Collinearity: a review of methods to deal with it and a simulation study evaluating their performance

Carsten F. Dormann[1,17]*, Jane Elith[2], Sven Bacher[3], Carsten Buchmann[4], Gudrun Carl[5], Gabriel Carré[6], Jaime R. García Marquéz[8], Bernd Gruber[1,16], Bruno Lafourcade[6], Pedro J. Leitão[9, 10], Tamara Münkemüller[6], Colin McClean[11], Patrick E. Osborne[12], Björn Reineking[13], Boris Schröder[14, 7], Andrew K. Skidmore[15], Damaris Zurell[4, 14] & Sven Lautenbach[1,18]

[1] Helmholtz Centre for Environmental Research-UFZ
Department of Computational Landscape Ecology
Permoserstr. 15
04318 Leipzig, Germany

[2] School of Botany
The University of Melbourne, Parkville
Victoria 3010, Australia

[3] University of Fribourg
Department of Biology
Unit of Ecology & Evolution
Chemin du Musée 10
1700 Fribourg, Switzerland

[4] University of Potsdam
Plant Ecology & Nature Conservation
Maulbeerallee 2
14469 Potsdam, Germany

[5] Helmholtz Centre for Environmental Research-UFZ
Department of Community Ecology
Theodor-Lieser-Str. 4
06120 Halle, Germany

[6] Laboratoire d'Ecologie Alpine, UMR-CNRS 5553
Université J. Fourier
BP 53, 38041 Grenoble Cedex 9, France

[7] Landscape Ecology
Emil-Ramann-Str. 6
85354 Freising, Germany

[8] Senckenberg Research Institute and Natural History Museum
Biodiversity and Climate Research Centre (LOEWE BiK-F)
Senckenberganlage 25
60325 Frankfurt/Main, Germany

[9] Geomatics Lab
Geography Department
Humboldt-University Berlin
Rudower Chaussee 16
12489 Berlin-Adlershof, Germany

[10] Centre for Applied Ecology
Institute of Agronomy
Technical University of Lisbon
Tapada da Ajuda
1349 - 017 Lisboa, Portugal

[11] Environment Department
University of York, Heslington
York YO10 5DD, UK

[12] Centre for Environmental Sciences
Faculty of Engineering and the Environment
University of Southampton, Highfield
Southampton SO17 1BJ, UK

[13] Biogeographical Modelling, BayCEER
University of Bayreuth
Universitätsstr. 30
95447 Bayreuth, Germany

[14] Institute of Earth and Environmental Sciences
University of Potsdam
Karl-Liebknecht-Str. 24/25
14476 Potsdam, Germany

[15] ITC
University of Twente
P.O. Box 6
7000 AA Enschede, The Netherlands

[16] Institute for Applied Ecology
Faculty of Applied Science
University of Canberra
ACT 2601 Australia

[17] Biometry and Environmental System Analysis
Tennenbacher Straße 4
University Freiburg
D - 79085 Freiburg, Germany

[18] University of Bonn
Institute of Geodesy & Geoinformation
Dept. Urban Planning & Real Estate Management
Nussallee 1
D-53115 Bonn, Germany

* corresponding author:
Email: carsten.dormann@biom.uni-freiburg.de
Tel: ++49 761 203-3749; fax: ++49 761 203-3751

Abstract

Collinearity refers to the non independence of predictor variables, usually in a regression-type analysis. It is a common feature of any descriptive ecological data set and can be a problem for parameter estimation because it inflates the variance of regression parameters and hence potentially leads to the wrong identification of relevant predictors in a statistical model. Collinearity is a severe problem when a model is trained on data from one region or time, and predicted to another with a different or unknown structure of collinearity. To demonstrate the reach of the problem of collinearity in ecology, we show how relationships among predictors differ between biomes, change over spatial scales and through time. Across disciplines, different approaches to addressing collinearity problems have been developed, ranging from clustering of predictors, threshold-based pre-selection, through latent variable methods, to shrinkage and regularisation. Using simulated data with five predictor-response relationships of increasing complexity and eight levels of collinearity we compared ways to address collinearity with standard multiple regression and machine-learning approaches. We assessed the performance of each approach by testing its impact on prediction to new data. In the extreme, we tested whether the methods were able to identify the true underlying relationship in a training dataset with strong collinearity by evaluating its performance on a test dataset without any collinearity. We found that methods specifically designed for collinearity, such as latent variable methods and tree based models, did not outperform the traditional GLM and threshold-based pre-selection. Our results highlight the value of GLM in combination with penalised methods (particularly ridge) and threshold-based pre-selection when omitted variables are considered in the final interpretation. However, all approaches tested yielded degraded predictions under change in collinearity structure and the "folk lore"-thresholds of correlation coefficients between predictor variables of $|r| > 0.7$ was an appropriate indicator for when collinearity begins to severely distort model estimation and subsequent prediction.

The use of ecological understanding of the system in pre-analysis variable selection and the choice of the least sensitive statistical approaches reduce the problems of collinearity, but cannot ultimately solve them.

Keywords

Partial least squares, penalisation, shrinkage, cluster analysis, principal component, variance inflation, condition number, machine-learning, dimensionality reduction, latent root regression

Introduction

Collinearity describes the situation where two or more predictor variables in a statistical model are linearly related (sometimes also called multicollinearity: Alin 2010). Many statistical routines, notably those most commonly used in ecology, are sensitive to collinearity (Belsley 1991, Chatfield 1995, Stewart 1987): parameter estimates may be unstable, standard errors on estimates inflated and consequently inference statistics biased. But even for less sensitive methods, two key problems arise under collinearity: variable effects cannot be separated and extrapolation is likely to be seriously erroneous (Meloun, et al. 2002, p. 443). This means, for example, that if we want to explain net primary productivity (NPP) using mean annual temperature and annual precipitation, and we find that temperature and precipitation are negatively linearly related, we will not be able to separate the effects of the two factors. Using one will partly explain the effect of the other. NPP might be limited only by precipitation but we may not be able to ascertain this relationship because temperature is collinear with precipitation: our model might contain both variables or perhaps only temperature. We will make incorrect inferences and prediction may be compromised. Suppose we want to predict the effect of climate change on NPP and our climate scenarios

indicate no change in precipitation but an increase in temperature. Since our regression wrongly includes temperature, we would erroneously predict a change in NPP.

Collinearity is a problem recognised by most introductory textbooks on statistics, where it is often described as a special case of model non-identifiability. As demonstrated in the example above, it cannot be solved: if two highly collinear variables are both correlated with Y, without further information the "true" predictor cannot be identified. Nevertheless, there are approaches for exploring it and working with it. Despite the relevance of the problem and the variety of available methods to address it, most ecological studies have not embraced measures to address collinearity (Graham 2003, Smith, et al. 2009). The main reasons for this are likely to be: (1) belief that common statistical methods are unaffected by collinearity; (2) uncertainty about which method to use; (3) unsuitability of a method given the type of data to be analysed; (4) lack of interpretability of results when using approaches that combine variables; or (5) inaccessible software. The issue is by no means restricted to ecology (e.g. Kiers and Smilde 2007, Mikolajczyk, et al. 2008, Murray, et al. 2006).

In this paper we aim at facilitating better understanding of collinearity and of methods for dealing with it, by reviewing and testing existing approaches and providing relevant software. The review is structured into five parts. In the first we reflect on when collinearity is, or is not, a problem. The second illustrates spatio-temporal variation in relationships between environmental variables that are commonly used as explanatory variables in regression analyses. The third part introduces the different methods we review, starting with diagnostics, through "pre-analysis clean-up methods" to methods that incorporate collinearity or are tolerant to the problem (see Supplementary material Appendix 1.1 for details on their implementation). In the fourth part we carry out a large simulation study to compare all reviewed methods. We provide complementary case studies on real data in Supplementary material Appendix 1.2. The fifth part discusses our findings with respect to the scattered literature on collinearity. Most importantly it provides advice for the appropriate choice of an

approach and supporting information for its application (e.g. parameterization). Finally we close with suggestions for further research.

Part I. When is collinearity a problem?

To avoid ambiguity, we first clarify the meaning and context of collinearity that we are studying here. We are considering collinearity in the context of a statistical model that is used to estimate the relationship between one response variable and a set of predictor ("independent" or "explanatory") variables. Examples include regression models of all types, classification and regression trees as well as neural networks. Ecologists might be interested in understanding the factors affecting some observed response, or they might want to fit a model (i.e. "train" it) and predict new cases. The impact of collinearity varies with application.

In all real world data, there is some degree of collinearity between predictor variables. Collinearity exists for several reasons. Most commonly, collinearity is intrinsic, meaning that collinear variables are different manifestations of the same underlying, and in some cases, immeasurable process (or latent variable). For example, we could try to explain the jumping distance of a collembolan by the length of its furca, its body length or its weight. Since they are all representations of body size, they will all be highly correlated. Collinearity also arises in compositional data (data where the whole set of information is described by relative quantities: Aichison 2003), such as soil fractions which sum to 100% and hence are not independent of each other. More sand necessarily means less clay or silt. Collinearity may also be incidental, meaning that variables may be collinear by chance, for example when sample size is low, because not all combinations of environmental conditions exist in the study area or when very many variables are involved (as in hyperspectral remote sensing data, e.g. Schmidt, et al. 2004). It is important to highlight that the best way to deal with incidental

collinearity is to avoid it by a well-designed sampling strategy that covers representative geographic and environmental space.

Perfect collinearity occurs if predictors are exact linear functions of each other (e.g. age and date of birth) and is simply a case of model misspecification - one variable needs to be omitted. Mathematically, collinearity in predictors is a case of extreme non-orthogonality and has several undesirable consequences in least squares regression (i.e. models of the form $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$, with response vector $\mathbf{Y}$, predictor set $\mathbf{X}$, parameter estimates $\mathbf{b}$ and residual error $\mathbf{e}$). In these, $\mathbf{b}$ is estimated as $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Since the columns of the design matrix $\mathbf{X}$ are nearly linearly dependent, $\mathbf{X}^T\mathbf{X}$ is nearly singular and the estimation equation for the regression parameters is ill-conditioned. Therefore, parameter estimates $\mathbf{b}$ will be unstable, i.e. small changes in the data may cause large changes in $\mathbf{b}$ (Dobson 2002, p. 94).

