Using distance-based eigenvector maps (DBEM) in multivariate variation partitioning

Part 2:
An adjusted bimultivariate redundancy statistic to replace the traditional but biased R-square

Pierre Legendre and Pedro Peres-Neto

Département de sciences biologiques

Université de Montréal

Pierre.Legendre@umontreal.ca

Pedro.Peres-Neto@uregina.ca

Special Session "*Spatial Statistics at Multiple Scales*", ESA-INTECOL 2005 Joint Meeting

1. Statement of the problem

Variation partitioning by canonical analysis (RDA, CCA) is routinely used to relate community composition data **(Y)** to tables of environmental variables and spatial base functions.

The canonical $R^2_{\mathbf{Y}|\mathbf{X}}$ is commonly used to estimate the variation of **Y** explained by the explanatory variables **X**. We will see that this estimate is biased.

No effort has been reported in the literature to find better estimators of the variation of **Y** explained by tables of explanatory variables **X**.

Such an estimator is needed:

- to compare fractions in variation partitioning,
- to compare different canonical models.

2. Statistical properties of the bimultivariate redundancy statistic, $R^2_{Y|X}$

Description of the statistic

In canonical redundancy analysis (RDA), the proportion of the species variation (in matrix **Y**) explained by a set of explanatory variables (in matrix **X**) is called:

- the bimultivariate redundancy statistic[1],
- the canonical coefficient of determination,
- or the canonical *R*-square.

In this talk, it is noted $R^2_{\mathbf{Y}|\mathbf{X}}$

[1] Miller, J. K., and S. D. Farr. 1971. Bimultivariate redundancy: a comprehensive measure of interbattery relationship. *Multivariate Behavioral Research* 6: 313-324.

Computation of $R^2_{\mathbf{Y}|\mathbf{X}}$

RDA can be computed in two steps:

• Compute the matrix of fitted values of the multivariate regression of **Y** on **X**:

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{Y}$$

• Compute a PCA of $\hat{\mathbf{Y}}$.

The bimultivariate redundancy statistic $R^2_{\mathbf{Y}|\mathbf{X}}$ is the total sum of squared deviations from the means (or *variation*) in $\hat{\mathbf{Y}}$ (SSR) divided by the total sum of squared deviations from the means (or *variation*) in **Y** (SST):

$$R^2_{\mathbf{Y}|\mathbf{X}} = \text{SSR} / \text{SST}$$

The PCA step of RDA is not necessary to compute $R^2_{\mathbf{Y}|\mathbf{X}}$.

Bias of $R^2_{\mathbf{Y|X}}$

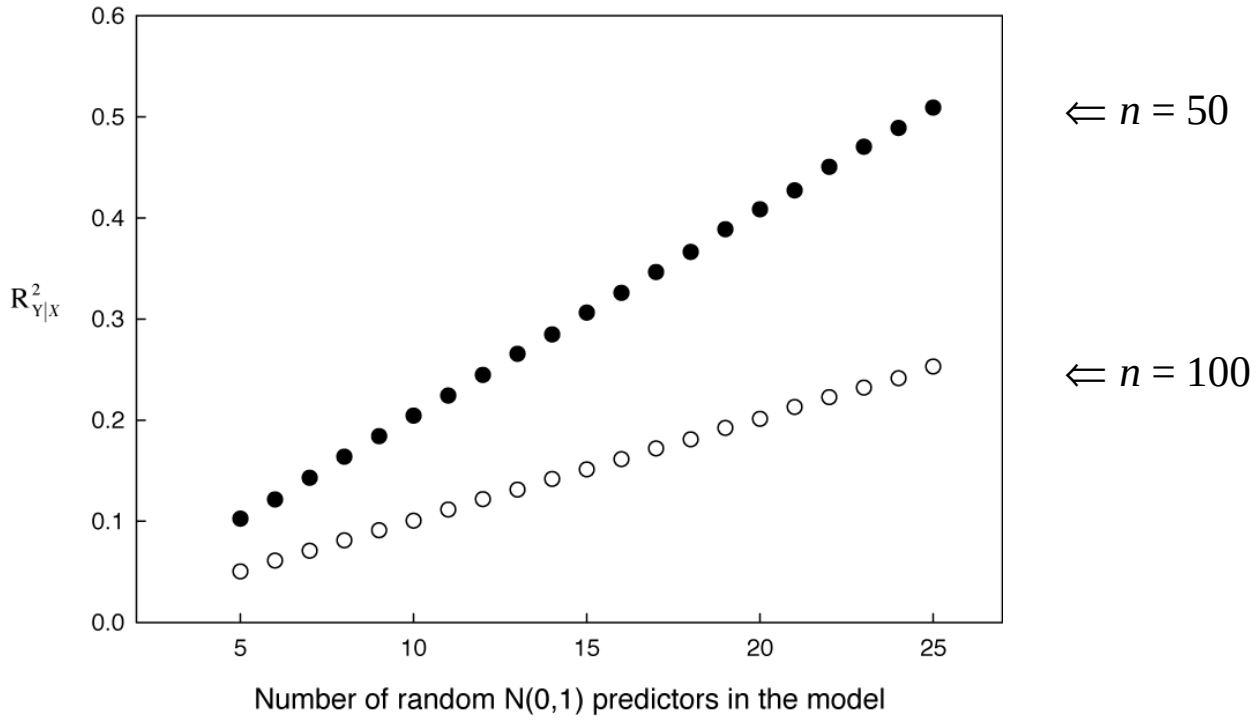


Fig. 1. Sample $R^2_{\mathbf{Y|X}}$ for **Y** (10 response variables) explained by random predictors; means after 1000 computer experiments.

