Aquat Ecol (2017) 51:131–143
DOI 10.1007/s10452-016-9604-1

Response variable selection in principal response curves using permutation testing

Nadia J. Vendrig · Lia Hemerik · Cajo J. F. ter Braak

Received: 3 May 2016 / Accepted: 30 September 2016 / Published online: 15 October 2016

Abstract Principal response curves analysis (PRC) is widely applied to experimental multivariate longitudinal data for the study of time-dependent treatment effects on the multiple outcomes or response variables (RVs). Often, not all of the RVs included in such a study are affected by the treatment and RV-selection can be used to identify those RVs and so give a better estimate of the principal response. We propose four backward selection approaches, based on permutation testing, that differ in whether coefficient size is used or not in ranking the RVs. These methods are expected to give a more robust result than the use of a straightforward cut-off value for coefficient size. Performance of all methods is demonstrated in a simulation study using realistic data. The permutation testing approach that uses information on coefficient size of RVs speeds up the algorithm without affecting its performance. This most successful permutation testing approach removes roughly 95 % of the RVs that are unaffected by the treatment irrespective of the characteristics of the data set and, in the simulations, correctly identifies up to 97 % of RVs affected by the treatment.

Keywords Principal response curves · multivariate analysis · variable selection · permutation testing · longitudinal data · multivariate time series

Electronic supplementary material The online version of this article (doi:10.1007/s10452-016-9604-1) contains supplementary material, which is available to authorized users.

Handling Editor: Piet Spaak.

N. J. Vendrig (✉) · L. Hemerik · C. J. F. ter Braak
Biometris, Wageningen University & Research,
P.O. Box 16, 6700 AA Wageningen, The Netherlands
e-mail: nadia.vendrig@wur.nl

L. Hemerik
e-mail: lia.hemerik@wur.nl

C. J. F. ter Braak
e-mail: cajo.terbraak@wur.nl

Introduction

In ecological research, the effect of a treatment is often assessed for several response variables (RVs) at several points in time. This results in multivariate longitudinal data, also called multivariate time series data. For instance, if we wish to assess how invertebrate communities in ditches change as a result of a single application of a certain pesticide, we would select a number of ditches (experimental sites), assign every ditch to a treatment of a dose of pesticide or a control treatment, and measure the abundances of the invertebrate species living in the ditches at several times before and after treatment. Abundance of invertebrates is influenced not only by our treatment but also by the moment of sampling due to external factors such as the time of year. Principal response curves analysis (PRC) (Van den Brink and Ter Braak 1998, 1999) removes these unwanted time effects; succinctly describes the

time-dependent overall response of the community to the treatment(s) relative to the control treatment; and indicates for each of the species whether their response is positively or negatively correlated to the overall response and to which extent.

PRC is a special case of redundancy analysis (RDA) used to describe experimental multivariate longitudinal data. It estimates differences among treatments on a collection of RVs over time and the extent to which the response of those individual RVs resembles the overall response. PRC has been widely applied in aquatic ecology and ecotoxicology (e.g., Hartgers et al. 1998; Cuppen et al. 2000; Roessink et al. 2006; Duarte et al. 2008; Verdonschot et al. 2015), terrestrial ecology and ecotoxicology (e.g., Heegaard and Vandvik 2004; Pakeman 2004; Britton and Fisher 2007; Moser et al. 2007), microbiology (e.g., Andersen et al. 2010; Fuentes et al. 2014) and soil science (e.g., Kohler et al. 2006; Cardoso et al. 2008).

The main results of PRC are two sets of coefficients visualized in two easily interpretable graphs. The first set consists of the dose-time coefficients (c_{dt}s) estimated for each combination of the treatment levels ($d = 1, \ldots, D$) and the time-points ($t = 1, \ldots, T$). The c_{dt}s represent the effect size of treatment d at time t relative to the reference treatment at the same time. Thus, by definition, $c_{dt} = 0$ for the reference treatment. The reference treatment is often the control treatment, but the choice of reference treatment does not affect the estimates of differences between treatments; it merely defines the baseline, i.e., relative to which treatment the results are presented. The c_{dt}s are depicted in the principal response curves, a line-plot of c_{dt}s against time grouped by treatment (Fig. 1). The second set of coefficients are the weights for the RVs (b_ks) estimated for each of the RVs ($k = 1, \ldots, K$). They represent the resemblance of RV k to the overall response pattern specified by the principal response curves (i.e., the c_{dt}s) and are typically depicted on a vertical bar alongside the line-plot. The further b_k is from zero, the more the response pattern of RV k resembles the overall response pattern (if $b_k > 0$) or the negative overall response pattern (if $b_k < 0$). A b_k of zero indicates that the expected value of RV k at time t does not differ between treatments or is uncorrelated with the overall response pattern.

The c_{dt}s and b_ks can be used to rank dose-time combinations or RVs, respectively. For instance, if $|c_{23}| > |c_{24}|$, the estimated treatment effect for treatment 2 is larger at time-point 3 than at time-point 4. The coefficients, however, do neither have a unit nor a direct interpretation. The coefficients are estimated under the assumption that $\pi_{tdk} = c_{dt} b_k$, where π_{tdk} is the difference in expected value of RV k, at time t between treatment d and the reference treatment. The expected value y_{tdk} of RV k, at time t under treatment d, is thus estimated as $y_{tdk} = a_{tk} + \pi_{tdk}$, where a_{tk} is the expected value in the reference group.

Standard PRC assumes that only one factor (e.g., treatment) is relevant, while other (environmental) factors are either as similar as possible or, if not, randomized by design of the experiment. PRC has also been applied to monitoring sites where this assumption is more problematic. It should be noted that it is possible to adjust for unwanted variation between sites if this variation is due to one or more measured environmental variables. The environmental variables can be included as covariates in addition to the factor time which is the default covariate in PRC. This possibility is not yet available in Vegan (Oksanen et al. 2015), a much used R-package that includes a PRC-function, but it is available in Canoco 5 (Šmilauer and Lepš 2014), a computer program for multivariate statistical analysis using ordination. An example is given in Fuentes et al. (2014).

When PRC is applied in aquatic ecology, the research interest typically is the response of a community as a whole to a treatment and the set of RVs thus typically consists of abundance data on all available species or taxa (e.g., all taxa of invertebrates) at an experimental site. RVs are included irrespective of their expected susceptibility to the treatment beforehand, and a large proportion of the included RVs could thus be unaffected by the treatment. PRC handles RVs that do not follow the response pattern (Noise-RVs) by assigning these RVs b_k-estimates close to zero which is advantageous in contrast to the use of, e.g., Bray-Curtis Similarity (Bray and Curtis 1957) which is calculated with equal weights for all RVs (Van den Brink and Ter Braak 1998). But although inclusion of Noise-RVs in PRC does not add bias to c_{dt}-estimates, their inclusion introduces extra noise into the data set which adds extra imprecision to the estimates and reduces power. It would be advantageous to be able to point out which RVs are Noise-RVs. Reducing the data set accordingly would not only improve c_{dt}-estimation, it would also improve comparability of results of PRC between studies. As of