In traditional regression models, parameter estimation is a key part of model fitting and interpretation. Models are often used for hypothesis testing, probing the statistical significance of the effect of predictors on the response. High collinearity between predictors means that variables in the collinear set share substantial amounts of information. Coefficients can be estimated, but with inflated standard errors (see Wheeler 2007, for spatial regression examples). Small changes in the data set can strongly affect results so the model tends to be unstable (high variance), and the relative importance of variables is difficult to assess. The inflated errors result in inaccurate tests of significance for the predictors, meaning that important predictors may not be significant, even if they are truly influential (see Ohlemüller, et al. 2008, for an example where three hypotheses are indistinguishable due to collinearity). Problems are exacerbated if stepwise selection methods are used (Harrell 2001, Meloun, et al. 2002), because if one, rather than another, collinear predictor is dropped from the model, the selection process may proceed on a wrong trajectory.

Some of the newer modelling methods, especially those in machine learning where parameter estimation methods are quite different or where recursive partitioning provides the

basis for fitting the model, do not attempt to provide interpretable parameter estimates and standard errors (Hastie, et al. 2001). Nevertheless, they share with traditional methods the problems that the model is sensitive to slight changes in the data set, and that, as a consequence of variable contributions being spread across collinear sets, it is difficult to interpret the final model and to separate the effects of collinear variables (e.g. Shana, et al. 2006).

There are some situations in which the effects of collinearity have limited impact. If the main use of the model is to predict new cases within the range of the sampled data (i.e., to interpolate), the model will do this reliably as long as the collinearity between variables remains constant (Harrell 2001). However, extrapolation beyond the geographic or environmental range of sampled data is prone to serious errors, because patterns of collinearity are likely to change. Obvious examples include use of statistical models to predict distributions of species in new geographic regions or changed climatic conditions (Araújo and Rahbek 2006, Thuiller 2004), and these motivated our interest here in the problem of predicting to changed collinearity structures.

What can we do about this? The most important step is to understand the problems of collinearity and to know your data well enough to be aware of patterns of collinearity in both training and prediction data sets. This paper aims to contribute substantially to this step and to compare methods for identifying and dealing with collinearity. Some other broad advice is relevant. In any regression-style model, the results will be most informative if predictors that are directly relevant to the response are used, i.e. proximal predictors are strongly preferable over distal ones (Austin 1980, Austin 2002). This general concept leads to careful consideration of candidate predictor sets in the light of ecological knowledge, rather than amassing whatever data can be found and challenging the model to make sense of it.

However, are collinear variables necessarily redundant? No. For example, a butterfly larva feeding on a plant will profit from warm temperatures because it accelerates its

development, but also because the plant provides more food, since photosynthesis is temperature-dependent. Thus, both direct and indirect temperature effects and their collinear representations "degree-days" and "leaf photosynthetic activity" are ecologically proximal predictors and sound components of the response "larval size" (see also Hamilton 1987, for a statistical argument). However, as we illustrate in our simulation experiment, collinearity will not be problematic if the correct form of the functional relationship is known. For the butterfly example the collinearity problem may be minimized by representing the functional relationship in a structurally more realistic way, e.g. using Bayesian methods (e.g. Gelman and Hill 2007, HilleRisLambers et al. 2006). However, collinearity may bias parameter estimation in Bayesian approaches and extrapolation to different collinearity structures would still not be sensible.

Part II. Spatio-temporal patterns in collinearity

Recent interest in predicting to new times and places raises the question of the impact of changing collinearity structures for these applications. Collinearity between environmental variables is not constant in space (see Fig. 1, which uses Pearson correlation coefficients as an approximate indicator of collinearity). As an additional twist, collinearity in biogeographical data may differ across spatial scales, making it difficult to elucidate at which spatial scale each environmental driver is acting (Battin and Lawler 2006, Murwira and Skidmore 2005, Wheeler 2007).

Consider, for example, the environmental information contained within a Landsat TM satellite image of the cereal-steppe habitats centred on Castro Verde in Baixo Alentejo, Portugal (Fig. 2). By applying principal components analysis to all seven wavelength bands (data re-sampled to 100 m pixels) we know that the correlation between any two components across the whole image must be zero. Yet within this, local correlations calculated within

moving windows of different sizes reveal complex and varying patterns of relatedness (Fig. 2).

Temporal variations in the relationships between climatic variables span time scales from daily over seasonal to decadal fluctuations. Fig. 3 shows that correlations may vary by 0.2 over decades, and even change their sign. Stronger correlations are less likely to vary as much, because they are causally linked and the causation does not change (e.g. the correlation between temperature and vapour pressure deficit, where the Pearson correlation coefficient *r* for these four stations is around 0.8, but the fluctuations only 0.1; data not shown).

Part III. Methods for dealing with collinearity

We do not think that the problem of collinearity can be solved, for logical reasons: without mechanistic ecological understanding, collinear variables cannot be separated by statistical means. Nevertheless, we might expect some approaches to be superior with regard to robust model fitting and prediction. As a general rule of thumb, a good strategy is to select variables that a) are ecologically relevant, b) are feasible to collect data on and c) are closer to the mechanism (in the sequence resource-direct-indirect-proxy variables: Austin 2002, Harrell 2001). Then, if the statistical method suggests deleting an ecologically reasonable or important variable, prominence should be given to ecology. Despite such careful selection, we might still end up with a set of collinear variables, either because there are several ecologically important variables for a phenomenon under study (e.g. chemical composition of forage), or because we do not yet know which of the predictors are important. The key challenge is now to extract or combine variables meaningfully, as explored in the following sections.

For technical details, types of response and predictor variables that can be used, key references and example studies in ecology please refer to the Supplementary material Appendix 1.1. Since the realm of regression methods is vast, we have focussed on methods

commonly used or likely to have promise; the review and following case study is not exhaustive. All code for data generation is available in the Supplementary material Appendix 2 and interested readers can apply it to any method we failed to cover. The Supplementary material Appendix 1.3 contains a short outline on a number of excluded approaches.

1. Detect it: diagnostics

When are variables collinear? The statistical literature offers several quantifications of collinearity (Table 1), with the most common being the pairwise correlation coefficient (r), the condition index (the square root of the ratio of each eigenvalue to the smallest eigenvalue of X) [1], the variance inflation factor (VIF) and its generalised version (gVIF: Fox and Monette 1992), and the variance decomposition proportions (VD, which gives more specific information on the eigenvectors' contribution to collinearity: Belsley, et al. 1980, Brauner and Shacham 1998). While these methods calculate one value per variable pair (with the exception of the VD where the number of calculated values equals the square of the number of variables), there are also approaches that estimate a single value to describe the degree of collinearity in the full dataset ("variable set indices"). Most commonly used are the determinant of the correlation matrix (det(**R**)) and the condition number (CN, the square root of the ratio of the largest and the smallest eigenvalue of **X**). Code for all of these is supplied in the Supplementary material Appendix 2.

The most useful class of indices depends on the complexity of the dataset. Variable-set indices are preferable when quickly checking for collinearity in datasets with large numbers

[1] There is some confusion in the literature regarding the terms condition index and the condition number. Sometimes, the condition index is defined as the ratio of the largest to the smallest eigenvalue instead of the condition number. We follow here the definitions given by Rawlings, J. O., Pantula, Sastry G., Dickey, David A. 1998. Applied Regression Analysis: A Research Tool - Springer..

of explanatory variables. Per-variable-indices give a more detailed picture of the number of variables involved and the degree of collinearity. Sometimes the per-variable-indices may indicate collinearity although the variable-set indices miss it.

2. Removing collinearity prior to analysis

The first assemblage of collinearity methods, and also the largest, comprises approaches that remove collinearity from the variable set or modify the variables set such that collinearity is removed before the analysis. This assemblage divides into two groups, which differ rather fundamentally in their approach. The first group of pre-analysis clean-up methods identifies which variables are clustering together and thus form a proxy-set (section 2.1). Once a cluster is identified, several ways to proceed are possible, and they are discussed below (section 2.2). The second group does not go through clusters to arrive at new data sets (section 2.3), but uses a variety of other methods to get from the collinear input to the non-collinear output data. Several of the methods presented below use correlation as an indicator for collinearity. We note that correlation and collinearity are not the same: collinearity means linearly related, whereas data with varying amounts of linear relatedness can have the same correlation coefficient. Nevertheless, high absolute correlation coefficients usually indicated high linear relatedness.

2.1 Identify clusters/proxy sets

There exist numerous methods to cluster variables, from which we selected the most common ones. At this point a conceptual decision arises: whether the response variable (**y**) should be used when identifying clusters. Harrell (2001) argues that the response should be ignored, because the clusters represent the grouping of explanatory variables *in relation to themselves*, not grouping of variables in their relation to the response. In the following we will explicitly mention whenever **y** is used as input.

Principal component analysis (PCA) is one of the most common ways to remove correlations in a variable set and to reduce collinearity (as correlation may serve as an indicator of collinearity). It can only be applied to continuous variables, though there are closely related ordination methods such as correspondence analysis that can deal with other kinds of variables. PCA produces orthogonal (i.e. perfectly uncorrelated) axes as output, so without clustering, the PC-axes may be used directly in subsequent analyses in place of the original variables. We discuss this approach later in the section "Latent variable modelling". To use PCA for clustering, the PCA should be applied to the correlation matrix (rather than the covariance matrix, which is distorted by the different scale of variables). Methods exist for applying clustering directly to the components or to rotations of them (Booth, et al. 1994). We only used the direct approach, as described in detail in the Supplementary material Appendix 1.1. The general idea is to work progressively through the PCA axes, study the loadings of the variables onto the axes, and identify groupings. Variables with absolute loadings larger than 0.32 form the "proxy groups" or clusters of interest (Booth, et al. 1994). The value 0.32 is chosen because it represents 10% of the variance for the variable being explained by the PC-axes (Tabachnick and Fidell 1989). Note that PCA is sensitive to outliers (extreme values), transformations, missing data and assumes multi-normal distributions. In practice, the technique is relatively robust when used for description (as opposed to hypothesis testing) so long as the data are continuous, not strongly skewed and without too many outliers. Other ordination techniques (PCoA, nMDS, (D)CA) can be employed analogously and may be more suitable for any given data. *K*-means clustering is equivalent to PCA-based clustering (Ding and He 2004, Zha, et al. 2001).

