Hydrobiologia **269/270**: 485–502, 1993.
H. van Dam (ed.), Twelfth International Diatom Symposium.

Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages

Cajo J. F. ter Braak [1,2] & Steve Juggins [3,4]
[1] *Agricultural Mathematics Group-DLO, Box 100, 6700 AC Wageningen, The Netherlands;* [2] *DLO-Institute for Forestry and Nature Research, Box 23, 6700 AA Wageningen, The Netherlands;* [3] *Environmental Change Research Centre, Department of Geography, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom;* [4] *Botanical Institute, University of Bergen, Allégaten 41, N-5007 Bergen, Norway*

Key words: diatoms, gradient analysis, indicator values, palaeo-environments, partial least squares regression, PLS, species-environment calibration, transfer function

Abstract

Weighted averaging regression and calibration form a simple, yet powerful method for reconstructing environmental variables from species assemblages. Based on the concepts of niche-space partitioning and ecological optima of species (indicator values), it performs well with noisy, species-rich data that cover a long ecological gradient (> 3 SD units). Partial least squares regression is a linear method for multivariate calibration that is popular in chemometrics as a robust alternative to principal component regression. It successively selects linear components so as to maximize predictive power. In this paper the ideas of the two methods are combined. It is shown that the weighted averaging method is a form of partial least squares regression applied to transformed data that uses the first PLS-component only. The new combined method, weighted averaging partial least squares, consists of using further components, namely as many as are useful in terms of predictive power. The further components utilize the residual structure in the species data to improve the species parameters ('optima') in the final weighted averaging predictor. Simulations show that the new method can give 70% reduction in prediction error in data sets with low noise, but only a small reduction in noisy data sets. In three real data sets of diatom assemblages collected for the reconstruction of acidity and salinity, the reduction in prediction error was zero, 19% and 32%.

Introduction

Current environmental problems, like acid rain and global warming, have increased interest in fossil species assemblages as indicators of the palaeo-environment (e.g. Battarbee & Charles, 1987; COHMAP Members, 1988) and, thus, in quantitative methods for reconstructing environmental variables from species assemblage data. In pollen-climate studies, popular methods include best modern analogue techniques (Overpeck *et al.*, 1985; Guiot, 1990; Prentice *et al.*, 1991) and multiple linear regression (Howe & Webb, 1983; Huntley & Prentice, 1988). In palaeolimnology, the weighted averaging method (Ter Braak & Van Dam, 1989) gained popularity, because it combines ecological plausibility (niche-space partitioning) with simplicity and empirical

predictive power (Oksanen *et al.*, 1988; Birks *et al.*, 1990a, b; Line & Birks, 1990; Dixit *et al.*, 1991; Walker *et al.*, 1991; Fritz *et al.*, 1991). The weighted averaging method seems less vulnerable to the curse of dimensionality (the high number of species) than the analogue method (*cf.* Hastie & Tibshirani, 1990). It does not assume linearity as multiple regression does, is not hindered by multicollinearity and is less sensitive to outliers. Comparatively, weighted averaging is at its best with noisy, species rich, compositional data, with species that may be absent in many of the samples, and a long ecological gradient (> 3 SD; Hill & Gauch, 1980; Ter Braak & Prentice, 1988). However, weighted averaging also has its weak points: it is known to be sensitive to the distribution of the environmental variable in the training set (Ter Braak & Looman, 1986), it considers each environmental variable separately, and it disregards residual correlations among species (*i.e.* correlations that remain after fitting the environmental variable and that are often caused by environmental variables which are not taken into account).

This paper presents an improvement of the weighted averaging method that utilizes the residual correlations in the species data. The improvement is based on partial least squares regression (PLS; Wold *et al.*, 1984; Stone & Brook, 1990). PLS is popular in chemometrics, in particular for extracting chemical information from near infra-red spectra (Martens & Naes, 1989). PLS gives often lower prediction error than the closely related method of principal component regression (PCAR). Both PLS and PCAR are biased regression methods that guard against multicollinearity among predictor variables through the selection of a limited number of orthogonal components, but PLS has a smarter way of selecting components as we will show. The number of components is estimated through cross-validation on the basis of empirical predictive power (*i.e.* prediction error of sum of squares, PRESS). Two main versions of PLS are currently in use, namely a univariate and a multivariate version, PLS1 and PLS2, respectively. Here we restrict discussion to PLS1, because the statistical benefit of PLS2 is limited for calibration (Martens & Naes, 1989). PLS has the same aim and about the same performance as ridge regression (Naes *et al.*, 1986).

In this paper we show that the weighted averaging method is equivalent to PLS regression on transformed data if the first component only is used. The improvement which we have termed Weighted Averaging Partial Least Squares regression (WA-PLS), consists of using further components, namely as many as are useful in terms of predictive power. The further components utilize the residual structure in the species data for improving the species parameters ('optima') in the final weighted averaging predictor. The mathematics is kept as simple as possible; mathematically oriented readers are referred to Ter Braak *et al.* (1993). We show using simulated data that the new method can give up to 70% reduction in prediction error in data sets with low noise, but that little reduction can be achieved with noisy data sets. In three diatom data sets we found zero, 19%, and 32% reduction in prediction error.

Theory

Notation

Let x denote the environmental variable to be calibrated on the basis of a (modern) training data set consisting of the environmental vector $\boldsymbol{x} = (x_i)$, with x_i the value of the environmental variable (*e.g.* pH) in site i (*e.g.* a lake), and the $n \times m$ matrix $\mathbf{Y} = (y_{ik})$ with y_{ik} the abundance of taxon k in site i ($y_{ik} \geq 0$) ($i = 1 \ldots n$ sites and $k = 1 \ldots m$ taxa). A '+' replacing a subscript denotes summation over that subscript, *e.g.* $y_{i+} = y_{i1} + y_{i2} + \cdots + y_{im}$. The estimated or inferred value of x in site i is denoted by $\hat{x}_i$. The index 0 indicates a (fossil) assemblage for which the environmental variable is to be inferred (x_0).

A PLS algorithm

This section describes an algorithm for PLS and highlights the main differences from principal components analysis (PCA) and PCA followed

by regression (PCAR). There are many algorithms for PLS (Helland, 1988), all of which result in the same technique. For better understanding of PLS, the algorithm presented below is similar to the two-way weighted summation algorithm of PCA of Ter Braak (1987: Table 5.6) and Ter Braak & Prentice (1988). The algorithm proceeds by successive extraction of components. The maximum number of components is equal to the minimum of m and $n-1$.

Step 0. Optionally preprocess the environment and species data (*e.g.* subtract means). Thereafter denote the data by $(\tilde{x}_i)$ and $(\tilde{y}_{ik})$.

Step 1. Take the environmental variable $(\tilde{x}_i)$ as initial site scores (r_i).

Do Steps 2 to 7 for each component:

Step 2. Calculate new species scores (b_k) by weighted summation of the site scores, *i.e.* $b_k = \Sigma_i \tilde{y}_{ik} r_i$.

Step 3. Calculate new site scores (r_i) by weighted summation of the species scores, *i.e.* new $r_i = \Sigma_k \tilde{y}_{ik} b_k$.

Step 4. For the first component go to Step 5. For second and higher component, make the new site scores (r_i) uncorrelated with the previous components by orthogonalization (Ter Braak, 1987: Table 5.6b)

Step 5. Standardize the new site scores (r_i) (Ter Braak, 1987: Table 5.6c).

Step 6. Take the standardized scores as the new component.

Step 7. Regress the environmental variable $(\tilde{x}_i)$ on the components obtained so far and take the fitted values as current estimates of $(\tilde{x}_i)$. Go to Step 2 with the residuals of the regression as the new site scores (r_i). (The stop-criterion, *i.e.* the choice of the number of components is discussed below).

Comparison with Table 5.6 of Ter Braak (1987) reveals three small but important differences between the algorithms of PLS and PCA. (1) PCA uses arbitrary initial site scores whereas PLS uses the environmental scores $\tilde{x}_i$ or the current residuals (Steps 1 and 7). (2) PCA requires an iterative process of Steps 2–5 to calculate each component, whereas in PLS each single execution of Steps 2–5 yields a new component (Step 6). (3) The regression for fitting the $(\tilde{x}_i)$ is part of the PLS algorithm (Step 7), whereas the regression is carried after the extraction of the principal components in principal component regression (PCAR). As a result of these differences, the components in PCAR are calculated irrespective of their predictive value for the environmental variable x, whereas they show maximum covariance with x in PLS (guaranteed by Steps 2 and 3; Stone & Brook, 1990). If we take as many PLS-components as species, PLS reduces to a multiple regression of x on all species variables. The number of components is an essential ingredient of PLS: the choice number minimizes the prediction error as estimated by cross-validation methods (Wold *et al.*, 1984; Martens & Naes, 1989). An example of cross-validation is given later on.