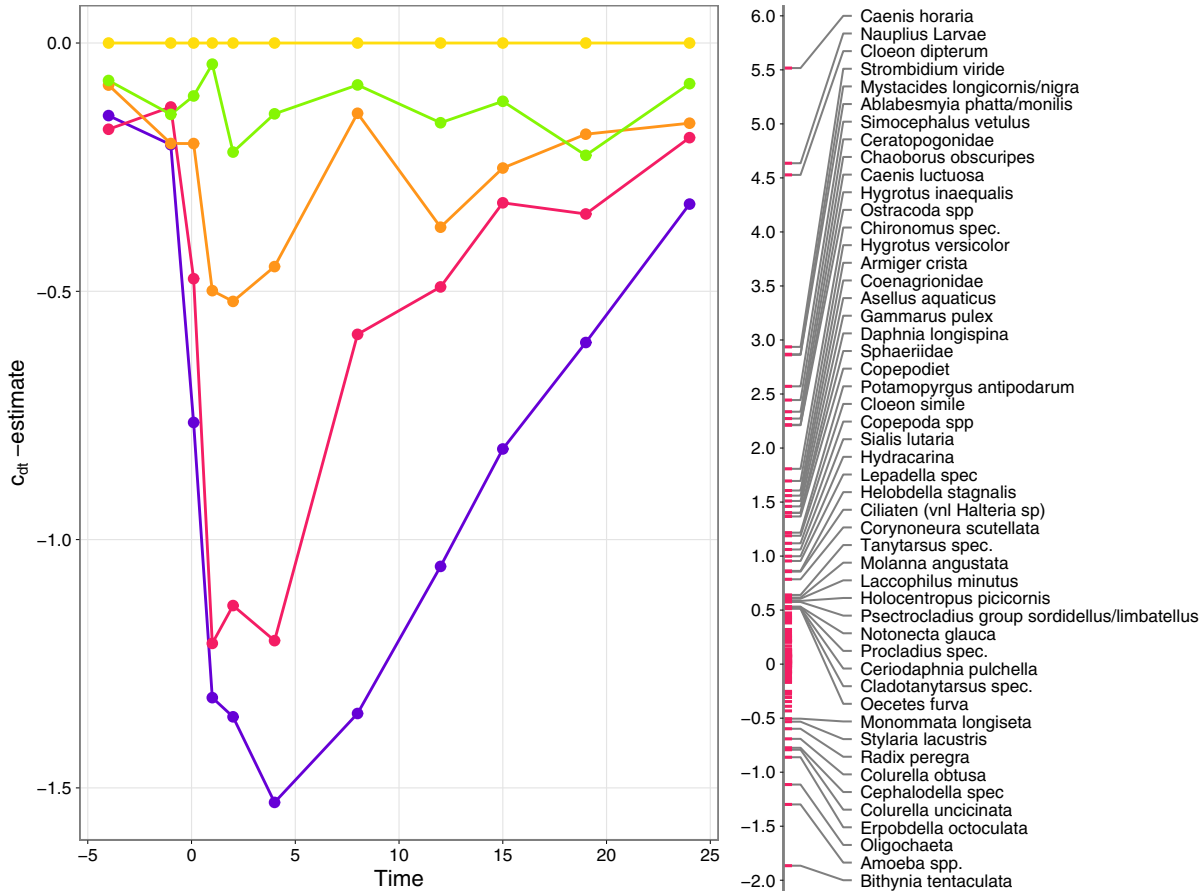


Fig. 1 Principal response curves (*left*) for the Pyrifos data (Van den Brink and Ter Braak 1999) for the different doses of Chlorpyrifos (0 *yellow circle*, 0.1 *green circle*, 0.9 *orange circle*, 6 *pink circle*, and 44 µg/L *purple circle*) with b_k-estimates (right). Only RVs with an absolute b_k-estimate above 0.5 are labeled. (Color figure online)

yet this is difficult because the coefficients have no unit, so only the shape of the principal response curves and the order of the species weights can be compared between studies. Reduction of the number of RVs in the analysis would also improve the readability of RV-weights graphs. At present, authors improve readability of the RV-weights graph by showing only RVs that exceed a certain threshold (mostly 0.5). Although effective in reducing the number of RVs, this practice is at best sub-optimal because b_k values (1) depend on the extent to which other RVs in the same data set are affected by the treatment, (2) are affected by the type of scaling used, and (3) are affected by the choice of standardization (see Online Resource 1 for details and illustrated examples on effect of these factors on b_k-estimates).

In this paper, we propose permutation testing approaches as an improved method for RV-selection in PRC. We further show that these approaches are robust to high residual correlation between RVs and to adding additional RVs with strong effect (very high b_k) or adding many RVs with no effect ($b_k = 0$) to the data set. We specifically show that information obtained from ranking RVs based on b_k scores of the full model can help accelerate the algorithm for variable selection without performance loss.

Materials and methods

Principal response curves analysis

PRC models the expected value of RV k at time t in treatment level d as the sum of three effects: (1) the expected value of the RV in the reference group a_{tk}, (2) the time-specific effect of treatment level (π_{tdk}), and

(3) an error term (ϵ_{ik}). The (multivariate) regression model for y_{ik}, i.e., the observed value of RV k in observation i (where $i = 1, \ldots, I$ with $I = T \cdot$number of experimental sites), is:

$$y_{ik} = \sum_{t=1}^{T} a_{tk} w_{it} + \sum_{t=1}^{T} \sum_{d=1}^{D} \pi_{tdk} z_{idt} + \epsilon_{ik} \quad (1)$$

where w_{it} and z_{idt} are indicator variables (0/1 or dummy variables) that indicate, respectively, whether (1) or not (0) observations are in the reference treatment and whether or not observations received dose d at time t. The general assumption of PRC is that $\pi_{tdk} = b_k c_{dt}$ which implies that b_k and c_{dt} can be estimated by partial RDA (i.e., reduced rank regression with concomitant variables) (Davies and Tso 1982) using Eq. 1. Note that, in contrast to what is written in Smilde et al. (2012) and in the appendix of Timmerman and Ter Braak (2008), a_{tk} is a free, unknown parameter of the model that is estimated by the partial RDA. Note that the estimation procedure also works with unbalanced data, as PRC fits in the regression framework which is more general than the ANOVA framework used by Smilde et al. (2012).

The estimates for c_{dt} and b_k are determined on an arbitrary scale because $c_{dt} b_k = \beta b_k * \frac{c_{dt}}{\beta}$, where β is an arbitrary scalar (i.e., any real number). As a result, the coefficients lack a unit and a direct interpretation and the scalar can be chosen such that it gives the coefficients the desired properties. In Canoco (Šmilauer and Lepš 2014), the first software package to include PRC, the default is to scale coefficients such that the mean square of b_k-estimates is 1 and we used this scaling in Fig. 1. The result is that, *ceteris paribus*, larger true treatment effects result in larger absolute estimates of c_{dt}. The b_k-estimates are expected to fall roughly between -3 and 3, independent of treatment effect. Therefore, when applying this scaling one could opt to select RVs based on a cut-off value of absolute b_k (usually 0.5).