Cluster analysis is the partitioning of a set of explanatory variables into subsets, i.e. clusters are based on distance between variables (Jain et al. 1999). Clustering can be performed bottom-up (agglomerative) or top-down (divisive). Unfortunately the results depend strongly on which of the many clustering algorithms and which of the many distance-

metrics are used (Lebart et al. 1995). The most commonly recommended ones are Ward-clustering based on the correlation matrix or a Hoeffding-clustering (Harrell 2001, Lebart et al. 1995), but new methods such as self-organising maps (Kohonen 2001) and other machine-learning algorithms may be superior (Hastie et al. 2009). Because hierarchical cluster analysis provides a full cluster tree, a distance-threshold has to be specified to form the actual clusters.

Iterative variance inflation factor analysis (iVIF) is a method based on the quantification of collinearity by VIF (Booth et al. 1994). VIFs are the diagonal elements of the inverse of the correlation matrix. Iterative VIF analysis works, essentially, by comparing the VIF values of a set of predictor variables with and without an additional explanatory variable. All the variables that show an increase of the VIF value above a certain threshold are grouped with the newly added variable into one cluster (proxy-set in the terms of Booth et al. 1994). The iterative formula guarantees that all variable combinations are tested. The method identifies different groups compared with a classification based on pairwise VIF values because it also considers the VIF of groups of more than two variables.

2.2 Dealing with clusters

Once clusters are identified there are several ways to handle them, the three most common being: 1) perform a PCA based on variables in the cluster and use the principal components (PCs); 2) represent the cluster by the variable closest to the cluster centroid; or 3) represent the cluster by the variables with highest univariate predictive value for the response.

PCA on cluster variables is the most common way to create “cluster scores” (Harrell 2001). As long as all principal components are used in the subsequent regression, the analysis will be unbiased (Frank Harrell, pers. comm. in R-help). Where subsets of PCs are chosen, the resulting bias may be tolerable if the selected axes explain most of the cluster inertia. The advantage is that this approach based on composite-axis-score integrates all variables of the cluster, but the disadvantage is that PCs will often be difficult to interpret.

Selecting a "central" variable from the cluster overcomes the interpretation problems, but inevitably introduces a bias by omitting certain variables (Fraley and Raftery 1998). The variables closest (e.g. in terms of Euclidean distance) to the multidimensional cluster centre is an obvious candidate.

Using the "best regressor" from the variables in a cluster has the disadvantage of using the response to determine which variables are selected. This circularity of using **y** in the analysis may inflate type I errors (Harrell 2001). However, since an exploratory data analysis commonly precedes the analysis anyway, the best-regressor-approach ("data snooping") may not distort the analysis too badly compared to completely ignoring collinearity.

Note that although some methods may seem more appropriate because they use "interpretable" variables rather than composite-axis-scores, this is deceptive: in whichever way we represent a cluster, the variable used represents *all* other variables of the cluster and should not be interpreted only at face value. Renaming the retained variable to reflect its multiple identities is a sensible precaution.

2.3 Cluster-independent methods

Two main options exist to bypass the identification of clusters and either directly use the collinear input variables during the analysis or to produce a less collinear set of predictors.

Select variables correlated $|r|<0.7$ is the most commonly applied method across different fields of science, albeit with various thresholds. This only has an unambiguous interpretation when a clear difference in ecological importance exists between correlated variables. Where this is not the case, nonlinear univariate pre-scans of each variable ("data snooping") can be used to determine the sequence of importance (see Murray and Conner 2009, for a review of methods using only linear approaches). Although a threshold of 0.7 is the most common, also more restrictive (e.g. 0.4 in Suzuki et al. 2008) and less restrictive (0.85 in Elith et al. 2006) thresholds have been used.

Sequential regression (Graham 2003) aims to create new purged explanatory variables by reciprocally subtracting the common variation from the less important variables. It linearly regresses explanatory variables against each other and uses the residuals to represent them. Note that while this approach is sometimes also called "residual regression" (Graham 2003), it is fundamentally different from the rightly criticised approach of "regression of residuals" (Freckleton 2002). In sequential regression the *predictors* are regressed, while in "regression of residuals" the residuals of the *independent variable* are used in a second-step regression. In practice, sequential regression comprises the following two steps: 1. Identify a sequence of importance for the explanatory variables. Preferably, this should be done through ecological reasoning. If the data are ecologically indistinguishable (e.g. concentration of trace minerals in the soil), nonlinear univariate regressions on the response variable can be used to determine the order of importance. 2. Calculate the independent contribution of each explanatory variable. The first (most important) variable will remain as it is. The second variable will be regressed against the first, and the residuals of this regression represent the independent contribution of the second variable after accounting for the effect of the first. The third variable will now be regressed against the first and the residuals of the second, and so forth. The resulting variables are orthogonal, but conditional. They cannot be interpreted without the previous variables. Also a standard stepwise model simplification cannot be used, because after removing a variable, all variables of lower importance have to be re-calculated. The interpretation of variables changes from "there is a positive effect of precipitation" to "there is a precipitation effect additional to the contribution it already made through its relationship with temperature". Conceptually, sequential regression is related to semi-partial correlation analysis (Bortz 1993) and path analysis, methods where variables can act through their relationships with other variables (Grace 2006).

3. Modelling with latent variables

Some methods are designed to incorporate collinear variables. The methods described in this section deal with collinearity by constructing so-called "latent" variables, i.e. unobserved variables which underlie the observed collinear variables. As a result of the methods used, most variance in the observed explanatory variables is concentrated in the first few new latent variables and usually the less important latent variables are discarded, leading to a reduction in dimensions. Methods differ in how the latent variables are derived, whether the response variable is included in this derivation and how many latent variables are extracted.

Principal component regression (PCR) simply uses the PCs as explanatory variables and is restricted to linear fits to those variables. Often only those PCs are used that cumulatively explain over 90% of the variance. Then a stepwise procedure simplifies the model further. Ridge principal component regression (Vigneau, et al. 1997) is a special case of PCR, where the PCs are not used in an ordinary regression model, but in a penalised regression model. For details on penalisation see section "Tolerant methods" below.

Partial Least Squares (PLS) iteratively modifies the loadings of the explanatory variables on a PCA in order to maximise the fit of the PCA regression onto the response variable **y** (Abdi 2003). It thus keeps the PLS-axes orthogonal, but they no longer represent maximum variance in **X**. The intention of this approach is that the chosen latent variables are relevant not only for **X**, but also for **y,** though Hastie et al. (2009) show that the variance in X still tends to dominate.

In ordinary PLS, rotations of principal components are fitted to the response variable. By changing the rotation in an iterative procedure, the best linear fit to the response is found. **Penalised Partial Least Squares** uses a non-linear fit, based on splines, to find the best rotation and hence best PLS-components (Krämer et al. 2007). PPLS can hence be seen as a combination of PLS and Generalised Additive Models (GAMs). However, GAMs are very

flexible models, which may overfit the data considerably (i.e. have high performance on training data, but low on test data). To overcome this problem, parameters are penalised, leading to a more robust model. This process is also discussed in the statistical literature as shrinkage or regularisation (Harrell 2001, Reineking and Schröder 2006). For more details on penalisation refer to the section "Tolerant methods" below.

Constrained Principal Component Analysis (CPCA: Vigneau, et al. 2002) works in a similar way to PLS, but is not iterative. To find the best rotation of **X** it requires the estimation of a tuning parameter, which balances fit to **y** against PCA-like maximisation of variance on consecutive axes (see Supplementary material Appendix 1.1 for details). Thus, while a PCA aims to represent variation in **X** with as few principal components as possible and PLS focuses on the fitting of **y**, CPCA balances these two objectives.

In **latent root regression** (Gunst et al. 1976; Webster et al. 1974) the response variable is included in a PCA with the predictors. This identifies important PCA-axes as those with a high loading of **y**. A possibility for selection of axes is to define certain thresholds for the eigenvalues and the loadings of **y** (Vigneau et al. 1996). Then the PCA is re-run, but only on the selected variables, deleting "the non-predictive collinearity" (Gunst et al. 1976). Citing Joliffe (2002, p. 180): "Thus, latent root regression deletes those PCs which indicate multi-collinearities, but only if the multi-collinearity appears to be useless for predicting *y*." Hawkins (1973) and Hawkins & Eplett (1982) keep the response variable when re-calculating the PCA, which we regard as incorrect. The decision about which eigenvalues count as large enough to retain their high-loading variables is somewhat arbitrary (Gunst and Mason 1980). A more elegant approach, in which linear combinations of predictors are formed sequentially and related to the dependent variable to determine their relevance for predictions, was introduced by Vigneau et al. (2002). The advantage and disadvantage of LRR is nicely described by Guerard & Vaught (1989, p. 349): "Latent root regression adds a biased term while eliminating the ill-conditioning. […] the bias term is small and the mean square error of

the latent root regression estimator is less than the mean square error of the ordinary least square estimator. Thus, LRR is preferred to OLS [ordinary least squares] analysis as long as the parameter vector is not parallel to the latent vector corresponding to the smallest latent root of the correlation matrix."

Dimension reduction (DR) is related, structurally, to factor analysis since it also produces new, orthogonal axes and tests for the number of dimensions required to represent the data set. However, DR also uses the response variable to do so. There are different DR-techniques: sliced-inverse regression (SIR: Li 1991), sliced average variance estimation (SAVE: Cook and Weisberg 1991), principal Hessian directions (PHD: Li 1992) and inverse regression estimation (IRE: Wen and Cook 2007). According to Weisberg (2008), the first three of these methods examine the inverse regression problem of **X** | **y**, rather than the forward regression problem of **y** | **X**. A major benefit of DR over the other latent variable approaches is that categorical variables can be analysed too. Axes loadings can be used in the same way as for PCA to construct clusters.

4. Tolerant methods

Some regression techniques may be more sensitive to collinearity than others. Recent developments in model selection methods have introduced new methods for balancing model complexity and fit. Although not necessarily designed to be tolerant of collinearity, they offer approaches that may be less sensitive. The approaches listed here fall into four different groups.