3. The adjusted bimultivariate redundancy statistic, $R^2_{\mathbf{Y}|\mathbf{X}adj}$

Computation of R^2_{adj}

In multiple regression, the adjusted coefficient of multiple determination, R^2_{adj} (Ezekiel 1930)[1], takes into account the numbers of degrees of freedom of the numerator and denominator of $F(R^2)$:

$$R^2_{adj} = 1 - (1 - R^2)\left[\frac{\text{total d.f.}}{\text{residual d.f.}}\right] = 1 - (1 - R^2)\left[\frac{n-1}{n-m-1}\right]$$

where m is the number of predictors.

Using simulated data (normal error), Ohtani (2000)[2] showed that R^2_{adj} is an unbiased estimator of the contribution of the explanatory variables **X** to the explanation of **y**. R^2_{adj} is a suitable statistic for comparing regression equations involving different numbers of objects and explanatory variables.

[1] Ezekiel, M. 1930. *Methods of correlation analysis.* John Wiley and Sons, New York.

[2] Ohtani, K. 2000. Bootstrapping R^2 and adjusted R^2 in regression analysis. *Economic Modelling* 17: 473-483.

Computation of $R^2_{\mathbf{Y|X}adj}$

The bimultivariate redundancy statistic $R^2_{\mathbf{Y|X}}$ of canonical analysis can be corrected in the same way to give the adjusted bimultivariate redundancy statistic $R^2_{\mathbf{Y|X}adj}$:

$$R^2_{\mathbf{X|Y}adj} = 1 - (1 - R^2_{\mathbf{X|Y}}) \left[\frac{n-1}{n-m-1} \right]$$

We will show that in RDA, for normally distributed continuous data or Hellinger-transformed species abundances, $R^2_{\mathbf{Y|X}adj}$ produces unbiased estimates of the real contributions of the **X** variables to the explanation of a response table **Y**.[1]

[1] Peres-Neto, P., P. Legendre, S. Dray and D. Borcard. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* (submitted).

