Available online at www.sciencedirect.com
SCIENCE DIRECT
Ecological Modelling 187 (2005) 524–536
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Constrained ordination analysis with flexible response functions

Mu Zhu [a,*], Trevor J. Hastie [b], Guenther Walther [b]

[a] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ont., Canada N2L 3G1*
[b] *Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA*

Received 17 April 2004; received in revised form 25 January 2005; accepted 27 January 2005
Available online 11 March 2005

Abstract

Canonical correspondence analysis (CCA) is perhaps the most popular multivariate technique used by environmental ecologists for constrained ordination; it is an approximation to the maximum likelihood solution of the Gaussian response model. In this article, we look at the constrained ordination problem from a slightly different point of view and argue that it is this particular point of view that CCA implicitly adopts. This gives us additional insights into the nature of CCA. We then exploit the new perspective to generalize the Gaussian response model to incorporate more flexible response functions. A real example is presented to illustrate the use of the more flexible model.

Keywords: Canonical correspondence analysis; Density estimation; Discriminant analysis; Likelihood ratio; Multinomial and Poisson likelihood; Random weights; Vegetation succession study

1. Introduction

Constrained ordination analysis has been widely used by environmental ecologists to study how different species respond to changes in environmental conditions. The Gaussian response model can be generally regarded as an ideal model for such problems (ter Braak, 1996; Yee, 2004); its parameters are typically estimated with canonical correspondence analysis (CCA), a multivariate technique first proposed by ter Braak (1986). Two facts about CCA are particularly relevant for our article. Firstly, as a variation of correspondence analysis, CCA is only an approximation to the maximum likelihood solution of the Gaussian response model (e.g., ter Braak, 1985; Yee, 2004). Secondly, from an algebraic point of view, CCA is known (e.g., Takane et al., 1991) to be equivalent to a number of other multivariate techniques such as optimal scoring (also known as dual scaling) and linear discriminant analysis (LDA), a popular technique for classification. The connection

* Corresponding author. Tel.: +1 519 888 4567; fax: +1 519 746 1875.
E-mail addresses: m3zhu@uwaterloo.ca (M. Zhu), hastie@stanford.edu (T.J. Hastie), gwalther@stanford.edu (G. Walther).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.01.049

between CCA and LDA can also be found elsewhere (e.g., ter Braak and Verdonschot, 1995); early ideas for analyzing ecological data with techniques of the LDA type can be found in, for example, Green (1971, 1974).

In this article, we first give a brief review of the constrained ordination problem, the classic Gaussian response model as well as the equivalence between the CCA and LDA algorithms (Section 2). Next, we present a probabilistic model for LDA in the context of ecological ordination (Section 3). Due to the equivalence between LDA and CCA, this model sheds important light on the nature of the widely-used CCA algorithm. We then generalize the LDA model to perform constrained ordination analysis with flexible response functions (Section 4). Finally, we give an illustrative example (Section 5).

2. Constrained ordination analysis

In a typical study, one has a data matrix $\mathbf{Y} = \{y_{ik}\}$, whose element y_{ik} records the abundance of species k at site i, and a covariate matrix $\mathbf{X} = \{x_{im}\}$, whose element x_{im} is a measurement of a particular environmental factor at site i, e.g., average temperature or the concentration of a certain chemical contained in the soil. We will use the vector notation $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ for a total of d environmental variables at site i.

$\mathbf{Y}$:

	Species		
	1	...	K
Site 1	y_{11}	...	y_{1K}
Site 2	y_{21}	...	y_{2K}
⋮	⋮	⋱	⋮
Site n	y_{n1}	...	y_{nK}

$\mathbf{X}$:

	Measurements		
	1	...	d
Site 1	x_{11}	...	x_{1d}
Site 2	x_{21}	...	x_{2d}
⋮	⋮	⋱	⋮
Site n	x_{n1}	...	x_{nd}

2.1. The general statistical model

Often the abundance measure y_{ik} is simply the count of species k at site i. Therefore y_{ik} is often assumed to follow an independent Poisson distribution; the corresponding rate parameter λ_{ik} is modeled as

$$\lambda_{ik} = f_k(\boldsymbol{x}_i).$$

The function f_k is called the *response function* for species k; it is a function of how species k responds to different environmental conditions. If the conditions at site i are favorable for species k, $\lambda_{ik} = f_k(\boldsymbol{x}_i)$ will be large, i.e., the species will be abundant at that site. The Poisson model gives rise to an explicit likelihood function for the data:

$$\begin{aligned}\text{likelihood} &= \prod_{k=1}^{K}\prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda_{ik}}\lambda_{ik}^{y_{ik}}}{y_{ik}!} \\ &= \prod_{k=1}^{K}\prod_{i=1}^{n} \frac{\mathrm{e}^{-f_k(\boldsymbol{x}_i)}(f_k(\boldsymbol{x}_i))^{y_{ik}}}{y_{ik}!}.\end{aligned}$$

Apart from a constant not depending on the response functions f_k ($k = 1, 2, \ldots, K$), the corresponding log-likelihood function is simply

$$\text{log-likelihood} = \sum_{k=1}^{K}\sum_{i=1}^{n} -f_k(\boldsymbol{x}_i) + y_{ik}\log\ f_k(\boldsymbol{x}_i). \tag{1}$$

Conventionally, the function f_k is often defined on a low-dimensional subspace, e.g., along a preferred direction $\boldsymbol{\alpha} \in \mathbb{R}^d$. In this case, $f_k(\boldsymbol{x}_i)$ is actually a ridge function $f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)$. The direction $\boldsymbol{\alpha}$ is called the *environmental gradient*. We shall write $z_i = \boldsymbol{\alpha}^T\boldsymbol{x}_i$ and use the following notations interchangeably: $f_k(\boldsymbol{x}_i)$, $f_k(z_i)$ and $f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)$.

2.2. The Gaussian response model

The so-called Gaussian response model assumes that the response function f_k has the form of a Gaussian density function along the environmental gradient (Fig. 1), i.e., f_k as a function of $z = \boldsymbol{\alpha}^T\boldsymbol{x}$ is log-quadratic:

$$\log(f_k(z_i)) = a_k - \frac{(z_i - u_k)^2}{2t_k^2} \quad \text{or}$$

$$\log(f_k(\boldsymbol{x}_i)) = a_k - \frac{(\boldsymbol{\alpha}^T\boldsymbol{x}_i - u_k)^2}{2t_k^2}. \tag{2}$$

The interpretation of the parameters are as follows:

a_k: The *maximum* of species k on the log-scale—e^{a_k} is the expected count of species k at its optimal environment.