The weighted averaging method

The weighted averaging method is based on the idea that species occupy different niches in environment space (Shelford, 1911; Whittaker, 1956) and that the niches can be characterized by their centres (u_k) and breadths (t_k). This characterization is particularly appropriate if the niches are closely packed along the environmental variable and follow unimodal, or even Gaussian, response curves, so that the centres and breadths are the optima and tolerances of these curves (Ter Braak & Barendregt, 1986). In practical applications of weighted averaging, it has rarely been found advantageous to use differential niche breadth estimates (Cumming *et al.*, 1991). Because it simplifies the formulae, niche breadths are disregarded until the Discussion section.

The weighted averaging (WA) method consists of three parts: WA regression, WA calibration and a deshrinking regression. The parts are motivated as follows. A species with a particular optimum will be most abundant in sites with x-values close to its optimum. This motivates

Part 1 (WA regression): Estimate species optima (u_k^*) by weighted averaging of the x-values of the sites, *i.e.* $u_k^* = \Sigma_i\, y_{ik}\ x_i/y_{+k}$.

Species present and abundant in a particular site will tend to have optima close to its x-value. This motivates

Part 2 (WA calibration). Estimate the x-values of the sites by weighted averaging of the species optima, *i.e.* $x_i^* = \Sigma_k\, y_{ik}\ u_k^*/y_{i+}$.

Because averages are taken twice, the range of the estimated x-values (x_i^*) is shrunken. The amount of shrinking can be estimated from the training set by regressing either (x_i^*) on (x_i) or (x_i) on (x_i^*) proposed by Ter Braak (1988) and Ter Braak & Van Dam (1989), respectively. Birks *et al.* (1990a) discuss the virtue of these two deshrinking methods. For establishing the link with PLS we need the latter, 'inverse' deshrinking regression. This method also has the attractive property of giving minimum root mean squared error in the training set. This motivates

Part 3 (deshrinking regression). Regress the environmental variable (x_i) on the preliminary estimates (x_i^*) and take the fitted values as the estimates of (x_i).

The final prediction formula for inferring the value of the environmental value from a fossil species assemblage is thus

$$\hat{x}_0 = a_0 + a_1 x_0^* = a_0 + a_1\, \Sigma_k\, y_{0k}\, u_k^*/y_{0+}$$
$$= \Sigma_k y_{0k}\, \hat{u}_k/y_{0+}$$

where a_0 and a_1 are the coefficients of the deshrinking regression and $\hat{u}_k = a_0 + a_1\, u_k^*$. The final prediction formula is thus again a weighted average, but one with updated species optima.

Definition of WA-PLS

It shown in the Appendix that, with a small amendment, the weighted averaging method is equivalent with PLS applied to transformed data, using the first component only. The amendment modifies the deshrinking regression to a weighted regression with weights proportional to the site total (y_{i+}). This amendment is prudent because the variance of a weighted average tends to be inversely related to the site total (Ter Braak & Barendregt, 1986: equations (5.6) and (7.4)). For data with constant site totals it is of course immaterial. This weighting is used in all subsequent statistical calculations, such as means, variances, prediction error sums of squares, regression, standardization and orthogonalization.

With the equality of the first component of PLS on transformed data and weighted averaging established, the questions are what the further PLS-components and the final predictor look like and, of course, whether it is an improvement. The latter question is answered later on by analyzing simulated and real data. To answer the first question, an explicit algorithm for the new method is given by integrating all the necessary data transformations in the PLS-algorithm. This is the method that we term WA-PLS:

Step 0. Centre the environmental variable by subtracting the weighted mean, *i e.* $x_i := x_i - \Sigma_i\, y_{i+} x_i/y_{++}$. This simplifies the formulae.

Step 1. Take the centred environmental variable (x_i) as initial site scores (r_i).

Do Steps 2 to 7 for each component:

Step 2. Calculate new species scores (u_k^*) by weighted averaging of the site scores, *i.e.* $u_k^* = \Sigma_i\, y_{ik}\, r_i/y_{+k}$.

Step 3. Calculate new site scores (r_i) by weighted averaging of the species scores, *i.e.* new $r_i = \Sigma_k\, y_{ik}\, u_k^*/y_{i+}$.

Step 4. For the first axis go to Step 5. For second and higher components, make the new site scores (r_i) uncorrelated with the previous components by orthogonalization (Ter Braak, 1987: Table 5.2b)

Step 5. Standardize the new site scores (r_i) (Ter Braak (1987: Table 5.2c).

Step 6. Take the standardized scores as the new component.

Step 7. Regress the environmental variable (x_i) on the components obtained so far using weights (y_{i+}/y_{++}) in the regression and take the fitted values as current estimates ($\hat{x}_i$). Go to Step 2 with the residuals of the regression as the new site scores (r_i). (The stop-criterium, *i.e.* the choice of the number of components is discussed below).

We see that the first component is a two-way weighted average for the original environmental variable. Further components are two-way weighted averages for the residual of this variable. In Step 7, a joint estimate $\hat{x}_i$ is obtained as a linear combination of the components of WA-PLS, each of which is a weighted average of species scores. The final prediction formula is thus again a weighted average, but one with updated species optima ($\hat{u}_k$). Intuitively, from Step 2, the optima of species that are abundant in sites with large residuals are likely to be updated.

As in PLS, the number of components is determined by cross-validation on the basis of prediction error sum of squares (see below). WA-PLS is expected either to equal or to outperform the original weighted averaging method depending on whether the optimal number of components is 1 or greater than 1.

Table 1 demonstrates the need for cross-validation on an artificial example with 100 sites and 131 species. With each additional WA-PLS component, the model fits the environmental variable better as measured by the root mean square of the errors (RMSE). But, the RMSE is not corrected for degrees of freedom; it is like the coefficient of determination R^2 in regression: the fit can be perfect ($R^2 = 1$ and $RMSE = 0$), for example with n sites and n-1 species, even if there is no relation between the environmental variable and the species at all. How untrustworthy the RMSE is can be demonstrated by applying the resulting transfer functions to a test set of 1000 sites. For each site in the test set a prediction of the environmental variable is made from its species data and compared with its known value of the environmental variable. The errors in the prediction are accumulated and expressed as the root mean square of the errors of prediction, RMSEP for short, to distinguish it from the untrustworthy RMSE in the training set. Table 1 shows that the RMSEP initially decreases but already starts to increase when more than two components are used. Thus, WA-PLS with many components fits the data perfectly, but has little predictive value. In other words, the optimum number of components should not be determined on the basis of the model fit or 'apparent' errors in the training set (RMSE), but on the prediction errors in a test set (RMSEP). But, in real applications large test sets are generally not available. Instead, the prediction errors in a test set are simulated by cross-validation. In the example with 100 sites, cross-validation by 'leave-one-out' means that WA-PLS is applied 100 times to a set of 99 sites, leaving out each site in turn. The transfer function based on these 99 sites is applied to the omitted site giving for this site a prediction and, by subtraction of the measured x-value, a prediction error. The sites so take in turn the rôle of a test set, each time of size 1. The prediction errors are accumulated to a 'leave-one-out' RMSE which is a consistent estimate of the true RMSEP. In the example, the 'leave-one-out' RMSE is least with three components (Table 1),

Table 1. Performance of WA-PLS in relation to the number of components (s): apparent error (RMSE) and prediction error (RMSEP) in simulated data (R = 1 from simulation series III). The estimated optimum number of components is 3 because three components give the lowest RMSEP in the training set. The last column is not available for real data. For further explanation see text.

s	Training set		Test set
	Apparent	Leave-one-out	
	RMSE	RMSEP	RMSEP
1	6.14	6.22	6.61
2	3.37	4.24	4.40
3	2.87	4.16*	4.57
4	2.22	4.65	4.94
5	2.01	4.65	5.11
6	1.82	4.50	5.62

hence, the number of components to use is three. In the example, the optimum number is actually 2 as judged from the RMSEP in the test set, but such a large test set is never available in real applications. The number of WA-PLS components is therefore always based on cross-validation. The final transfer function is based on all training sites.

