ln 0$$

  $$= S_n(p_1, p_2, \ldots, p_n)$$

- There are n degenerate distributions.

  $\Delta_1 = (1, 0, \ldots, 0)$

  $\Delta_2 = (0, 1, \ldots, 0)$

  $\vdots$

  $\Delta_n = (0, 0, \ldots, 1)$

  For all f these $S_n(p) = 0$.

- Concavity

  Let $\phi(p) = -p \ln p$

  $\phi'(p) = -\ln p - 1$

$$\phi\,''(p) = -\frac{1}{p}$$

Thus ϕ (p) is a concave function of p. And because the sum of concave functions is also concave S_n (p) is a concave function of (p_1, p_2,…., p_n).

Optimization

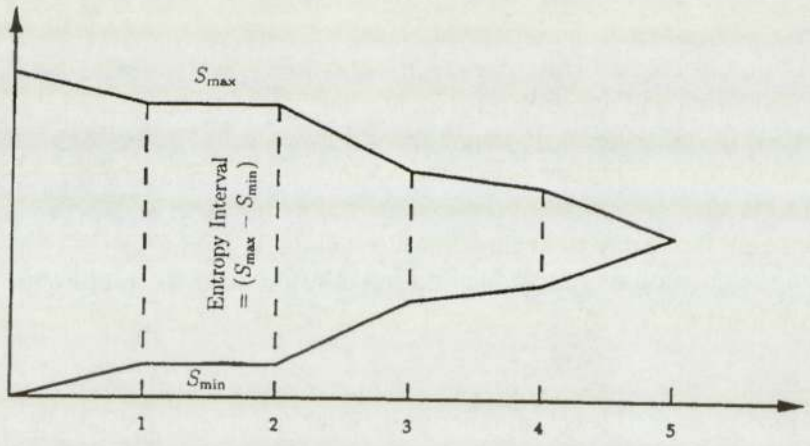
Optimization includes maximization and minimization as well as simultaneous maximization of one function and minimization of another function. Most applications are frequently concerned with constrained optimization. In the sense of a business, the situation is that the company normally wants to maximise its production in the face of constraints such as capital investment available, physical resources, manpower etc. Despite all the mathematical techniques available in entropy optimization, Lagrange's method of constrained optimization will be the preferred method for entropy optimization.

The reason we optimize entropy is that it provides a way of picking out probability distributions of desired level of uncertainty bounded by the available information. This means that the uncertainty is being reduced by obtaining more and more information. The reason for entropy optimization is best described in the die example.

Example: Rolling a die.

1. The number of unknowns is the number of faces – *n*.
    - Being told that the number of faces is 6, reduces the uncertainty.
2. We now have many different distributions, $(p_1, p_2, \ldots\ldots, p_n)$.
    - We are now only limited to probability distributions $(p_1, p_2, \ldots\ldots, p_6)$, where $\sum_{i=1}^{n} p_i = 1$.
3. If in addition we are told that the mean number of points on the die is 4.5
    - $p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 = 4.5$
    - Our choice of distributions is now restricted to those satisfying $\sum p_i = 1$ and $p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 = 4.5$, and our uncertainty is further reduced.
4. If in addition we are also told that:
    - * $1^2p_1 + 2^2p_2 + 3^2p_3 + 4^2p_4 + 5^2p_5 + 6^2p_6 = 15$.
    - Our choice of probability distributions is now further reduced to those satisfying this constraint, our uncertainty is also further reduced.

Variation of S_{max} and S_{min} with constraints.



We can carry on getting more and more information which will decrease our uncertainty. We can go on to get a unique set of values of p_1 through till p_6, the uncertainty for these values is completely removed. At the earlier stages, we have an infinity of probability distributions available satisfying the constraints. At each stage, out of all the distributions consistent with the constraints till then, one will have a maximum uncertainty, S_{max}, and the other extreme will have a minimum uncertainty, S_{min}. Since at each stage the set of probability distributions is a subset of the probability distributions available at the earlier stage, S_{max} decreases and S_{min} increases. At the sixth stage on the graph, there is only one probability distribution that gives maximum and minimum uncertainty.

It is worth noting that the information given by the constraints does not reduce uncertainty about the outcome, but these constraints are only reducing the uncertainty about the values of $p_1, p_2, \ldots\ldots, p_6$. The only way to remove the uncertainty about the outcome is to carry out the experiment.

Principles of entropy maximisation

The principle of entropy maximising is a method for analyzing the available information in order to determine a unique probability distribution. We should use all the information given to us and scrupulously avoid using any information not given to us. Out of all the probability distributions consistent with a given set of constraints, we will choose the one that has maximum uncertainty.The two principles that I will be following are the following:

Out of all the probability distributions satisfying given constraints, choose the distribution that is closest to the uniform distribution.

The maximisation of uncertainty can be considered by this principle, because if we first find the most uncertain distribution subject only to the constraint,

$p_1, p_2, \ldots, p_n$ where $p_1 \geq 0, p_2 \geq 0, \ldots\ldots\ldots, p_n \geq 0, \quad \sum_{i=1}^{n} p_i = 1.$

The only other information we have is that the sum of the probabilities is unity and that is all the information we have about these probabilities. This means that there is no reason to choose different values of $p_1, p_2, \ldots, p_n$, which means;

$$p_1 = p_2 = \ldots = p_n = \frac{1}{n}.$$

Therefore we choose the probability distribution

$$U = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

This is called the uniform distribution. This distribution has maximum uncertainty out of all the probability distributions for *n* outcomes. Any other probability distribution can only be used if we have any additional information, we would not have in this case. This principle is also called Laplace's principle of insufficient reason, because the only information we have is $p_i \geq 0$, and the sum of the probabilities is unity.

There are many situations where this principle would not be sufficient because of additional information available. If we have information suggesting different starting values for probabilities, we should use this information. The additional information is usually in the form of a priori probability distribution. That means the chosen probability distribution would satisfy the given constraints and be closest to the known priori distribution.

Out of all the probability distributions satisfying the given constraints, choose the distribution that is closest to the given a priori distribution.

This is also known as the principle of minimum directed divergence or the principle of minimum discrimination information or more appropriately the minimum cross entropy principle.

Combining the two principles, we have a general principle:

Out of all the probability distributions satisfying given moment constraints together, choose the distribution that is closest to the given a priori probability distribution, in the case of no a priori distribution, choose the distribution that is closest to the uniform distribution.

In the future

In this project so far I have looked at certain measures of entropy suggested by mathematicians. To increase my knowledge on the subject I will look at more measures of entropy suggested by other mathematicians and the context in which they are used. I will also look at the applications in which these measures of entropy, with the same principles for maximisation of entropy are used. This will provide useful when I focus on a particular problem myself.

After having looked at certain measures of entropy and the applications in which they are used, I will go onto focus on a particular problem. I will discuss the difference in using different measures of entropy on the chosen application and apply the chosen methods on the particular application.