Taxonomic analysis of marine phytoplankton

Bill Whiten[1] Barry McDonald[2] Chris Drovandi[3]

(Received 14 August 2010; revised 4 September 2011)

Abstract

Samples of sea water contain phytoplankton taxa in varying amounts, and marine scientists are interested in the relative abundance of each taxa. Their relative biomass can be ascertained indirectly by measuring the quantity of various pigments using high performance liquid chromatography. However, the conversion from pigment to taxa is mathematically non trivial as it is a positive matrix factorisation problem where both matrices are unknown beyond the level of initial estimates. The prior information on the pigment to taxa conversion matrix is used to give the problem a unique solution. An iteration of two non-negative least squares algorithms gives satisfactory results. Some sample analysis of data indicates prospects for this type of analysis. An alternative more computationally intensive approach using Bayesian methods is discussed.

http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/3391
 Published September 11, 2011. ISSN 1446-8735.

Contents

1 Introduction

Phytoplankton are small organisms, typically 0.001 to 0.5 mm, that live in the ocean and form the base of the oceanic food chain. They come in many different forms and, in spite of being tiny, are by mass the major life form in the oceans. They use sunlight as an energy source to photosynthesise organic compounds from carbon dioxide, producing oxygen as a by-product. Because

of their importance in the oceanic ecosystem, the Australian Antarctic Division collects samples across the Southern Ocean to determine the phytoplankton types present and study their behaviour.

Using microscopic methods to identify the phytoplankton taxa is both slow and tedious. An alternative is to look for pigments that identify the different taxa of phytoplankton. The pigments can be identified by high performance liquid chromatography (HPLC). It is usually known which pigments are present in each taxa but the proportions vary and some pigments are shared by several taxa. The Antarctic Division has a program CHEMTAX [5] that estimates the relative abundance of the taxa and the proportion of each pigment in the various taxa. CHEMTAX uses trial steps and steepest descent to locate a solution to this problem, but the Antarctic Division is interested in possibilities for further development.

The MISG problem was to convert the data matrix of samples by pigments S^0, to the product of two matrices: samples by taxa C providing the proportions of the taxa in each sample, and a transform matrix of taxa by pigments F giving the pigment proportions in each taxa.

There is prior information available for the values in the taxa by pigments matrix. In particular, the position of zero entries in this matrix are generally known. Estimated means and standard deviations are supplied for the non-zero elements. The Australian Antarctic Division supplied a sample data set for use at the MISG.

Sections 2 and 3 define the conversion problem as an optimisation problem. The following two Sections 4–5 cover some possible general methods of solution to the optimisation problem, and then cover the specific method used in this article. Section 6 describes some data, and then Section 7 discusses properties of the solution. It is shown that the algorithm converges reliably, and the effect of a scale factor (or standard deviation) for the taxa by pigment matrix is examined. It is shown that bootstrap methods can be used to estimate the accuracy of the calculated values, while the differences between the observed and calculated values provides information on the quality of fit. Section 7.4

examines residuals and the type of information that can be obtained from them.

Section 8 describes an alternative approach to the analysis of the data using Bayesian techniques. These methods have the potential to provide more extensive statistical information about parameter estimates and taxa proportions.

2 Conversion from taxa to pigments

Although the problem is the conversion from pigments to taxa, the basic equation is a conversion from taxa to pigments. If the proportion of each taxa present in a particular sample and the amount of pigment in each taxa are known, then the amount of each pigment is calculated using the following vector matrix product and the transformation matrix F:

$$\mathbf{s} = \mathbf{c}\mathsf{F}\,, \tag{1}$$

where

$\mathbf{s}$ is the row vector of amounts (mass/volume) of each pigment,

$\mathbf{c}$ is the row vector of amounts (mass/volume) of each taxa,

F is a matrix with each row giving the amounts of each pigment in a taxa (mass of pigment / mass of taxon).

In the case where there are multiple samples the row vectors $\mathbf{s}$ and $\mathbf{c}$ become matrices S and C with each row corresponding to a sample, so that equation (1) becomes

$$\mathsf{S} = \mathsf{C}\mathsf{F}\,. \tag{2}$$

All three matrices are restricted to contain only positive or zero values.

As all of the taxa contain chlorophyll-A, it is often convenient to rescale the rows of masses to be relative to the amount of chlorophyll-A present.

For convenience it is assumed that the last column of S and F contain the chlorophyll-A values. An example of such an F matrix is given in Table 1. The normalisation is done by multiplying by a diagonal matrix D containing the inverse of the chlorophyll-A column in S. A second normalisation of the matrix F uses the diagonal matrix E containing the inverses of the chorophyll-A column of F so that equation (2) can be written as

$$(\mathsf{DS}) = (\mathsf{DCE}^{-1})(\mathsf{EF}). \tag{3}$$

Each row of the matrix DCE^{-1} sums to one (this follows from the unit columns of DS and EF) and gives the proportions of each taxa relative to their production of chlorophyll-A.

3 Conversion from pigments to taxa

The values of particular interest are the proportions of the various taxa present, given in the rows of matrix C. These need to be obtained from measured values of the pigments present, given in the rows of matrix S^0. This is an inverse of the calculation given in the previous section, and can only be solved when there are at least as many pigments as taxa. In addition to the measured values S^0, there is some knowledge of the transformation matrix F. In particular, the locations of the zero values are generally known and approximate values are available for the non-zero values. However, it is known that the non-zero elements of F are not constant, but vary with the conditions, such as the amount of sunlight and nutrients available.

The essential part of this problem is finding the matrices C and F such that their product is close to the measured values of the pigments S^0. However, this is not sufficient to make the values of C and F unique, as given one solution, other solutions can be generated as CZ^{-1} and ZF with any matrix Z that does not generate negative elements in these products.

To get a unique solution the values of the F matrix are required to be close to an initial estimate F^0. The initial F^0 together with the standard deviations

TABLE 1: Estimated F matrix for samples collected at 0-15m depth. Compare to Table 1 of Wright et al. [10]: the names of pigments and taxa are the same as in Table 1 of Wright et al.