Test data

Simulations

The simulations focus on whether WA-PLS can improve on WA in data sets with a single underlying environmental variable. Simulations with two underlying variables are reported in Ter Braak *et al.* (1993).

The simulated training set consists of 100 sites with x sampled from a lognormal distribution limited to the interval [0, 100]. The sample mean and standard deviation are 32 and 25, respectively. The geometric mean is 22. The test set consists of 1000 equidistant sites on the interval [0, 100]. The mean is thus 50. Because the mean in the training set is much lower, we can expect negative bias and bad performance in the upper range of x.

There are three series of simulations which differ in the way the species abundance data are generated. In series I, 50 species respond to x according to Gaussian response curves (Ter Braak & Van Dam, 1989). The abundance value y_{ik} is read off the Gaussian response curve for species k at the value x_i. Series I has three parts. In series Ia, the Gaussian curves have equal maximum (10) and equal tolerance (t) and optima that are equidistant on the interval $[-3t, 100 + 3t]$, well embracing the sampling interval of the training set. Four values of t (100, 50, 25 and 12.5) are taken to assess the influence of beta-diversity (length of gradient) on the performance of WA-PLS. These values correspond to gradient lengths of 1, 2, 4 and 8 Standard Deviation (SD) units (Hill & Gauch, 1980; SD = (sample range)/t). In parts Ib and Ic, SD is 4, the optima are uniformly distributed (instead of being equidistant) and the maxima are lograndom between 5 and 20 (*i.e.* their logarithm is uniform between ln(5) and ln(20)). In part Ib, $t = 25$, whereas in part Ic, t is uniformly distributed between 10 and 35. The performance is measured by RMSE (weighted by the species total) of the estimated optima ($\hat{u}_k$) and by the bias (estimated – true value) in $\hat{u}$ and $\hat{x}$ as a function of the true value.

Simulation series II and III are more realistic. Species response curves vary randomly in shape among species, are skew and, compared to the Gaussian curve, longer-tailed and flatter topped. Moreover, qualitative and quantitative noise is added to generate the data. In series II the amount of qualitative noise (percentage absence) is varied and in series III the length of gradient. For these series the COMmunity PAttern Simulator (COMPAS; Minchin, 1987) is used.

The details are as follows. In COMPAS, each species curve is a unimodal, generalized beta function with five parameters which can be varied randomly between species. They are: the range (r) over which the curve rises and falls, the modal coordinate or optimum (u), the maximum (A_u), and two shape parameters, α and γ, that govern the skewness and kurtosis. The length of gradient can then be defined in Range (R) units as (sample range)/$\bar{r}$, where $\bar{r}$ is the mean range of the species curves. The gradient length was held constant in series II (R = 1, approximately 5 SD) and varied in series III (R = 0.5, 1, 2). The following settings are used: 150 species response curves are generated with r uniform between 50/R and 150/R, u uniform between −75/R and 175/R (thus well embracing the sampling interval), α and γ independent and uniform between 0.5 and 3.5 (giving skew and flat-topped curves) and A_u uniform between 10 and 50 (series II) and lograndom between 10 and 80 (series III). The abundance data are generated by sampling from the 150 response curves, with the addition of qualitative and quantitative noise. Quantitative noise is derived from the Poisson distribution. Qualitative noise is added by randomly replacing abundance values by zeroes. The probability that such a replacement does not happen is specified by the same beta function in which A_u is replaced by P_u the

maximum probability of occurrence of the species. In series II, the probability curves are made very flat-topped by also multiplying the shape parameters α and γ by 0.2; P_u is held constant within a data set but varied between simulations to give a range of noise levels. The noise level is expressed as the complement of P_u, *i.e.* the minimum percentage absence. In series III the qualitative noise is held constant: in each simulation, P_u is lograndom between 0.25 and 1 (mean noise level *ca.* 0.5), and α and γ are those of the quantitative response curve. In summary, each abundance value y_{ik} is thus a count that is Poisson distributed with an expected value specified by reading off its response curve at the value x_i and an extra probability of absence determined by the complement of its probability of occurrence curve at the value x_i.

Real data

WA-PLS is also applied to three surface sediment diatom/water chemistry data sets. Two data sets relate to lake-acidification studies. The first was developed as part of the Surface Waters Acidification Programme (SWAP) and was used to derive the weighted averaging-based transfer function that provided diatom-based pH reconstructions for all sediment core studies in the SWAP project (Birks *et al.* 1990a). The data set contains 167 samples from Norway, Sweden and the United Kingdom. pH values range from 4.3 to 7.3 (mean 5.6). The distribution of pH is skew, with approximately 50% collected from lakes of pH 4.5–5.5. Further details of the data set can be found in Stevenson *et al.* (1991).

The second training set was developed at Bergen University by H. J. B. Birks, J. F. Boyle & F. Berge (unpublished), and consists of 92 samples from lakes in southern and central Norway. The data set was developed to provide transfer functions for inferring pH, DOC and labile aluminium, and was designed to give an even coverage of samples along these gradients. Samples are more or less uniform over the range of pH 4.3–8.3 (mean 5.8), except for a concentration of samples (34%) between pH 4.5–5.0.

The third training set was collected as part of a study into the palaeoecology of the Thames Estuary (UK) and was used to develop a weighted averaging based transfer function to provide salinity reconstructions for the estuary over the last 2000 years. The data set consists of 135 samples collected from 17 sites evenly distributed along the salinity gradient from the tidal head to the lower estuary, 72 km downstream. Salinity (calculated as the annual mean half-tide value at each site) ranges from 0.081 to 17.1 g l^{-1}, and was log-transformed for all analyses, using the transformation $\log_{10}$ (salinity-0.08). The Thames data set is fully described by Juggins (1992).

The diatom data for all 3 training sets are expressed as percentages of the total valve count. Rare taxa were excluded from each set by retaining only those which achieved a relative abundance of greater than 1.0 percent in any single sample. This gave totals of 277, 150 and 110 taxa for the SWAP, Bergen and Thames data sets, respectively.

Data analysis

The number of species present in a site was usually below 50, except in the simulations without qualitative noise. For weighted averaging methods, the number of species in a site is not very important. For example, if there are three species present with abundances 100, 1, and 1, respectively, the first species takes nearly all the weight so that the effective number of species is close to 1. A good measure for the effective number is Hill's (1973) N_2 measure of diversity, which is the reciprocal of Simpson's diversity index (cf. Hill, 1979: 28 and Ter Braak, 1990). We report the median and range of the effective numbers of species per site and of the effective number of occurrences per species (defined analogously). To further characterize the training sets in data-analytical terms, a detrended canonical correspondence analysis was carried out using the program CANOCO 3.1 (Ter Braak, 1986; 1990) with x as the only environmental variable. Detrending-by-segments was used. The

length of gradient of the first axis of this analysis is the gradient length of x in SD-units. We also report the gradient length of the second, unconstrained axis, and the first two eigenvalues.

WA and WA-PLS were applied to the training sets. For comparison, standard PLS was also applied in series III. We report the results of PLS on standardized log-transformed abundance data ($\ln(y_{ik}+1)$), which tended to give better results than PLS without standardization or without log-transformation. The calculations were carried out by using the program CALIBRATE (Juggins & Ter Braak, 1992). The number of components was determined by leave-one-out. Up to six components were tried, since initial trials showed that the optimal number of components was always less than six. The resulting transfer function was applied to the test set. The performance of WA, WA-PLS and, in series III, PLS was measured by the root mean squared error of prediction of x (RMSEP). For the training set, RMSEP was estimated by leave-one-out. The gain of WA-PLS over WA was expressed as 1 – RMSEP(WA-PLS)/RMSEP(WA). Further aspects of performance were the average bias and maximum bias in the prediction in the test set. For estimation of the maximum bias, the sampling interval (0, 100) was subdivided into 10 equal intervals, the bias per interval calculated and the (signed) maximum of the 10 values calculated. The 5% and 95% envelopes of the error (estimated – true value) were also calculated for each of the 10 intervals.

