Appendix 1

Two-way canonical analysis of variance for paired observations

Pierre Legendre, mars 2004

The analysis of variance problem at hand is difficult because it involves two crossed factors, the observations are paired for one of the factors, and the response data are multivariate (15 taxa). Canonical redundancy analysis (RDA; Rao 1964) can be used to compute a multifactorial and multivariate analysis of variance (Legendre & Anderson 1999; ter Braak & Smilauer 2002, Section 3.7.6). RDA is the direct extension of multiple regression to the modelling of multivariate response data (Legendre & Legendre 1998).

This section shows how the factors and pairing variables must be coded for this type of problem. We will use an example data set presented by Zar (1999, p. 162) to illustrate the paired-sample *t*-test and show that RDA can be used to obtain the same test, provided that the factors are coded in an appropriate way. We will then modify Zar's example to include a second factor crossed with the first one, thus creating a univariate analogue of the multivariate data studied in the main paper.

Comparing the means of two groups of paired observations

The original example of Zar compares the means of the lengths of the hind and fore legs of 10 deer using a paired-sample *t*-test (Table A1.1). The paired-sample statistic for that example is $t = 3.41379$ with 9 degrees of freedom and an associated two-tailed parametric probability $P = 0.0077$ (the numerical results reported by Zar (1999), are imprecise and differ slightly from these values). The mean length of the hind legs is 144.7 cm and that of the fore legs is 141.4 cm. The

mean difference is thus 3.3 cm. The test indicates that, if the sample data are representative of the population, the hind legs are significantly longer than the fore legs in that population of deer.

Using coded pairing variables, we will recompute the *t*-test by (1) multiple regression and (2) canonical redundancy analysis (RDA). The dummy variables representing the main factor "Hind/Fore leg" as well as the pairing of the legs for each deer are shown in Table A1.1. "Hind/Fore leg" is coded as +1 and –1 for a reason that will be explained in the next section. For the simple example of the present section, it could have been coded as 0 and 1, with the same result.

The tests of significance of the factor "Hind/Fore leg" is obtained by regressing "Leg length" on "Hind/Fore leg", using as covariables the nine dummy variables shown in Table A1.1 which pair the hind and fore legs pertaining to the same animals. In regression analysis, one only has to include all variables in the analysis and look for the test of the parameter associated with "Hind/Fore leg" in the multiple regression results. The regression results are found in Table A1.2. The *t*-statistic for the test of significance of the partial regression coefficient of the term "Hind/Fore leg", as well as the P-value, are identical to those of the *t*-test for paired observations reported in the previous paragraph. Thus, inclusion of the P1-P9 dummy variables as covariables in the multiple regression effectively produced a test for the parameter of the "Hind/Fore leg" variable equivalent to a *t*-test for paired observations.

The exact same results can be obtained by canonical redundancy analysis (RDA), using the program Canoco (ter Braak & Smilauer 2002). "Leg length" can be analysed using "Hind/Fore leg" as the explanatory variable, and variables P1 to P9 as covariables. The statistic for the test of the explanation of "Leg length" provided by "Hind/Fore leg" is $F = 11.654$, which is the square of $t = 3.41379$. Canoco tests the *F*-statistic by permutation; the P-value estimated in

that way remains valid when the normality assumptions of the parametric test are not met; it is also entirely appropriate for small samples, as in the present example (Legendre & Legendre 1998, Section 1.2; Anderson & Legendre 1999). For the deer leg problem, the P-value provided by Canoco was 0.0114 after 999 random permutations of the residuals under the reduced model. This is very close to the parametric probability reported in a previous paragraph.

The canonical analysis solution also provided a statistic, called *trace* in the Canoco program, which estimates the proportion of the variance of "Leg length" explained by "Hind/Fore leg": *trace* = 17.855%. This is equivalent to a partial R^2 statistic. Since the covariables P1-P9 are orthogonal to the explanatory variable "Hind/Fore leg" in this particular example, *trace* is equal to the coefficient of determination (R^2) of the simple linear regression of "Leg length" on "Hind/Fore leg": $R^2 = 0.17855$ or 17.855%. The *F*-statistic, which is the ratio of two *independent* portions of the dependent variable's variation, is computed as follows from the R^2 values (Legendre & Legendre 1998; ter Braak & Smilauer 2002; see Fig. A1.1, leaving out for the moment the contributions of "Sex" and "Interaction"):

$$F = \frac{R^2_{\text{explanatory variables}} \,/\, m}{(1 - R^2_{\text{explanatory variables}} - R^2_{\text{covariables}}) \,/\, (n - 1 - m - q)} \qquad (1)$$

$$F = \frac{0.17855 \,/\, 1}{(1 - 0.17855 - 0.68356) \,/\, (20 - 1 - 1 - 9)} = 11.654$$

with the number of data rows $n = 20$, the number of explanatory variables $m = 1$ and the number of covariables $q = 9$ (Table A1.1).