This approach, which we will refer to as Naive RV-selection (Naive RVS), has some pitfalls. We wish to distinguish RVs affected by the treatment (Effect-RVs) from RVs that are uncorrelated to the overall response pattern. Such RVs are either unaffected by the treatment (Noise-RVs) or contribute to minor response patterns. In a situation with only Noise-RVs however, due to scaling, some Noise-RVs will get a b_k-estimate above the cut-off value. Vice versa, scaling causes the b_k-estimate of an Effect-RV to be lower when a very strongly affected Effect-RV is in the data set than when that strongly affected RV is not in the data set. As a result, including a very strongly affected RV to the data set could result in b_k-estimates of other RVs to drop below the cut-off value. Another pitfall is that Naive RVS has little value when coefficients are scaled differently. Coefficients could for instance be scaled such that mean square of $\widetilde{c}_{dt}$s is 1, where $\widetilde{c}_{dt}$s are a centered version of the c_{dt}s. In Vegan (Oksanen et al. 2015) the default option scales the coefficients differently with both the b_ks and c_{dt}s showing effect sizes. For any of these scaling-methods, choosing a cut-off value in advance does not make sense.

Response variable selection protocols

Ideally, an RVS protocol would make perfect predictions and thus remove all the Noise-RVs from the model and keep all the Effect-RVs in the model. Such a result is not feasible in practice. Therefore, we aim at achieving an optimal, yet realistic method for RVS, in which every Noise-RV has a $1 - \alpha$ probability to be removed from the model (e.g., $\alpha = 0.05$) while keeping as many Effect-RVs in the model as possible. With this aim there is no need to correct for multiplicity in statistical testing of RVs (such as Bonferroni) in the RVS protocols that we propose.

For any RV k, the hypothesis that its expected value is independent from the treatment (i.e., whether or not $b_k = 0$) can be tested by calculating a permutation p value and comparing it to α. A permutation p value for RV k is obtained by performing 500 permutations in which time series of observations from RV k on the same experimental unit (e.g., ditch, plot, or site) are permuted between treatments (including the control treatment). We estimate b_k in PRC on non-permuted data and on all 500 permuted data sets. The permutation p value is the proportion of the 501 estimated b_ks (including the b_k from non-permuted data) greater than or equal to the estimated b_k from PRC with non-permuted data, if the estimated b_k from the PRC with non-permuted data is positive. If the estimated b_k from PRC with non-permuted data is negative, the proportion equal or lower is used. The number of 500 is large enough to provide sufficient power at $\alpha = 0.05$ and is still acceptable in terms of computing time.

As an alternative to Naive RVS, we propose four RVS protocols based on permutation testing (in short: permutation RVS protocols) that all incorporate permutation p value calculation as described above. All four permutation RVS protocols are backward procedures, indicating that they start with the whole set of RVs and predict which of those are Noise-RVs that can be removed from the model and which are Effect-RVs that should be kept.

Two-Step RVS The most thorough permutation RVS protocol is the Two-Step RVS. In this protocol, we calculate a permutation p value for all RVs in the data set. If any of the permutation p values is higher or equal to α, the RV with the highest permutation p value is removed from the model. Thereafter, we repeat the procedure with the remaining RVs and keep repeating until only RVs with a permutation p value lower than α remain. The advantage of this elaborate approach is that it accounts for RVs being correlated. The pitfall is that it is computationally intensive because many permutation p values need to be calculated (e.g., for $K = 200$; as many as $0.5(K^2 + K) = 20,100$).

Screening RVS We could do with a less computationally intensive protocol if it would be reasonable to assume that the permutation p value of an RV is independent of the other RVs in the data set. This simpler protocol, called the Screening RVS protocol, calculates a permutation p value once for each RV in the data set using the full model. All RVs with permutation p values higher or equal to α are removed from the model at once.

Stepwise RVS Importantly, estimated b_ks of Noise-RVs are expected to be closer to zero than estimated b_ks of Effect-RVs. Thus, to incorporate this information, a third RVS approach uses an even less computationally intensive procedure. This protocol, called the Stepwise RVS protocol, performs PRC on the data set, selects the RV with the estimated b_k closest to zero, and calculates a permutation p value for that RV. If that permutation p value is higher or equal to α, it removes the RV from the model. If it is not, it keeps the RV in the model and calculates the permutation p value of the RV with the estimated b_k second closest to zero. Once an RV is kept in the model, its permutation p value is not calculated again. Stepwise RVS is computationally less intensive than Screening RVS because the PRC-procedure, which is performed 501 times per permutation p value, gets faster with a smaller number of RVs in the model. In Stepwise RVS, permutation p values are calculated using PRC on the reduced model with increasingly less RVs as the procedure progresses, whereas, in Screening RVS, all permutation p values are calculated using PRC on the full set of RVs.

Stepwise Stop RVS When we are willing to assume that all RVs with an absolute estimated b_k under a certain threshold are Noise-RVs, we can make an even faster version of theStepwise RVS protocol: the Stepwise Stop RVS protocol. This protocol is the same as the Stepwise RVS protocol, except that it stops entirely when the first permutation p value lower than α is encountered.

Simulation study

We evaluated the performance of the four permutation testing protocols and Naive RVS in a simulation study. The data used in this simulation study were modeled after the so-called Pyrifos data set. The Pyrifos data set, used as example throughout this paper, consists of log-transformed abundance data obtained from a toxicological experiment in outdoor experimental ditches, explained in detail by van Wijngaarden et al. (1996) and Van den Brink et al. (1996). In the experiment, experimental ditches were randomly allocated to the reference treatment or a dose of insecticide chlorpyrifos. The RVs are abundances of species of invertebrates. In this simulation study, we generated data from scenarios inspired by the Pyrifos-experiment. In the Pyrifos-like data scenario, an experiment was conducted in which the effects of three levels of treatment (reference, low and high dose) were measured on four independent locations per treatment at five different time-points. The Pyrifos-like data contain abundance data of 100 RVs, 50 of which are Noise-RVs which are unaffected by the treatment ($b_k = 0$) and 50 are Effect-RVs which have a low, medium, high or reversed low treatment effect ($b_k = 1,2,3,$ or -1). Covariance between time-points is auto-regressive and covariance between RVs resembles covariance in the Pyrifos data set. Error terms were simulated using a multivariate normal distribution. We back-transformed the sum of the structural effect and the error term to the abundance-scale, used it as expected value for a random draw

from a Poisson-distribution, and log-transformed the result (for more details: Online Resource 2).

To provide additional experimental outcomes that approximated the range of treatment effects in the literature, we also generated data based on 17 data scenarios similar to the Pyrifos-like data scenario with one or two parameters manipulated. We manipulated the composition of the set of Effect-RVs, the number of Noise-RVs, the number of ditches, the amount of covariance between RVs, and the treatment-effect size. For an overview, see Table 1.

For each of the 18 data scenarios, 100 data sets were generated which were centered before analysis (Centering). We also analyzed each data set after standardizing data per RV (Standardization) resulting in another 18 simulation scenarios. Standardization in addition to Centering is useful when it is of interest whether RVs are affected by a treatment (positively, negatively, or not at all) and not so much what the size of the difference in effect between RVs is. For Naive RVS, coefficients were scaled such that mean squares of b_k are 1 as this is the only scaling that is sensible for this protocol. Scaling of coefficients does not affect the RV-selection in the permutation RVS protocols.