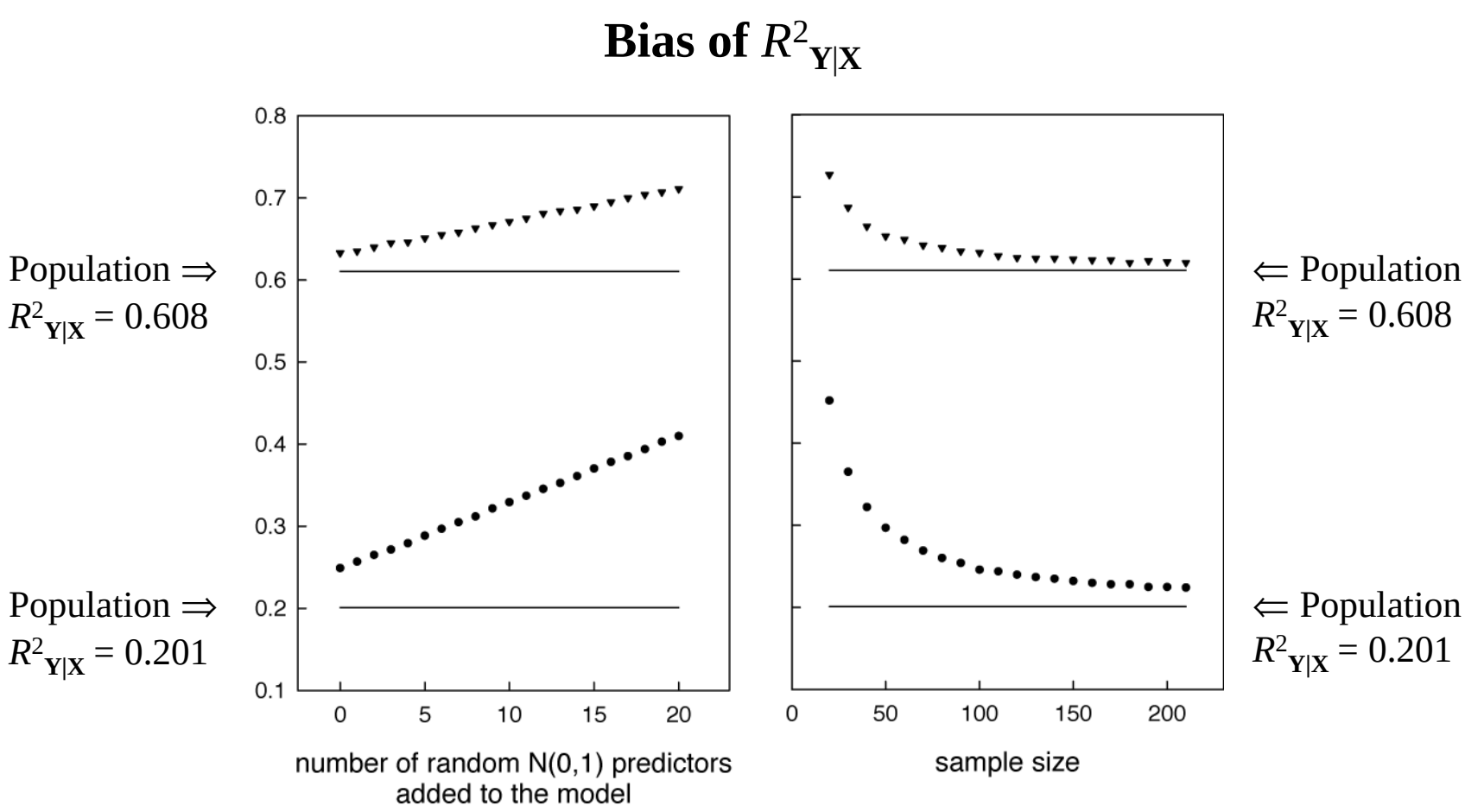
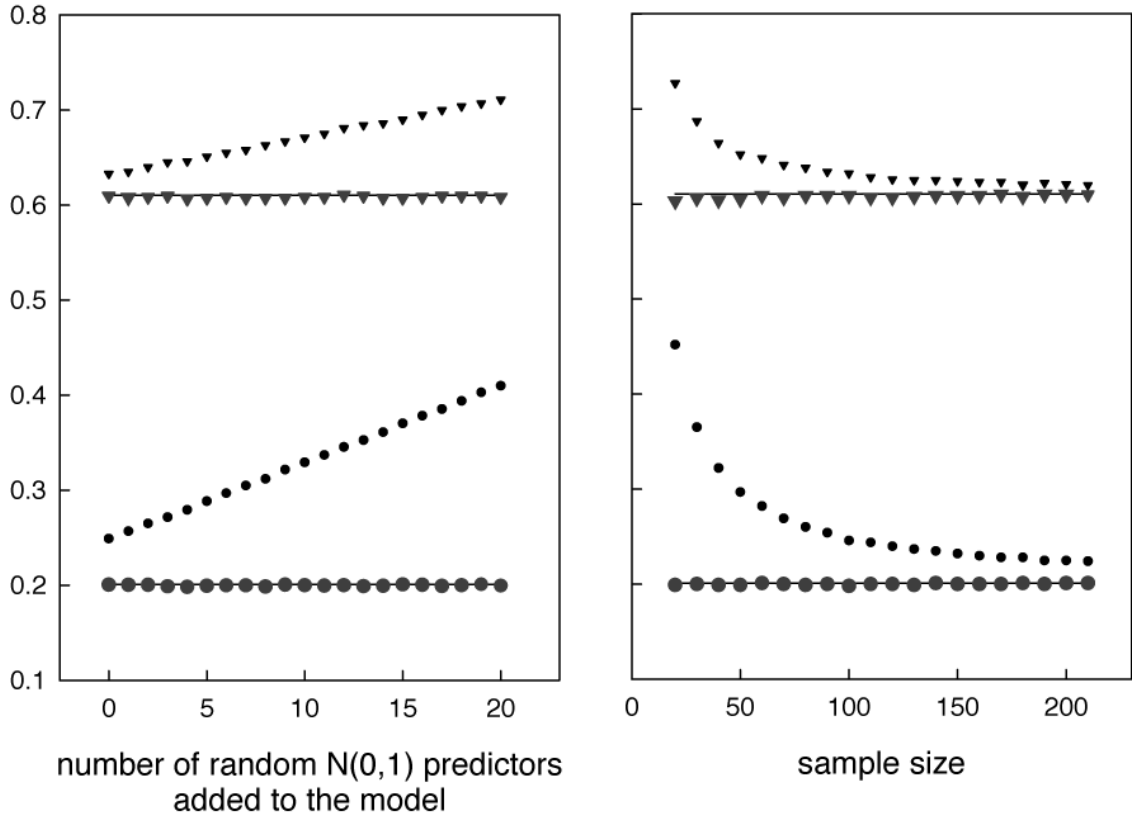


Fig. 2a. Influence of random predictors added to 6 active predictors, and of sample size, on $R^2_{Y|X}$ (means after 1000 computer experiments). Left panel: $n = 100$; right: $20 \leq n \leq 210$.

Behaviour of $R^2_{\mathbf{Y|X}adj}$

Population ⇒
$R^2_{\mathbf{Y|X}} = 0.608$

Population ⇒
$R^2_{\mathbf{Y|X}} = 0.201$



⇐ Population
$R^2_{\mathbf{Y|X}} = 0.608$

⇐ Population
$R^2_{\mathbf{Y|X}} = 0.201$

Fig. 2b. Black symbols: $R^2_{\mathbf{Y|X}}$. Red symbols: $R^2_{\mathbf{Y|X}adj}$.