$\boldsymbol{\alpha}$: The *environmental gradient*—it is a vector in $\mathbb{R}^d$ that assigns each site i an environmental *score*, z_i, according to its environment conditions $\boldsymbol{x}_i$. According to ecological theory (e.g., MacArthur and Levins, 1967), over the course of evolution, species tend

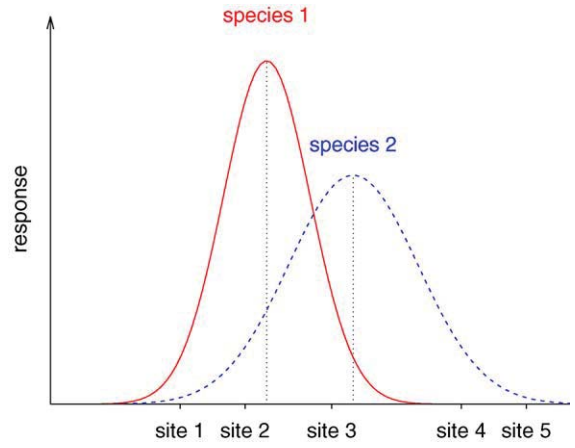

Fig. 1. Illustration of the Gaussian response curve. This graph depicts a situation where the environmental conditions at site 2 are close to being optimal for species 1, whereas site 3 is close to being optimal for species 2.

to develop maximally separated niches over limiting environmental resources. The optimal $\boldsymbol{\alpha}$, therefore, should be a direction in which the species' response functions are the most different.

u_k: The *optimum* of species k—species k is most abundant (i.e., λ_{ik} reaches its maximum) at site i if site i receives an environmental score equal to u_k, i.e., if $z_i = \boldsymbol{\alpha}^T \boldsymbol{x}_i = u_k$.

t_k: The *tolerance* of species k—if t_k is large, then the response function is relatively flat, indicating that the species can survive well under a great variety of environmental conditions, i.e., it has high tolerance; likewise, if t_k is small, then the response function drops sharply as z_i moves away from the optimal point u_k, indicating that the species has low tolerance for different environmental conditions.

2.3. The CCA algorithm

The parameters in the Gaussian response model can be estimated by directly maximizing the log-likelihood function (1), something that, surprisingly, has been devoid in the literature until quite recently (see Yee, 2004). Traditionally, canonical correspondence analysis (CCA) has been the *de facto* algorithm for estimating the parameters under the extra assumption that the species have an equal tolerance parameter, i.e., $t_k = t$ for all k. Notice that the scales of z_i, u_k and t_k are arbitrary; this can be resolved, for example, by requiring $\sum_{k=1}^{K} t_k^2/K = 1$ (ter Braak, 1985). The equal tolerance assumption, therefore, can be explicitly stated as $t_k = t = 1$. It was made clear in the original CCA paper (ter Braak, 1986) that one can first estimate the scores $\boldsymbol{x}_i$ ($i = 1, 2, \ldots, n$) and then simply regress z_i onto $\boldsymbol{x}_i$ to estimate the environmental gradient $\boldsymbol{\alpha}$. By differentiating the log-likelihood function (1) with f_k being Gaussian (2) and $t_k = 1$ for every k, we can obtain the normal equations for estimating a_k, z_i and u_k:

$$a_k = \log \frac{\sum_{i=1}^{n} y_{ik}}{\sum_{i=1}^{n} \exp(-(z_i - u_k)^2/2)}, \tag{3}$$

$$z_i = \frac{\sum_{k=1}^{K} y_{ik} u_k}{\sum_{k=1}^{K} y_{ik}} - \frac{\sum_{k=1}^{K} (z_i - u_k) f_k(z_i)}{\sum_{k=1}^{K} y_{ik}}, \tag{4}$$

and

$$u_k = \frac{\sum_{i=1}^{n} y_{ik} z_i}{\sum_{i=1}^{n} y_{ik}} - \frac{\sum_{i=1}^{n} (z_i - u_k) f_k(z_i)}{\sum_{i=1}^{n} y_{ik}}. \tag{5}$$

Under certain mild conditions, it can be argued (ter Braak, 1985) that both Eqs. (4) and (5) are dominated by the leading term[1] so one can simply approximate them with the well-known reciprocal averaging equations of correspondence analysis:

$$z_i = \frac{\sum_{k=1}^{K} y_{ik} u_k}{\sum_{k=1}^{K} y_{ik}} \quad \text{and} \quad u_k = \frac{\sum_{i=1}^{n} y_{ik} z_i}{\sum_{i=1}^{n} y_{ik}}.$$

In matrix form, these equations can be written as follows:

$$\mathbf{z} \propto \mathbf{R}^{-1}\mathbf{Y}\mathbf{u} \quad \text{and} \quad \mathbf{u} \propto \mathbf{C}^{-1}\mathbf{Y}^T\mathbf{z},$$

where $\mathbf{R} = \mathrm{diag}\{y_{i\cdot}\}$, $\mathbf{C} = \mathrm{diag}\{y_{\cdot k}\}$, $y_{i\cdot} = y_{i1} + y_{i2} + \cdots + y_{iK}$ and $y_{\cdot k} = y_{1k} + y_{2k} + \cdots + y_{nk}$. If the score

[1] Alternatively, one can easily see that the term $(z_i - u_k) f_k(z_i)$ cannot be too large because, under the Gaussian assumption, $f_k(z_i)$ can only be large if $z_i - u_k$ is close to zero whereas if $z_i - u_k$ is far from zero $f_k(z_i)$ is necessarily small.

z_i is constrained to be $z_i = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i$ (or $\mathbf{z} = \mathbf{X}\boldsymbol{\alpha}$), we obtain the main equations for CCA:

$$\mathbf{X}\boldsymbol{\alpha} \propto \mathbf{R}^{-1}\mathbf{Y}\mathbf{u} \tag{6}$$

$$\mathbf{u} \propto \mathbf{C}^{-1}\mathbf{Y}^{T}\mathbf{X}\boldsymbol{\alpha}. \tag{7}$$

Multiplying (6) by $(\mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{R}$ on both sides gives

$$\boldsymbol{\alpha} \propto (\mathbf{X}^{T}\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y}\mathbf{u}. \tag{8}$$

Clearly, estimates of $\mathbf{u}$ and $\mathbf{z}$ can easily be obtained from the above equations once $\boldsymbol{\alpha}$ is estimated. In what follows, we shall always focus on the estimation of $\boldsymbol{\alpha}$. Combining Eqs. (8) and (7), we get

$$\boldsymbol{\alpha} \propto \boldsymbol{\Gamma}_{\alpha}\boldsymbol{\alpha} \quad \text{where} \quad \boldsymbol{\Gamma}_{\alpha} \equiv (\mathbf{X}^{T}\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}, \tag{9}$$

which means $\boldsymbol{\alpha}$ can be estimated as the eigenvector of the matrix $\boldsymbol{\Gamma}_{\alpha}$.

2.4. An equivalent algorithm: LDA

Takane et al. (1991) established that CCA is mathematically equivalent to Fisher's reduced-rank linear discriminant analysis (LDA), an algorithm widely used in classification. Here, we only present a highly condensed summary of how this equivalence can be understood. The data matrices in a typical classification problem are shown below.