Taxa	Chl c_3	Chl c_1	Peri	Fuco	Neo	Pras
Prasinophytes	0	0	0	0	0.079	0.096
Chlorophytes	0	0	0	0	0.074	0
Cryptophytes	0	0	0	0	0	0
Diatoms A	0	0.15	0	0.90	0	0
Diatoms B	0.036	0	0	0.86	0	0
Dinoflagellates A	0	0	0.84	0	0	0
Haptophytes-H	0.20	0	0	0.08	0	0
Haptophytes-L	0.12	0	0	0.01	0	0

Taxa	Violax	19'-Hex	Allox	Lutein	Chl b	Chl a
Prasinophytes	0.049	0	0	0.006	0.60	1
Chlorophytes	0.037	0	0	0.21	0.16	1
Cryptophytes	0	0	0.22	0	0	1
Diatoms A	0	0	0	0	0	1
Diatoms B	0	0	0	0	0	1
Dinoflagellates A	0	0	0	0	0	1
Haptophytes-H	0	0.23	0	0	0	1
Haptophytes-L	0	1.23	0	0	0	1

of its elements (given in the matrix G), are the prior knowledge about the transformation matrix F. Section 6 discusses the availability of the data needed for this calculation.

Similarly, the measure of closeness to the measured pigment data S^0 is defined using a matrix T of standard deviations values. Then C and F are found from the minimisation, with respect to the elements of C and the non-zero

elements of F, of the sum of squares

$$\sum_{i,j}\left(\frac{\sum_k c_{i,k}f_{k,j} - s^0_{i,j}}{t_{i,j}}\right)^2 + \sum_{k,j\in\Phi}\left(\frac{f_{k,j} - f^0_{k,j}}{g_{k,j}}\right)^2 \tag{4}$$

where i indexes samples, j indexes pigments, k indexes taxa, and

$c_{i,k}$ is the i,k element of the matrix C,

$s^0_{i,j}$ is the i,j element of the matrix S^0, the measured values of pigment per volume,

$t_{i,j}$ is the standard deviation of the i,j element of the matrix S^0,

$f_{k,j}$ is the k,j element of the transform matrix F,

$f^0_{k,j}$ is the k,j element of the initial estimate matrix F^0,

$g_{k,j}$ is the standard deviation of the k,j element matrix F^0,

Φ is the set of k,j values corresponding to non-zero elements in F.

The measured values S^0 of pigment concentrations are used to define the error terms in the minimisation as then the errors are more likely to be closer to a Gaussian distribution than the scaled values of equation (3), and the standard deviations can be estimated directly from the data values. In particular, this makes the samples with low pigment values less likely to cause difficulties by biasing the results.

We wish to find values of the matrices C and F that give the product CF close to the measured values of the pigments S^0. The measure of closeness is described by a matrix T of standard deviations values. The estimates of the transformation matrix F^0 together with the corresponding standard deviations G provide the initial knowledge about F.

This formulation of the problem differs from that used in CHEMTAX. In particular, the prior information about the F values is given as a mean and standard deviation, and standard deviations have been introduced for the

errors in the measured data S^0. This means that the optimality criterion is different and the results will not be identical with those from the CHEMTAX program.

4 Methods of solution

To examine the possible methods of solution it is easiest to first consider methods of solution of the simpler problem of minimising the sum of the squared elements of the matrix

$$\mathsf{S}^0 - \mathsf{CF} \tag{5}$$

with respect to the values in C and the non-zero values of F. The elements of C and F all need to be non-negative. So this is a positive matrix factorisation problem.

The first possible solution method is to use a general purpose optimisation algorithm. This loses any advantage that may be available from the special structure of the problem. As the problem has many variables this approach may not be the most efficient solution method.

If F is known, minimising (5) is a linear regression for C. Without the condition that the elements are positive C can be found from $\mathsf{S}^0\mathsf{F}^\mathsf{T}(\mathsf{F}\mathsf{F}^\mathsf{T})^{-1}$. The restriction that C does not contain negative values makes this a non-negative regression problem for each row of C. There is a standard solution procedure for non-negative regression given by Lawson and Hanson [3]. This reduces the problem to finding the non-zero elements of F which can be done using a general purpose optimiser or a constrained nonlinear least squares program. Typically the number of non-zero elements in F is much smaller than in C so this makes a large reduction in the number of dimensions in the problem.

It is also possible to solve for F, if C is known, using the non-negative least squares algorithm, since this is a linear least squares problem for each column

of F. Thus, given an initial estimate for F, a solution can be found by using non-negative least squares to alternatively solve for C and then for F. As each step minimises the sum of squares, this iteration is expected to converge.

Lee and Seung [4] proposed an iterative algorithm that progressively improves approximate solutions for C and F, in the least squares problem (5). Their formula to update C is

$$c^{\dagger}_{i,k} = c_{i,k}[SF^{T}]_{i,k}/[CFF^{T}]_{i,k} \tag{6}$$

where the † indicates the updated value. This iteration, given positive initial values, will maintain a positive solution. We note that $SF^{T} = CFF^{T}$ at the solution to the least squares problem for C, and this iteration solves the non-negative linear least squares problem for C. Lee and Seung [4] prove that each application of this formula reduces the sum of squares and the stationary point is the solution of the least squares problem for C given F. Lee and Seung [4] use a similar formula for improving the estimate for F:

$$f^{\dagger}_{k,j} = f_{k,j}[C^{T}S]_{k,j}/[C^{T}CF]_{k,j}. \tag{7}$$

As well as maintaining positive values, any zero values in F are maintained as zero.

The two steps of the Lee and Seung [4] iteration use only simple matrix operations, that are available as basic linear algebra algorithms [9], optimised for minimum run times on various processors, and thus the algorithm steps can be implemented very efficiently. In trials of the above algorithms with the phytoplankton data the more efficient calculation gives a large advantage over more accurate steps using the non-negative least squares algorithm. However, both iterations converge slowly and require a large number of iterations on the phytoplankton data.

The result from the positive matrix factorisation algorithms is not unique. The initial values determine which result is obtained. This problem is addressed in the next section. Trials indicate there is no benefit in normalising one or both of the factors during the iterations, as this can be done after the iteration has converged, saving some calculation steps during the iterations.

5 Extension to the phytoplankton problem

The phytoplankton minimisation problem (4) is a bilinear sum of squares and iterations similar to the last section can be derived. By introducing augmented matrices $[\mathsf{C}^\mathsf{T}\mathsf{I}]$, $[\mathsf{S}^{0\mathsf{T}}\mathsf{F}^{0\mathsf{T}}]$, and $[\mathsf{T}^\mathsf{T}\mathsf{G}^\mathsf{T}]$, the expression (4) can be converted to the form (5). However, it is simpler to work directly with expression (4).