Fig. A1.1 shows why, in the case of paired observations, a test involving the pairing variables as covariables is necessarily equally or more powerful than an unpaired *t*-test or an anova for unpaired data. In a test for unpaired data, the denominator of the *F*-statistic would not

include the variation accounted for by the pairing variables (68.356% in this example), nor the degrees of freedom associated with these variables (9). The *F*-statistic would be:

$$F = \frac{0.17855 / 1}{(1 - 0.17855) / (20 - 1 - 1)} = 3.912$$

This value, which is the square of the statistic one would obtain in a *t*-test for independent data (*t* = 1.97802), leads to a higher probability of the data under H_0 (parametric P = 0.0634, two-tailed test, 18 degrees of freedom) than the corresponding *t*-test for paired observations (P = 0.0077, two-tailed test, 9 degrees of freedom). The *t*-test for independent observations would not reject H_0 at the 5% significance level in this example, whereas the *t*-test for paired observations does. The latter is more powerful because it is more capable of identifying that a small difference between the means of the groups of observations is significantly larger than what is expected from random variation.

Two-way analysis of variance for paired observations

The next challenge was that there are not one, but two factors in the benthos data analysed in the main paper: the two geographic zones and the two counting methods. We could have analysed the two zones separately, but there was real interest in determining if there were significant differences in the infauna between the two zones. So, the design of the analysis was: two crossed fixed factors: the geographic zones called “Zones”, and the counting methods called “Methods”. The observations were paired for the second factor since the *same cores* were analysed and counted using the two methods. For illustration of the statistical method of analysis, the Zar example was turned into a two-factor problem by pretending that animals 1 to 5 were males and animals 6 to 10 were females (Table A1.1). A possible interaction between “Hind/Fore

leg" and "Sex" can be studied through an "Interaction" variable obtained by computing the product of the dummy variables representing the main factors.

In Table A1.1, the variables "Hind/Fore leg" and "Sex", which are crossed, are coded using orthogonal dummy variables. The codes for each one sum to 0 and their scalar product (or dot product) is 0; these variables are said to be *orthogonal*. For balanced study designs, the row-by-row products of variables coded in this way produce an interaction variable which is orthogonal to the two crossed variables. Thanks to this property, the fractions of variation explained by the main factors and the interaction term do not overlap in Fig. A1.1. If we had used ordinary dummy variables coded by 0's and 1's, the interaction variable would not have been orthogonal to the main factors and their circles would have overlapped in Fig. A1.1. Various methods can be used to code an experimental factor into orthogonal dummy variables; an easy method is described in Appendix C of Legendre & Anderson (1999).

The relationships among the explanatory variables were complex in this example. Whereas the variables coding for "Hind/Fore leg", "Sex" and "Interaction" are orthogonal to one another (their correlation is 0), the pairing dummy variables P1-P9 explain "Sex" completely in a multiple regression ($R^2 = 1$). Table A1.1 shows that the dummy variables P1 to P5 separate the males from the females, so that variables P1 to P5 are sufficient to explain "Sex" entirely. A Venn diagram of the coefficients of determination (*trace* statistics, or R^2) of the explanatory variables on the "Leg length" response variable, drawn in the spirit of the variation partitioning diagrams pioneered by Borcard et al. (1992) and Legendre (1993), is presented in Fig. A1.1. In other studies where the pairing affects the two crossed factors, if the factors are orthogonal to each other (as it is the case if the design is balanced), they are also orthogonal to the set of pairing

dummy variables. In such a case, the surfaces explained by the two main factors, the interaction, and the pairing variables do not overlap in the Venn diagram.

Because, in the present design, "Sex" was entirely explained by (and was thus collinear to) the pairing variables, we computed the regression residuals of P1-P9 on "Sex" before proceeding with the analysis. The rank of the covariance matrix of the 9 residual variables was 8; projection on "Sex" had resulted in the loss of one dimension. The 9 residual variables were subjected to principal component analysis (PCA). The object scores along the 8 axes (Axis 1 to Axis 8) that had non-zero positive eigenvalues were linearly independent of "Sex"; they became the new explanatory variables in the analysis of variance that follows.

The tests of significance of the main effects and interaction are tests of the partial regression coefficients of these terms in a multiple regression of "Leg length" against the two main effects and interaction variables, when the principal components are also included as covariables in the regression analysis.