Results

Simulated data

Series Ia concerns data sets in which weighted averaging of true optima of Gaussian curves is fully efficient compared to maximum likelihood (Ter Braak & Barendregt, 1986). However, the true optima are not available, they must be estimated. The full WA method does not estimate them reliably (Table 2a, component 1), especially when the gradient is short. For both long and short gradients, further components of WA-PLS achieved a substantial decrease in the error in estimating the optima (Table 2a). Figure 1a shows that WA (component 1) overestimates the small optima and underestimates the large ones. Further components of WA-PLS remove this bias. In plots of $\hat{u}$ against u (not shown), the further components are seen to 'stretch out the ends'. Figure 1b shows that WA (component 1) gives biased predictions of x. Further components of WA-PLS decrease this bias. In simulations with other distributions of x in the training set, similar patterns of bias were found for $\hat{u}$. The patterns of bias in $\hat{x}$ were different, but all bias vanished when more components were added.

Table 2. Simulation series I: effect of the length of gradient in SD-units (Ia), differential heights (Ib and Ic) and widths (Ic) of the Gaussian response curves on the RMSE of the optima ($\hat{u}$) estimated by WA-PLS with 1 to 4 components for noiseless data. For further details see text.

(a) Part Ia

SD	Components			
	1	2	3	4
1	13.4	3.5	3.6	3.8
2	11.7	5.4	1.3	0.9
4	8.5	5.5	2.2	1.2
8	4.8	3.1	2.2	1.4

(b) SD = 4

Part	Components			
	1	2	3	4
Ib	7.9	6.0	4.5	7.5
Ic	10.5	9.6	9.6	10.0

In parts Ib and Ic of series I the Gaussian response curves are more variable. WA of true optima is not efficient then and does not give perfect predictions, not even for noiseless data. In these cases WA-PLS does not recover the true optima (Table 2b): after an initial decrease, the error in $\hat{u}$ starts to increase when component 4 is added. Although the bias in the optima does not vanish with higher components (Fig. 1c, e), the bias in the prediction of x does (Fig. 1d, f).

The training sets of simulation series II are

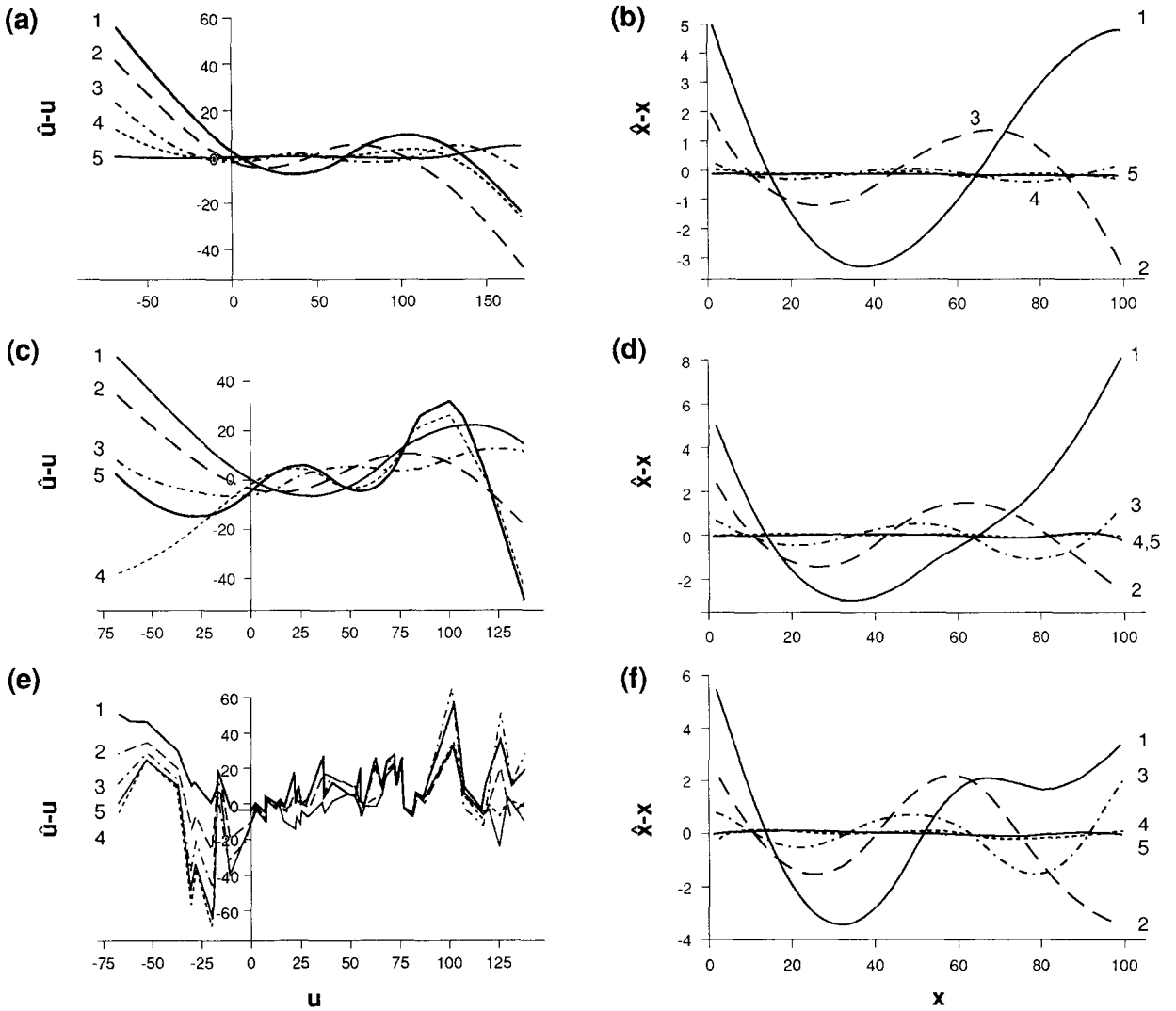

Fig. 1. Simulation series I (SD = 4): Bias in $\hat{u}$ (a, c, e) and $\hat{x}$ (b, d, f) as a function of the true value of u and x, respectively, for an increasing number of components (s = 1, 2...5) in WA-PLS. From top to bottom, the variability among the Gaussian response curves increases: equal maxima and tolerance in (a) and (b), random maxima and equal tolerance (Table 2b: Part Ib) in (c) and (d) and random maxima and tolerances (Table 2b: Part Ic) in (e) and (f).

summarized in Table 3. With an increase in noise level, the percentage absence increases and, thus, the effective number of species per sample and the effective number of occurrences per species decrease. The gradient length of the second axis of a detrended canonical correspondence analysis increases with the noise level from *ca* 1 to 5 SD. This is remarkable because the data sets are one-dimensional by simulation. Table 4a shows that, compared with WA, WA-PLS greatly reduces the prediction error in low noise data sets (up to 72%), but the reduction decreases with noise level. For very high noise levels (p = 75 in Table 4) WA-PLS actually performs slightly worse than WA on the test set (−3%). The gain as estimated by leave-one-out in the training set is somewhat over-optimistic, when compared with the gain achieved in the test set.

Table 4b shows that the average bias in WA and WA-PLS reconstructions is small in view of

Table 3. Data summary of the training sets in simulation series II. (p: level of qualitative noise; N_2 for samples: effective number of species per sample; N_2 for species: effective number of occurrences per species; λ_1, λ_2 and SD_1, SD_2: eigenvalue (λ) and length of gradient (SD) of detrended canonical correspondence analysis).

p	0	10	25	50	75
N_2 for samples					
Minimum	41	30	25	12	4
Median	46	38	31	21	10
Maximum	50	43	38	28	20
N_2 for species					
Minimum	1	1	1	1	1
Median	35	28	25	15	9
Maximum	92	80	68	48	23
λ_1	0.62	0.65	0.65	0.64	0.66
λ_2	0.03	0.06	0.07	0.17	0.36
SD_1	4.7	5.0	5.0	4.9	5.0
SD_2	0.9	1.1	1.5	4.9	4.5

the difference in mean of −18 between the training and the test sets. The maximum bias (Table 4c) tends to be smaller in WA-PLS than in WA, except at the highest noise level.

In applications, the amount of noise in the fossil data may differ from that in the modern training set. Table 5 shows that if the training set has a low noise level, the error depends strongly on the noise in the test set.