As before, non-negative least squares can be applied to each row of C and each column of F, omitting the elements known to be zero. Extending the Lee and Seung [4] iteration involves considering the derivatives of (4) with respect to (wrt) C and F:

$$\text{wrt } c_{p,r}, \quad 2\sum_j \frac{(\sum_k c_{p,k}f_{k,j})f_{r,j}}{t_{p,j}^2} - 2\sum_j \frac{s_{p,j}^0 f_{r,j}}{t_{p,j}^2}, \tag{8}$$

$$\text{wrt } f_{r,q}, \quad 2\sum_i \frac{c_{i,r}(\sum_k c_{i,k}f_{k,q})}{t_{i,q}^2} + 2\frac{f_{r,q}}{g_{i,q}^2} - 2\sum_i \frac{c_{i,r}s_{i,q}^0}{t_{i,q}^2} - 2\frac{f_{r,q}^0}{g_{i,q}^2}. \tag{9}$$

The zeroes of these derivatives define the solution to the least squares problem (4). Further, when the derivative is positive, the variable differentiated with respect to ($c_{p,r}$ or $f_{r,q}$) is too high, and when the derivative is negative the variable is too low. The ratio of the positive and negative terms in the derivative can be used to improve an estimate of the variable:

$$c_{p,r}^\dagger = c_{p,r}\frac{\sum_j s_{p,j}^0 f_{r,j}/t_{p,j}^2}{\sum_j(\sum_k c_{p,k}f_{k,j})f_{r,j}/t_{p,j}^2}, \tag{10}$$

$$f_{r,q}^\dagger = f_{r,q}\frac{\sum_i c_{i,r}s_{i,q}^0/t_{i,q}^2 + f_{r,q}^0/g_{i,q}^2}{\sum_i c_{i,r}(\sum_k c_{i,k}f_{k,q})/t_{i,q}^2 + f_{r,q}/g_{i,q}^2}. \tag{11}$$

Unlike the iterations given in Section 4, the prior information for F, namely F^0 and its standard deviation G, ensures that this iteration will in general converge to a unique result. If the initial values are positive, then the calculated values remain positive; and if an element in the initial value of F is

zero, then that element will remain zero. Values in F can be held at a fixed value by giving them a very small standard deviation. If it is desirable to give more weight to the calculation of F or some elements of F, then this can be done by reducing the standard deviations $g_{i,q}$.

Similar to the algorithm of Lee and Seung [4], the formulas (10) and (11) can be written in terms of matrix operations that have been optimised for minimum run times, and hence the steps in this iteration also run rapidly. The sum of squares (equation (4)) is reduced by each iteration, but convergence requires several thousand iterations on the phytoplankton data. An occasional application of the much slower non-negative least squares algorithm to update the value of C has been found to reduce the time needed to reach convergence; however, it was found that the non-negative least squares approach alone takes a significantly longer time to converge on the phytoplankton data using Matlab code.

6 The phytoplankton data

Sample phytoplankton data was made available to the MISG for evaluation of analysis techniques. Details of the data collection and estimated abundances using CHEMTAX were given by Wright et al. [10]. The data consists of 1114 samples taken at various depths over an area of the southern ocean. The samples were analysed by HPLC for twelve pigments, and an initial estimate of the conversion matrix from eight taxa to the twelve pigments was provided.

The previous analysis using the CHEMTAX program was performed using the normalised (with respect to chlorophyll) pigment values without the use of standard deviations. It is thought that more realistic results are obtained if standard deviations are used to indicate the accuracy of the data pigment matrix S^0, which contains values that vary by more than three orders of magnitude. Unfortunately, repeat values are not available and so standard deviations of the HPLC data had to be estimated. The reported accuracy of

the HPLC instrument is about one percent of the reading with a detection limit of about 0.0003, hence a standard deviation of $\mathsf{T} = 0.01\mathsf{S}^0 + 0.0003$ is used in the following demonstrations of the analysis method. These values need to be compared with the actual residuals that occur after fitting as in Section 7.4.

The reconstruction of taxa content depends only on the sample that passed through the HPLC unit, and is not affected by sampling variations that might have occurred before the HPLC process. It would be useful to have directly measured standard deviations from the HPLC unit as these often include extra variation not included in estimates. In particular, the error involved in the matrix factorisation in this article is only one possible source of variation, and is likely to be small compared to the sampling errors involved in measuring seawater from the Southern Ocean. Repeatability measures from earlier stages in the data collection would help in the interpretation of the data.

The values F^0 and the associated standard deviations G are estimates that are provided as prior knowledge by the user. Here the standard deviations act as inverse weighting factors that determine how large a change in the values is reasonable. The actual amounts of pigments in each taxa depend on the amounts of nutrients and sun light that were available to the sample. Most of the taxa can be grown in the laboratory and the pigments measured; however, it is difficult to determine the normal range of pigment concentrations in the laboratory. Nor can it be assumed with certainty that all the phytoplankton sampled in the open ocean that survived the processes of freezing, transportation and laboratory growth, produced the same pigment ratios that are seen in the ocean. As the matrix factorisation does depend on the prior values, they need to be specified carefully, and the associated errors examined after the model has been fitted, as for example in Section 7.4.

In the following the standard deviations G are set proportional to the initial estimates F^0. Section 7.2 looks at how the proportionality factor for these standard deviations affects the fit to the data.

7 Properties of the solution

7.1 Convergence

The algorithm for calculating C and F is iterative starting from some initial values. To be useful it needs to converge reliably to the same answer regardless of the initial values used. To test this, elements of the initial C are randomly generated between 0.1 and 1.1, and used as the initial values. A large number of iterations are used to get a high accuracy. Figure 1 plots the ratio of standard deviation to the mean, against the mean value for each of the values in the matrices C and F. The values obtained are repeatable to about eight decimal places which is as good as can be expected for finding the minimum of a sum of squares using 16 digit floating point arithmetic.

7.2 Scale of standard deviations for F^0

The standard deviations $g_{i,j}$ in equation (4) determine how close the estimated values $f_{i,j}$ are to the user supplied estimates $f^0_{i,j}$. The effect of this choice is examined by fitting the C and F values with different scaling values multiplying the values of $g_{i,j}$. For this the values of $g_{i,j}$ are initially set equal to a factor times $f^0_{i,j}$, except for the last column of F which consists of ones and is given very small standard deviations. Figure 2 shows how the error in the S values changes with the factor used for standard deviations of F (that is, $g_{i,j}$). While the factor is greater than 0.01 there is little effect on the root mean square errors for S. That is, the initial estimates F^0 are not being given so much weight that they prevent the estimates of C (and hence of S) from finding their optimal values to minimise (8). However, small $g_{i,j}$ force F to stay too close to F^0 and this constrains C such that the errors in (8) become large. To provide an amount of stabilisation that ensures that the result does not excessively depend on randomness in the data it is normal to choose a factor near the upward bend in Figure 2. From this plot the factor is chosen to