Penalised regressions account both for the number of parameters p in a model and their absolute estimates β: $\text{model complexity} = \sum_{j=1}^{p} \left|\beta_j\right|^{\lambda}$. The degree of penalisation differs between approaches: In *ridge regression* λ=2 (also called "L2-norm": Hoerl and Kennard 1970), in *LASSO regression* λ=1 ("L1-norm": Tibshirani 1996) and in OSCAR (see below) λ

is optimised using the L_1-norm together with the pairwise $L\infty$-norm (Bondell and Reich 2007). The combination of L_1 and L_2 norms is called the elastic net (Zou and Hastie 2005) and is similar to OSCAR (Bondell and Reich 2007). Depending on the form of the penalty, the regression coefficients are shrunk and/or selected. While all methods mentioned lead to shrinkage of the regression coefficients towards zero, ridge regression performs neither selection nor grouping, while LASSO selects but does not group parameters. Shrinkage of the coefficients towards zero leads to an estimation bias, but also to a smaller prediction error due to decreased variance (Hastie, et al. 2009).

Octagonal shrinkage for clustering and regression (OSCAR) provides the user with clusters based on a regression of all variables against the response (Bondell and Reich 2007). Because both response and explanatory variables are standardised before the analysis, only normally distributed responses and continuous explanatory variables can be employed. OSCAR requires specification of two control parameters (the penalisation of the L_1 norm and the penalization of the pair-wise L_∞ norm), which should be optimised, making OSCAR a rather computer-intensive method.

Machine-learning methods are a vibrant area of research in ecology (Elith, et al. 2006, Hastie, et al. 2009), and we only present four methods, chosen for their interest to ecologists. Our machine-learning methods are build around Classification and Regression trees (*Boosted Regression Trees*, BRT: Friedman et al. 2000, and *randomForest*: Breiman 2001) or very flexible, high-order, multidimensional polynomials or splines (*Support Vector Machines*, SVM: Fan et al. 2005, and *Multivariate Adaptive Regression Splines*, MARS: Friedman 1991). Details of these methods can be found in the Supplementary material Appendix 1.1.

Collinearity-weighted regression (CWR) is a new idea developed during this study by CFD, TM and BR. The method down-weights those data points that most strongly contribute to the collinearity pattern in the regression of the response variable against the

explanatory variables (**X**). This is likely to be most useful in situations where outliers are incidental and (partly) responsible for strong collinearity.

Part IV: Comparison of methods on simulated data

To compare methods for dealing with collinearity, we simulated data sets with a range of predictor collinearity and with five different functional relationships between the response, **y**, and the (collinear) predictors, **X**. We then explored the predictive performance of the methods on test data sets with five different collinearity structures. In the following sections we describe our simulation and analysis. Further details can be found in Supplementary material Appendix 1.2.

Data simulation

For our simulation experiment, we created training and test data sets that had 1000 cases and 21 explanatory variables (predictors). These were grouped into four clusters (A-D) of five variables each plus a single uncorrelated variable. Collinearity was restricted to within clusters, imitating collinearity among climatic variables, among land-cover variables and so forth. The parameter "decay" controlled the degree of collinearity with high values of "decay" meaning low collinearity (for details on data simulation see Supporting material Appendix 1.2). The 21st predictor was always created as uncorrelated with all others. All **X** were then standardised.

For all training and test data sets, the response variable was calculated as a function of predictors **X** plus random normal noise (sd = 0.5). We simulated five different relationships of increasing complexity:

1. $f_1 = 25+\mathbf{X}_{Ai}$, i.e. one linear predictor from cluster A;
2. $f_2 = 25+\mathbf{X}_{Ai}+\mathbf{X}_{Aj}+\mathbf{X}_{Bk}+\mathbf{X}_{Bl}+\mathbf{X}_{Cm}+\mathbf{X}_{21}$, i.e. many linear, of which some are collinear ($\mathbf{X}_{Ai}$ and $\mathbf{X}_{Aj}$, $\mathbf{X}_{Bk}$ and $\mathbf{X}_{Bl}$);

3. $f_3 = 25+\mathbf{X}_{Ai}-\mathbf{X}_{Ai}^2+\mathbf{X}_{Bj}-\mathbf{X}_{Bj}^2+\mathbf{X}_{Ck}-\mathbf{X}_{Ck}^2$, i.e. three non-linear (quadratic), without collinearity between any of the variables;
4. $f_4 = 25+\mathbf{X}_{Ai}+\mathbf{X}_{21}+\mathbf{X}_{Ai}\mathbf{X}_{21}$, i.e. two interacting but uncorrelated predictors;
5. $f_5 = 25+\mathbf{X}_{Ai}+\mathbf{X}_{Ai}^2+\mathbf{X}_{21}+\mathbf{X}_{21}^2+\mathbf{X}_{Ai}\mathbf{X}_{21}$, i.e. two non-linear predictors without collinearity, plus an interaction between their linear terms.

In the above formulation, $\mathbf{X}_1$ to $\mathbf{X}_5$ belong to cluster A, $\mathbf{X}_6$ to $\mathbf{X}_{10}$ to B, $\mathbf{X}_{11}$ to $\mathbf{X}_{15}$ to C and $\mathbf{X}_{16}$ to $\mathbf{X}_{20}$ to D. Values of *i*, *j*, *k*, *l*, *m* are randomly drawn, because cluster analysis often identified the alphanumerically first variable as representative of a cluster, thereby biasing the results.

For each of the five functional relationships, we created training datasets with varying degrees of collinearity within clusters by choosing eight different levels for "decay" (0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2 and 0.5). Then, we produced five different test data sets for each training data set to mimic changes in collinearity structure that may occur through time and space. In "test same" the same collinearity structure as in the training data was used (i.e. the data generation algorithm was identical to the training data but with different random seeds). In "test more" and "test less" the decay was half and twice that of the training data, respectively, simulating an increase or decrease in the collinearity of the predictors. The fourth test set, "test non-linear", simulated a non-linear change in the collinearity structure, where only high values of the two variables become less collinear (see code in Supplementary material Appendix 1.2 for details). Finally, "test none" was generated as 21 completely independent predictors with mean = 0 and standard deviation = 1.

Each of the five functional relationships was simulated for each of the eight levels of within-cluster collinearity, yielding 40 different sets. These were replicated 100 times to provide a total of 4000 data sets of different collinearity and **y**-**X**-complexity. Seeds for the random number generator were used to allow full reproducibility of these data. Data generation code is available, together with implementation code for all methods, in Supplementary material Appendix 2.

Simulation analysis

First, collinearity diagnostics were computed for all 4000 data sets: the determinant of the correlation matrix, κ, the condition number, minimum, mean and maximum eigenvalue of the correlation matrix, number of proxy sets (sensu Booth et al. 1994) and number of large variance inflation factors (see section Diagnostics and Table 1 for details, and Supplementary material Appendix 2 for code). Then, we analysed each data set with all collinearity methods. Since some methods have multiple options, a total of 55 different approaches were employed (see Supplementary material Appendix 1.1, Table A1, and example analysis therein).

For each model, we predicted the response ($\hat{\mathbf{y}}$) for all test data sets, and compared this with the "true" **y**-values. Model quality was assessed as Pearson's coefficient of determination (R^2), Root Mean Squared Error ($RMSE = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}$), slope and intercept of the calibration curve (regression of true vs. predicted values). Distributions of R^2, slope and intercepts were heavily skewed by extreme outliers. Hence we report RMSE-values here, which appear to be less sensitive in this context. Qualitatively, R^2, slope and intercept yield the same results (see Supplementary material Appendix 1.2).

Finally, we explored the detail of the effects of collinearity on model selection by targeting two of the simulated data sets. We also ran two case studies with real data (distributions of black grouse in Europe, and drivers of global bird diversity) to illustrate the effects of collinearity on model selection across methods. These are presented in Supplementary Material Appendix 1.2.

Results

In total, 4000 data sets, analysed by 23 different methods were produced. We carried out two pre-analyses on all data sets. The first compared model selection procedures based on the small sample-size corrected Akaike's Information Criterion (AICc) and Bayesian Information

Criterion (BICc); the second compared three different ways to represent clusters (see Supplementary material Appendix 1.2 for details). Based on the results we represented a cluster by its central variable, and kept all four clustering methods. We used BICc-derived models except for CPCA, PLS and PPLS. As references we used three models: firstly, a correctly specified linear model (i.e. a GLM with Gaussian errors, from here on referred to as GLM), where we only estimated the parameters (ML true); secondly a backward stepwise simplified GLM (starting with linear and quadratic terms and first-order interactions for all 21 predictors and BICc setting); and, thirdly, a GAM with cubic splines and shrinkage (i.e. reduction in spline flexibility Wood 2006) applied to all predictors (see Fig. 6 for all methods remaining).

Model validation under collinearity

In the analysis, we focussed on three aspects affecting the performance of a method, as assessed by the Root Mean Squared Error (RMSE) on different test data: 1. degree of collinearity present in the data (X-axis in Fig. 4 and 6); 2. complexity of the functional relationship used for simulation (five subfigures of Fig. 6); and 3. change in collinearity structure from training to test data set (five line types within each panel in Fig. 6). As absolute reference, we used the RMSE of a correctly specified model (first panel with formula for simulation; here all error is due to the noise imposed in the simulations).

Summarised across all functional relationships and model types, we did not detect a trend of degeneration of model fit on the test-same, test-more or test-less data with increasing collinearity (Fig. 4). When the collinearity structure changed non-linearly or was completely lost, however, model fit decreased substantially and became much more variable as collinearity increased (test non-linear and test none, Fig. 4).

As a first rough guide on which statistical approaches worked best, we analysed the shortlisted 23 methods plus the reference ‘ML true’ across all functional relationships (Fig.

5). When evaluated using the test data with the same collinearity structure, most methods performed very well in terms of RMSE. A moderate loss of performance was observed for PPLS, PCA-based clustering and BRT when the collinearity structure changed slightly (test-less). This trend was aggravated under non-linear changes of collinearity, where also variability started to substantially increase for several methods (among them GLM and several latent variable approaches). Using the test data without collinearity (test none), however, the verdict came clearly in favour of the select07/04 methods, ridge, lasso, DR, GAM and MARS. Other methods were also similar in their median performance but exhibited much larger variability (GLM, seqreg, machine learning methods). Neither latent variables (except DR) nor clustering approaches could compete. This may differ between functional relationships, so we subsequently analysed this in more detail.