Result 1 – The adjusted bimultivariate redundancy statistic, $R^2_{\mathbf{Y|X}adj}$, is a quasi-unbiased estimator of the population $R^2_{\mathbf{Y|X}}$ in RDA.

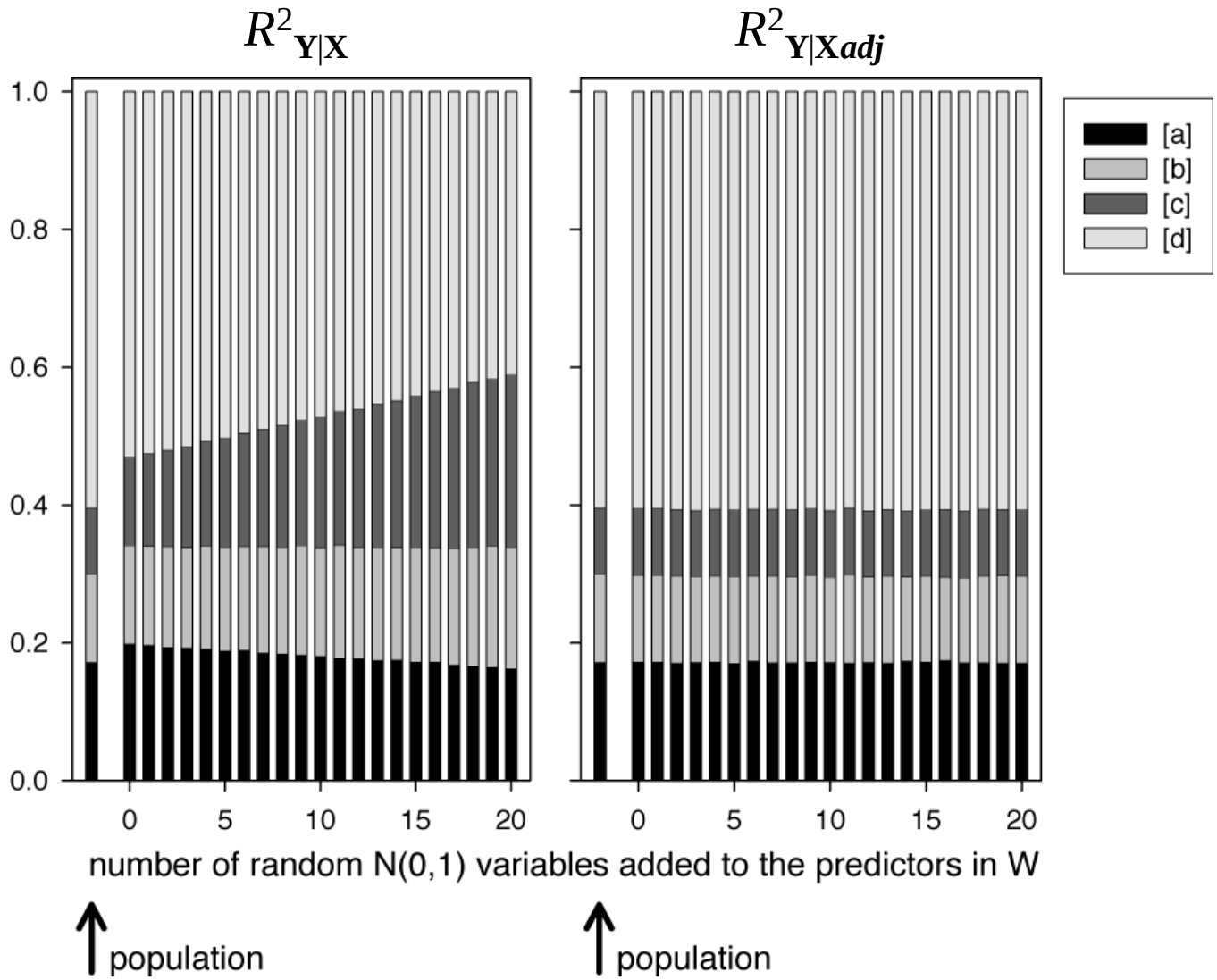
For estimation, the adjusted statistic should always be preferred to the unadjusted sample $R^2_{\mathbf{Y|X}}$.