$\mathbf{Y}$:

	Classes		
	1	...	K
Obs 1	1	...	0
Obs 2	0	...	1
⋮	⋮	⋱	⋮
Obs n	1	...	0

$\mathbf{X}$:

	Predictors		
	1	...	d
Obs 1	x_{11}	...	x_{1d}
Obs 2	x_{21}	...	x_{2d}
⋮	⋮	⋱	⋮
Obs n	x_{n1}	...	x_{nd}

Here, we see that the matrix $\mathbf{X}$ is the same as before whereas the matrix $\mathbf{Y}$ is slightly different; it is a binary matrix of class indicators. The LDA algorithm finds the most important direction $\boldsymbol{\alpha} \in \mathbb{R}^d$ for classification by maximizing the Rayleigh quotient

$$\frac{\boldsymbol{\alpha}^{T}\mathbf{B}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{T}\mathbf{W}\boldsymbol{\alpha}}, \tag{10}$$

where

$$\mathbf{B} \propto \sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})^{T} \quad \text{and}$$

$$\mathbf{W} \propto \sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^{T} \tag{11}$$

are the *between*-class and *within*-class sum-of-squares matrices; the maximizing solution is well-known to be the leading eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ (see, e.g., Mardia et al., 1979). Hastie et al. (1995, Section 3) showed that, when $\mathbf{Y}$ is an indicator matrix (the standard situation for classification), finding the eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{B}$ is equivalent to finding the eigenvectors of the matrix

$$\boldsymbol{\Psi}_{\alpha} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^{T}\mathbf{X},$$

which can be easily seen to resemble the matrix $\boldsymbol{\Gamma}_{\alpha}$ from CCA (see Section 2.3).

To see that they are exactly the same, note that, in the classification problem, there is only one observation for each row of $\mathbf{X}$, whereas in the ecological ordination problem, each row of $\mathbf{X}$ corresponds to a different site and there are multiple species at each site. That's why each row of $\mathbf{X}$ must be properly weighted by the total species count $y_{i\cdot} = y_{i1} + y_{i2} + \cdots + y_{iK}$ and the quantity $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ must be replaced with $\mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{X}$. Conversely, starting from $\boldsymbol{\Gamma}_{\alpha}$ it is also easy to see that in the classification case when each row of $\mathbf{X}$ belongs to just one of the K classes, the weight matrix $\mathbf{R} = \mathbf{I}$ is the identity matrix, so $\boldsymbol{\Gamma}_{\alpha} = \boldsymbol{\Psi}_{\alpha}$. See Takane et al. (1991) for a more formal argument.

3. A probabilistic model for LDA

We now present a probabilistic model for LDA in the context of constrained Gaussian ecological ordination.

3.1. Data collection

Recall that $\boldsymbol{x} \in \mathbb{R}^d$ is a vector of d environmental measurements. The data contained in the matrices $\mathbf{X}$ and $\mathbf{Y}$ are typically collected in the following way: First, n geographical sites are selected and various environmental measurements are taken at each site, giving rise to a sample $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. Along the environmental gradient $\boldsymbol{\alpha}$, we get a univariate sample, which we shall write as $S(\boldsymbol{\alpha}) = \{z_1, z_2, \ldots, z_n\}$, where $z_i = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i$. We can think of each z_i as having been sampled from a certain distribution $g(z)$ (e.g., ter Braak, 1996, Chapter 1):

$$z_1, z_2, \ldots, z_n \sim g(z). \tag{12}$$

Once the sites have been determined, the species are counted at each site; y_{ik} is the count of species k at site i and is assumed to follow a Poisson distribution with rate $\lambda_{ik} = f_k(z_i)$. This means we have

$$E(y_{ik}|z_i) = f_k(z_i). \tag{13}$$

3.2. A weighted sample model

For each species k, we now think of y_{ik}—its abundance measure at site i—as the weight it puts on the observation $\boldsymbol{x}_i$. As a result, we obtain a weighted sample $S_k = \{(\boldsymbol{x}_i, w_{ik}); i = 1, 2, \ldots, n\}$; the weight on each $\boldsymbol{x}_i$ is $w_{ik} = y_{ik}$. More specifically, we can think of S_k to have been generated as follows: each observation $\boldsymbol{x}_i$ from the sample S is replicated y_{ik} times to form the weighted sample S_k.

Now let $p_k(\boldsymbol{x})$ denote the density function of the weighted sample S_k; notice that $p_k(\boldsymbol{x})$ can be estimated directly from S_k. We shall use the notation $p_k^{(\alpha)}$ to denote the marginal distribution of p_k along the environmental gradient $\boldsymbol{\alpha}$, i.e., $p_k^{(\alpha)}$ is the distribution of the univariate sample $S_k(\boldsymbol{\alpha}) = \{(z_i, w_{ik}); i = 1, 2, \ldots, n\}$. We wish to characterize $p_k^{(\alpha)}$ and make a connection between $p_k^{(\alpha)}$ and the response function f_k.

Because of the weights, it is not clear how to determine this distribution, or even how this distribution is defined in the first place. We will use the following trick: while it is hard to determine this distribution directly, we can nevertheless still construct the empirical distribution of the weighted sample $S_k(\boldsymbol{\alpha})$. Suppose we can prove that for every interval I the empirical measure of I converges almost surely to $\int_I p(z)\mathrm{d}z$ for some density $p(z)$. Then, applying the law of large numbers backwards and using the fact that a distribution is uniquely determined by the values it gives to intervals, we can conclude that the sample must have been generated by a distribution with density $p(z)$. Note that the final statement is not asymptotic; the detour via the law of large numbers is just a convenient way to characterize the distribution of the weighted sample $S_k(\boldsymbol{\alpha})$.

Theorem 1. *Suppose $z_1, z_2, \ldots, z_n$ are an i.i.d. sample from $g(z)$ and that, given z_i, w_i is a non-negative random variable with expected value $f(z_i)$, i.e., $E(w_i|z_i) = f(z_i)$, where f is a non-negative, measurable function that satisfies $0 < \int f(z)g(z)\mathrm{d}z < \infty$. Furthermore, suppose that the w_is are independent. Then for every interval I the empirical measure of I with respect to the weighted sample $\{(z_i, w_i); i = 1, 2, \ldots, n\}$ converges almost surely to $\int_I p(z)\mathrm{d}z$ with $p(z) \propto g(z)f(z)$.*

Since the weight $w_{ik} = y_{ik}$, it follows from Eq. (13) and Theorem 1 above that

$$p_k^{(\alpha)}(z) \propto g(z) f_k(z) \tag{14}$$

for every $k = 1, 2, \ldots, K$. This implies that if we choose to model the function $p_k^{(\alpha)}$ instead of f_k and estimate it directly from the data $S_k(\boldsymbol{\alpha})$, we can recover the response function $f_k(z)$ up to a function $g(z)$ that is common for all k. A proof of a more general version of Theorem 1 that allows for multivariate z_i is given in Appendix A.