The training sets of simulation series III are summarized in Table 6. Again, detrended canonical correspondence analysis detects a strong second gradient. Table 7 compares WA, WA-PLS and standard PLS for this series. Over the tested range of gradient lengths, WA-PLS outperforms both WA and PLS in terms of RMSE and maximum bias. If the gradient is long (R = 2, SD = 11), standard PLS is clearly inappropriate as judged by the error in the test set. Figure 2 shows the case of intermediate gradient length (R = 1, SD = 5.6) in more detail. As the bias and error bands in Fig. 2a show, WA overestimates for low values of x, underestimates for intermediate values and overestimates for high values. WA-PLS also gives biased predictions, but the bias and error are less. In the middle of the training set (x between 15 and 50) PLS gives prediction errors that are comparable or somewhat larger than WA-PLS. However, the error bands

Table 4. Simulation series II: effect of level of qualitative noise (p) on the RMSEP (a) and the average (b) and maximum bias (c) in WA and WA-PLS in the training set of 100 lognormal distributed sites (t) and the test set of 1000 equidistant samples (e). Between brackets is the estimated optimal number of components in the training set. This number is used in the test set. For further details see text.

p	Set	WA	WA-PLS	Gain
(a) RMSEP				
0	t	1.89	0.53 (4)	72%
	e	1.91	0.65	66%
10	t	2.48	1.42 (3)	42%
	e	2.32	1.43	38%
25	t	2.81	1.97 (3)	30%
	e	2.56	2.14	16%
50	t	2.98	2.61 (3)	12%
	e	3.44	3.26	5%
75	t	4.88	4.69 (2)	5%
	e	5.24	5.41	−3%
(b) Average bias				
0	e	0.35	0.09	
10	e	0.17	−0.23	
25	e	−0.16	−0.48	
50	e	−0.05	−0.25	
75	e	−1.86	−2.05	
(c) Maximum bias over the range (0, 100)				
0	e	2.81	−0.45	
10	e	3.22	−0.83	
25	e	3.81	−2.06	
50	e	2.90	−2.06	
75	e	−5.01	−9.10	

Table 5. Simulation series II: effect of level of qualitative noise (p) in the test set on the RMSEP in WA and WA-PLS. The training set has noise level 10.

p	WA	WA-PLS	Gain
0	1.80	0.92	49%
10	2.32	1.43	38%
25	2.61	1.97	24%
50	3.46	3.01	13%

Table 6. Data sumary of the training sets in simulation series III. (R: Range unit). For legend see Table 3.

R	0.5	1.0	2.0
N_2 for samples			
Minimum	13	6	5
Median	21	14	10
Maximum	29	26	14
N_2 for species			
Minimum	1	1	1
Median	17	9	8
Maximum	84	57	48
λ_1	0.38	0.72	0.92
λ_2	0.18	0.29	0.36
SD_1	2.9	5.6	11.5
SD_2	2.6	4.0	3.6

widen at the ends of the scale, most notably the upper end. This error pattern in PLS is also present in similar graphs (not shown) for R = 0.5 and R = 2. The example of Table 1 is case 'R = 1' in Table 7. The RMSEP's quoted in Table 7 for WA and WA-PLS are easily retraceable in Table 1.

Table 7. Simulation series III: effect of the length of gradient (in R-units) on the RMSEP (a), average (b) and maximum (c) bias in WA, WA-PLS and standard PLS. For legend see Table 4.

R	Set	WA	WA-PLS	PLS
(a) RMSEP				
0.5	t	7.19	5.66 (2)	6.99 (1)
	e	7.82	6.12	7.03
1.0	t	6.22	4.16 (3)	5.01 (2)
	e	6.61	4.57	6.15
2.0	t	3.68	2.92 (3)	6.25 (2)
	e	2.82	2.70	8.09
(b) Average bias				
0.5	e	−0.50	−1.44	0.41
1.0	e	1.06	0.08	0.83
2.0	e	0.52	0.51	1.13
(c) Maximum bias over the range of (0, 100)				
0.5	e	9.21	4.19	7.58
1.0	e	−9.26	−5.00	4.58
2.0	e	3.33	2.78	6.83

Real data

Table 8 summarizes the three diatom/chemistry data sets. Despite the large total number of taxa (*ca* 100–300) individual samples are generally species poor in comparison with the simulated data sets of series II and III. The median N_2 of 8–13 species per sample, indicates, as expected, a high degree of noise. The SWAP and Bergen pH-related data sets have large secondary gradients, as revealed by detrended canonical correspondence analysis. For the SWAP data set the second unconstrained DCCA axis is larger than the first, reflecting its greater diversity of lake

Table 8. Data summary for the real-data training sets. For legend see Table 3.

	SWAP	Bergen	Thames
N samples	167	92	135
N taxa	277	150	110
N_2 for samples			
Minimum	1.6	1.4	3.4
Median	11.3	8.1	12.5
Maximum	29.7	21.6	32.3
N_2 for species			
Minimum	1.0	1.0	1.1
Median	7.8	6.1	26.1
Maximum	82.6	51.0	99.4
λ_1	0.50	0.73	0.44
λ_2	0.39	0.33	0.13
SD_1	3.5	4.3	2.7
SD_2	4.1	2.9	1.9
Variable	pH	pH	log S* − 0.08)
Minimum	4.3	4.3	−3.0
Mean	5.6	5.8	−1.1
Median	5.3	5.4	−1.2
Maximum	7.3	8.3	1.2
Stand. dev.	0.77	1.15	1.31

* S = salinity in g l^{-1}.

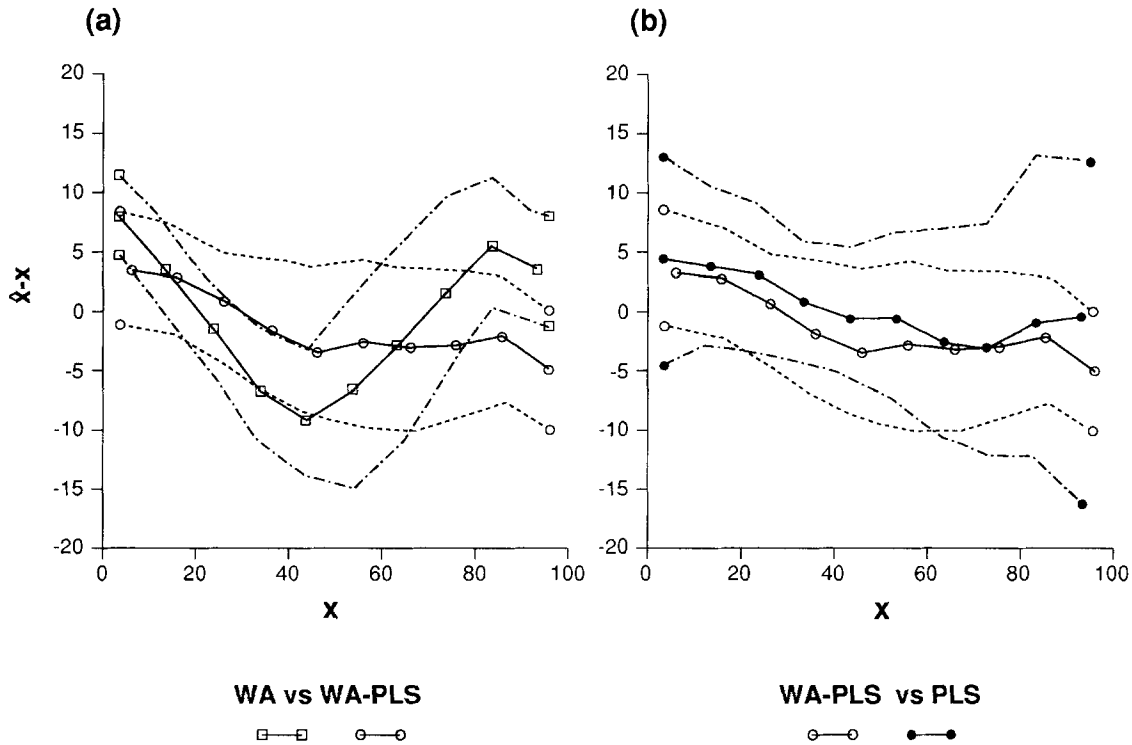

Fig. 2. Simulation series III (R = 1.0): Bias in $\hat{x}$ (solid line) and 5%- and 95%-envelopes (dashed) of prediction error ($\hat{x}$-x) as function of the true value (x) for WA-PLS (open circles) compared to WA (squares) in (a) and to PLS (solid circles) in (b).

types and corresponding non-pH related variation in diatom composition. The Thames Estuary data, by comparison, are essentially one-dimensional, reflecting their collection from an ecological system dominated by a single strong environmental gradient.