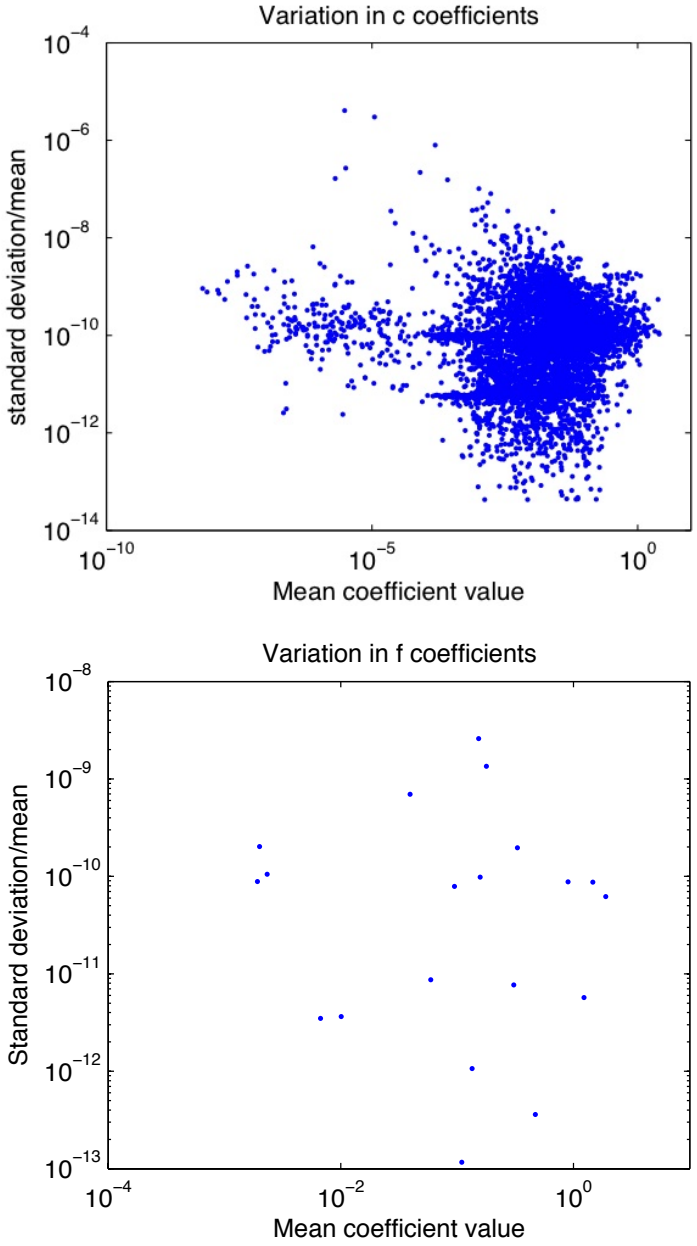

FIGURE 1: Convergence of values of C and F given random starts.

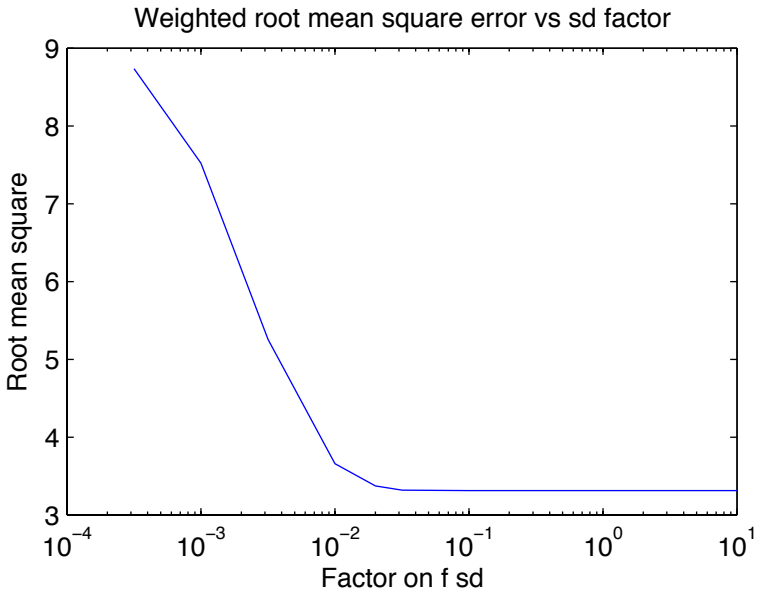

FIGURE 2: Effect of standard deviation size for F on the size of errors in S.

be 0.01 giving standard deviations of $0.01\mathsf{f}^0_{i,j}$.

It is of interest to compare the estimates produced by our method to those produced by CHEMTAX. Wright et al. [10, Table 1] give both initial estimates F^0 of the pigment/ChlA ratios, and final estimates F for samples collected at 0–15 m depth. Our Table 1 shows our estimates for the same data subset. For most taxa our method gives very similar estimates to CHEMTAX, except for Diatoms A and Haptophytes-H. These particular ratios were picked out for discussion by Wright et al. [10] because the CHEMTAX results were rather different to the initial estimates in F^0. For these cases our figures are mostly closer to the initial estimates, except for the 19'-Hex pigment for Haptophytes-H where our estimate is about half that of CHEMTAX. We conclude that the method in this article can be considered to be producing reasonably similar results to those of CHEMTAX.

7.3 Accuracy of results

To determine the accuracy of the calculated values a bootstrap procedure [1], where the input data is perturbed to determine the effect on the output values, is used. The first procedure examined is a parametric bootstrap. The values in S^0 are replaced by log Gaussian random variables with mean equal to the value in S^0 and the standard deviation $0.01S^0 + 0.003$, which is the estimated accuracy of the HPLC instrument. This is repeated ten times and the results are shown in Figure 3.

Figure 4 shows the results from a nonparametric bootstrap. This is done by ten repeats of randomly selecting, with replacement, rows of S^0 to create a new S^0 to be used in the fitting. The nonparametric bootstrap (Figure 4) indicates a slightly lower accuracy for the F coefficients and a larger variation in accuracy for the smaller C coefficients compared with the parametric bootstrap (Figure 3). These differences probably reflect that the actual data have more sources of variation than the formula $0.01S^0 + 0.003$, that is based on only the errors in analytical measurement, provides. However, both bootstrap approaches indicate an accuracy of about two decimal places in the F matrix for the larger coefficients, and one to two decimal places for larger elements in the C matrix. The relative accuracy in the values drops as the coefficients become smaller.

The results from the bootstrap analysis can be divided up in different ways. Figure 5 shows the division into the separate taxa for the C elements. This figure subdivides the data plotted in Figure 3. It can be seen that some taxa occur in higher proportions than others, and some taxa are determined more accurately than others.