3.3. The Gaussian response function

The classic Gaussian response model assumes that the response function f_k is Gaussian (Section 2.2). If we choose to model $p_k^{(\alpha)}(z)$ instead, it is easy to see that we can make a direct connection to the Gaussian response model if we make the following two assumptions:

A1: The distribution $p_k(\boldsymbol{x})$ is a (multivariate) Gaussian with mean vector $\boldsymbol{\mu}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}_k$; and

A2: $g(z)$ is a uniform distribution on a suitable range of z.

The first assumption immediately implies that the marginal distribution $p_k^{(\alpha)}$ is a univariate Gaussian with mean $\boldsymbol{\alpha}^T\boldsymbol{\mu}_k$ and variance $\boldsymbol{\alpha}^T\boldsymbol{\Sigma}_k\boldsymbol{\alpha}$. Because of (14), the second assumption then implies we must have

$$u_k = \boldsymbol{\alpha}^T\boldsymbol{\mu}_k \quad \text{and} \quad t_k^2 = \boldsymbol{\alpha}^T\boldsymbol{\Sigma}_k\boldsymbol{\alpha}.$$

This means under the assumptions A1 and A2, modeling f_k as Gaussian is the same as modeling p_k as Gaussian and the location and scale parameters are the same for f_k and $p_k^{(\alpha)}$.

The assumption that $g(z)$ is uniform means the sites are sampled in such a way that different environmental conditions along the environmental gradient are equally represented. Of course, it is generally hard to guarantee such uniformity in reality, but below we will introduce a criterion for estimating the environmental gradient $\boldsymbol{\alpha}$ that is invariant to $g(z)$ (Section 4.2) so this assumption is actually not required at all.

3.4. Estimation with LDA

With each y_{ik} now regarded as a weight on $\boldsymbol{x}_i$, how should the parameters be estimated? Under the assumption A1, the parameters of p_k, namely $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, can be estimated directly from the sample S_k by (weighted) maximum likelihood. For any given $\boldsymbol{\alpha}$, we've already seen that $u_k = \boldsymbol{\alpha}^T\boldsymbol{\mu}_k$ and $t_k^2 = \boldsymbol{\alpha}^T\boldsymbol{\Sigma}_k\boldsymbol{\alpha}$, so the main statistical problem is the estimation of $\boldsymbol{\alpha}$, the environmental gradient.

Recall that the optimal $\boldsymbol{\alpha}$ should be a direction in which the species' response functions are the most different (Section 2.2). This means we can choose $\boldsymbol{\alpha}$ to make the response functions f_k ($k = 1, 2, \ldots, K$) as different as possible. Eq. (14) implies that this is the same as choosing $\boldsymbol{\alpha}$ to make $p_k^{(\alpha)}$ ($k = 1, 2, \ldots, K$) as different as possible since the function g is common for all k.

As in Section 2.3, we first assume $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for every k; note this is the same as saying that the species have the same tolerance parameter, $t_k^2 = t^2 = \boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha}$ for any $\boldsymbol{\alpha}$. Since a Gaussian distribution is completely characterized by its mean and its variance, to make a number of Gaussian distributions as different as possible under the assumption $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, it suffices to focus on the separation of the means. A well-known criterion for measuring the difference among $p_k^{(\alpha)}(k = 1, 2, \ldots, k)$ under such circumstances is given by Fisher (1936):

$$\frac{\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha}}, \tag{15}$$

where

$$\boldsymbol{\Delta} = \sum_{k=1}^{k} \pi_k(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^T$$

is the *between-species* covariance matrix; $\bar{\boldsymbol{\mu}}$ is the overall mean of $\boldsymbol{x}$ across all species and π_k, the probability or relative frequency of species k such that $\pi_1 + \pi_2 + \ldots + \pi_K = 1$. In practice, both $\boldsymbol{\Delta}$ and $\boldsymbol{\Sigma}$ must be estimated empirically from data. Common estimates are the sample *between*-species and *within*-species sum-of-squares matrices, $\mathbf{B}$ and $\mathbf{W}$, as given in (11). In other words, the environmental gradient $\boldsymbol{\alpha}$ can be estimated by the LDA algorithm.

In addition, Hastie and Tibshirani (1996) showed that $u_k = \boldsymbol{\alpha}^T\boldsymbol{\mu}_k$ is actually the maximum likelihood estimate for the mean of p_k under the rank constraint that the means of p_k ($k = 1, 2, \ldots, K$) lie in a rank-one subspace; the rank-one subspace is spanned by the LDA solution $\boldsymbol{\alpha}$. Therefore, in the context of this weighted sample model, LDA can be seen to provide a direct maximum likelihood solution to the constrained ordination problem if the response functions are assumed to be Gaussian and the tolerance parameters t_k are assumed to be equal among all species.

3.5. Implication

Let us summarize our discussion so far. While CCA is an approximation to the maximum likelihood solution for the classic Gaussian response model, LDA actually gives an exact maximum likelihood solution for the alternative weighted sample model. Because CCA and LDA are mathematically equivalent and since CCA is generally accepted as the standard method for solving the constrained ordination problem, the research community has, in effect, been implicitly relying on the weighted sample model we presented above for quite a long time. An important feature of this alternative model is that the environmental gradient is explicitly defined as the direction where the species' response functions differ the most. Below, we focus on how this alternative weighted sample model can be further exploited to obtain an easy generalization of the constrained ordination problem to allow the use of more flexible response functions.

4. Generalization

There have been growing interests in the ecological community to conduct constrained ordination analysis using more flexible models (e.g., Johnson and Altman, 1999; Heegaard, 2002). The equal tolerance assumption, for example, is restrictive, although it does make the ordination diagrams easier to read. For some problems, the assumption that the response function is Gaussian may also not be appropriate. For example, a species does not always respond symmetrically to the extreme conditions on either side of its optimum; some species may even exhibit multimodal responses, perhaps because sub-species (either identified or unidentified) have adapted to different environmental conditions and occupied different niches.

4.1. A likelihood ratio criterion

For maximal flexibility, we would like to work with very general response functions f_k ($k = 1, 2, \ldots, K$). In view of our weighted sample model and the relation (14), this means we want to work with very general probability distribution or density functions p_k ($k = 1, 2, \ldots, K$). Zhu and Hastie (2003) provide the general principles needed in order to estimate the environmental gradient $\boldsymbol{\alpha}$ when each p_k is an arbitrary probability distribution function; the principles exploit the likelihood ratio interpretation of Fisher's LDA criterion (15). In particular, for any given direction $\boldsymbol{\alpha}$, a likelihood-ratio criterion is used to measure the difference among $p_k^{(\alpha)} (k = 1, 2, \ldots, k)$:

$$\mathrm{LR}(\boldsymbol{\alpha}) = \log \frac{\prod_{i=1}^{n} \prod_{k=1}^{k} \left(p_k^{(\alpha)}(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\right)^{y_{ik}}}{\prod_{i=1}^{n} \prod_{k=1}^{k} \left(p^{(\alpha)}(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\right)^{y_{ik}}}. \tag{16}$$

Here $p^{(\alpha)}$ is the distribution of $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}$ ignoring the species labels. It is easy to see that criterion (16) will be large when there is an advantage for having a separate model for each species and small otherwise. Hence, the environmental gradient $\boldsymbol{\alpha}$ can simply be estimated as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^d}{\arg\max} \quad \mathrm{LR}(\boldsymbol{\alpha}).$$

In practice, the computation of $\mathrm{LR}(\boldsymbol{\alpha})$ requires estimating the functions $p_k^{(\alpha)}$ and $p^{(\alpha)}$ for any given $\boldsymbol{\alpha}$; $p_k^{(\alpha)}$ can be estimated using the sample $S_k(\boldsymbol{\alpha})$ and $p^{(\alpha)}$ can be estimated using a pooled sample, in this case $\{(z_i, w_{i\cdot}); i = 1, 2, \ldots, n\}$ with $w_{i\cdot} = w_{i1} + w_{i2} + \cdots + w_{iK}$.