Table 9 summarizes the results of WA-PLS. For the SWAP data set the prediction error, as judged by the leave-one-out RMSEP, is minimal for the second WA-PLS component. However, the reduction in prediction error from the first to second component is small (0.310 to 0.302 pH

Table 9. The performance of WA-PLS applied to the three diatom data sets in relation to the number of components (s) in terms of apparent RMSE and leave-one-out RMSEP. (* = selected model).

Dataset	SWAP		Bergen		Thames	
s	RMSE	RMSEP	RMSE	RMSEP	RMSE	RMSEP
1	0.276	0.310*	0.353	0.394	0.341	0.354
2	0.232	0.302	0.256	0.318*	0.238	0.279
3	0.194	0.315	0.213	0.330	0.196	0.239*
4	0.173	0.327	0.192	0.335	0.166	0.224
5	0.153	0.344	0.174	0.359	0.153	0.219
6	0.134	0.369	0.164	0.374	0.140	0.219
Reduction in prediction error (%)		0		19		32

units). Therefore we would use the first component only for reconstruction. For the SWAP data set WA-PLS offers no improvement over WA.

For the Bergen data set the second WA-PLS component yields the lowest RMSEP (Table 9), so a two-component model is selected, giving a 19% reduction in prediction error over WA. Figure 3 shows that the first component (*i.e.* WA) overestimates low values of pH, and underestimates high values (compare Fig. 1b). The second

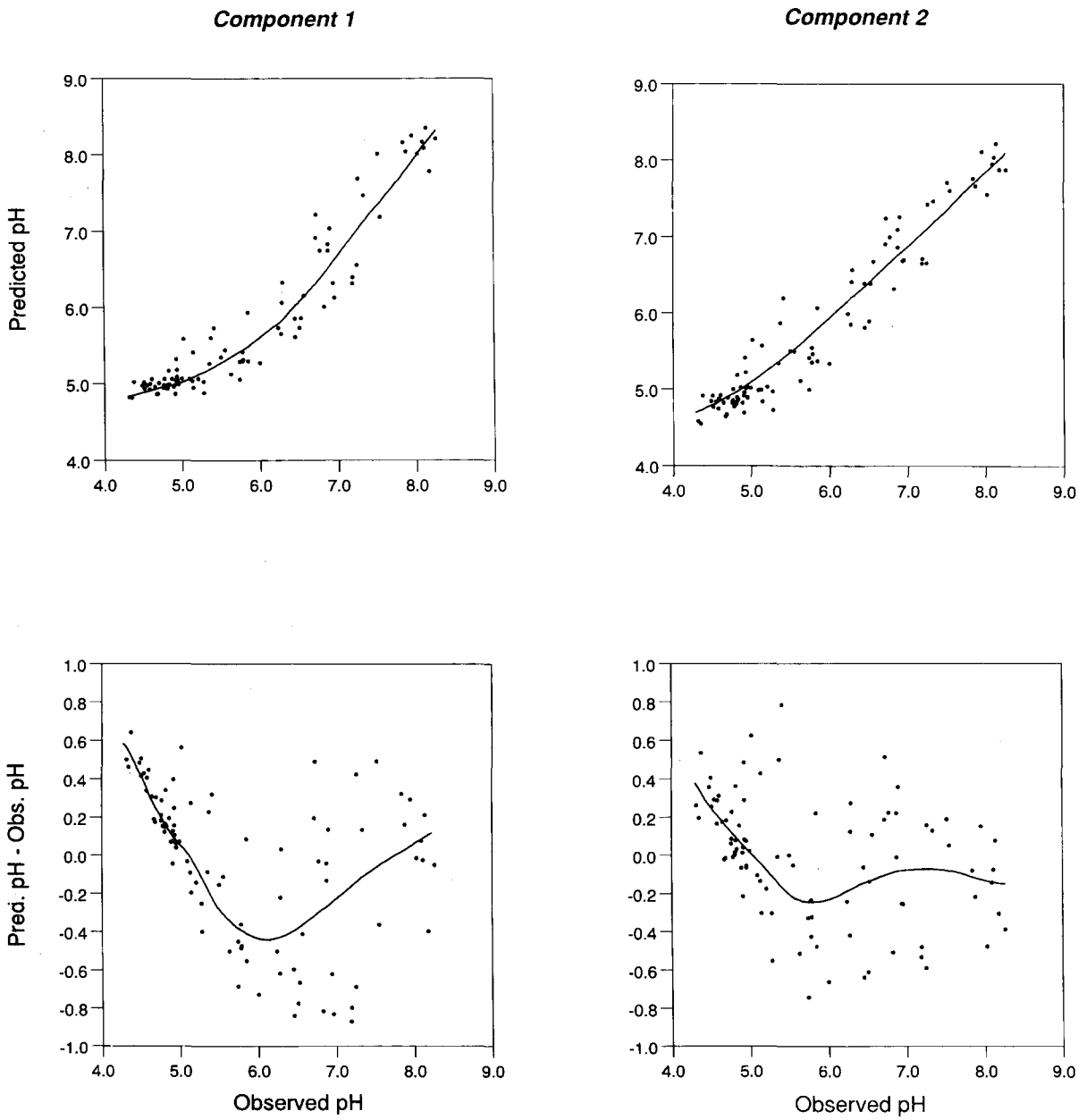

Fig. 3. Bergen data set: predicted pH and bias as a function of observed pH for components 1 and 2 in WA-PLS. Solid lines represent Cleveland's Loess scatterplot smooth (1979).

component removes some bias, and 'straightens out' the predicted values, although there is still a tendency to overestimate at low pH.

For the Thames data the prediction error is minimal with five components of WA-PLS. Because reduction in error is small after three components, we decided to retain three components only, giving a reduction in prediction error over WA of 32%. WA overestimates at low values of salinity (Fig. 4), except for some samples that are at the head of the estuary. These samples have a entirely freshwater diatom flora and many