7.4 Residual distributions

The difference between the measured values S^0 and the predicted values CF provide residuals that can be examined to determine how accurately the

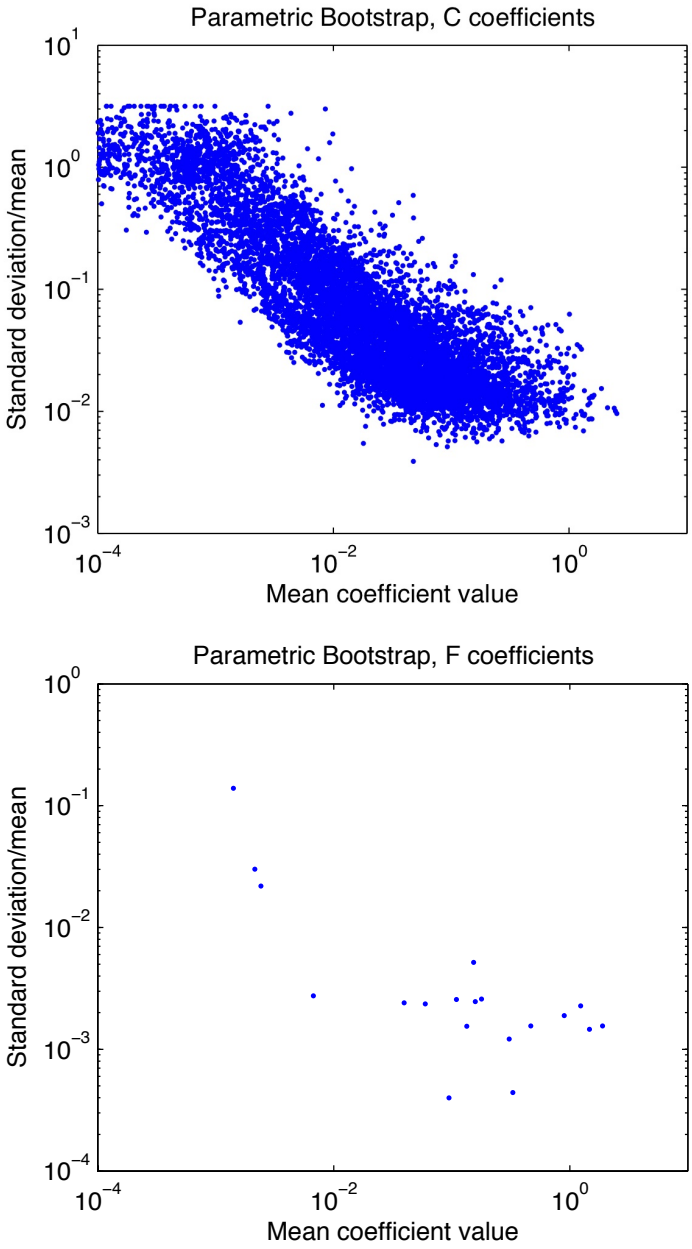

FIGURE 3: Parametric bootstrap results for C and F.

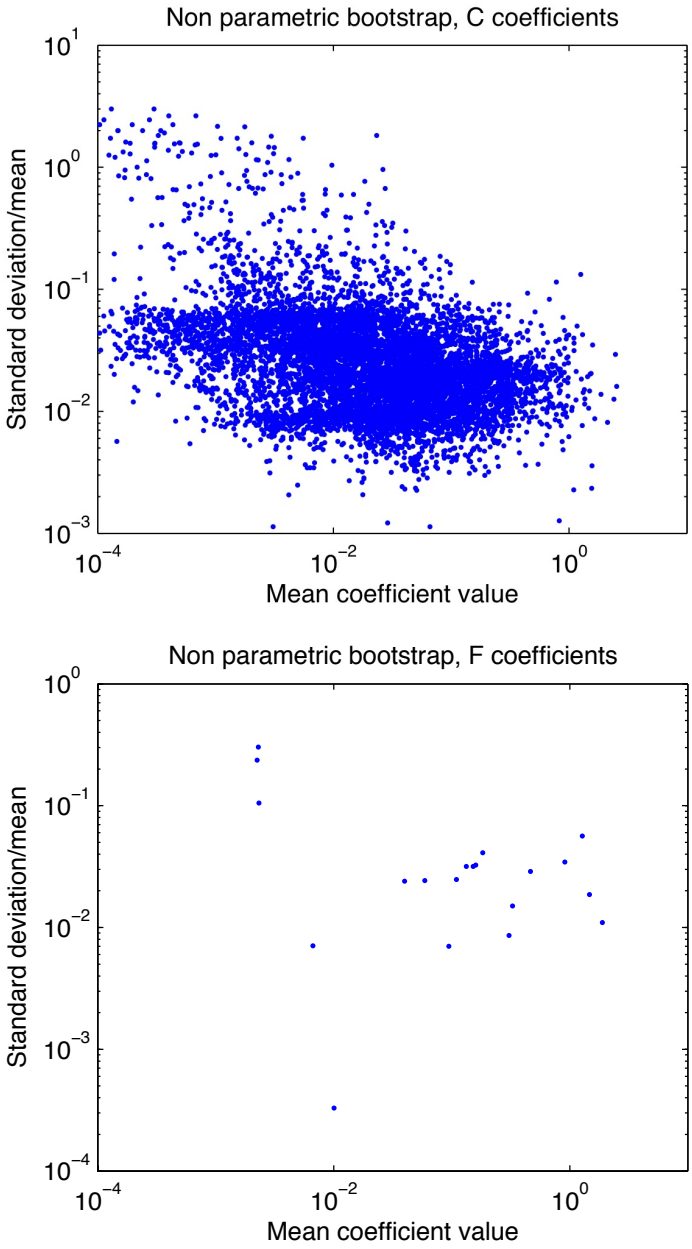

FIGURE 4: Nonparametric bootstrap results for C and F.