Fig. 6 shows the effect of increasing collinearity on prediction accuracy (in terms of RMSE) of all models on the different test data sets. We found that collinearity affects model performance negatively for most methods and functional relationships (increasing RMSE towards the right in the panels of Fig. 6). Collinearity effects were generally non-linear, and almost all methods proved tolerant under weak collinearity (CN below 10). A threshold of CN = 30 (indicated in Fig. 6) was clearly too high for most methods analysed here. Notable exceptions from this pattern are PCA-based clustering and SVM, which increased in performance with collinearity (albeit PCA-based clustering starting from a very poor fit).

The results across all collinearity test structures are complex (Fig. 6). They can be first summarised by looking for a general pattern of low and consistent RMSE across all condition numbers, excluding the hardest case that of prediction to completely changed collinearity (the long dashed line). Consistently well-performing methods include select04/07, GAM, ridge, lasso, MARS and DR. Some other methods were consistent, but at a higher RMSE level (Hoeffding/Ward & Spearman/average clustering, seqreg, LRR, OSCAR, randomForest, BRT and SVM).

When investigating, by eye, the performance under severe collinearity (i.e. to the right of the CI=30 line), we found most methods outperformed GLM-like approaches. In particular clustering, penalised and machine-learning approaches yielded lower-error models. However, several of the purpose-built latent variable techniques were only marginally better than GLM, delaying the degeneration of model performance from a condition number of 10 (for GLM) to 30 (LRR, DR, CPCA, PLS, PPLS). Two other noteworthy results are that the GAM also did well at high collinearity, while the commonly used Principal Component Regression showed no improvement on the GLM.

The performance of methods changed only slightly across levels of functional complexity (Fig. 6). Trends became more pronounced as the underlying functions became non-linear, and at a level of functional complexity that might be typical for an ecological regression model (two quadratic terms and an interaction), clustering methods in particular suffered from poor model fits. Also three of the four penalised approaches were unable to regularise the model sufficiently and thus only the ridge was still performing very well.

The most striking pattern we observed was the performance under changing collinearity structure. Since we generally have little idea of how environmental variables are correlated over time or space, this will not help us decide which method to use. Generally, few of the methods were able to correctly predict under the most difficult combination of high collinearity in training data and complete loss of the collinearity structure in the test data (as reported in the right tail of the long dashed lines in Fig. 6). Methods where the RMSE for this combination stayed lowest were select04/07, ridge and MARS, with GAM, randomForest, BRT, SVM, lasso and OSCAR working fine up to a condition number of approximately 150-200 (2.2 on the log-scale of Fig. 6).

Some methods deserve specific comment. The *PCA-based clustering* was useful only under highest collinearity. Under normal circumstances, using the most central variable in a cluster is likely to mislead variable identification. However, using the principal component of

each cluster was even worse (see Supplementary material Appendix 1.2, Fig. A3). *Select04* and *select07* were yielding nearly identical results in all runs. This is probably due to the way we generated our data, where correlations within a cluster are very high, and both thresholds ($|r| = 0.4$ and $|r| = 0.7$) hence led to near-identical selection of predictors. *Ridge* penalisation failed to converge for the quadratic model (function 3) without collinearity (see also Tips and Tricks in the Discussion). *PPLS* was the most unreliable approach, despite combining the strengths of PLS, GAM and penalisation. Finally, *CWR* yielded very similar results to the GLM, only slightly outperforming GLMs under high collinearity. Again, this is probably due to the way we generated our data as collinearity between variables was modelled as intrinsic and not incidental due to outliers which is the main (proposed) application domain of *CWR*.

For each group of approaches, our simulations suggest the following most promising methods: from the control group, GAM; from clustering, Hoeffding/Ward or Spearman/average; from the latent variable approaches, DR; from the penalised approaches, ridge; and from the machine learning group: MARS, randomForest and BRT.

Part V. Discussion

Collinearity cannot be "solved" if we have no additional information. If two highly collinear predictors $\mathbf{X}_1$ and $\mathbf{X}_2$ are correlated with $\mathbf{Y}$, then there is no logical way to glean information about which of the two is "really" behind the correlation. The problem collinearity poses is thus similar to the fact that correlation is not causation: when one variable (or more) is correlated with a response then there need not be any causal relationship between them. This is also the general philosophy behind the latent variable approach. Here one assumes that all variables measured are reflections or proxies for an underlying "true", but unobserved and possibly unobservable, variable. Our review and comparison of methods therefore cannot find a solution to the problem of collinearity, because without additional information the truth cannot be extracted (see similar conclusions in Kiers and Smilde 2007). Still, there are

modelling approaches that are more sensitive to collinearity than others. The aim of this study was to evaluate which methods can be relied on most if no additional information is available.

Our analysis is but one in a list of studies comparing different approaches to collinearity (Aucott, et al. 2000, Graham 2003, Kiers and Smilde 2007, Morlini 2006, Smith, et al. 2009), albeit the most extensive. In particular, our study is, to our knowledge, the only one where non-linear and interacting effects are used in data simulation, which is vitally important for ecological data.

General comments

The methods compared vary widely in their ecological **interpretability**. Latent variable methods (including PCA and PLS) leave the analyst to interpret correlation over several variables. These methods provide no guidance about which of the highly correlated variables may actually be the best candidate for an underlying causal relationship. This, however, can also be seen as a virtue. GLM, GAM, CWR, select07/04 and sequential regression all identify a reduced set of predictors, but their statistical support may only be marginally better than that of those variables collinear with them but omitted during the analysis. Therefore, there exists a high risk of over-interpreting fitted models and relative variable importance and the (relative) simplicity of these analyses is thus paid for by a higher risk of failing to identify an important variable (type II error). If collinearity is incidental, a second, independent data set is likely to have a different collinearity structure and may assist ecological interpretation.

Another aspect of interpretability is the **meaning of the derived variables**. In sequential regression, a predictor could be the residuals of three or more consecutive regressions. Interpreting this effect requires careful wording and head-scratching, for example: “There was an additional effect of slope after accounting for slope-related variability in temperature, precipitation and altitude.” While this may sound convoluted, it is actually much closer to our intuitive understanding than the variable “slope” itself. When we

note "slope" by itself to be significant, we should really only mean the slope, and *not* the fact that sloped sites are drier or higher up in the mountains. So, sequential regression may sometimes be easier to interpret than its iterative derivation would suggest. Since sequential regression also fares rather well in our comparison, we recommend it as one of the methods worthy of further exploration.

Simulations and case studies

From our simulation study we drew four main conclusions:

1. When the **correct form of the functional relationship** is known, collinearity does not harm the fitting and therefore prediction to changing collinearity structures. In Fig. 6, the true model (top left) always yielded a near-perfect fit. This is why mechanistic modellers try to build their models from ecophysiological or population biological principles (see, e.g., Kearney and Porter 2008). Whenever unique model structure is not given, collinearity is likely to bias estimates.
2. The simple and common strategy to **use the better univariate predictor of a set of two collinear variables** (select07 and select04) fared surprisingly well (in line with Smith, et al. 2009). It may well be that this can be attributed, in part, to the design of our simulations, where within each cluster only one variable was causally linked to the response (except for function number 2).
3. Most collinearity **approaches worked reasonably well under moderate collinearity** (i.e. condition number < 10): GLM, GAM, sequential regression, most latent variable methods (PCR, LRR, DR, CPCA, PLS), LASSO, PPLS and machine-learning methods (randomForest, BRT and MARS). Only a few methods failed even under mild collinearity: PCA-based clustering, PPLS and SVM (see section "Tips and tricks" for hints why that may be).

4. Under severe collinearity (condition number > 30), **changes in collinearity structure** (different line types in Fig. 6) were even more worrisome than effects of collinearity *per se*. In particular, non-linear changes in collinearity (where high correlation changed more than low ones) and the complete loss of any collinearity proved detrimental for most methods. Even methods that worked nicely under similar collinearity structure (e.g. seqreg, clustering or the latent variable methods) broke down, indicating that in fact the right predictors or correct parameter estimates were *not* identified by the models.

Our case studies (Supplementary material Appendix 1.2) covered additional issues of whether correct predictors were selected, and investigated performance under small sample size, extremely heterogeneous collinearity, categorical variables, non-normal response variables, and highly-skewed predictors. The results varied with the study, from consistency across several methods in selection of particular variables, to apparently random selection of one variable or another, to selection of all variables and giving small importance to each. For the real data, we do not know the truth, but the results are interesting as demonstrations of the tendencies of different methods.

Caveats

Our analysis cannot be comprehensive. Although it is the most extensive comparison of methods, and contains a large set of varying functional relationships, collinearity levels and test data sets, there will always be cases that fundamentally differ from our simulations. During the selection of case studies we noted in particular two situations we did not consider in the simulations: small data sets and collinearity that did not occur in clusters. Additionally, we shall briefly discuss some other points which are relevant for generalisations from our findings.

Small data sets (where the number of data points is in the same order as the number of predictors) generally do not allow the inclusion of all predictors in the analysis. An ecology-driven pre-selection for importance may reduce or increase collinearity. If we apply univariate (possibly non-linear) pre-scans or machine-learning-based pre-selection, we confound collinearity with variable selection. We chose to exclude these examples from this study to avoid confusion of these two topics, although they clearly are related. Selecting the correct variable to retain in a model is more error-prone under collinearity (Faraway 2005), and the resulting reduced data set will also be biased (see Elith et al. (2010) and Synes & Osborne (2011), for more details).

In our simulations, we grouped the 21 predictors into four clusters of five variables each, and a separate, uncorrelated variable. Within-cluster collinearity was usually much higher than between-cluster collinearity. This led to a bimodal distribution of correlation coefficients (with a low and a high peak). In contrast, in our real-world examples (Appendix 2), the distribution of correlation coefficients was unimodal, with only very few high correlations and many low ones ($|r| < 0.4$). Separating variables into clusters is intrinsically less meaningful in such data sets. Similarly, latent variables have high loadings by many variables and are less interpretable. Finally, the lack of differences between select07 and select04 can be attributed to our grouping structure: if they were not correlated with $|r| > 0.7$, they were often also not correlated at $|r| > 0.4$.