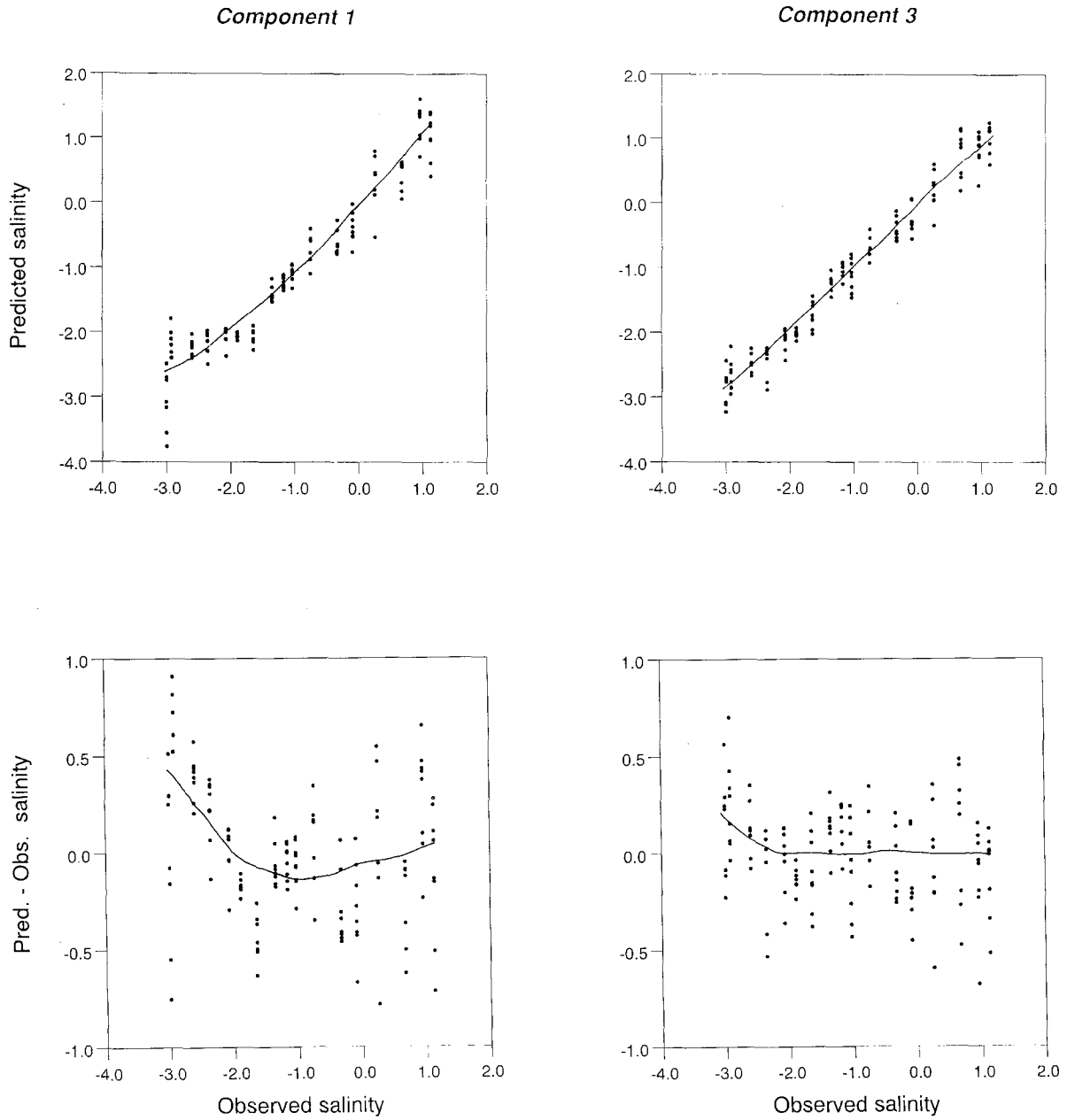

Fig. 4. Thames data set: predicted salinity and bias as a function of observed salinity for components 1 and 3 in WA-PLS. Solid lines represent Cleveland's Loess scatterplot smooth (1979). Salinity in g l^{-1} and transformed as $\log_{10}$ (salinity − 0.08).

of the taxa are poorly represented elsewhere in the estuary. The salinity optima of the constituent taxa are well estimated by WA regression but the subsequent deshrinking is somewhat overzealous for these samples. Figure 4 shows that the reduction in prediction error is largely due to the improved fit at the lower values of salinity.

Discussion

This is the first paper to test the WA method by simulation. The simulations and real data show that WA can give biased predictions. The patterns of error apparent in Figs 3 and 4 have been noticed in other applications of WA with inverse deshrinking (e.g. Birks *et al.*, 1990b; Hall & Smol, 1992). In WA with classical deshrinking the initial inferred values (x_i^*) are regressed on observed (x_i), so the residuals are orthogonal to the predicted values ($\hat{x}_i$), and so uncorrelated with the original observed values. With inverse deshrinking the residuals are orthogonal to the original inferred values (x_i^*), not the observed ones, and plots of (final) inferred against observed reveal this bias. Since, by comparison with classical deshrinking, inverse deshrinking 'pulls' the predicted values towards the mean of the training set, this inevitably leads to overestimation at low, and underestimation at high values. Our improved method, WA-PLS, exploits the patterns in the error to update the transfer function, so reducing the error and the pattern in the bias. However, since inverse deshrinking is implicit in the method, it does not remove this source of bias completely. We plan to investigate the use of smoothing splines (*cf.* Wold, 1992) to improve the deshrinking. The simulations also point out that noise may prevent the method from achieving a real improvement. So far, application of the method to real data showed either no improvement or a modest reduction in prediction error (19–32%), but never a large improvement. The new method is worth trying, especially when WA gives suspect patterns in the error in the training set.

The WA method was designed as an approximation to maximum likelihood calibration using a Gaussian model (Ter Braak & Van Dam, 1989). Nevertheless WA can be improved for two reasons. The first is that the approximation is not an ideal one: WA regression and WA calibration are fraught with edge effects giving nonlinear distortions that are well known from correspondence analysis (Hill & Gauch, 1980). The linear deshrinking regression does not solve these problems but WA-PLS does (Fig. 1). The second reason is that there are likely to be additional environmental variables that govern the species assemblage. The structure that results from these variables is not used at all in WA. WA thus assumes that environmental variables other than the one of interest have negligible influence (Ter Braak, 1988; Birks *et al.*, 1990a). WA-PLS does use the additional structure in a similar way as the multiple linear regression approach to calibration (see Lorber *et al.*, 1987). Consequently, in WA-PLS environmental variables other than the one of interest may influence the species assemblage but for optimal performance their joint distribution in the fossil set should be the same as in the training set (Brown, 1979).

Although WA tends to perform less well than WA-PLS, both methods perform reasonably with the simulated data, despite the complexity of and variability in the underlying response curves. The magnitude of the error should be compared to the sampling range of 100. In all simulations, the correlation between the WA reconstruction and the true value exceeds 0.95. The bias is small compared to difference in mean (-18) between the training set and the test set.

Because WA-PLS uses the residual structure in the species data to improve upon WA, we had hoped that WA-PLS would improve more over WA the noisier the data. However, Table 4 shows otherwise. An explanation might be that in the simulations unstructured noise is added. If the noise is due to other environmental gradients, it has more structure for WA-PLS to use. Ter Braak *et al.* (1993) confirmed this conjecture by simulation of data sets with two underlying gradients. In these simulations WA-PLS halved the prediction error of WA!

Table 5 shows in terms of practical reconstruc-

tion that even with a reliable training set, the reconstruction error may be high if the fossil data set is noisy. Conversely, if the training set is very noisy, a reliable fossil set can, however, not give precise reconstructions.

So far we have disregarded differential niche breadths. The weighted averaging method can, however, be modified to take account of differential niche breadths (t_k) (Ter Braak & Barendregt, 1986; Ter Braak & Van Dam, 1989). The modification requires estimation of the (t_k) (e.g. Ter Braak & Van Dam, 1989), but no new algorithm. It is sufficient to transform the species data to (y_{ik}/t_k^2). The same modification is open for use in WA-PLS.

Perhaps in hindsight it is not surprising that the algorithm for WA-PLS is as similar to the two-way weighted averaging algorithm of correspondence analysis (CA) (Ter Braak, 1987: Table 5.2) as the PLS-algorithm is to the two-way summation algorithm of PCA. The comparison justifies the assertion that WA-PLS relates to CA as PLS does to PCA. As such WA-PLS is the natural extension of transfer function methods based on CA regression (Roux, 1979; Rousseau, 1991), canonical CA of environmental classes (Gasse & Tekaia, 1983; Roux *et al.*, 1991), as well as those based on WA.

Acknowledgements

We acknowledge H. van der Voet (GLW-DLO, Wageningen) for discussions that led to WA-PLS, H. J. B. Birks (Botanical Institute, Bergen) for stimulus and continued interest in improving reconstruction methods, H. J. B. Birks, F. Berge (Botanical Institute, Bergen) and J. F. Boyle (Department of Geography, Liverpool) for allowing us to use their unpublished Bergen data set and P. W. Goedhart, M. Stapel, H. J. B. Birks, H. van Dam and P. Legendre for comments on the manuscript. Part of the work reported here was conducted at the Botanical Institute, University of Bergen while S. J. was in receipt of a Royal Society Fellowship, and he would like to thank H. J. B. Birks and other members of the institute for their kind hospitality and friendship during his visit.

Appendix

In this Appendix it is shown that the amended weighted averaging method is equivalent with a PLS on transformed data using the first component only. The amendment is that Part 3 of WA uses a weighted regression with weights proportional to (y_{i+}) rather than an unweighted regression. The data transformation consists of a transformation of the environmental variable and of the species data in Step 0 of the PLS-algorithm:

$$\tilde{x}_i = y_{i+}^{1/2} x_i \text{ and } \tilde{y}_{ik} = y_{ik}(y_{i+}y_{+k})^{-1/2}.$$

The proof is by elementary algebra, as follows. Step 1 sets $r_i = \tilde{x}_i = y_{i+}^{1/2} x_i$. By insertion in Step 2 of PLS,

$$b_k = \Sigma_i \tilde{y}_{ik} r_i = \Sigma_i y_{ik}(y_{i+}y_{+k})^{-1/2} y_{i+}^{1/2} x_i$$
$$= y_{+k}^{-1/2} \Sigma_i y_{ik} x_i = y_{+k}^{1/2} u_k^*,$$

which shows the equivalence of Step 2 and WA regression. Step 3 gives

$$\text{new } r_i = \Sigma_k \tilde{y}_{ik} b_k = \Sigma_k y_{ik}(y_{i+}y_{+k})^{-1/2} y_{+k}^{1/2} u_k^*$$
$$= y_{i+}^{-1/2} \Sigma_k y_{ik} u_k^* = y_{i+}^{1/2} x_i^*,$$

which shows the equivalence with WA calibration. Step 4 is skipped for the first component. Step 5 is a rescaling that, in this simple case, can more easily be taken care of by the regression of Step 7. Then, Step 7 is a regression of $y_{i+}^{1/2} x_i$ on $y_{i+}^{1/2} x_i^*$. This is equivalent to a deshrinking regression of x_i on x_i^* with weights y_{i+}. This concludes the proof.