Variation partitioning

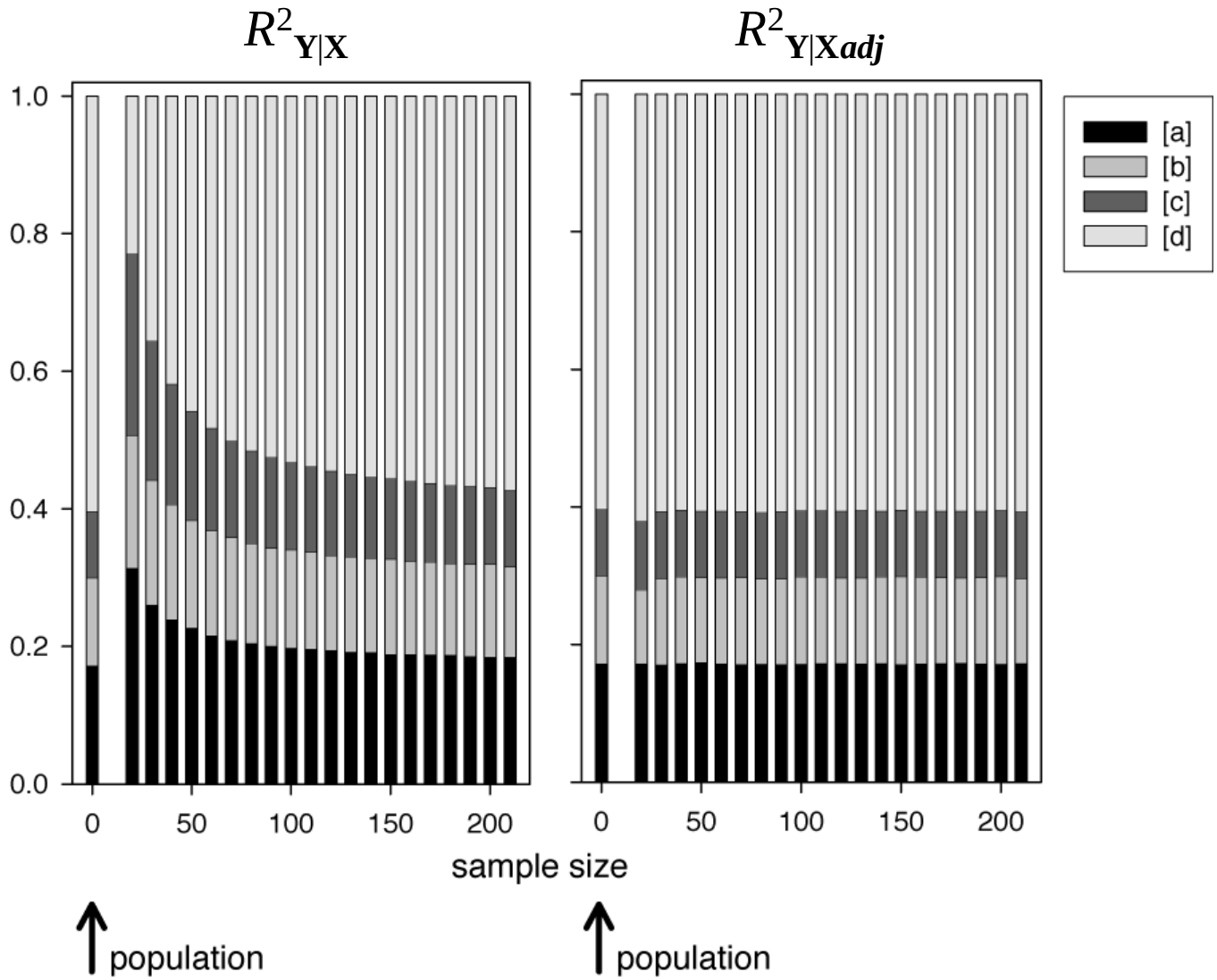
- Compute $R^2_{\mathbf{Y|X}}$ for each simple RDA (do *not* use partial RDA's).
- Calculate the corresponding $R^2_{\mathbf{Y|X}adj}$ for each RDA.
- Calculate the elementary fractions by adding and subtracting.

Fractions of variation	Probability (999 perm.)	Proportion of variation of Y (R^2)	Adjusted R^2
Table **X**: [a+b]	0.005*	0.450	0.347
Table **W**: [b+c]	0.001*	0.734	0.639
X and **W**: [a+b+c]	0.001*	0.784	0.627
[a]	0.549	~~0.051~~	–0.012
[b]	*Cannot be tested*	~~0.399~~	0.359
[c]	0.011*	~~0.334~~	0.280
Residuals=[d]		~~0.216~~	0.373
[a+b+c+d]		1.000	1.000

$R^2_{Y|X}$ and $R^2_{Y|Xadj}$ in variation partitioning (continuous case)



$R^2_{Y|X}$ and $R^2_{Y|Xadj}$ in variation partitioning (continuous case)



Result 2 – In variation partitioning, the fractions of variation [a], [b], [c], and [d], calculated from the adjusted bimultivariate redundancy statistic, $R^2_{\mathbf{Y|X}adj}$, represent quasi-unbiased estimates of the population fractions.

Fractions of variation calculated from $R^2_{\mathbf{Y|X}adj}$ statistics are correct estimates of the importance of each explanatory table in the model.

These results were obtained for simulated continuous response variables with normal error.

Variation partitioning of species-like data

Continuous variables from the previous simulations were transformed as follows to make them resemble community composition data:

1. Generate a matrix **Y** of N(0,1) data.
2. Standardize the variables in **Y** $\Rightarrow$ $\mathbf{Y}_{std} = [y_{std}]$
3. Subtract 0.5 to provide a large number of zeros (47%);

   give a standard deviation of 1.2 to each variable;

   exponentiate to make the distributions highly asymmetrical:

$$y' = \exp(1.2(y_{std} - 0.5))$$

4. Round the values to the lower integers.

Matrix $\mathbf{Y}' = [y']$ was used in simulations.