All our predictors were continuous variables. Including categorical predictors would exclude several methods from our analysis (some of the clustering and most of the latent variable methods). Collinearity between categorical and continuous variables is very common (see e.g., Harrell (2001)'s example analysis of Titanic survivors, where large families were exclusively in class 3). We expect collinearity with categorical predictors to be similarly influential as with continuous variables.

Our response variable was constructed with normally distributed error. Binary data (often for example used in species distribution modelling, e.g. case study 2 in Supplementary material Appendix 1.2) are intrinsically poorer in information and we would hence expect the errors in predictive performance for such simulations to be considerably higher. Still, the overall pattern of decreasing prediction accuracy with increasing collinearity should be similar.

We only investigated a single strength of the environment-response relationship. For much weaker determinants, results may well differ (see Kiers and Smilde 2007, for a study varying the noise level). Penalisation and variable selection would then cause an elimination of more predictors, and potentially suffer a higher loss of model performance than the other methods. Latent variable methods, on the other hand, may increase in relative performance, since they embrace all predictors without selecting among them. Similarly, machine-learning approaches could be better under these circumstances.

Despite these caveats, our analysis confirmed several expectations and common practices. In particular, the rule-of-thumb not to use variables correlated at $|r| > 0.7$ (approximately equivalent to a condition number of 10) sits well with our results (at least for similar collinearity structures in the test data – i.e. the same, more and less scenarios). We have no evidence that latent variable methods are particularly useful in ecology for dealing with collinearity: they did not outperform the traditional GLM or select07 approach. And, finally, tree-based models are *no* more tolerant of collinearity than regression-based approaches (compare BRT or randomForest with ridge or GAM).

The choice of which method to use will obviously be determined by more than its ability to predict well under collinearity. From our analysis we conclude that methods specifically designed for collinearity are not *per se* superior to the traditional select07-approach or machine-learning methods (in line with the findings of Kiers and Smilde 2007). In fact, latent variable methods are actually not any better but are more difficult to interpret,

since all variables are retained in the new latent variable. Penalised methods, in contrast, worked especially well (particularly ridge) and should possibly be more widely used in descriptive ecological studies.

Tricks and tips

In this section we briefly share our experience with some of the methods, particularly the choice of parameters. Please refer to the Supplementary material Appendix 1.1 for more detailed implementation information.

Clustering methods and **latent variable approaches**: Clustering is highly affected by pre-selection of variables. Omitting a variable may break a cluster in two, resulting in a very different cluster structure. Fewer variables generally mean better-defined clusters. A crucial point when using cluster-derived variables is to recognise that non-linear relationships will not be properly represented, *unless* the new, cluster-derived variables are also included as quadratic terms and interactions. In the ecological literature, PCA-regression, cluster-derived and latent variables are almost always only included as linear, additive elements. In a pilot analysis of the same data, this resulted in a near-complete failure of these methods. The new variables can best be thought of as alternative variables, and then processed as one would normally do in a GLM, with interactions and quadratic (or even higher-order) terms . Furthermore, latent variable approaches do not provide easily-extractable importance values for variables.

Choice of clustering method: We compared three different methods for processing clusters (Supplementary material Appendix 1.2 Fig. B3). While using univariate pre-scans was the best method, this has consequences with respect to the true error estimates. Because the response was used repeatedly, the errors given for the final model are incorrect and have to be replaced e.g. by bootstrapped errors (Harrell 2001). Therefore our choice and recommendation is to use the “central” variable from each cluster.

LASSO and **ridge**: In the implementation we used (see Supplementary material Appendix 1.1), interactions could not be included. For both approaches, we used a combination of L_1- and L_2-norm penalisation (as recommended by Goeman 2009). This requires that the optimum penalisation for the L_2 and L_1-norm (i.e. the penaliser *not* used by the method), respectively, must be sought *before* running the model. For example, when we run a LASSO (= L_1-norm), we first find the optimum value of the L_2-norm penalisation, and then run the LASSO itself. An alternative that allows simultaneous optimisation of L_1- and L_2-norm, called the elastic net (Zou and Hastie 2005), was slightly inferior to both methods, and much slower (data not shown), though we note that newer and reputedly faster versions have since been released (Friedman, et al. 2010). Both LASSO and ridge require fine-tuning in order to unfold their great potential. For our simulated data, this approach worked nicely. For the more data-limited case studies, manual fine-tuning of the penalisation values was required.

RandomForest and **Boosted Regression Trees (BRT)**: Both methods build on many regression trees, but use different approaches to develop and average trees (bagging vs. boosting). While the performance on test data was very similar, randomForest consistently over-fitted on training data. This means that the model fit on the training data was not a good indicator of its performance on the test data. When using either of the methods for projections to a scenario (where no validation is possible), both methods are likely to yield qualitatively similar predictions, but one might erroneously put more confidence in the (over-fitted) randomForest model. There is no obvious way to correct for this other than by (external) cross-validation.

BRT and **MARS** were also found to benefit from a combination with PLS in the presence of collinearity (Deconinck, et al. 2008). In fact, MARS has been claimed to be sensitive to collinearity, but less so when combining it with PCA (De Veaux and Ungar 1994). Whether this evidence is more than anecdotal remains to be seen (Morlini 2006). Our

simulations show MARS to perform very well and not to suffer from collinearity, although there is no guarantee that it selects the correct predictors and hence should be used with caution (Fig. B1).

Final remarks

Within the limits of our study, we derive the following recommendations:

1. Because collinearity problems cannot be "solved", interpretation of results must always be carried out with due caution.
2. None of the purpose-built methods yielded much higher accuracies than those that "ignore" collinearity. We still regard their supplementary use as helpful for identifying the structure in the predictor data set.
3. Select07/04 yields high accuracy results and identifies collinearity but use with consideration of the omitted variables – e.g., rename the retained variable to reflect its role as standing for two (or more) of the original variables. Because our study was simplistic with respect to the collinearity structure (four well-separated clusters of predictors), select07/04 may have profited unduly. Future studies should explore this further.
4. Avoid making predictions to new collinearity structures in space and/or time, even under moderate changes in collinearity. In the absence of a strong mechanistic understanding, predictions to new collinearity structures have to be treated as unreliable.
5. Given the problems in predicting to changed correlations, it is clearly necessary that collinearity should be assessed in both training and prediction data sets. We suggest to use pairwise diagnostic tools here (e.g. correlation matrix, VIF, cluster diagrams).

Which method to choose is determined by more than each method's ability to withstand collinearity. When using mixed models, for example in a nested design, several methods

(including most latent variable methods and some machine-learning ones) are inappropriate, because they do not allow for the specification of the correct model structure. Collinearity is but one of a list of things that analysts have to address (Harrell 2001, Zuur et al. 2009), albeit an important one.

A number of research questions are unanswered and deserve further attention:

1. *How much change in correlation can be tolerated?* Further research is necessary to define rules of thumb for when the collinearity structure has changed too much for reliable prediction, and how to define the extent and grain at which to assess collinearity.
2. *How to detect and address "non-linear" collinearity (concurvity)*: Collinearity describes the existence of *linear* dependence between explanatory variables. As such, Pearson's r-correlation indices are usually used to indicate how collinear two variables are. Using a non-parametric measure of correlation, such as Spearman's ρ or Kendall's τ, will measure any monotonous relationships, but no approach for detecting and dealing with "concurvity" (Buja et al. 1989, Morlini 2006) more generally is currently available.
3. *Guidance on the relevance of asymmetric effects of positive and negative correlations*: Mela & Kopalle (2002) report that different diagnostic tests for collinearity may yield different results. In particular, positive correlations between predictors tend to cause less bias than negative correlations. Additionally, the former may *deflate* variance, rather than inflate it. However, this issue apparently has not found its way into any relevant scientific paper in any discipline (perhaps with the sole exception of Echambadi et al. 2006), so it is difficult to judge its practical relevance.

In conclusion, our analysis of a wide variety of methods used to address the issue of collinear predictors shows that simple methods, based on rules of thumb for critical levels of collinearity (e.g. select07), work just as well as built-for-purpose methods (such as penalised

models or latent variable approaches). Under very high collinearity, penalised methods are somewhat more robust, but here the issue of changes in collinearity structure also becomes graver. For predictions, our results indicate sensitivity to the way predictors correlate: small changes will affect predictions only moderately, but substantial changes lead to a dramatic loss of prediction accuracy.

Acknowledgements & author contributions

CFD acknowledges funding by the Helmholtz Association (VH-NG-247) and the German Science Foundation (4851/220/07) for funding the workshop “Extracting the truth: Methods to deal with collinearity in ecological data” from which this work emerged. JE acknowledges the Australian Centre of Excellence for Risk Analysis and Australian Research Council (grant DP0772671). JGM was financially supported by the research funding programme “LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts. PJL acknowledges funding from the Portuguese Science and Technology Foundation FCT (SFRH/BD/12569/2003). BR acknowledges support by the “Bavarian Climate Programme 2020” within the joint research centre FORKAST.

We thank Thomas Schnicke and Ben Langenberg for supporting us to run our analysis at the UFZ high performance cluster system. We further acknowledge the helpful comments of four anonymous reviewers.

CFD designed the review and wrote the first draft. CFD and SL created the data sets and ran all simulations. SL, CFD and DZ analysed the case studies. GuC, CFD, SL, JE, GaC, BG, BL, TM, BR and DZ wrote much of the code for implementing and operationalising the methods. PEO, CMC, PJL and AKS analysed the spatial scaling pattern of collinearity, SL that of biome patterns and CFD the temporal patterns. All authors contributed to the design of

the simulations, helped write the manuscript and contributed code corrections. We should like to thank Christoph Scherber for contributing the much-used stepAICc-function.

Literature

Abdi, H. 2003. Partial Least Squares (PLS) regression. - In: M. Lewis-Beck, et al. (eds), Encyclopedia of Social Sciences Research Methods. Sage, pp. 792-795.

Aichison, J. 2003. The Statistical Analysis of Compositional Data. - The Blackburn Press.

Alin, A. 2010. Multicollinearity. - WIREs Computational Statistics

Araújo, M. B. and Rahbek, C. 2006. How does Climate Change affect biodiversity? - Science 313: 1396-1397.