Performance of the RVS protocols was evaluated using sensitivity and specificity. Sensitivity is the number of Effect-RVs kept in the model divided by the total number of Effect-RVs in the data set. Specificity is the number of Noise-RVs removed from the model divided by the total number of Noise-RVs in the data set. Permutation method is expected to have a specificity of 0.95 with $\alpha = 0.05$, indicating that 5 % of saved RVs could in fact be Noise-RVs. In the ideal situation, sensitivity would be 1, indicating that all effect-RVs are identified. In practice, we would expect sensitivity to increase with increasing power, e.g., with larger effect size or more observations.

There is a trade-off between specificity and sensitivity which becomes apparent when comparing both Stepwise RVS procedures. All RVs removed in the Stepwise Stop RVS procedure are also removed in the Stepwise RVS procedure. In the Stepwise RVS procedure, some additional RVs could be removed. Stepwise Stop RVS thus always keeps the same or more Effect-RVs in the model than Stepwise RVS and thus has an equal or higher sensitivity. Stepwise Stop RVS always removes the same number or less Noise-RVs from the model than Stepwise RVS and thus has an equal or lower specificity.

The overall quality of RVS protocols was evaluated with the Matthews correlation coefficient (M_c) (Matthews 1975) which is a correlation coefficient between a prediction and the reality:

$$M_c = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})} \quad (2)$$

where TP (true positives) is the number of kept Effect-RVs, TN (true negatives) is the number of removed Noise-RVs, FP (false positives) is the number of kept Noise-RVs, and FN (false negatives) is the number of removed Effect-RVs. The M_c ranges between -1 and 1 where 1 indicates perfect prediction (i.e., all Noise-RVs removed, all Effect-RVs kept), 0 indicates prediction no better than random, and -1 indicates

Table 1 Overview of data scenarios in the simulation study with three treatments, incl. control, at five time-points

Data scenario	Description
Pyrifos-like	As described in "Simulation study" section (4 replications, 50 effect-RVs, 50 Noise-RVs)
More ditches	As Pyrifos-like, with 4 additional ditches per treatment (8 total)
Most ditches	As Pyrifos-like, with 8 additional ditches per treatment (12 total)
Weak effect-RVs	As Pyrifos-like, with effect-RVs consisting of 38 RVs with $b_k = 1$ and 12 RVs with $b_k = -1$
Strong effect-RVs	As pyrifos-like, with 12 additional strong effect-RVs with $b_k = 10$
One Noise-RV	As Pyrifos-like, with only 1 Noise-RV
Many Noise-RVs	As Pyrifos-like, with 150 additional Noise-RVs (200 total)
No covariance	As Pyrifos-like, except there is no covariance between RVs
More covariance	As Pyrifos-like, with 40 % higher correlation between RVs
<name of data scenario>$^+$	All nine data scenarios described above, with a larger treatment effect ($c_{dt}{}^+ = 4c_{dt}$)

total disagreement between prediction and reality (i.e., all Noise-RVs kept, all Effect-RVs removed).

The effect of RVS on model fit was evaluated in terms of difference in residual mean squared error ($RMSE_{diff}$). RMSE of the reduced model ($RMSE_{reduced}$) was compared to RMSE of the reduced set of RVs calculated using fitted values from the full model ($RMSE_{full}$).

After evaluating performance of the RVS protocols, we applied the best protocol to the Pyrifos data as a case study. In order to better compare the shapes of PRC on the full and the reduced data set, we scaled such that the population variance of all available case scores $\{x_i = c_{dt}z_{idt}\}$ was 1. For balanced data, this corresponds to setting the mean square of $\widetilde{c}_{dt}$s to 1. The scaling such that the mean square of b_k is 1 always results in higher b_k-estimates and lower c_{dt}-estimates when comparing results before to after removing Noise-RV, because Noise-RVs typically have low b_k-estimates. All data simulations and analyses were performed in R 3.1.0. The scripts to replicate the case study are available in Online Resource 3.

Results

General results

In our simulation study, we assessed sensitivity, specificity, and M_c of the Two-Step, Screening, Stepwise, and Stepwise Stop permutation RVS protocols and Naive RVS. The aim was to find an RV-selection method that is 0.95 specific while being as sensitive as possible. Computing time of the Two-Step RVS protocol was extremely long. Analysis of one data set generated using the Pyrifos-like data scenario took on average 2 h and 24 min, whereas Screening RVS took 3 min 50 s, Stepwise RVS took 2 min and 48 s, and Naive RVS took less than a second. Therefore, Two-Step RVS was run on 12 rather than 100 data sets per scenario. The results thereof gave no reason to assume that Two-Step RVS outperformed Screening or Stepwise RVS. On the contrary, based on confidence intervals around the mean, we found that mean specificity in the Two-Step RVS was different from 0.95 in 7 out of 36 simulation scenarios, whereas for Screening and Stepwise RVS, also based on 12 iterations, mean specificity was different from 0.95 in respectively 3 and 0 out of 36 data scenarios. As a result, we decided to base results of the Two-Step RVS on 12 iterations and not report the results in text.

Based on 100 data sets per scenario, we concluded that Screening and Stepwise RVS hardly differed in specificity and sensitivity. Per scenario, the difference between methods in mean specificity ranged from -0.020 to 0.030 and the difference in mean sensitivity ranged from -0.011 to 0.006. The Stepwise Stop RVS protocol did not meet the requirement of being 0.95 specific. The 95 % confidence interval of mean specificity excluded 0.95 in all of the 36 simulation scenarios. Therefore, we will only report on results from Stepwise RVS in text which we will compare to results from Naive RVS. Full results for all methods and all simulation scenarios can be found in Online Resource 4 in Table 1–4.

The overall quality of prediction M_c of both Stepwise RVS and Naive RVS (from 0.25 to 0.92) was moderately to highly positive except in the Weak Effect-RVs data scenarios (due to very low power) and One Noise-RV data scenarios (due to specificity of either 0 or 1) for both Stepwise and Naive RVS, and in Many Noise-RVs data scenarios using Naive RVS. $RMSE_{diff}$, the difference between $RMSE_{full}$ and $RMSE_{reduced}$, was not large and did not differ much between the RVS protocols, indicating that removing RVs from the model with RV-selection did not influence model predictions for RVs kept in the model much. In the data scenarios with Pyrifos-like treatment effect, $RMSE_{diff}$ ranged from -0.142 to 0.066 and in the data scenarios with increased treatment effects (such as Pyrifos-like^{+}) $RMSE_{diff}$ ranged from -0.341 to 0.068.