$R^2_{\mathbf{Y|X}}$ and $R^2_{\mathbf{Y|X}adj}$ with species-like data

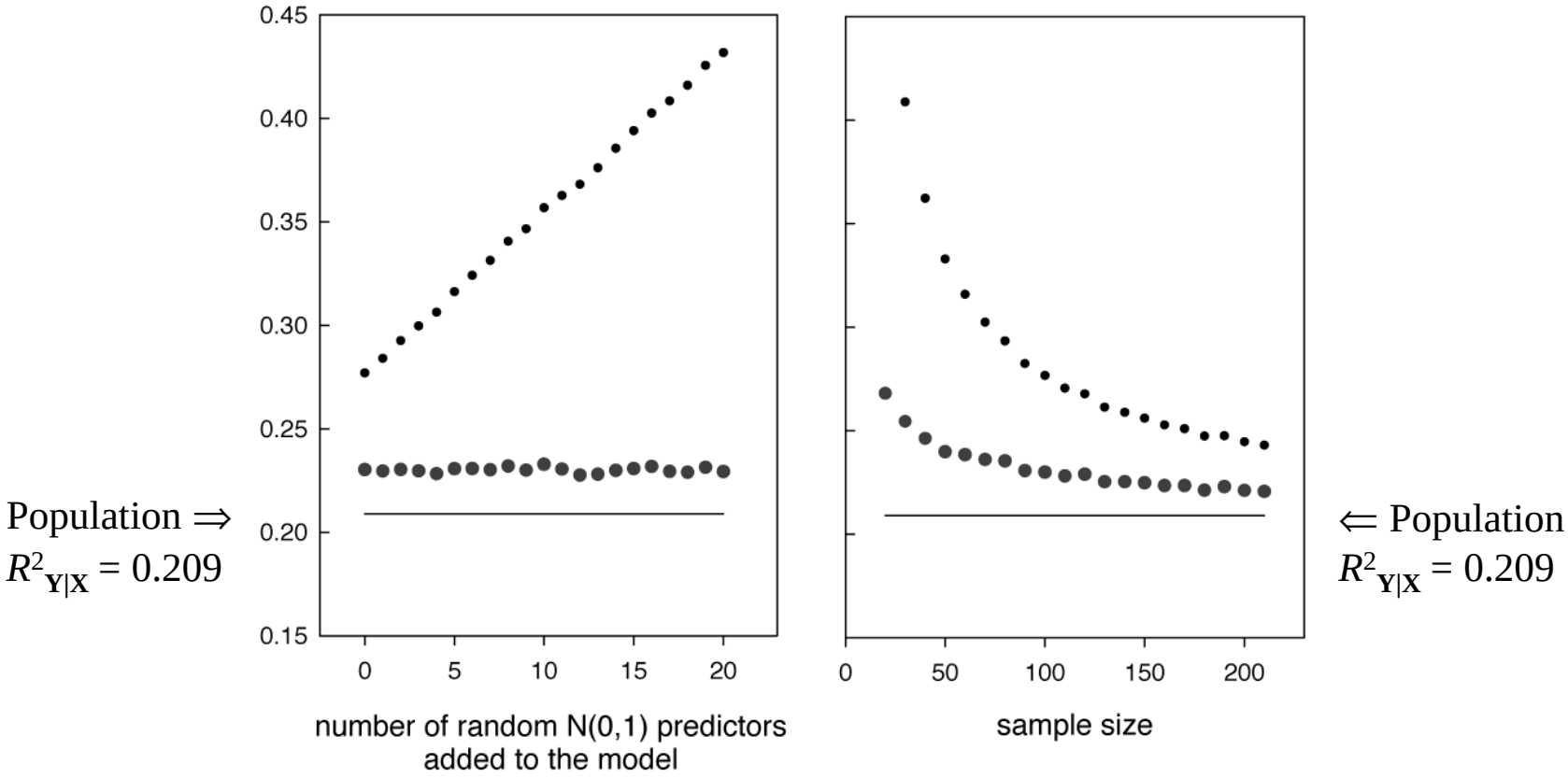


Fig. 4. Species-like data: influence of random predictors added to 6 active predictors, and of sample size, on $R^2_{\mathbf{Y|X}}$ and $R^2_{\mathbf{Y|X}adj}$ (means after 1000 experiments). Left panel: $n = 100$; right: $20 \leq n \leq 210$.

Variation partitioning of Hellinger-transformed species-like data

The integer species-like data were Hellinger-transformed[1]:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$$

where y_{i+} is the sum of abundances at site i.

Matrix $\mathbf{H} = [y'_{ij}]$ was used in the simulations instead of $\mathbf{Y}'$.

[1] Legendre, P. and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271-280.