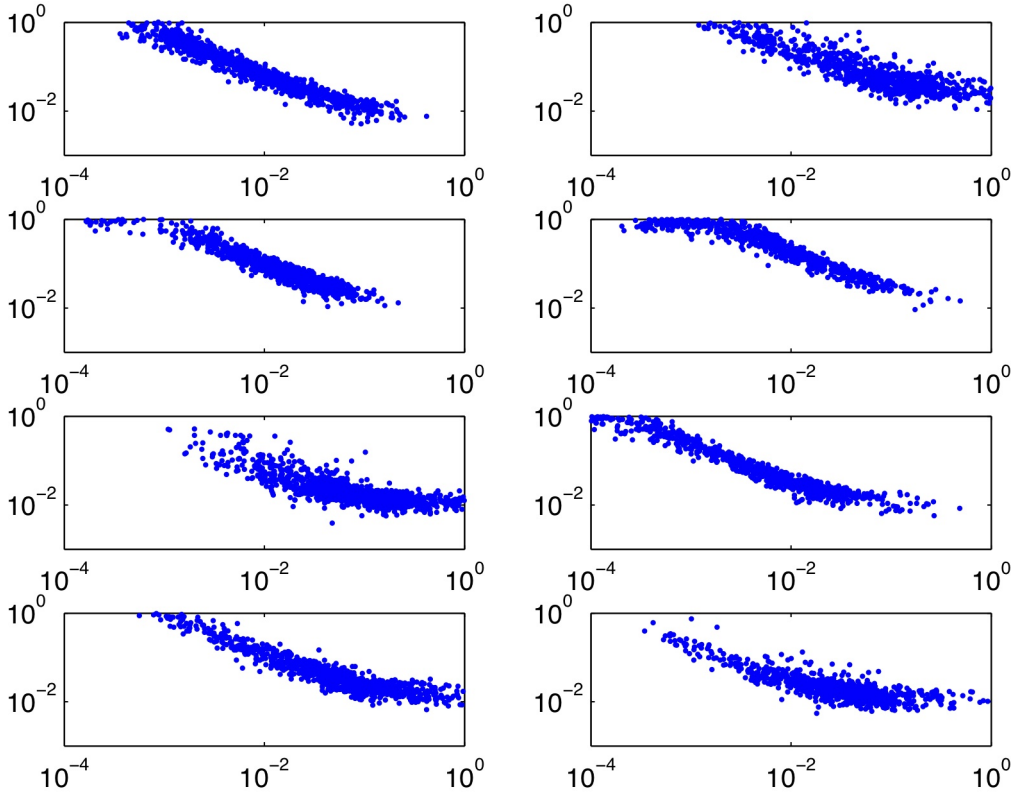

FIGURE 5: Parametric bootstrap results for C divided according to taxa. Here similar to the previous figures, the x-axis is the mean value and the y-axis the standard deviation/mean.

data has been reproduced. As an overall measure, the standard deviations are compared with the root mean square of the residuals for the different parts of the data. The residuals for particular taxa are examined by location and depth. Figure 6 shows the residuals for one particular pigment, ChlC1, which is uniquely associated with the phytoplankton taxon Diatom A. The changing convexity of the normal probability plot suggests the residuals have a mixture distribution of multiple components and there are occasional extreme outliers associated with phytoplankton blooms that occur near the Antarctic sea-ice boundary where the waters are relatively rich in nutrients especially iron. The analysis by Wright et al. [10] shows that the abundance and relative proportions of different taxa vary considerably with latitude, depth and longitude (the latter a weaker effect relating to topography and ocean currents). One might hope that these systematic effects are fully confined to the fitted values of $\mathsf{S}(=\mathsf{CF})$, but the fitted line plots of the $\log_{10}(|\text{residuals}|)$ indicate that these factors influence the residuals as well (P-value < 0.001 for each relationship). A comparison of the standard deviations of the residuals with the mean assumed standard deviation $\mathsf{T} = 0.01\mathsf{S}^0 + 0.0003$ indicates that T could be too large or too small for particular taxa by an order of magnitude. Since T is used solely for weighting purposes this just means that some taxa may have slightly more influence on the final fit than they 'should' (in the light of their variability) but the same is also true of the CHEMTAX algorithm. A full analysis of the residuals is beyond the scope of this article, but these results suggest that there is room for further improvement in the model.

The preceding sections focussed on the mathematical task of converting from S^0 and F^0 to estimates C and F. The elements of S^0 and F^0 are assumed to be measured accurately (to small standard deviation) and each sample (row of S^0) is treated as an independent observation. The approach is applicable to settings other than phytoplankton. It does not make any (statistical) distributional assumptions about the sampling errors, nor does it pay attention to skewness in the errors. We now turn to a statistical approach that addresses these latter issues.

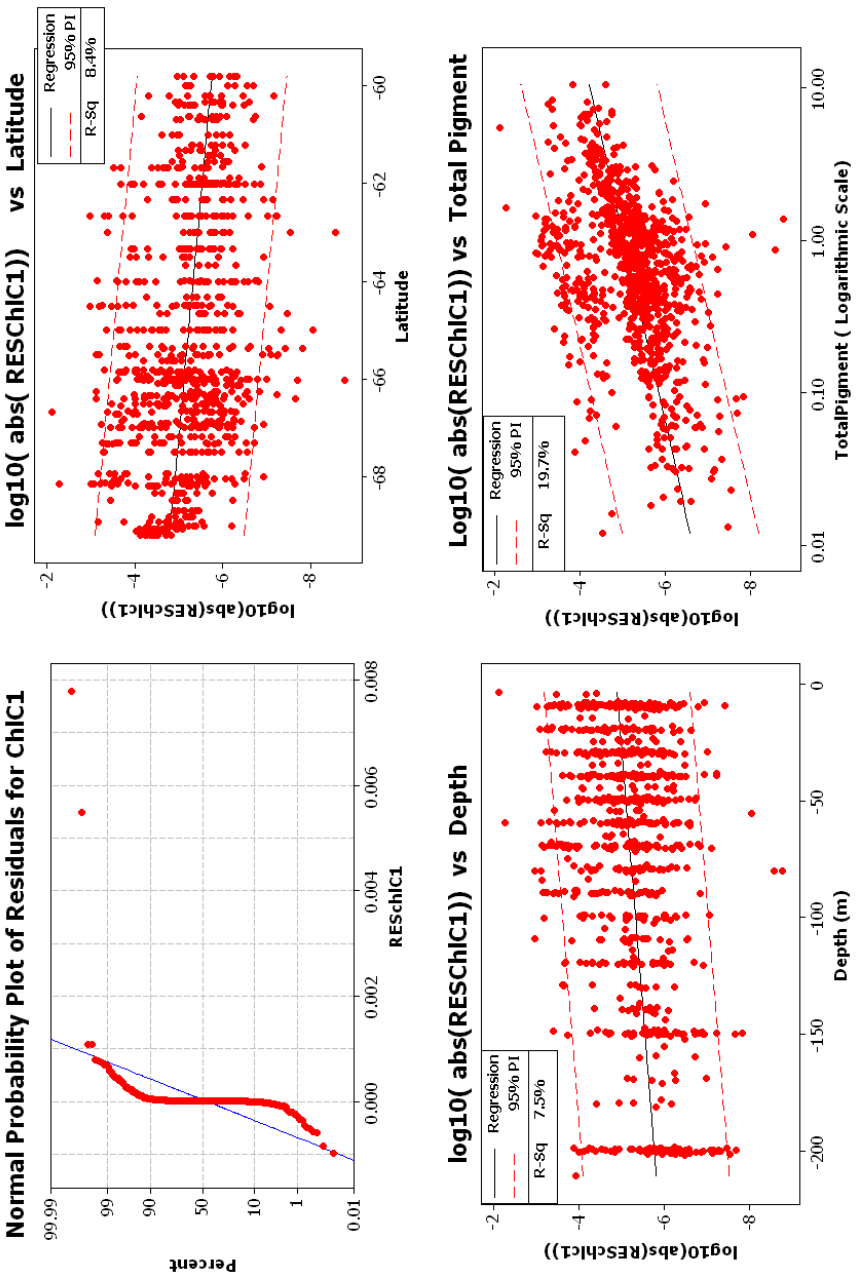

FIGURE 6: Residuals for pigment Chlc1.