Comparing mean M_c within the same simulation scenario, M_c of Stepwise RVS was higher than Naive RVS in all but 5 out of 36 simulation scenarios (difference from -0.05 to 0.25, mean = 0.05). The main difference in performance of both methods lies in the trade-off between specificity and sensitivity. Stepwise RVS was more successful than Naive RVS in identifying the vast majority of Noise-RVs, as judged from the mean specificity results per simulation scenario. Mean specificity of Stepwise RVS was consistently high (from 0.87 to 0.95) and its 95 % confidence interval included 0.95 in 23 out of 36 simulation scenarios, whereas mean specificity of Naive RVS was highly varying (from 0.37 to 1) and its 95 % confidence interval never included 0.95. For both Stepwise RVS and Naive RVS, mean specificity

approached 0.95 more closely with increasing power. In Stepwise RVS, the 95 % confidence interval included 0.95 more often in data scenarios with larger treatment effect (16 out of 18) than in data scenarios with Pyrifos-like treatment effect (7 out of 18). For Naive RVS, mean specificity of scenarios with was higher than of scenarios without larger treatment effects (e.g., compare Pyrifos^{+} to Pyrifos-like), the difference ranged from 0.08 to 0.43 (mean 0.31). Mean specificity also increased with increasing sample size (difference between Pyrifos-like, More Ditches, and Most Ditches data scenarios; Online Resource 4; Fig. 1). Mean sensitivity is highly variable for both Stepwise (from 0.17 to 0.97) and Naive RVS (from 0.35 to 0.95). For Stepwise RVS, mean sensitivity increases when the analysis has more power (due to larger treatment effects or increased sample size). Such a straightforward relationship could not be found for Naive RVS. Mean sensitivity between simulation scenarios with and without larger treatment effects did not increase in all cases and was not clearly affected by increasing the sample size.

Standardization rather than only Centering did not affect results of Stepwise RVS regarding specificity (difference −0.06 to 0.0006) and sensitivity (from −0.006 to 0.017) to great extent. For Naive RVS, Standardization in addition to Centering resulted in lower mean specificity (from −0.02 to −0.25; mean −0.10) and higher mean sensitivity (from 0.01 to 0.32; mean 0.11).

Results of Stepwise RVS are more robust to changes in the composition of the set of RVs than results of Naive RVS. Mean specificity and sensitivity changed less than 0.05 point after adding additional strong Effect-RVs to the Pyrifos-like data set (Strong Effect-RVs; Fig. 2) and after removing or adding Noise-RVs (One Noise-RV and Many Noise-RVs; Online Resource 4, Fig. 2). Note that we calculated specificity and sensitivity of the Strong Effect-RVs data scenario without including results on the additional strong Effect-RVs as to better compare results to the Pyrifos-like data scenario. Using Naive RVS, specificity increased and sensitivity decreased comparing Pyrifos-like to Strong Effect-RVs simulations scenarios. Comparing the One Noise-RV to the Many Noise-RVs data scenario, specificity decreased and sensitivity slightly increased. These changes are smaller when using Standardization in addition to Centering.

We found that both Stepwise and Naive RVS do not differ in performance between the No Covariance, Pyrifos-like, and More Covariance data scenarios (Online Resource 4, Fig. 3). This indicates that covariance in the residuals is not reflected in the b_k-estimates which confirms that PRC deals with this issue well.

Case study

Stepwise RVS on the Pyrifos data reduced the set of RVs from 178 to 38 species (Fig. 3). The shape of the principal response curves was mildly affected (Fig. 4). In general, the shape after RVS seems slightly smoother and the unexpected W-shape around Time = 2 of the 6 µg/L dose before RVS has disappeared.

When scaling such that mean square of b_k is 1, species with an absolute b_k-estimate over 0.5 in the full model were more likely to be in the reduced model (26 out of 50; 52 %) than species with an absolute b_k-estimate under 0.5 (12 out of 128; 9.4 %).

Discussion

The main reason to apply response variable selection (RVS) in PRC is to be able to distinguish between those species that do follow the principal response and those that do not. Standard PRC usually gives small coefficients to species of the latter group. By setting these coefficients actually to zero, that is, by removing these species, the noise in the data caused by these species is removed from the estimation of the principal response curves. The result is a better estimate of the true response when there were many Noise-RVs and as visibly suggested in the case study where the response curves were smoother after RVS.

One may argue that PRC after selection of response variables is a PRC of a subset of the species only and no longer the PRC of the whole community. We argue that it is still the PRC of the whole community, but one in which non-responding species received a zero coefficient. This differential weighing of species was already an advantage of PRC over similarity analysis (Van den Brink and Ter Braak 1998), but is an even bigger advantage in PRC with Stepwise RVS.

We found no differences in performance between the Two-Step, Screening, and Stepwise RVS protocols. In Two-Step RVS, RVs were removed from the model