Test of the "Interaction"—From the regression analysis results (Table A1.3), the statistic associated with "Interaction" is $t = -1.70941$, P = 0.1257 (parametric probability). The same result was obtained from a redundancy analysis (RDA) of the response variable "Leg length" by the explanatory dummy variable coding for "Interaction", with "Hind/Fore leg", "Sex", and the 8 principal components representing pairing residuals (Axis 1-Axis 8) as covariables ($F = 2.922$, P = 0.1222 after 9999 permutations of residuals of the reduced model in Canoco). With $m = 1$ and $q = 10$, we find, using in Eq. 1 the R^2 values from Fig. A1.1:

$$F = \frac{0.03689 / 1}{(1 - 0.17855 - 0.15757 - 0.52599 - 0.03689) / (20 - 1 - 1 - 10)} = \frac{0.03689}{0.10100 / 8} = 2.922$$

"Hind/Fore leg", "Sex", and Axis 1-Axis 8 are used as covariables in this analysis. $F = t^2$ since, in this example, the number of degrees of freedom of the numerator of the F-statistic is $m = 1$. If there were more than 2 classes in one of the main factors, "Interaction" would be coded by more than one variable and the anova could only be done by partial multiple regression for a single response variable, or partial canonical analysis for single or multiple response variables. In the present example, a non-significant interaction indicates that the effects of the classes of each factor do not significantly differ between the classes of the other factor. We can thus proceed with the analysis of the main factors in the two-way context.

Because "Sex" is entirely explained by, and collinear to, the pairing variables P1-P9, an equivalent method is to exclude "Sex" from the analysis and analyse "Leg length" by the explanatory dummy variable "Interaction", with covariables "Hind/Fore leg" and the original 9 dummy variables P1-P9 representing pairing. If one uses a program, such as Canoco, capable of handling collinearity among the covariables, "Sex" can remain among the covariables in the analysis. All these forms produce the same results for the test of "Interaction".

Test of the "Hind/Fore leg"—We expect the test of the factor "Hind/Fore leg" to either have the same power or be more powerful in a two-way anova than the t-test of that factor alone reported in the previous section. The reason is that the variability explained by "Interaction" was not included in the denominator of the F-statistic in the previous one-factor analysis of variance (Fig. A1.1). If "Interaction" contributes to lowering the value of the F-statistic more than a random factor would, it may reduce the P-value of the test.

In the regression analysis (Table A1.3), the t-statistic associated with this factor is 3.76070, P = 0.0055 (parametric probability). This result shows slightly more power than the t-test reported in the previous section (P = 0.0077). An equivalent result was obtained from a

redundancy analysis (RDA) of the response variable "Leg length" by the explanatory dummy variable coding for "Hind/Fore leg", with "Sex", "Interaction", and the eight principal components representing pairing (Axis 1-Axis 8) as covariables (F = 14.143, P = 0.0085 after 9999 permutations of residuals of the reduced model in Canoco). $F = t^2$ since, in this example, the number of degrees of freedom of the numerator of the F-statistic is m = 1.

Because "Sex" is entirely explained by, and collinear to, the pairing variables P1 to P9, an equivalent method is to exclude "Sex" from the analysis and analyse the response variable "Leg length" by the explanatory dummy variable coding for "Hind/Fore leg", with covariables "Interaction" and the original 9 dummy variables representing pairing.

Test of the "Sex"—In the regression analysis table, the t-statistic associated with this factor is –3.53276, P = 0.0077 (parametric probability). The same result was obtained from a redundancy analysis (RDA) of the response variable "Leg length" by the explanatory dummy variable coding for "Sex", with "Hind/Fore leg", "Interaction" and the 8 principal components representing pairing (Axis 1-Axis 8) as covariables (F = 12.481, P = 0.0099 after 9999 permutations of residuals of the reduced model in Canoco). $F = t^2$ since, in this example, the number of degrees of freedom of the numerator of the F-statistic is m = 1.

Discussion

The results reported in the previous sections show that, in the case of a single response variable, multiple regression or canonical redundancy analysis (RDA) can be used to obtain tests equivalent to a paired-sample t-test (for a single factor) or a paired-sample two-way analysis of variance (for two crossed factors). For multivariate response data tables, RDA is the instrument of choice to conduct such tests since it can handle multivariate response data tables and allows

incorporation in the analysis of dummy variables coding for the main factors, the interaction, as well as pairing variables tailored to the problem under study. For the multivariate case, RDA offers the additional advantage of allowing the representation of the manova results in the form of biplots displaying the responses of the dependent variables to the manova factors. RDA is used in the main paper to analyse multivariate data of the same form as the univariate example presented in this Appendix.