8 A Bayesian approach

A Bayesian approach to this problem was considered by Meersche et al. [6]. They recommend their approach for relatively small numbers of samples compared to those considered here. A statistical solution may be appropriate in this application since the matrix S^0 is measured with sampling error and the parameters consist of the unknown elements in C and F. In theory, a statistical approach would allow us to quantify the accuracy of C and F, where these are regarded as estimates of unobserved mean taxa and pigment proportions for each location. Unfortunately, such direct measures of uncertainty are not available in the approaches detailed above. A Bayesian approach is particularly attractive since there is prior information on the parameters of the F matrix. Furthermore, there are various constraints on the parameters that can be easily handled in a Bayesian framework via the choice of appropriate prior probability distributions.

Bayesian analysis proceeds via the posterior distribution of C and F when S^0 is given, that is, $\Pr(\mathsf{C}, \mathsf{F} \mid S^0)$, which is proportional to the likelihood (of the model for the observed data) multiplied by the prior (contains information about the model parameters before data is collected). The posterior distribution thus contains the combined information about the parameters held in the prior and the observed data. For notational simplicity we assume that each element of the matrix F is unknown but it is straightforward to extend the results below to the case where some elements of F are fixed. The posterior is

$$\Pr(\mathsf{C}, \mathsf{F} \mid S^0) \propto \Pr(S^0 \mid \mathsf{C}, \mathsf{F}) \Pr(\mathsf{C}, \mathsf{F}), \tag{12}$$

where $\Pr(S^0 \mid \mathsf{C}, \mathsf{F})$ is the likelihood of S^0 given C and F, and $\Pr(\mathsf{C}, \mathsf{F})$ is the prior distribution of C and F. Meersche et al. [6] assume that C and F are independent a priori, producing $\Pr(\mathsf{C}, \mathsf{F}) = \Pr(\mathsf{C}) \Pr(\mathsf{F})$. Furthermore, the rows of C are independent, yielding $\Pr(\mathsf{C}) = \prod_{s=1}^{\mathsf{N}} \Pr(\mathbf{c_s})$, where $\mathbf{c_s}$ corresponds to the sth row of C and N is the number of samples. The only prior information on such rows is that the sum must equal one. This

constraint can be implemented using a Dirichlet distribution for each row. The 'no information' aspect can be handled by setting all the parmeters of the Dirichlet distribution equal to one. From this we have $\Pr(\mathbf{c_s}) \propto 1$ and the posterior of interest simplifies to

$$\Pr(C, F \mid S^0) \propto \Pr(S^0 \mid C, F) \Pr(F). \tag{13}$$

The non-zero unknown elements of the matrix F contain prior information in the form of a mean, $E(F_{t,p}) = e_{t,p}$, and variance, $\mathrm{Var}(F_{t,p}) = v_{t,p}$, where $F_{t,p}$ denotes the random variable of the ratio for the tth taxon and pth pigment, and $e_{t,p}$ and $v_{t,p}$ are given. Furthermore, there is a positive constraint on such parameters. Meersche et al. [6] choose to give each element a Gamma prior distribution, $F_{t,p} \sim \mathrm{Gamma}(\alpha_{t,p}, \beta_{t,p})$ (although other models such as the log Normal distribution seem just as plausible). The parameters $\alpha_{t,p}$ and $\beta_{t,p}$ can be computed by solving simultaneously $E(F_{t,p}) = \alpha_{t,p}/\beta_{t,p} = e_{t,p}$ and $\mathrm{Var}(F_{t,p}) = \alpha_{t,p}/\beta^2_{t,p} = v_{t,p}$ for $\alpha_{t,p}$ and $\beta_{t,p}$. This result arises by considering the expectation and variance of a Gamma random variable. Given the above and assuming independence amongst the elements of F, the following prior is obtained:

$$\Pr(F) = \prod_{t,p \in \Phi} \mathrm{Gamma}(F_{t,p}; \alpha_{t,p}, \beta_{t,p}). \tag{14}$$

where Φ is as in section 3. Now that the prior distribution is derived, the model for the data, $\Pr(S^0 \mid C, F)$, is required to complete the specification. It is assumed that the variance of each sample in S^0, $\mathrm{Var}(S^0_{s,p}) = v_{s,p}$ is known. Furthermore, given the parameters, Meersche et al. [6] assume that the mean is given by the sth row of C multiplied by the pth column of F, $E(S^0_{s,p}) = \sum_{t=1}^{T} C_{s,t} F_{t,p} = \mu_{s,p}$. To handle the positive constraint Meersche et al. [6] give $S^0_{s,p}$ a Gamma distribution with parameters $\alpha_{s,p}$ and $\beta_{s,p}$ such that

$$\alpha_{s,p}/\beta_{s,p} = \mu_{s,p} \quad \text{and} \quad \alpha_{s,p}/\beta^2_{s,p} = v_{s,p}\,. \tag{15}$$

Unfortunately, the posterior does not have a recognisable distribution. A popular approach to overcome this is to produce approximate samples from the

posterior distribution using Markov chain Monte Carlo (MCMC) techniques [7]. In this approach a Markov chain is constructed whose stationary distribution is given by the joint posterior of the parameters. Inference on the marginal distributions proceeds by ignoring the samples of other parameters since the algorithm automatically integrates them out via Monte Carlo integration.

Meersche et al. [6] implemented a particular type of MCMC algorithm known as the random walk Metropolis–Hastings sampler [2]. This approach proceeds as follows. Given a current value of the parameter θ (here $\theta = (\mathsf{C}, \mathsf{F})$), we propose a θ^* from $\Pr(\theta^* \mid \theta)$ and accept the proposal with probability

$$r = \min\left(1, \frac{\Pr(\mathbf{y} \mid \theta^*)\Pr(\theta^*)\Pr(\theta \mid \theta^*)}{\Pr(\mathbf{y} \mid \theta)\Pr(\theta)\Pr(\theta^* \mid \theta)}\right), \tag{16}$$

where $\Pr(\mathbf{y} \mid \theta)$ is the likelihood of data $\mathbf{y}$ and $\Pr(\theta)$ is the prior distribution.