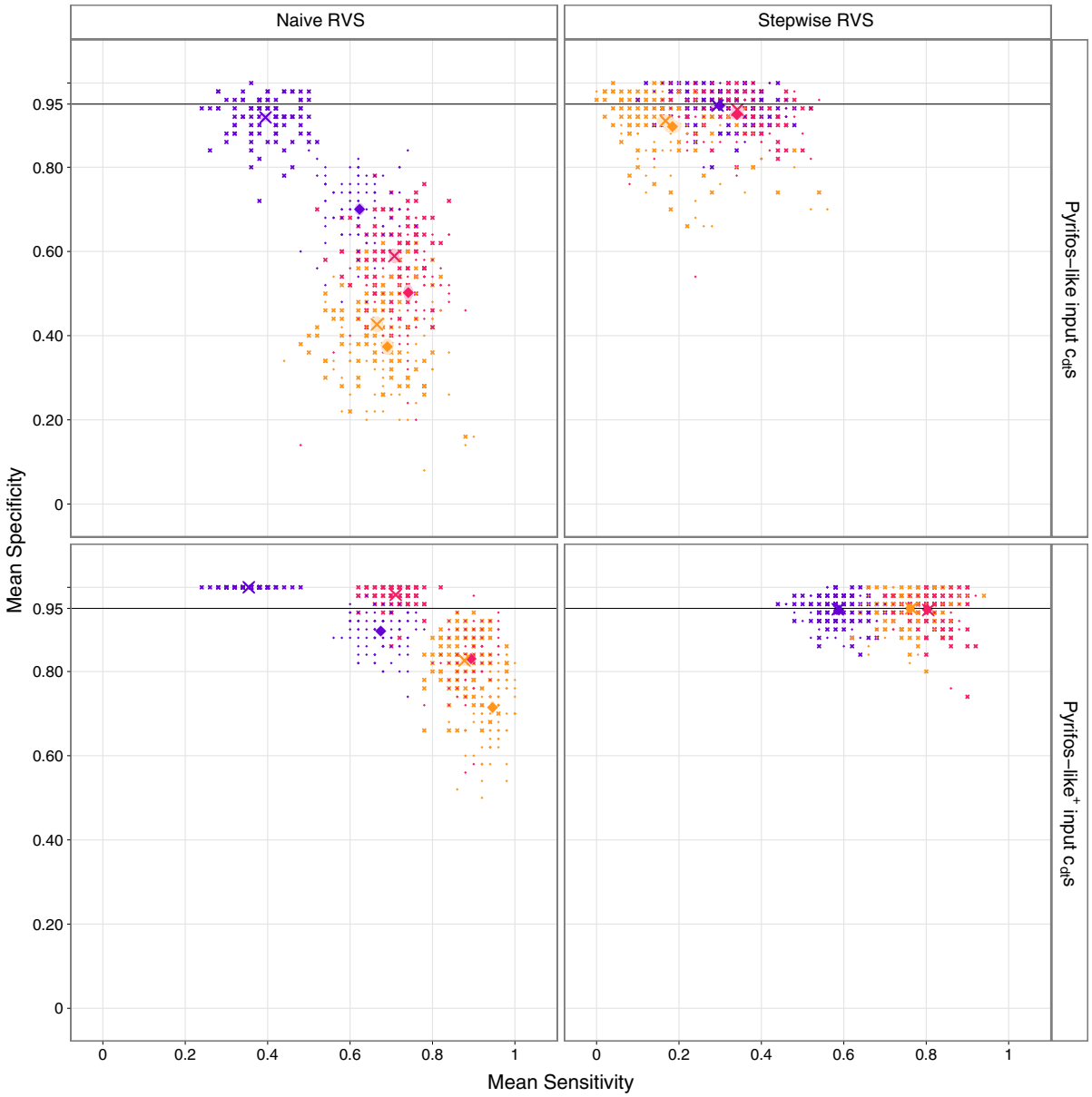
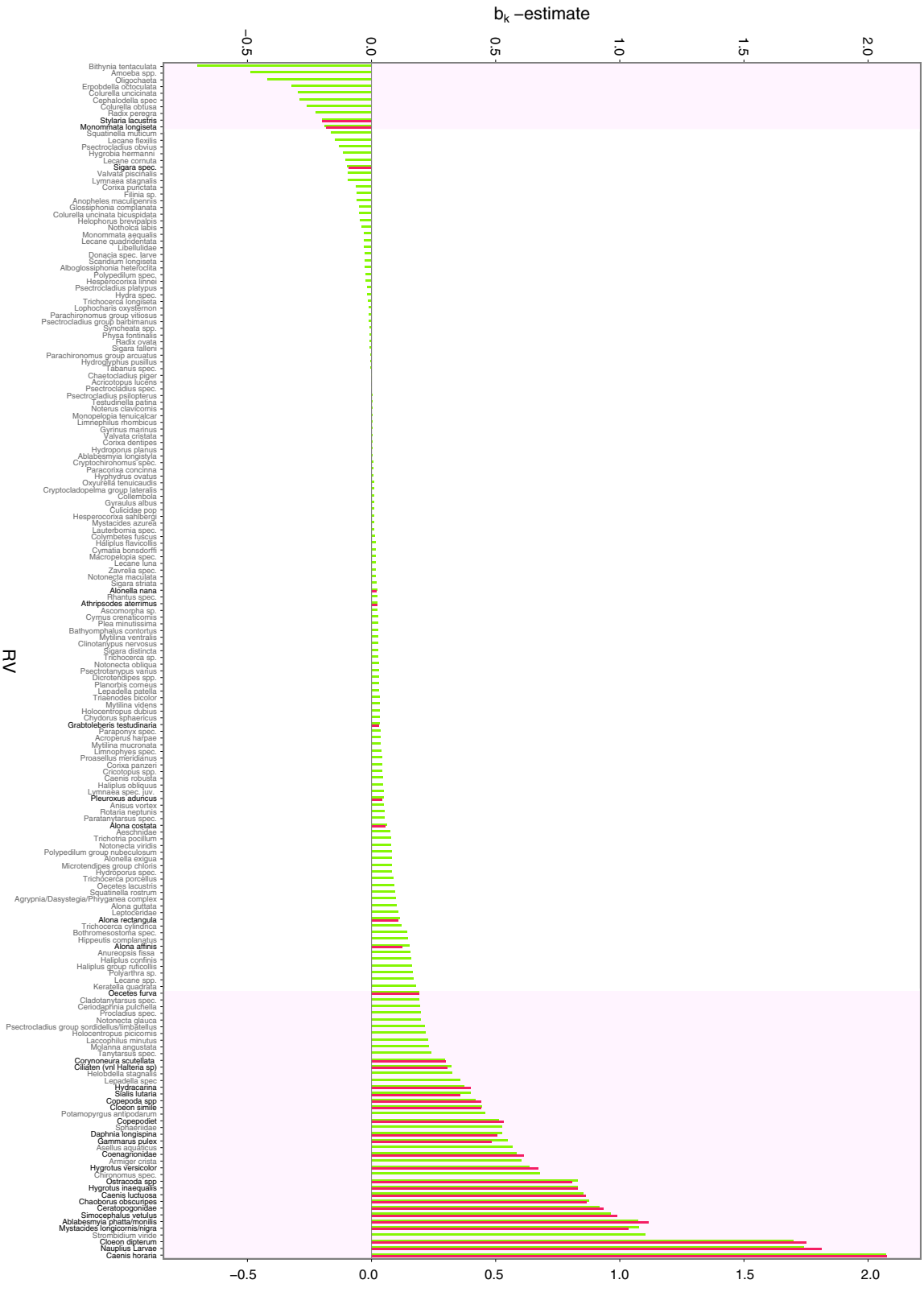


Fig. 2 Specificity and sensitivity of Naive and Stepwise RVS when applied to standardized (*points*) or centered (*crosses*) data generated using the Pyrifos-like/Pyrifos-like^{+} (*pink circle top row/bottom row*), strong effects RVs/strong effects RVs^{+} (*purple circle*), and weak effects RVs/weak effects RVs^{+} (*orange circle*) data scenarios. Mean specificity and sensitivity over 100 simulations are represented by *large symbols*, and specificity and sensitivity per simulation are represented by *small symbols*. *Ellipses* indicate the 95 % confidence region of the mean of the estimates. As the confidence regions are small the *ellipses* are difficult to see. (Color figure online)

one at a time, based on permutation *p* values that were recalculated every time an RV was removed from the model. In Screening RVS, permutation *p* values were calculated once for every RV using the full model. As Two-Step RVS did not yield better results than Screening RVS, we concluded that calculating permutation p values based on models with increasingly less Noise-RVs did not enhance performance. This conclusion was supported by the finding that adding additional Noise-RVs to or removing Noise-RVs from the data did not affect specificity and sensitivity of permutation RVS protocols.