4.2. An invariance property

An important observation here is that the sampling distribution g does not affect the estimation of $\boldsymbol{\alpha}$. This is because, by (14), $p_k^{(\alpha)}(z) \propto g(z) f_k(z)$ and likewise $p^{(\alpha)}(z) \propto g(z) f(z)$ so $g(z)$ appears in both the numerator and the denominator of (16), which implies the criterion $\mathrm{LR}(\boldsymbol{\alpha})$ is *invariant* with respect to the sampling distribution g. In other words, the assumption A2 that $g(z)$ is uniform is not required for estimating the environmental gradient.

This is not to say that the assumption A2 is entirely irrelevant. If $g(z)$ is uniform, we obtain a further simplification that allows us to estimate the response function $f_k(z)$ directly with $p_k^{(\alpha)}(z)$; the (unknown) proportionality constant is not important here since we are interested mostly in the shape of f_k. Likewise, in the classic Gaussian response model, assuming $g(z)$ to be uniform allows us to estimate the optimum of species k directly with $u_k = \boldsymbol{\alpha}^T \boldsymbol{\mu}_k$ (see Section 3.4).

If $g(z)$ is not uniform, however, we can still estimate it relatively easily from the (unweighted) sample $S(\boldsymbol{\alpha}) = \{z_1, z_2, \ldots, z_n\}$ and use it to estimate each response function f_k up to an inconsequential scaling factor, i.e., $f_k(z) \propto p_k^{(\alpha)}(z)/g(z)$.

4.3. Connection with LDA

We have already seen in Section 3.4 that if p_k is $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, the environmental gradient $\boldsymbol{\alpha}$ can be estimated by maximizing Fisher's LDA criterion (15). The LR criterion (16) presented above is, in fact, related to the LDA criterion (15). In particular, if we model p_k as a multivariate Gaussian distribution with different means but a common covariance matrix and p as a Gaussian distribution that is the same for all species $k = 1, 2, \ldots, K$, then maximizing $\mathrm{LR}(\boldsymbol{\alpha})$ is equivalent to maximizing (15) and performing LDA (Zhu and Hastie, 2003).

4.4. Connection with the Poisson likelihood

Interestingly, this approach of estimating the environmental gradient using (16) is not entirely unrelated to the one that works directly with the original Poisson log-likelihood (1). To see this, note that $p^{(\alpha)}(z)$, the distribution of z ignoring the species labels, is just a mixture distribution

$$p^{(\alpha)}(z) = \sum_{k=1}^{K} \pi_k p_k^{(\alpha)}(z), \tag{17}$$

where π_k is the relative frequency of species k. Using (17), the argument contained in Appendix B shows that

$$\frac{p_k^{(\alpha)}(z)}{p^{(\alpha)}(z)} = \frac{f_k(z)}{\pi_k \sum_{j=1}^{K} f_j(z)}. \tag{18}$$

Plugging (18) back into (16), we get

$$\mathrm{LR}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\ \log\left(\frac{f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}{\sum_{j=1}^{K} f_j(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\ \log\ \pi_k. \tag{19}$$

On the other hand, if $y_{ik} \sim$ Poisson (λ_{ik}) and the y_{ik}s are independent, then it is well-known (see, e.g., Ross, 1997) that conditional on the sum $y_{i1} + y_{i2} + \cdots + y_{iK} = n_i$, each y_{ik} has a multinomial distribution with probability p_{ik}, where

$$p_{ik} = \frac{\lambda_{ik}}{\sum_{j=1}^{K}\lambda_{ij}}.$$

Apart from a constant not depending on the parameters, the corresponding log-likelihood for this (conditional) multinomial distribution is nothing but

$$\begin{aligned}\text{log-likelihood} &= \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\ \log\left(\frac{\lambda_{ik}}{\sum_{j=1}^{K}\lambda_{ij}}\right)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\ \log\left(\frac{f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}{\sum_{j=1}^{K} f_j(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}\right).\end{aligned} \tag{20}$$

since $\lambda_{ik} = f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)$ (see Section 2.1). Clearly, (19) is the same as (20) apart from an additive constant that does not depend on $\boldsymbol{\alpha}$.

The argument above shows that estimating $\boldsymbol{\alpha}$ by maximizing (16) is equivalent to maximizing the conditional multinomial log-likelihood (20) for independent Poisson observations. This gives us altogether three estimation methods to choose from: maximizing the unconditional Poisson log-likelihood (1), maximizing the conditional multinomial log-likelihood (20), and maximizing the log-likelihood-ratio criterion (16).

Yee (2004) takes the first approach of maximizing the Poisson log-likelihood (1) directly by fitting a rank-reduced vector generalized linear model (Yee and Hastie, 2003), using an extra quadratic term when the response functions are Gaussian. The advantage of the second approach using the conditional multinomial log-likelihood (20) is that, by conditioning on the total counts of all species at the given sites, it encourages "competition" among the species, thereby making the biological notion of the environmental gradient more explicit. The general algorithm proposed by Yee (2004) can be used to maximize the multinomial log-likelihood (20) as well, but the computation there is more complex. Typically, multinomial models are fitted using an iteratively reweighted least-squares (IRLS) algorithm (McCullagh and Nelder, 1989). With more than two species, the response variable must be coded as a vector and a non-diagonal weight matrix is needed for each observation, hence precluding the possibility of any simplified algorithms (see Hastie et al., 2001, Section 4.4.1).

By contrast, the third approach of maximizing the log-likelihood-ratio criterion (16) is equivalent to but conceptually much simpler than maximizing the multinomial log-likelihood (20) directly. For fixed $\boldsymbol{\alpha}$, we estimate $p_k^{(\alpha)}$ separately for each species. Moreover, the mixture distribution (17) in the denominator can be estimated in the same way from a pooled sample (see Section 4.1). There is no need to form and keep track of complicated weight matrices.

4.5. Nonparametric modelling: some details

Another attractive aspect of the log-likelihood-ratio approach is that it is easy to model $p_k^{(\alpha)}$ using nonparametric techniques and therefore perform constrained ordination analysis with fully flexible response functions. In this section, we give some details of how this can be done.