Aucott, L. S., et al. 2000. Regression methods for high dimensional multicollinear data. - Communications in Statistics - Simulation and Computation 29: 1021-1037.

Austin, M. P. 1980. Searching for a model for use in vegetation analysis. - Vegetatio 42: 11-21.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. - Ecol. Mod. 157: 101-118.

Battin, J. and Lawler, J. J. 2006. Cross-scale correlations and the design and analysis of avian habitat selection studies. - Condor 108: 59-70.

Belsley, D. A., et al. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. - John Wiley and Sons.

Belsley, D. A. 1991. Conditioning Diagnostics: Collinearity and Weak Data Regression. - Wiley.

Bondell, H. D. and Reich, B. J. 2007. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. - Biometrics 64: 115-121.

Booth, G. D., et al. 1994. Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. - In: US Dept. of Agriculture, Forest Service, p. 12.

Bortz, J. 1993. Statistik für Sozialwissenschaftler. - Springer

Brauner, N. and Shacham, M. 1998. Role of range and precision of the independent variable in regression of data. - American Institute of Chemical Engineers Journal 44: 603-611.

Breiman, L. 2001. Random forests. - Machine Learning 45: 5-32.

Buja, A., et al. 1989. Linear smoothers and additive models. - Annals of Statistics 17: 453-555.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference (with discussion). - J. R. Statist. Soc. A 158: 419-466.

Cook, R. D. and Weisberg, S. 1991. Discussion of Li (1991). - J. Am. Stat. Assoc. 86: 328-332.

De Veaux, R. D. and Ungar, L. H. 1994. Multicollinearity: A tale of two non-parametric regressions. - In: P.Cheeseman and R. W. Oldford (eds), Selecting Models from Data: AI and Statistics IV. Springer, pp. 293-302.

Deconinck, E., et al. 2008. Boosted regression trees, multivariate adaptive regression splines and their two-step combinations with multiple linear regression or partial least squares to predict blood-brain barrier passage: a case study. - Analytica Chimica Acta 609: 13-23.

Ding, C. and He, X. 2004. K-means clustering via Principal Component Analysis. - Proceedings of the International Conference of Machine Learning 225-232.

Dobson, A. J. 2002. An Introduction to Generalized Linear Models. - Chapman & Hall

Douglass, D. H., et al. 2003. Test for harmful collinearity among predictor variables used in modeling global temperature. - Climate Research 24: 15-18.

Echambadi, R., et al. 2006. Encouraging best practice in quantitative management research: An incomplete list of opportunities. - Journal of Management Studies 43: 1803-1820.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. - Ecography 29: 129-151.

Elith, J., et al. 2010. The art of modelling range-shifting species. - Methods in Ecology & Evolution 1: 330-342.

Fan, R.-E., et al. 2005. Working set selection using second order information for training SVM. - Journal of Machine Learning Research 6: 1889-1918.

Faraway, J. J. 2005. Linear Models with R. - Chapman & Hall/CRC.

Fox, J. and Monette, G. 1992. Generalized collinearity diagnostics. - J. Am. Stat. Assoc. 87: 178-183.

Fraley, C. and Raftery, A. E. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. - The Computer Journal 41: 578-588.

Freckleton, R. P. 2002. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. - J Anim Ecol 71: 542-545.

Friedman, J., et al. 2010. Regularization paths for Generalized Linear Models via coordinate descent. - Journal of Statistical Software 33: 1-22. URL http://www.jstatsoft.org/v33/i01/.

Friedman, J. H. 1991. Multivariate adaptive regression splines. - Annual Statistics 19: 1-141.

Friedman, J. H., et al. 2000. Additive logistic regression: a statistical view of boosting. - Annals of Statistics 28: 337-407.

Gelman, A. and Hill, J. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. - Cambridge University Press.

Goeman, J. 2009. penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-23. - http://CRAN.R-project.org/package=penalized

Grace, J. B. 2006. Structural Equation Modeling and Natural Systems. - Cambridge University Press.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. - Ecology 84: 2809-2815.

Guerard, J. and Vaught, H. T. 1989. The handbook of financial modeling: the financial executive's reference guide to accounting, finance, and investment models. - Probus.

Gunst, R. F., et al. 1976. A comparison of least squares and latent root regression estimators. - Technometrics 18: 75-83.

Gunst, R. F. and Mason, R. L. 1980. Regression Analysis and its Application: A Data-Oriented Approach. - Marcel Dekker.

Hair, J. F., Jr., et al. 1995. Multivariate Data Analysis. - Macmillan Publishing Company.

Hamilton, D. 1987. Sometimes $R^2 > r^{2yx1} + r^{2yx2}$. Correlated variables are not always redundant. - American Statistician 41: 129-132.

Harrell, F. E., Jr. 2001. Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression, and Survival Analysis. - Springer.

Hastie, T., et al. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. - Springer.

Hastie, T., et al. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. - Springer.

Hawkins, D. M. 1973. On the investigation of alternative regression by principal components analysis. - Appl. Statist. 22: 275-286.

Hawkins, D. M. and Eplett, W. J. R. 1982. The Cholesky factorization of the inverse correlation or covariance matrix in multiple regression. - Technometrics 24: 191-198.

HilleRisLambers, J., et al. 2006. Effects of global change on inflorescence production: a Bayesian hierarchical analysis. - In: J. S. Clark and A. E. Gelfand (eds), Hierarchical Modelling for the Environmental Sciences. Oxford University Press, pp. 59-73.

Hoerl, A. E. and Kennard, R. W. 1970. Ridge regression: biased estimation for non-orthogonal problems. - Technometrics 12: 55-67.

Jain, A. K., et al. 1999. Data clustering: a review. - ACM Computing Surveys 31 264 - 323.

Johnston, J. 1984. Econometric Methods. - McGraw-Hill Publishing Company.

Joliffe, I. T. 2002. Principal Component Analysis. - Springer.

Kearney, M. and Porter, W. P. 2008. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. - Ecology Lett. 12: 334-350.

Kiers, H. A. L. and Smilde, A. K. 2007. A comparison of various methods for multivariate regression with highly collinear variables. - Statistical Methods and Applications 16: 193-228.

Kohonen, T. 2001. Self-Organizing Maps. - Springer.

Krämer, N., et al. 2007. Penalized partial least squares with applications to B-splines transformations and functional data. - preprint available at http://ml.cs.tu-berlin.de/~nkraemer/publications.html

Lebart, L., et al. 1995. Statistique Exploratoire Multidimensionelle. - Dunod.

Li, K. C. 1991. Sliced inverse regression for dimension reduction (with discussion). - J. Am. Stat. Assoc. 86: 316-342.

Li, K. C. 1992. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. - J. Am. Stat. Assoc. 87: 1025-1034.

Mela, C. F. and Kopalle, P. K. 2002. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. - Applied Economics 34: 667-677.

Meloun, M., et al. 2002. Crucial problems in regression modelling and their solutions. - Analyst 127: 433-450.

Mikolajczyk, R. T., et al. 2008. Evaluation of logistic regression reporting in current obstetrics and gynecology literature. - Obstetrics & Gynecology 111: 413-419.

Morlini, I. 2006. On multicollinearity and concurvity in some nonlinear multivariate models. - Statistical Methods and Applications 15: 3-26.

Murray, C. J. L., et al. 2006. Eight Americas: Investigating mortality disparity across races, counties, and race-counties in the United States. - PLoS Medicine 3: e260.

Murray, K. and Conner, M. M. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. - Ecology 90: 348-55.

Murwira, A. and Skidmore, A. K. 2005. The response of elephants to the spatial heterogeneity of vegetation in a Southern African agricultural landscape. - Landscape Ecology 20: 217-234.

Ohlemüller, R., et al. 2008. The coincidence of climatic and species rarity: high risk to small-range species from climate change. - Biology Letters 4: 568-72.

Rawlings, J. O., Pantula, Sastry G., Dickey, David A. 1998. Applied Regression Analysis: A Research Tool -Springer.

Reineking, B. and Schröder, B. 2006. Constrain to perform: regularization of habitat models. - Ecol. Mod. 193: 675-690.

Schmidt, K. S., et al. 2004. Mapping coastal vegetation using an expert system and hyperspectral imagery. - *Photogrammetric Engineering and Remote Sensing* 70: 703-716.

Shana, Y., et al. 2006. Machine learning of poorly predictable ecological data. - Ecol. Mod. 195: 129-138.

Smith, A., et al. 2009. Confronting collinearity: comparing methods for disentangling the effects of habitat loss and fragmentation. - Landscape Ecology 24: 1271-1285.

Stewart, G. W. 1987. Collinearity and least squares regression. - Stat. Sci. 2: 68-100.

Suzuki, N., et al. 2008. Developing landscape habitat models for rare amphibians with small geographic ranges: a case study of Siskiyou Mountains salamanders in the western USA. - Biodiv. Conserv. 17: 2197-2218.

Synes, N. W. and Osborne, P. E. 2011. Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. - Global Ecol. Biogeogr. in press:

Tabachnick, B. and Fidell, L. 1989. Using Multivariate Statistics. - Harper & Row Publishers.

Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. - Global Chan. Biol. 10: 2020-2027.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. - J. Roy. Statist. Soc. B 58: 267-288.

Vigneau, E., et al. 1996. Application of latent root regression for calibration in near-infrared spectroscopy. - Chemometrics and Intelligent Laboratory Systems 35: 231-238.

Vigneau, E., et al. 1997. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. - Journal of Chemometrics 11: 239-249.

Vigneau, E., et al. 2002. A new method of regression on latent variables. Application to spectral data. - Chemometrics and Intelligent Laboratory Systems 63: 7-14.

Webster, J. T., et al. 1974. Latent root regression analysis. - Technometrics 16: 513-522.

Weisberg, S. 2008. dr: Methods for dimension reduction for regression. - R package version 3.0.3.

Wen, X. and Cook, R. D. 2007. Optimal sufficient dimension reduction in regressions with categorical predictors. - Journal of Statistical Inference and Planning 137: 1961-1979.

Wheeler, D. C. 2007. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. - Env. Plann. A 39: 2464-2481.

Wood, S. N. 2006. Generalized Additive Models. - Chapman & Hall/CRC.