b_k –estimate
RV
−0.5
0.0
0.5
1.0
1.5
2.0
Bithynia tentaculata
Amoeba spp.
Oligochaeta
Erpobdella octoculata
Colurella uncicinata
Cephalodella spec.
Colurella obtusa
Radix peregra
Stylaria lacustris
Monommata longiseta
Squatinella mutticum
Lecane flexilis
Psectrocladius obvius
Hygrobia hermanni
Lecane cornuta
Sigara spec.
Valvata piscinalis
Lymnaea stagnalis
Corixa punctata
Filinia sp.
Anopheles maculipennis
Glossiphonia complanata
Colurella uncinata bicuspidata
Helophorus brevipalpis
Notholca labis
Monommata aequalis
Lecane quadridentata
Libellulidae
Donacia spec. larve
Scaridium longiseta
Alboglossiphonia heterocilita
Polypedilum spec.
Hesperocorixa linnei
Psectrocladius platypus
Hydra spec.
Trichocerca longiseta
Lophocharis oxysternon
Parachironomus group vitiosus
Psectrocladius group barbimanus
Syncheata spp.
Physa fontinalis
Radix ovata
Sigara falleni
Parachironomus group arcuatus
Hydroglyphus pusillus
Tabanus spec.
Chaetocladius piger
Acricotopus lucens
Psectrocladius spec.
Psectrocladius psilopterus
Testudinella patina
Noterus clavicornis
Monopelopia tenuicalcar
Limnephilus rhombicus
Gyrinus marinus
Valvata cristata
Corixa dentipes
Hydroporus planus
Ablabesmyia longistyla
Cryptochironomus spec.
Paracorixa concinna
Hyphydrus ovatus
Oxyurella tenuicaudis
Cryptocladopelma group lateralis
Collembola
Gyraulus albus
Culicidae pop
Hesperocorixa sahlbergi
Mystacides azurea
Lauterborniа spec.
Colymbetes fuscus
Haliplus flavicollis
Cymatia bonsdorfi
Macropelopia spec.
Lecane luna
Zavrelia spec.
Notonecta maculata
Sigara striata
Alonella nana
Rhantus spec.
Athripsodes aterrimus
Ascomorpha sp.
Cymus crenaticornis
Plea minutissima
Bathyomphalus contortus
Mytilina ventralis
Clinotanypus nervosus
Sigara distincta
Trichocerca sp.
Notonecta obliqua
Psectrotanypus varius
Dicrotendipes spp.
Planorbis corneus
Lepadella patella
Triaenodes bicolor
Mytilina videns
Holocentropus dubius
Chydorus sphaericus
Grabtoleberis testudinaria
Paraponyx spec.
Acroperus harpae
Mytilina mucronata
Limnophyes spec.
Proasellus meridianus
Corixa panzeri
Cricotopus spp.
Caenis robusta
Haliplus obliquus
Lymnaea spec. juv.
Pleuroxus aduncus
Anisus vortex
Rotaria neptunis
Paratanytarsus spec.
Alona costata
Aeschnidae
Trichotria pocillum
Notonecta viridis
Polypedilum group nubeculosum
Alonella exigua
Microtendipes group chloris
Hydroporus spec.
Trichocerca porcellus
Oecetes lacustris
Squatinella rostrum
Agrypnia/Dasystegia/Phryganea complex
Alona guttata
Leptoceridae
Alona rectangula
Trichocerca cylindrica
Bothromesostoma spec.
Hippeutis complanatus
Alona affinis
Anureopsis fissa
Haliplus confinis
Haliplus group ruficollis
Polyarthra sp.
Lecane spp.
Keratella quadrata
Oecetes furva
Cladotanytarsus spec.
Ceriodaphnia pulchella
Procladius spec.
Notonecta glauca
Psectrocladius group sordidellus/limbatellus
Holocentropus picicornis
Laccophilus minutus
Mglanna angustata
Tanytarsus spec.
Corynoneura scutellata
Ciliaten (vnl Halteria sp)
Helobdella stagnalis
Lepadella spec
Hydracarina
Sialis lutaria
Copepoda spp
Cloeon simile
Potamopyrgus antipodarum
Copepodiet
Sphaeriidae
Daphnia longispina
Gammarus pulex
Asellus aquaticus
Coenagrionidae
Armiger crista
Hygrotus versicolor
Chironomus spec.
Ostracoda spp
Hygrotus inaequalis
Caenis luctuosa
Chaoborus obscuripes
Ceratopogonidae
Simocephalus vetulus
Ablabesmyia phatta/monilis
Mystacides longicornis/nigra
Strombidium viride
Cloeon dipterum
Nauplius Larvae
Caenis horaria

◀**Fig. 3** b_k-Estimates for the Pyrifos data set (Van den Brink and Ter Braak 1999) before (*light green bars*) and after RV-selection using Stepwise RVS (*dark pink bars*) (scaled such that mean square of $\tilde{c}_{dt}$s is 1). Abbreviated names of the species are printed in *black* if kept and printed in *gray* if removed from the model. *Shaded areas* represent which RV would be kept when using Naive RVS (with the appropriate scaling). (Color figure online)

We concluded that permutation p values of RVs were independent of other RVs in the data set because the performance of Screening RVS did not differ from the other protocols. Furthermore, we found that adding additional residual covariance did not affect the quality of b_k-estimates. So we confirmed that PRC is robust against between-species covariance in the residual, even though residual covariance between species is ignored in estimating the PRC coefficients as PRC uses simple least-squares. This is in contrast to what we would expect when selecting predictors rather than RV, such as in multiple regression. In that situation, one would expect coefficients, and thus their p values, and model predictions to be altered as a result of selection.

Performance of Stepwise RVS did not differ from performance of Screening RVS except for being computationally less intensive. It is less intensive, as calculating permutation p values is faster in data sets with a smaller number of RVs, and Stepwise RVS calculates permutation p values using an ever smaller set of RVs. The order of deleted RVs was determined based on estimated b_k, which is a reasonable indicator of effect size. The Stepwise Stop RVS protocol was computationally even less intensive than Stepwise RVS. This method, however, does not meet the goal of 0.95 specificity. Therefore, Stepwise RVS was selected as the preferred permutation RVS protocol.

Stepwise RVS combined a stable high specificity with a sensitivity that increased with power. Its performance was unaffected by the number of Noise-RVs in the data set, additional covariance in the residuals, adding additional strong Effect-RVs,

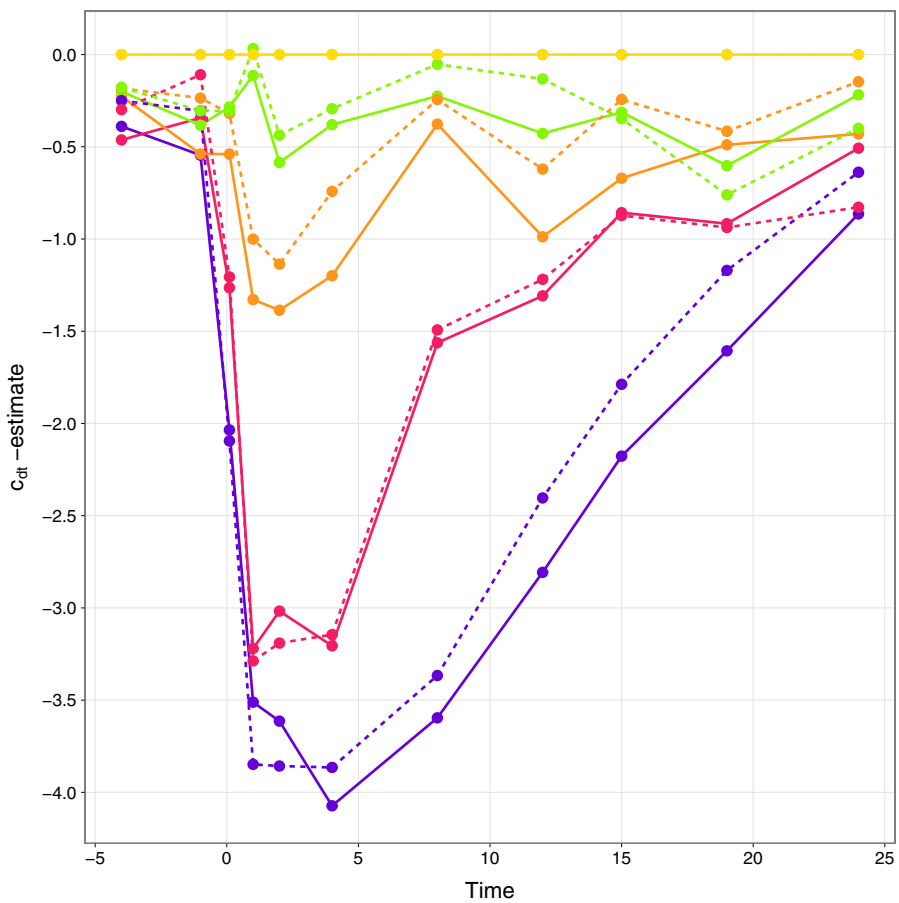


Fig. 4 Principal response curves for the Pyrifos data (Van den Brink and Ter Braak 1999) before (*solid line*) and after RV-selection using Stepwise RVS (*dashed line*) for the different doses of Chlorpyrifos (0 *yellow circle*, 0.1 *green circle*, 0.9 *orange circle*, 6 *pink circle*, and 44 µg/L *purple circle*) scaled such that mean square of $\tilde{c}_{dt}$s is 1. Note that the shape of the PRC before RV-selection is identical to the shape in Fig. 1. (Color figure online)