In practice, we use an iterative method such as Newton–Raphson to maximize $\mathrm{LR}(\boldsymbol{\alpha})$; this means throughout the computation, we always compute $\mathrm{LR}(\boldsymbol{\alpha})$ and its derivatives at a given $\boldsymbol{\alpha}$. To do so, we only need be able to estimate the marginal probability distributions $p_k^{(\alpha)}$ $(k = 1, 2, \ldots, K)$ and their derivatives for a given $\boldsymbol{\alpha}$. This can be done using standard tools for univariate density estimation (for more details, see Zhu and Hastie, 2003).

It is often the case that more than one environmental gradient is estimated. Since both CCA and LDA amount to solving an eigenvalue problem, the solutions are simply given by the leading eigenvectors. Using the two leading eigenvectors from CCA, for example, one can make an informative biplot of the data to show the

species–environmental relationship. Although with arbitrary p_k it is no longer possible to treat this as an eigenvalue problem, it is still possible to get ordered multiple solutions as before. Note that in LDA, the second eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ can be obtained as

$$\arg\max_{\boldsymbol{\alpha}^T\mathbf{W}\boldsymbol{\alpha}_1=0} \frac{\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{W}\boldsymbol{\alpha}},$$

i.e., one simply maximizes the same criterion with an added constraint that the solution is now orthogonal (with respect to the metric $\mathbf{W}$) to the previous one; this can also be achieved by first projecting the data onto the orthogonal complement of $\boldsymbol{\alpha}_1$ and maximizing the same criterion without the constraint. Such a process can be repeated recursively to solve for all the subsequent eigenvectors. These ideas are readily applicable in the more general case, i.e., one can maximize the general criterion LR($\boldsymbol{\alpha}$) recursively, transforming the data at each step. Details can be found in Zhu and Hastie (2003).

5. Example: a vegetation succession study

As a simple illustration, we analyze a data set studied by ter Braak (1987). A total of 68 vegetation species were sampled twice at 63 different sites along a rising seashore, once in 1978 and once in 1984, resulting in a 126×68 $\mathbf{Y}$ matrix. In order to investigate whether the vegetation succession tracked the land uplift of 0.5 cm per year, ter Braak (1987) used CCA with two "environmental variables," elevation and year. That is, the matrix $\mathbf{X}$ is 126×2, and the variable along the "environmental" gradient is: $\alpha_1 \times \text{elevation} + \alpha_2 \times \text{year}$. He then compared the ratio between α_1 and α_2, which can be interpreted as changes in vegetation per year, with the known uplift of 0.5; the ratio he obtained was 0.76.

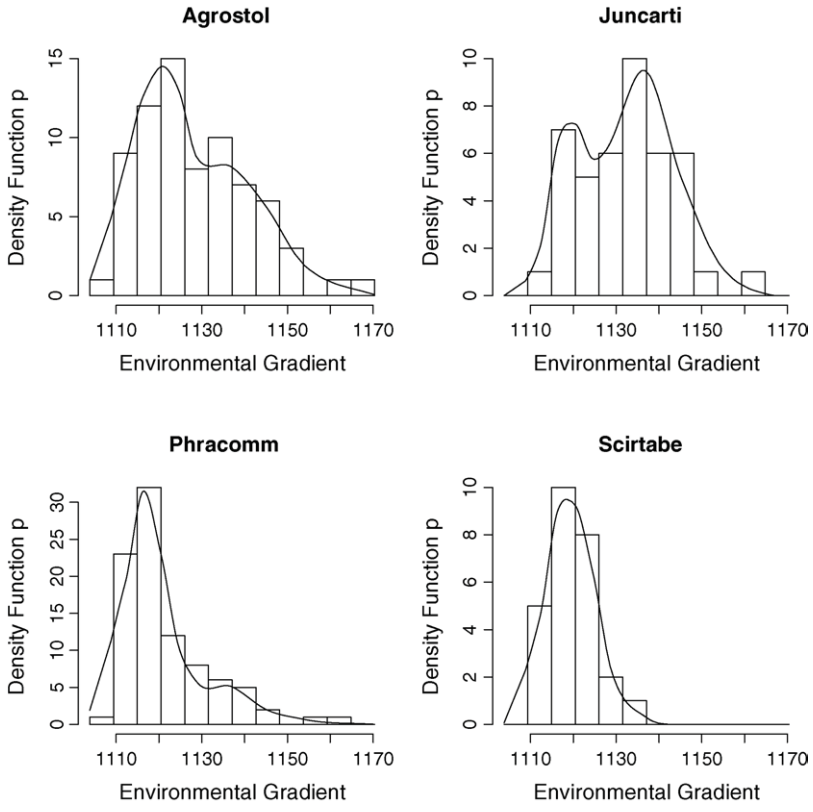

Fig. 2. Estimated density functions $p_k^{(\alpha)}$ of four randomly selected species superimposed onto the histograms of the respective weighted samples $S_k(\boldsymbol{\alpha})$ along the estimated environmental gradient.

Here we analyze this data without any parametric assumptions on the response function f_k and hence p_k. First, we notice that, out of the 126 sites, some species were observed at fewer than five sites in total. One can't possibly estimate a function nonparametrically with fewer than five observations. Therefore, in our analysis, we include only those species that were observed at a minimum of five different sites. There are 38 such species in total. We re-analyze the reduced data set with CCA and obtain a ratio of 0.82, not significantly different from ter Braak's original result. This is an indication that these relatively rare species do not affect the gradient in any significant way.

We then estimate the gradient by maximizing the generalized criterion LR($\boldsymbol{\alpha}$) using a Newton–Raphson algorithm. We use the `locfit` library (Loader, 1999) for nonparametric density estimation. Using a smoothing parameter of 0.5 in the `locfit` routine, the estimated variable along the gradient is $0.8256 \times \text{elevation} + 0.5642 \times \text{year}$. The ratio of interest based on this gradient is 0.68, which is a little smaller than the result from CCA and closer to the known uplift of 0.5.

Fig. 2 shows the estimated density functions $p_k^{(\alpha)}$ of four randomly selected species for illustration. Also shown are the histograms of their counts along the gradient. Fig. 3 shows the corresponding response functions f_k of the same four species after accounting for the sampling distribution g; an estimate of the function g is displayed in Fig. 4 alongside with the histogram of $S(\boldsymbol{\alpha}) = \{z_1, z_2, \ldots, z_n\}$. The same `locfit` routine is used to estimate g. The functions $p_k^{(\alpha)}$ have been rescaled in both plots so that they can appear side-by-side with the histograms (Fig. 2) and the functions f_k (Fig. 4).