$R^2_{Y|X}$ and $R^2_{Y|Xadj}$ with Hellinger-transformed species-like data

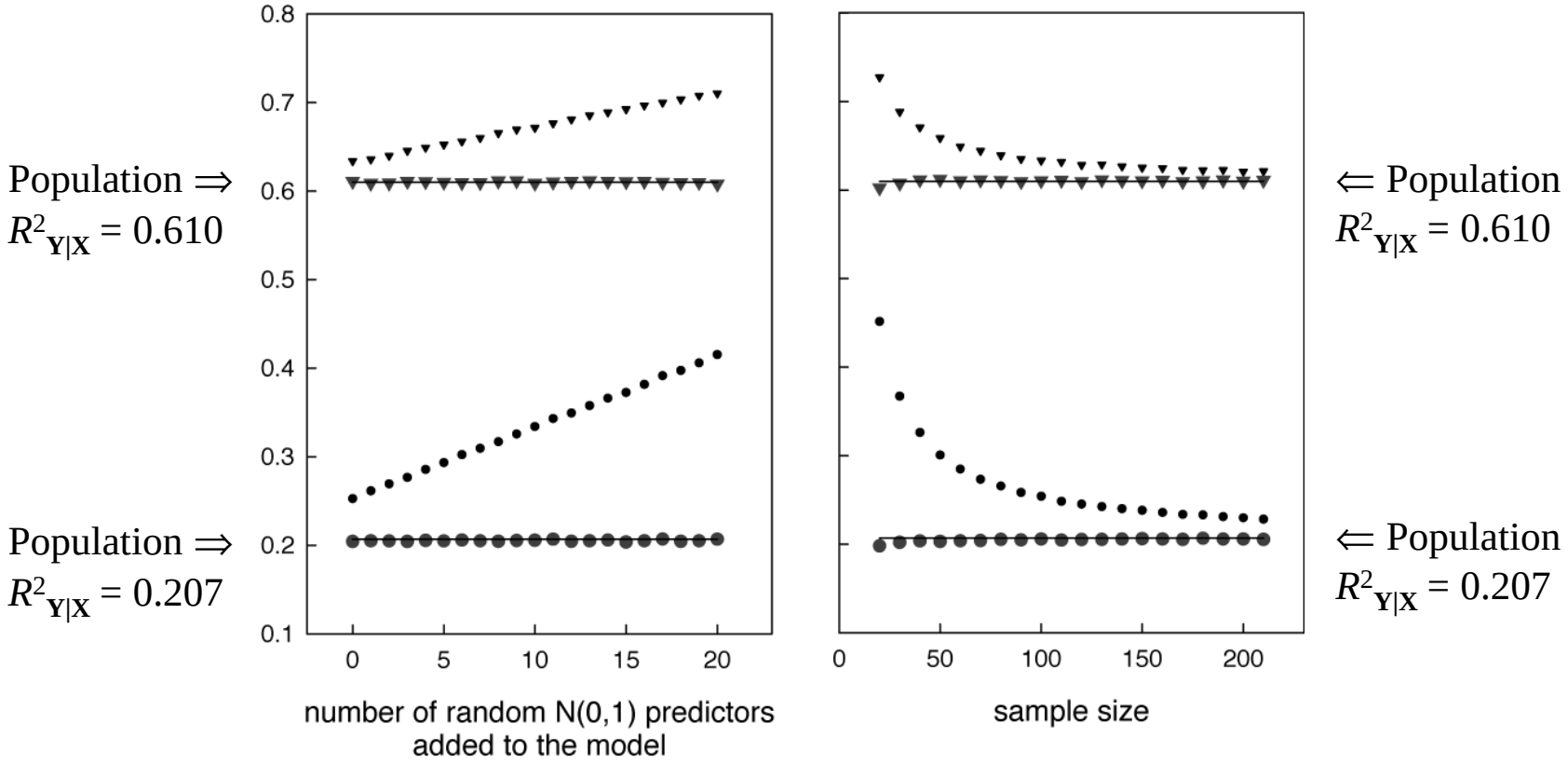
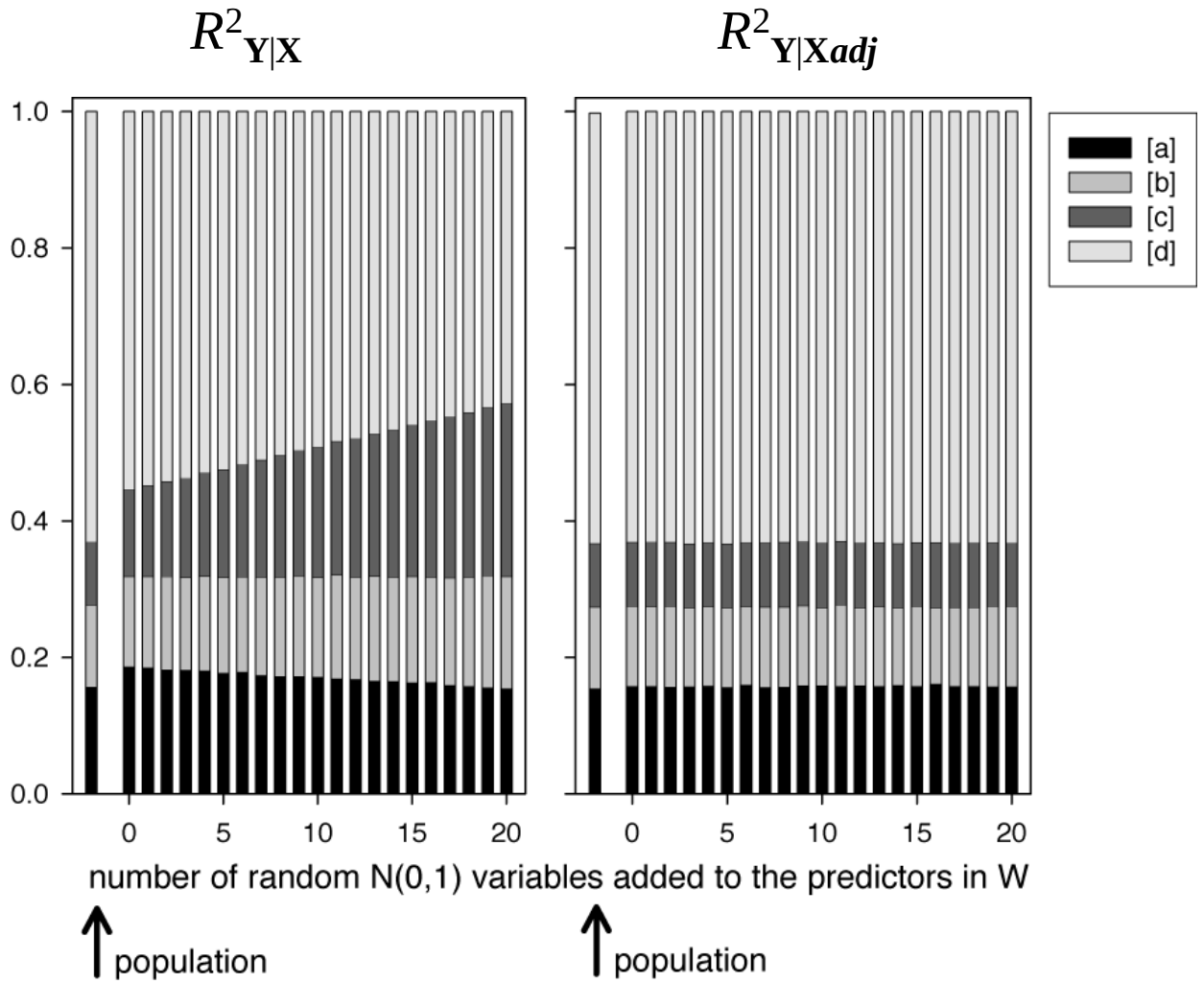