Meersche et al. [6] use a multivariate proposal such that the elements of F are proposed from a Normal random walk and each row of C is proposed from a Dirichlet jump distribution. This approach relies on the proposal parameters being tuned to achieve a desired acceptance probability. Too large jumps will often fall in regions of negligible posterior probability and result in a very low acceptance rate. Too small jumps will mean the acceptance probability is very high but the samples will be highly correlated and it will take an excessively long time for the full posterior support to be visited in the correct proportions.

Meersche et al. [6] provided some recommendations for the tuning of parameters of the proposal distribution, but they did not mention an additional problem, the curse of dimensionality. Roberts and Rosenthal [8] show that the optimal acceptance probability decreases with the number of parameters. In high dimensional problems such as this one, such an approach would appear difficult in practice. To compensate for the required low acceptance rate, small jumps will be used to increase the acceptance probability, ensuring that unacceptably long runs of the chain may be required to be confident that the full posterior space is explored appropriately.

An alternative scheme that may overcome such issues involves a Metropolis–Hastings within a Gibbs sampler. In the Gibbs set up, we derive full conditionals for each parameter, which is the probability distribution of the parameter of interest given all the other parameters and the data. Thus we attempt to update only one parameter at a time while fixing the rest (but block updates are still possible of course). If we cannot sample from the full conditionals directly, then we use a Metropolis–Hastings sampler to obtain approximate samples from the full conditionals. Furthermore, for the parameters with a substantial amount of prior information, it may be more appropriate to use an independent proposal distribution from the prior. This is an efficient update if the data do not provide much extra information about the parameter, implying that effectively independent draws are being generated from the posterior. Such approaches require further investigation.

9 Conclusions

The taxonomic analysis problem is a positive matrix factorisation problem, which can be expressed as a nonlinear least squares minimisation problem with non-negative constraints. The solution can be made unique by including prior information on one of the factors.

It is proposed that the best results are obtained by fitting to the original data before scaling and using standard deviations to ensure the error terms are weighted according to their accuracy. The use of standard deviations and prior information means that the results will not be identical to the previous program developed for this problem. It would be helpful if in future some repeat measurements are taken so that actual data standard deviations can be used.

The error terms, being bilinear, can be divided into two non-negative least squares problems, that can be solved alternatively with iterations that progressively reduce the target sum of squares. In Matlab the standard non-negative

least squares algorithm proved significantly slower than an adaptation of an algorithm proposed by Lee and Seung [4]. However, an occasional application of the non-negative least squares algorithm was found to speed the convergence.

Tests indicate a slow but satisfactory convergence to an accurate minimum. Bootstrap techniques can be used to obtain an estimate of the accuracy of the results. The taxa concentrations can be examined to determine variation with location and depth. The residual errors contain information that can be investigated to determine model deficiencies.

Bayesian methods provide an alternative analysis method that provide significantly more statistical information on the distribution of each of the taxa concentrations. As the taxon by sample matrix contains a large number of samples, convergence of the Bayesian methods is much slower than the optimisation approach. The method proposed by Meersche et al. [6] seems incapable of handling the size of data matrices that are needed in some applications of the problem, and a Bayesian method that takes advantage of the problem structure seems more appropriate.

Acknowledgements Simon Wright,[1] as the industry representative, patiently provided the details of the problem. Petras Potgieter located the Lee and Seung algorithm [4]. Others that assisted with problem formulation and solution are James Caffrey, Andrew Stacey, Lynne McArthur, Kaye Marion, Adil Bagirov, Nadia Sukhorukova, and Julien Ugon. We gratefully acknowledge the assistance received during this project.

[1] mailto:Simon.Wright@aad.gov.au

References

[1] Efron, B., and Tibirani, R. J., 1993. *An introduction to the Bootstrap*, Chapman & Hall, New York. M134

[2] Hastings, W. K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97–109. doi:10.1093/biomet/57.1.97 M142

[3] Lawson C. L., and Hanson R. L. 1995. *Solving Least Squares Problems*, SIAM, Philadelphia, Ch 23. M126

[4] Lee D. D., and Seung H. S. 2001. Algorithms for Non-negative Matrix Factorization, *Advances in Neural Information Processing Systems*, 13:556–562. http://luthuli.cs.uiuc.edu/~daf/courses/Optimization/Papers/lee01algorithms.pdf M127, M128, M129, M144

[5] Mackey, M. D., Mackey, D. J., Higgins, H. W. and Wright, S. W., 1996. CHEMTAX—A program for estimating class abundances from chemical markers: application to HPLC measurement of phytoplankton, *Marine Ecology - Progress Series*, 144:266–283. doi:10.3354/meps144265 M121

[6] van den Meersche, K., Soetaert, K., and Middleburg, J. J., 2008. A Bayesian computational estimator for microbial taxonomy based on biomarkers, *Limnol. Oceanogr. Methods* 6: 190–199. doi:10.4319/lom.2008.6.190 M140, M141, M142, M144

[7] Robert, C. P. and Casella, G., 2004. *Monte Carlo statistical methods*, Springer, New York. M142

[8] Roberts, G. O. and Rosenthal, J. S., 2001. Optimal Scaling for Various Metropolis–Hastings Algorithms, *Statistical Science* 16:351–367. doi:10.1214/ss/1015346320 M142

[9] Wikipedia, 2011. *Basic linear algebra subprograms*, http://en.wikipedia.org/wiki/Basic_Linear_Algebra_Subprograms M127

[10] Wright, S. W., van den Enden, R. L., Pearce, I., Davidson, A. T. Scott, F. J., and Westwood, K. J., 2010. Phytoplankton community structure and stocks in the Southern Ocean (30-80E) determined by CHEMTAX analysis of HPLC pigment signatures *Deep-Sea Research II* 57:758–778. doi:10.1016/j.dsr2.2009.06.015 M124, M129, M133, M138

Author addresses

1. **Bill Whiten**, University of Queensland (SMI, JKMRC), Brisbane 4072, Australia.
   mailto:W.Whiten@uq.edu.au
2. **Barry McDonald**, Massey University (IIMS), Auckland 0632, New Zealand.
   mailto:B.McDonald@massey.ac.nz
3. **Chris Drovandi**, QUT (Mathematical Sciences), Brisbane 4000, Australia.
   mailto:C.Drovandi@qut.edu.au