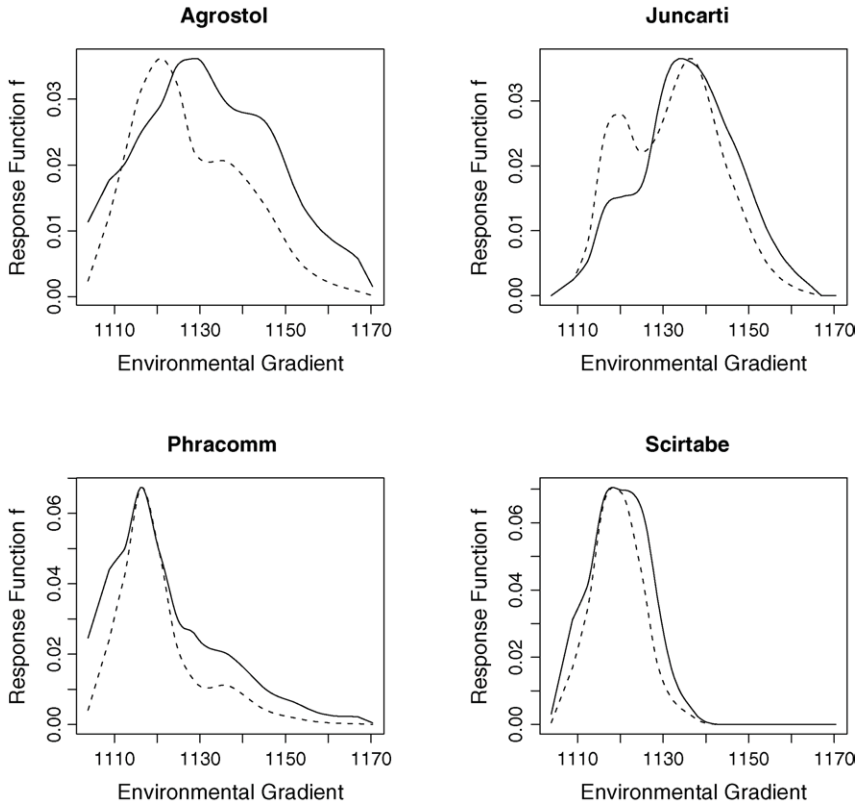

Fig. 3. Estimated response functions along the estimated environmental gradient of four randomly selected species. The solid line is the response function f_k; the dashed line is the density function $p_k^{(\alpha)}$ from Fig. 2.

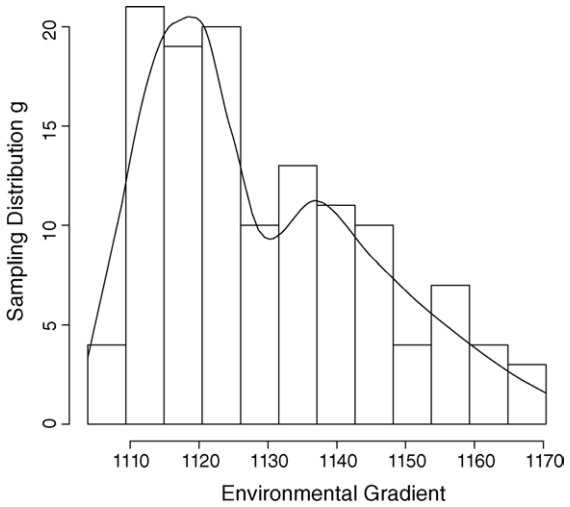

Fig. 4. Estimated sampling distribution $g(z)$ along the estimated environmental gradient.

Remark. It is clear that if $\boldsymbol{\alpha}' = k\boldsymbol{\alpha}$ for some scalar multiple k, then $\mathrm{LR}(\boldsymbol{\alpha}') = \mathrm{LR}(\boldsymbol{\alpha})$, so it suffices to restrict $\boldsymbol{\alpha}$ to be a unit vector ($\|\boldsymbol{\alpha}\| = 1$). In this particular example, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^{\mathrm{T}}$ is a unit vector in $\mathbb{R}^2$. One can, therefore, write $\boldsymbol{\alpha} = (\cos\theta, \sin\theta)$ and plot $\mathrm{LR}(\cdot)$ as a function of θ (Fig. 5).

As expected, the function is periodic as θ goes from 0 to 2π, since $\mathrm{LR}(\boldsymbol{\alpha}) = \mathrm{LR}(-\boldsymbol{\alpha})$. Therefore it suffices to focus on the interval $[0, \pi]$. We can clearly see from the plot that the maximum occurs at around $\theta = \arctan 0.68 \approx 0.60$. We also see another local maximum at around $\theta = 1.5$. This can be avoided by trying different starting values, a typical approach when using iterative methods to optimize a function. Generally, if there are too many local solutions, it is often necessary to use a larger smoothing parameter so as to smooth out the objective function (Zhu and Hastie, 2003). Plots such as those in Fig. 2 provide a good diagnostic tool for judging whether the smoothing parameter is well chosen. Notice also from Fig. 5 that the objective function is fairly flat near its maximum. This means the corresponding confidence interval for the ratio of interest would be fairly wide.

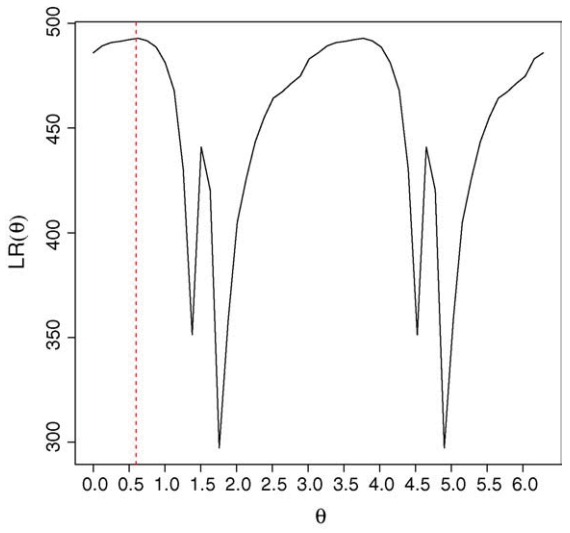

Fig. 5. The function $\mathrm{LR}(\cdot)$ as a function of θ as θ goes from 0 to 2π.

6. Summary

We have presented an alternative probabilistic model for the ecological ordination problem, one that leads directly to the LDA algorithm. Due to the equivalence between LDA and CCA, this model gives additional insights into the nature of CCA as a computational tool. In particular, LDA and hence CCA can be seen to find the environmental gradient explicitly as the direction where the species' response functions differ the most. Based on this alternative model, we have developed a simple generalization of canonical Gaussian ordination that allows the use of more flexible response functions. It is also shown that our approach is equivalent to maximizing a conditional multinomial likelihood function, which, by conditioning on the total species count at every site, implicitly encourages competition among the species.

Throughout this article, we have concentrated on the case where y_{ik} is a Poisson random variable. Sometimes only the presence or absence of a species is recorded, in which case y_{ik} would be a Bernoulli rather than a Poisson random variable. However, Theorem 1 is valid for general distributions of y_{ik}, which implies that the approach we have proposed for estimating the environmental gradient based on maximizing the LR criterion (16) is more generally applicable. What is not clear in those cases is whether maximizing LR (16) is still equivalent to maximizing the conditional likelihood when y_{ik} is not Poisson.