Fig. 5. Hellinger-transformed species-like data: influence of random predictors added to 6 active predictors, and of sample size, on $R^2_{Y|X}$ and $R^2_{Y|Xadj}$ (means after 1000 computer experiments). Left panel: n = 100; right: $20 \leq n \leq 210$.

Variation partitioning of Hellinger-tr. community composition data



Result 3 – For *Hellinger-transformed community composition data*, the adjusted bimultivariate redundancy statistic, $R^2_{\mathbf{Y|X}adj}$, is a quasi-unbiased estimator of the population $R^2_{\mathbf{Y|X}}$ in RDA.

It also correctly estimates the fractions of variation [a], [b], [c], and [d] in variation partitioning.

For *raw simulated species-like data*, $R^2_{\mathbf{Y|X}adj}$ overestimates the population $R^2_{\mathbf{Y|X}}$.

$\Rightarrow$ For estimation or variation partitioning, community composition data containing many zeros should always be transformed prior to the analysis.

The End

Testing the significance of the bimultivariate redundancy statistic

Miller (1975)[1] has shown that in the normal case, if the variables in **Y** are standardized before computation of RDA, the *F*-statistic associated to $R^2_{\mathbf{Y}|\mathbf{X}}$

$$F = \frac{R^2_{\mathbf{Y}|\mathbf{X}} / m}{(1 - R^2_{\mathbf{Y}|\mathbf{X}}) / (n - m - 1)}$$

is distributed like the Fisher-Snedecor *F*-distribution with mp and $p(n-m-1)$ degrees of freedom, where p is the number of response variables and m is the number of explanatory variables.

If the variables in **Y** are not standardized, the *F*-statistic is not distributed like the Fisher-Snedecor *F*-distribution. This is the case even in the normal case. A permutation test is required.

[1] Miller, J. K. 1975. The sampling distribution and a test for the significance of the bimultivariate redundancy statistic: a Monte Carlo study. *Multivariate Behavioral Research* 10: 233-244.

Selection of explanatory variables

Significance of adjustment (R^2) of a PCNM model is tested using the full set of PCNM variables, without selection of any kind.

Before representing the fractions of variation in biplots, it is customary to proceed to forward selection of the explanatory variables. Parsimonious models are obtained for the biplots. Selection is carried out separately for the environmental variables and the orthogonal PCNM base functions.

We know that forward selection is too liberal, incorporating too many variables in the model. Simulation work is in progress (Peres-Neto, Legendre & Dray) to find additional criteria that will limit the incorporation of explanatory variables to those that really have an effect on the response table.