References

Battarbee, R. W. & D. F. Charles, 1987. The use of diatom assemblages in lake sediments as a means of assessing the timing, trends, and causes of lake acidification. Progr. Phys. Geogr. 11: 552–580.

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson & C. J. F. Ter Braak, 1990a. Diatoms and pH reconstruction. Phil. Trans. r. Soc. Lond. B 327: 263–278.

Birks, H. J. B., S. Juggins & J. M. Line, 1990b. Lake surface-water chemistry reconstructions from palaeolimnological data. In B. J. Mason (ed.), The Surface Waters Acidification Programme. Cambridge University Press, Cambridge: 303–313.

Brown, G. H., 1979. An optimization criterion for linear inverse estimation. Technometrics 21: 575–579.

Cleveland, W. S., 1979. Robust locally-weighted regression and smoothing scatterplots. J. am. Statist. Assoc. 74: 829–836.

COHMAP Members, 1988. Climatic changes of the last 18000 years: observations and model simulations. Science 241: 1043–1052.

Cumming, B. F., J. P. Smol & H. J. B. Birks, 1991. The relationship between sedimentary chrysophyte scales (Chrysophyceae and Synurophyceae) and limnological characteristics in 25 Norwegian lakes. Nord. J. Bot. 11: 231–241.

Dixit, S. S., A. S. Dixit & J. P. Smol, 1991. Multivariable environmental inferences based on diatom assemblages from Sudbury (Canada) lakes. Freshwat. Biol. 26: 251–266.

Fritz, S. C., S. Juggins, R. W. Battarbee & D. R. Engstrom, 1991. Reconstruction of past changes in salinity and climate using a diatom-based transfer function. Nature 352: 706–708.

Gasse, F. & F. Tekaia, 1983. Transfer functions for estimating paleoecological conditions (pH) from East African diatoms. In J. Merilăinen, P. Huttunen & R. W. Battarbee (eds), Palaeolimnology. Development in Hydrobiology 15. Dr W. Junk Publishers, The Hague: 85–90. Reprinted from Hydrobiologia 103.

Guiot, J., 1990. Methodology of the last climatic cycle reconstruction in France from pollen data. Palaeogeogr. Palaeoclimatol. Palaeoecol. 80: 49–69.

Hall, R. I. & J. P. Smol, 1992. A weighted-averaging regression and calibration model for inferring total phosphorus concentration from diatoms in British Columbia (Canada) lakes. Freshwat. Biol. 27: 417–434.

Hastie, T. & R. Tibshirani, 1990. Generalized Additive Models. Chapman and Hall, London.

Helland, I. S., 1988. On the structure of partial least squares regression. Commun. Statist.-Simula. 17: 581–607.

Hill, M. O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54: 427–432.

Hill, M. O., 1979. DECORANA – A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Ecology and Systematics. Cornell University, Ithaca, New York, 55 pp.

Hill, M. O. & H. G. Gauch, 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42: 47–58.

Howe, S. & Webb, T. III, 1983. Calibrating pollen data in climatic terms: improving the methods. Quat. Sci. Rev. 2: 17–51.

Huntley, B. & I. C. Prentice, 1988. July temperatures in Europe from pollen data, 6000 years before present. Science 241: 687–690.

Juggins, S., 1992. Diatoms in the Thames estuary, England: Ecology, palaeoecology, and salinity transfer function. Bibl. diatomol. 25: 1–216.

Juggins, S. & C. J. F. Ter Braak, 1992. CALIBRATE – a program for species-environment calibration by [weighted-averaging] partial least squares regression. Unpublished computer program, Environmental Change Research Centre, University College London, 20 pp.

Line, J. M. & H. J. B. Birks, 1990. WACALIB version 2.1 – a computer program to reconstruct environmental variables from fossil assemblages by weighted averaging. J. Paleolimnol. 3: 170–173.

Lorber, A., L. E. Wangen & B. R. Kowalski, 1987. A theoretical foundation for the PLS algorithm. J. Chemometr. 1: 19–31.

Martens, H. & T. Naes, 1989. Multivariate calibration. Wiley, Chichester, 419 pp.

Minchin, P. R., 1987. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio 71: 145–156.

Naes, T., C. Irgens & H. Martens, 1986. Comparison of linear statistical methods for calibration for NIR instruments. Appl. Statist. 35: 195–206.

Oksanen, J., E. Laara, P. Huttunen & J. Merilainen, 1988. Estimation of pH optima and tolerances of diatoms in lake sediments by the methods of weighted averaging, least squares and maximum likelihood, and their use for the prediction of lake acidity. J. Paleolimnol. 1: 39–49.

Overpeck, J. T., T. Webb III & I. C. Prentice, 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. Quat. Res. 23: 87–108.

Prentice, I. C., P. J. Bartlein & T. Webb III, 1991. Vegetation and climate change in eastern North America since the last glacial maximum. Ecology 72: 2038–2056.

Rousseau, D. D., 1991. Climatic transfer function from Quaternary molluscs in European loess deposits. Quat. Res. 36: 195–209.

Roux, M., 1979. Estimation des paléoclimats d'après l'écologie des foraminifères. Cah. Anal. Données 4: 61–79.

Roux, M., S. Servant-Vildary & M. Servant, 1991. Inferred ionic composition and salinity of a Bolivian Quaternary lake, as estimated from fossil diatoms in the sediments. Hydrobiologia 210: 3–18.

Shelford, V. E., 1911. Ecological succession: stream fishes and the method of physiographic analysis. Biol. Bull. (Woods Hole) 21: 9–34.

Stevenson A. C., S. Juggins, H. J. B. Birks, D. S. Anderson, N. J. Anderson, R. W. Battarbee, F. Berge, R. B. Davis, R. J. Flower, E. Y. Haworth, V. I. Jones, J. C. Kingston, A. M. Kreiser, J. M. Line, M. A. R. Munro & I. Renberg, 1991. The surface waters acidification project Palaeolimnology programme: modern diatom/lake-water chemistry data-set. ENSIS, London, 86 pp.

Stone, M. & R. J. Brooks, 1990. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. J. R. Statist. Soc. B 52: 237–269.

Ter Braak, C. J. F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167–1179.

Ter Braak, C. J. F., 1987. Ordination. In R. H. G. Jongman,

C. J. F. Ter Braak & O. F. R. Van Tongeren (eds), Data analysis in community and landscape ecology. Pudoc, Wageningen: 91–173.

Ter Braak C. J. F., 1988. CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). Report LWA-88-02. Agricultural Mathematics Group, Wageningen, 95 pp.

Ter Braak C. J. F., 1990. Update notes: CANOCO version 3.1. Microcomputer Power, Ithaca, NY, 35 pp.

Ter Braak, C. J. F. & L. G. Barendregt, 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. Math. Bio. 78: 57–72.

Ter Braak, C. J. F. & C. W. N. Looman, 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65: 3–11.

Ter Braak, C. J. F. & I. C. Prentice, 1988. A theory of gradient analysis. Adv. Ecol. Res. 18: 271–317.

Ter Braak, C. J. F. & H. van Dam, 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178: 209–223.

Ter Braak, C. J. F., S. Juggins, H. J. B. Birks & H. van der Voet, 1993. Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. Chapter 25 in G. P. Patil & C. R. Rao (eds), Multivariate Environmental Statistics. North-Holland, Amsterdam.

Walker, I. R., R. J. Mott & J. P. Smol, 1991. Allerød-Younger Dryas lake temperatures from midge fossils in Atlantic Canada. Science 253: 1010–1012.

Whittaker, R. H., 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26: 1–80.

Wold, S., 1992. Nonlinear partial least squares modelling. II Spline inner relation. Chemometrics and Intelligent Laboratory Systems 14: 71–84.

Wold, S., A. Ruhe, H. Wold & W. J. Dunn III, 1984. The collinearity problem in linear regression: the partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5: 735–743.