Acknowledgments

Mu Zhu is partially supported by the Natural Science and Engineering Research Council of Canada. Trevor J. Hastie is partially supported by grant DMS-0204612 from the National Science Foundation and grant ROI-EB0011988-08 from the National Institutes of Health. Guenther Walther is partially supported by grant DMS-9875598 from the National Science Foundation and grant R33-H268522-01 from the National Institutes of Health. The authors would like to thank Cajo J.F. ter Braak for his insightful comments and for providing the data set for the vegetation succession study in Section 5. Finally, the authors would also like to thank the editor and two referees for their comments and encouragement.

Appendix A. Proof of Theorem 1

We will prove the theorem for multivariate $z_i \in \mathbb{R}^d$. The empirical measure puts weights $\frac{w_i}{\sum_{j=1}^n w_j}$ on the points z_i, $i = 1, \ldots, n$. Let $I \subset \mathbb{R}^d$ be a rectangle. We need to show

$$\sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} 1(z_i \in I) \to \frac{\int_I g(z)f(z)\mathrm{d}z}{\int_{\mathbb{R}^d} g(z)f(z)\mathrm{d}z}$$

almost surely. (21)

Note that, while conditional on z_i each w_i may have a different distribution, unconditionally the w_is are i.i.d. with mean $E(w_i) = E\{E(w_i|z_i)\} = \int_{\mathbb{R}^d} g(z)f(z)\mathrm{d}z < \infty$. Since w_i is non-negative, we have $E|w_i| = E(w_i) < \infty$. Hence, the strong law of large numbers applies and gives

$$\frac{1}{n}\sum_{i=1}^{n} w_i \to \int_{\mathbb{R}^d} g(z)f(z)\mathrm{d}z \quad \text{almost surely.}$$

Likewise, the quantities $w_i 1(z_i \in I)$ $(i = 1, 2, \ldots, n)$ are also non-negative and i.i.d. with mean

$$E\{w_i 1(z_i \in I)\} = E\{1(z_i \in I)E(w_i|z_i)\} = \int_I g(z)f(z)\mathrm{d}z < \infty.$$

By the strong law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} w_i 1(z_i \in I) \to \int_I g(z)f(z)\mathrm{d}z \quad \text{almost surely.}$$

Eq. (21) now follows.

Appendix B. Derivation of equation (18)

We now establish that Eq. (18) is true when Eq. (17) holds. First, we give an expression for π_k. Let w_k be the count of species k, then conditional on an environmental score z, w_k is a Poisson random variable with rate $f_k(z)$, which means $E(w_k|z) = f_k(z)$. This implies the expected count (or population size) of species k is

$$s_k \stackrel{\text{def}}{=} E(w_k) = E(E(w_k|z)) = \int E(w_k|z)g(z)\mathrm{d}z = \int f_k(z)g(z)\mathrm{d}z. \quad (22)$$

If we write $s = \sum_{k=1}^K s_k$, then the relative frequency of species k, π_k, is nothing but

$$\pi_k = \frac{s_k}{s}. \quad (23)$$

Empirically π_k can be estimated with $\hat{\pi}_k = y_{\cdot k}/y_{\cdot\cdot}$ where $y_{\cdot k} = \sum_{i=1}^n y_{ik}$ and $y_{\cdot\cdot} = \sum_{i=1}^n \sum_{k=1}^K y_{ik}$. From (22) it is now clear what the missing proportionality constant in (14) is and we get

$$p_k^{(\alpha)}(z) = \frac{1}{s_k} g(z) f_k(z). \quad (24)$$

Eq. (17) then gives

$$p^{(\alpha)}(z) = \sum_{k=1}^{K} \frac{\pi_k}{s_k} g(z) f_k(z) \quad (25)$$

Dividing (25) into (24) gives us

$$\frac{p_k^{(\alpha)}(z)}{p^{(\alpha)}(z)} = \frac{\frac{1}{s_k} g(z) f_k(z)}{\sum_{j=1}^{K} \frac{\pi_j}{s_j} g(z) f_j(z)} = \frac{\frac{1}{s_k} f_k(z)}{\sum_{j=1}^{K} \frac{\pi_j}{s_j} f_j(z)} = \frac{\frac{1}{s_k} f_k(z)}{\frac{1}{s}\sum_{j=1}^{K} f_j(z)} = \frac{f_k(z)}{\pi_k \sum_{j=1}^{K} f_j(z)},$$

where we have used the identities

$$\frac{\pi_j}{s_j} = \frac{1}{s} \quad \text{and} \quad \frac{s}{s_k} = \frac{1}{\pi_k},$$

which both follow from (23). Hence Eq. (18) is established.

References

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

Green, R.H., 1971. A multivariate statistical approach to the Hutchinsonian niche: bivariate mollucs of central Canada. Ecology 52, 543–556.

Green, R.H., 1974. Multivariate niche analysis with temporally varying environmental factors. Ecology 55, 73–83.

Hastie, T.J., Tibshirani, R.J., 1996. Discriminant analysis by Gaussian mixtures. J. R. Stat. Soc. Ser. B 58, 155–176.

Hastie, T.J., Buja, A., Tibshirani, R.J., 1995. Penalized discriminant analysis. Ann. Stat. 23 (1), 73–102.

Hastie, T.J., Tibshirani, R.J., Friedman, J.H., 2001. The Elements of Statistical Learning: Data-Mining, Inference and Prediction. Springer-Verlag.

Heegaard, E., 2002. The outer boarder and central border for species–environmental relationships estimated by nonparametric generalized additive models. Ecol. Model. 157 (2–3), 131–139.

Johnson, K.W., Altman, N.S., 1999. Canonical correspondence analysis as an approximation to Gaussian ordination. Environmetrics 10, 39–52.

Loader, C., 1999. Local Regression and Likelihood. Springer-Verlag.

MacArthur, R.H., Levins, R., 1967. The limiting similarity, convergence and divergence of co-existing species. Am. Nat. 101, 377–385.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall.

Ross, S.M., 1997. Introduction to Probability Models, sixth ed. Academic Press.

Takane, Y., Yanai, H., Mayekawa, S., 1991. Relationships among several methods of linearly constrained correspondence analysis. Psychometrika 56 (4), 667–684.

ter Braak, C.J.F., 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41, 859–873.

ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67 (5), 1167–1179.

ter Braak, C.J.F., 1987. The analysis of vegetation–environment relationships by canonical correspondence analysis. Vegetation 69, 69–77.

ter Braak, C.J.F., 1996. Unimodal Models to Relate Species to Environment. DLO-Agricultural Mathematics Group, Wageningen.

ter Braak, C.J.F., Verdonschot, P.F.M., 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat. Ecol. 57 (3), 255–289.

Yee, T.W., 2004. A new technique for maximum-likelihood canonical Gaussian ordination. Ecol. Monogr. 74 (1), 685–701.

Yee, T.W., Hastie, T.J., 2003. Reduced-rank vector generalized linear models. Stat. Model. 3, 15–41.

Zhu, M., Hastie, T.J., 2003. Feature extraction for nonparametric discriminant analysis. J. Comput. Graph. Stat. 12 (1), 101–120.