Tests of significance of the fixed factor in a mixed model (i.e., an anova with a fixed and a random factor), and of any factor in a model II anova (i.e., an anova for two random factors), require the use of the interaction mean square in the denominator, instead of the residual mean square. How to obtain a correct test is described in Section 3.7.6 of the Canoco manual (ter Braak & Smilauer 2002). For paired observations, the pairing variables must be included in the analysis as covariables, as shown in the present example.

References

Anderson, M. J. and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Stat. Comput. Simul.* 62: 271-303.

Borcard, D., P. Legendre and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73: 1045-1055.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 74: 1659-1673.

Legendre, P. and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69: 1-24.

Legendre, P. and L. Legendre. 1998. *Numerical ecology, 2nd English edition.* Elsevier Science BV, Amsterdam.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhyaá A* 26: 329-358.

ter Braak, C. J. F. and P. Smilauer. 2002. *Canoco reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5).* Microcomputer Power, Ithaca, New York.

Zar, J.H. (1999) *Biostatistical analysis, 4th edition.* Prentice Hall, Upper Saddle River, New Jersey.

Table A1.1. Example data set: lengths of the hind and fore legs of ten deers (Zar 1999). Variables P1-P9 identify the fore and hind legs pertaining to the same animals. The example was modified by creating a second variable "Sex" which is crossed with "Hind/Fore leg". The dummy variable coding for "Interaction" is the product of the main factors' ("Hind/Fore leg" and "Sex") dummy variables.

Deer No	Length (cm)	Pairing dummy variables P1	P2	P3	P4	P5	P6	P7	P8	P9	Hind/Fore leg	Sex	Interaction
Hind legs													
1	142	1	0	0	0	0	0	0	0	0	1	1	1
2	140	0	1	0	0	0	0	0	0	0	1	1	1
3	144	0	0	1	0	0	0	0	0	0	1	1	1
4	144	0	0	0	1	0	0	0	0	0	1	1	1
5	142	0	0	0	0	1	0	0	0	0	1	1	1
6	146	0	0	0	0	0	1	0	0	0	1	–1	–1
7	149	0	0	0	0	0	0	1	0	0	1	–1	–1
8	150	0	0	0	0	0	0	0	1	0	1	–1	–1
9	142	0	0	0	0	0	0	0	0	1	1	–1	–1
10	148	0	0	0	0	0	0	0	0	0	1	–1	–1
Fore legs													
1	138	1	0	0	0	0	0	0	0	0	–1	1	–1
2	136	0	1	0	0	0	0	0	0	0	–1	1	–1
3	147	0	0	1	0	0	0	0	0	0	–1	1	–1
4	139	0	0	0	1	0	0	0	0	0	–1	1	–1
5	143	0	0	0	0	1	0	0	0	0	–1	1	–1
6	141	0	0	0	0	0	1	0	0	0	–1	–1	1
7	143	0	0	0	0	0	0	1	0	0	–1	–1	1
8	145	0	0	0	0	0	0	0	1	0	–1	–1	1
9	136	0	0	0	0	0	0	0	0	1	–1	–1	1
10	146	0	0	0	0	0	0	0	0	0	–1	–1	1

Table A1.2. Regression analysis of the variable "Leg length" by the dummy variables coding for "Hind/Fore leg" and the 9 dummy variables (P1-P9) pairing the hind and front legs of the 10 animals. The *t*-statistic and P-value for "Rear/Front leg" are in bold.

Variable	Regression coefficient	Standard regression coefficient	*t*-statistic	P-value
Intercept	147.00000	0.00000	96.17686	< 0.0001
Hind/Fore leg	1.65000	0.42256	**3.41379**	**0.0077**
P1	–7.00000	–0.53780	–3.23844	0.0102
P2	–9.00000	–0.69146	–4.16371	0.0024
P3	–1.50000	–0.11524	–0.69395	0.5052
P4	–5.50000	–0.42256	–2.54449	0.0315
P5	–4.50000	–0.34573	–2.08186	0.0671
P6	–3.50000	–0.26890	–1.61922	0.1399
P7	–1.00000	–0.07683	–0.46263	0.6546
P8	0.50000	0.03841	0.23132	0.8222
P9	–8.00000	–0.61463	–3.70108	0.0049

Table A1.3. Regression analysis of the variable "Leg length" by the dummy variables coding for "Hind/Fore leg", "Sex", "Interaction", and the principal components (Axis 1-Axis 8) derived from the residuals of the regression of dummy variables P1-P9 on "Sex". The *t*-statistic and P-value for "Rear/Front leg", "Sex", and "Interaction" are in bold.