and the choice of Centering or Standardization of the data. In contrast, Naive RVS was highly variable in specificity and sensitivity and was affected by number of Noise-RVs in the data set and adding additional strong Effect-RVs. Because true b_k of RVs in data from practice are unknown, so is the performance of Naive RVS in terms of specificity and sensitivity. We therefore advise Stepwise RVS as the preferred method for RVS in PRC over Naive RVS. We see Stepwise RVS in PRC as an easy applicable and interpretable tool to enhance the insight in the response to treatment of a community over time.

Acknowledgments The research leading to these results has received funding from the Dutch Fund for Economic Structure Reinforcement (FES), under Grant Agreement Number 0908 (the "NeuroBasic PharmaPhenomics project").

References

Andersen R, Grasset L, Thormann MN, Rochefort L, Francez AJ (2010) Changes in microbial community structure and function following Sphagnum peatland restoration. Soil Biol Biochem 42(2):291–301. doi:10.1016/j.soilbio.2009.11.006

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of Southern Wisconsin. Ecol Monogr 27(4):325–349. doi:10.2307/1942268

Britton AJ, Fisher JM (2007) Interactive effects of nitrogen deposition, fire and grazing on diversity and composition of low-alpine prostrate Calluna vulgaris heathland. J Appl Ecol 44(1):125–135. doi:10.1111/j.1365-2664.2006.01251.x

Cardoso PG, Raffaelli D, Lillebø AI, Verdelhos T, Pardal MA (2008) The impact of extreme flooding events and anthropogenic stressors on the macrobenthic communities' dynamics. Estuar Coast Shelf Sci 76(3):553–565. doi:10.1016/j.ecssj.2007.07.026

Cuppen JG, Van den Brink PJ, Camps E, Uil KF, Brock TC (2000) Impact of the fungicide carbendazim in freshwater microcosms. I. Water quality, breakdown of particulate organic matter and responses of macroinvertebrates. Aquat Toxicol 48(2–3):233–250. doi:10.1016/S0166-445X(99)00036-3

Davies PT, Tso MKS (1982) Procedures for reduced-rank regression. Appl Stat 31(3):244. doi:10.2307/2347998

Duarte S, Pascoal C, Alves A, Correia A, Cássio F (2008) Copper and zinc mixtures induce shifts in microbial communities and reduce leaf litter decomposition in streams. Freshw Biol 53(1):91–101. doi:10.1111/j.1365-2427.2007.01869.x

Fuentes S, Van Nood E, Tims S, Heikamp-de Jong I, Ter Braak CJF, Keller JJ, Zoetendal EG, De Vos WM (2014) Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. ISME J 8(8):1621–1633. doi:10.1038/ismej.2014.13

Hartgers EM, Aalderink GH, Van den Brink PJ, Gylstra R, Wiegman JWF, Brock TCM (1998) Ecotoxicological threshold levels of a mixture of herbicides (atrazine, diuron and metolachlor) in freshwater microcosms. Aquat Ecol 32(2):135–152. doi:10.1023/A:1009968112009

Heegaard E, Vandvik V (2004) Climate change affects the outcome of competitive interactions? An application of principal response curves. Oecologia 139(3):459–466. doi:10.1007/s00442-004-1523-5

Kohler F, Gillet F, Gobat JM, Buttler A (2006) Effect of cattle activities on gap colonization in Mountain Pastures. Folia Geobot 41(3):289–304

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta Protein Struct 405(2):442–451. doi:10.1016/0005-2795(75)90109-9

Moser T, Römbke J, Schallnass HJ, Van Gestel CAM (2007) The use of the multivariate principal response curve (PRC) for community level analysis: a case study on the effects of carbendazim on enchytraeids in terrestrial model ecosystems (TME). Ecotoxicology 16(8):83–573. doi:10.1007/s10646-007-0169-6

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2015) vegan: Community Ecology Package. http://cran.r-project.org/package=vegan

Pakeman RJ (2004) Consistency of plant species and trait responses to grazing along a productivity gradient: a multi-site analysis. J Ecol 92(5):893–905. doi:10.1111/j.0022-0477.2004.00928.x

Roessink I, Crum SJH, Bransen F, van Leeuwen E, van Kerkum F, Koelmans AA, Brock TCM (2006) Impact of triphenyltin acetate in microcosms simulating floodplain lakes. I. Influence of sediment quality. Ecotoxicology 15(3):267–293. doi:10.1007/s10646-006-0058-4

Šmilauer P, Lepš J (2014) Multivariate analysis of ecological data using CANOCO 5, 2nd edn. Cambridge University Press, Cambridge. doi:10.1017/CBO9781139627061

Smilde AK, Timmerman ME, Hendriks MMWB, Jansen JJ, Hoefsloot HCJ (2012) Generic framework for high-dimensional fixed-effects ANOVA. Brief Bioinform 13(5):524–535. doi:10.1093/bib/bbr071

Timmerman ME, Ter Braak CJ (2008) Bootstrap confidence intervals for principal response curves. Comput Stat Data Anal 52(4):1837–1849. doi:10.1016/j.csda.2007.05.032

Van den Brink PJ, Ter Braak CJF (1998) Multivariate analysis of stress in experimental ecosystems by principal response curves and similarity analysis. Aquat Ecol 32(2):163–178. doi:10.1023/A:1009944004756

Van den Brink PJ, Ter Braak CJF (1999) Principal response curves: analysis of time-dependent multivariate responses

of biological community to stress. Environ Toxicol Chem 18(2):138–148. doi:10.1002/etc.5620180207

Van den Brink PJ, Van Wijngaarden RP, Lucassen WG, Brock TC, Leeuwangh P (1996) Effects of the insecticide Dursban 4E® (active ingredient chlorpyrifos) in outdoor experimental ditches: II. Invertebrate community responses and recovery. Environ Toxicol Chem 15(7):1143–1153. doi:10.1002/etc.5620150719

van Wijngaarden RPA, Van den Brink PJ, Crum SJH, Brock TCM, Leeuwangh P, Oude Voshaar JH (1996) Effects of the insecticide dursban® 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: I. Comparison of short-term toxicity between the laboratory and the field. Environ Toxicol Chem 15(7):1133–1142. doi:10.1002/etc.5620150718

Verdonschot RCM, Van Oosten-Siedlecka AM, Ter Braak CJF, Verdonschot PFM (2015) Macroinvertebrate survival during cessation of flow and streambed drying in a lowland stream. Freshw Biol 60(2):282–296. doi:10.1111/fwb.12479