Variable	Regression coefficient	Standard regression coefficient	*t*-statistic	P-value
Intercept	143.05000	0.00000	326.04120	< 0.0001
Hind/Fore leg	1.65000	0.42256	**3.76070**	**0.0055**
Sex	–1.54999	–0.39694	**–3.53276**	**0.0077**
Interaction	–0.75000	–0.19207	**–1.70941**	**0.1257**
Axis 1	–2.30585	–0.18674	–1.66196	0.1351
Axis 2	–1.84359	–0.14930	–1.32877	0.2206
Axis 3	–0.26967	–0.02184	–0.19436	0.8507
Axis 4	2.57978	0.20892	1.85937	0.1000
Axis 5	–4.37359	–0.35419	–3.15228	0.0136
Axis 6	6.16895	0.49959	4.44626	0.0021
Axis 7	–0.60997	–0.04940	–0.43964	0.6718
Axis 8	6.00003	0.21730	1.93398	0.0892

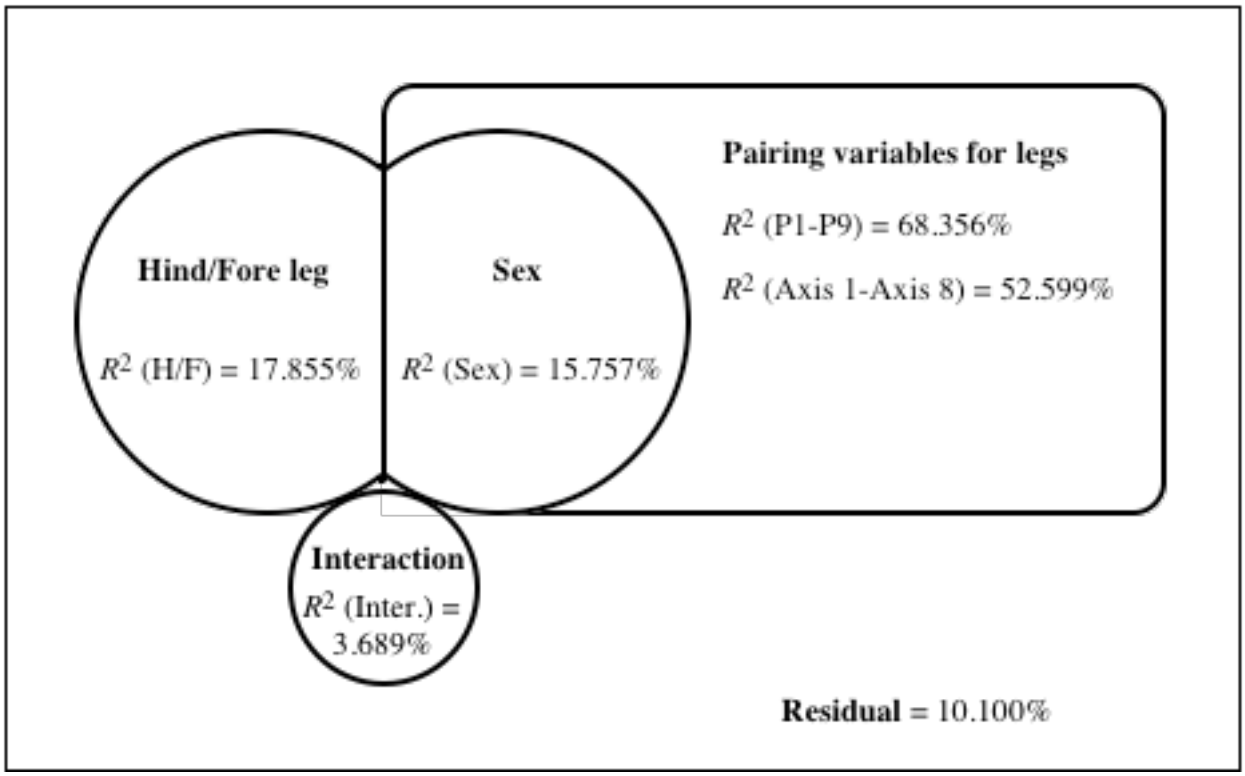

Fig. A1.1 – Venn diagram showing how the variation of the response variable (outer rectangle) is partitioned among the explanatory variables "Hind/Fore leg", "Sex", "Interaction, and the pairing variables for the legs. The pairing variables (68.356%) accounts entirely for the variation explained by "Sex" (15.757). The variation uniquely explained by the pairing variables and not by "Sex" is 68.356 – 15.757 = 52.599%. 10.100% of the variation of the response variable is not explained by the explanatory variables used in the model.