Zha, H., et al. 2001. Spectral relaxation for K-means clustering. - Neural Information Processing Systems 14 1057-1064.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. - Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67: 301-320.

Zuur, A. F., et al. 2009. A protocol for data exploration to avoid common statistical problems. - Methods in Ecology & Evolution 1: 3-14.

Supplementary material (Appendix EXXXXX at www.oikosoffice.lu.se/appendix).

Appendix 1: Method details, additional results and case studies.

Appendix 2: R-code for all methods, data sets for the case studies and a simulation example data set.

Table 1. Collinearity diagnostics: indices and their critical values.

Method	Description	threshold
Absolute value of correlation coefficients (|r|)[1]	If pairwise correlations exceed a threshold collinearity is high; Suggestion for thresholds: 0.5-0.7	>0.7
Determinant of correlation matrix (D)	Product of the eigenvalue; If D is close to 0 collinearity is high, if D is close to 1 there is no collinearity in the data	NA
Condition index (CI)[2]	Measure of severity of multi-collinearity associated with *j*th eigenvalues; The CIs of a correlation matrix are the square-roots of the ratios of the largest eigenvalue divided by the one in focus; All CIs equal or larger than 30 (or between 10 and 100?) are 'large' and critical	>30
Condition number (CN)	Overall summary of multi-collinearity: highest condition index	>30
Kappa (K)	Square of CN	5
Variance-decomposition proportions (VD)[1, 4]	Variance proportions of *i*th variable attributable to the *j*th eigenvalue; no variable should attribute more than 0.5 to any one eigenvalue	
Variance inflation factor (VIF)[4, 5]	$1/(1-r_i^2)$ with r_i^2 the determination coefficient of the prediction of all other variables for the *i*th variable; Diagonal elements of R^{-1}, with R^{-1} the inverse of the correlation matrix (VIF=1 if orthogonal); Values > 10 (r_i^2>0.9) indicates variance over 10 times as large as case of orthogonal predictors	>10
Tolerance	1/VIF	<0.1

1: (Booth, et al. 1994); 2: (Belsley, et al. 1980, Douglass, et al. 2003, Johnston 1984); 4: (Belsley 1991, p. 27-28); 5: (Hair, et al. 1995)

1055

Figure captions

Fig. 1. Changing collinearity structure of climate variables between eco-zones. Correlation matrix of the following six bioclimatic variables (www.worldclim.org): mean annual temperature, temperature seasonality (standard deviation), mean temperature of coldest quarter, annual precipitation, precipitation of driest month, precipitation seasonality (coefficient of variation). The upper triangular part of the matrix shows Pearson correlation coefficients, while the lower part shows Spearman coefficients. The diagonal elements are one by definition and displayed in grey.

Fig. 2. Correlations between environmental characteristics change with spatial resolution. Moving window Pearson correlations between principal components 1 and 2 of a Landsat TM scene for southern Portugal (pixel size 100 x 100 m). Window size increases from 500 x 500 m (top), through 1.1 x 1.1 km (middle) to 2.1 x 2.1 km (bottom). For the full image (i.e. a single window) the correlation is zero.

Fig. 3. Smoothed time-series of the correlation between mean daily temperature and precipitation for four US-American cities. Systematic seasonal variation was removed by Loess decomposition (contributing about twice as much as the long-term trend depicted here). Moving window width is 30 days (Data courtesy to Peter E. Thornton, Oak Ridge National Laboratories: http://www.daymet.org).

Fig. 4. Root Mean Square Errors across all simulations for the eight different levels of collinearity and using different collinearity structures for validation. Small linear changes, both increasing and decreasing absolute correlation (more/less), have little effect and are depicted together. Grey line indicates RMSE of the fit to the training data.

Fig. 5. Root Mean Square Errors across all simulations for the different methods and using different collinearity structures for validation, sorted by median. Top: Same correlation structure, bottom: none. Grey lines refer to RMSE on training data. Note that sequence of models is different in each panel. Test data "more" was very similar to those of "less", hence only the latter is shown.

Fig. 6. Relative prediction accuracy on test data for an ideal model (ML true) and 23 collinearity methods as a function of collinearity in the data set. In each panel, solid/short-hatched/dotted/dash-dotted/long-hatched locally-weighted smoothers (lowess) depict model predictions on same/more/less/non-linear/no correlation data sets accordingly (not discernable in function 5 for select07 and select04 because they yield nearly identical values). *X*-axis is log(Condition Number), depicted logarithmically. That is, x-values are in fact double-log-ed CNs (one log for the fact that CN is a ratio, the second because we chose logarithmic scaling of collinearity decay rates when generating the data). Data are scaled relative to simulated truth: an R^2 of 1 indicates as perfect prediction as possible. Vertical line (at CN = 30) indicates the rule-of-thumb threshold for CN beyond data set collinearity is deemed problematic.

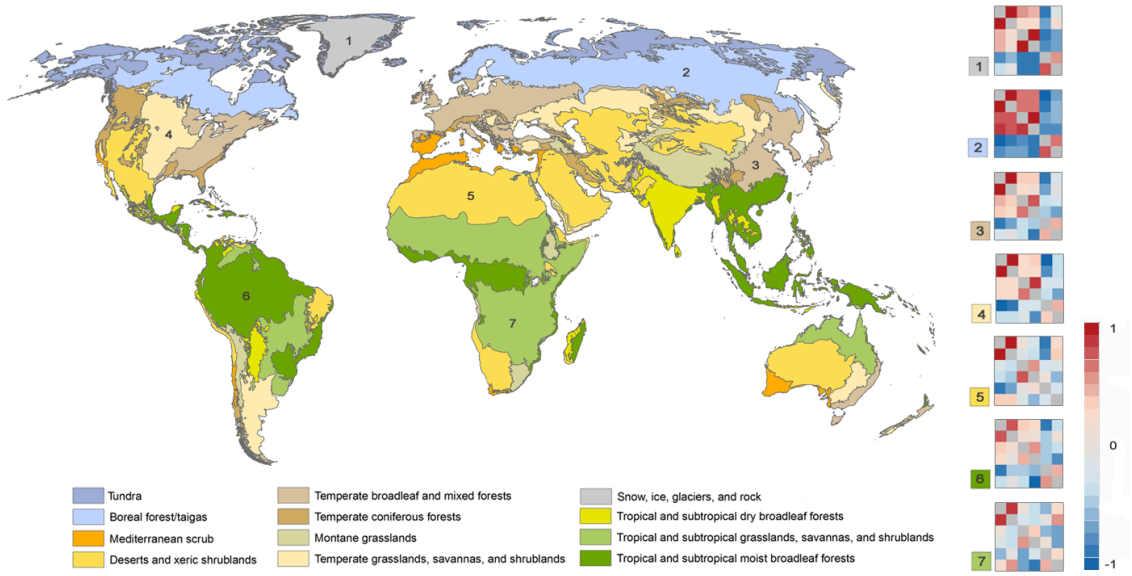
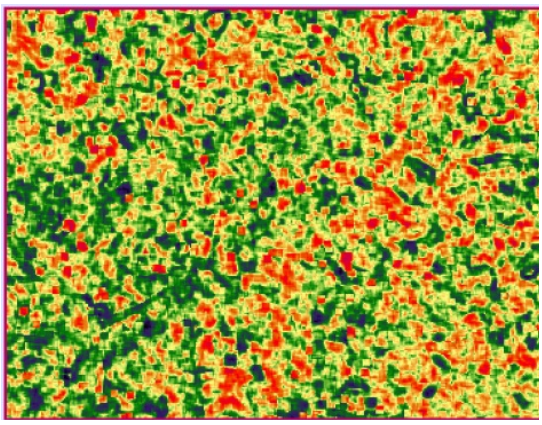


Fig. 1. Changing collinearity structure of climate variables between eco-zones. Correlation matrix of the following six bioclimatic variables (www.worldclim.org): mean annual temperature, temperature seasonality (standard deviation), mean temperature of coldest quarter, annual precipitation, precipitation of driest month, precipitation seasonality (coefficient of variation). The upper triangular part of the matrix shows Pearson correlation coefficients, while the lower part shows Spearman coefficients. The diagonal elements are one by definition and displayed in grey.



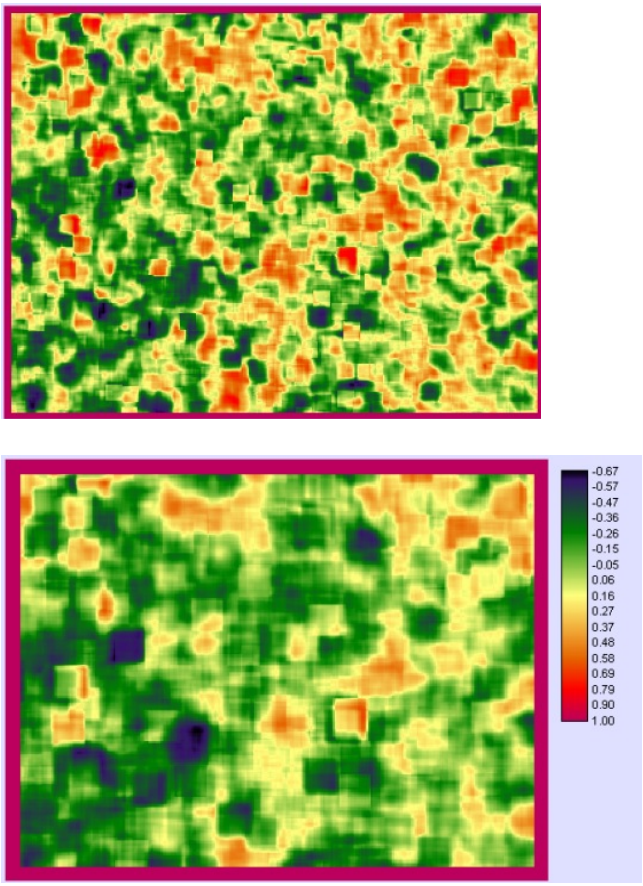


Fig. 2. Correlations between environmental characteristics change with spatial resolution. Moving window Pearson correlations between principal components 1 and 2 of a Landsat TM scene for southern Portugal (pixel size 100 x 100 m). Window size increases from 500 x 500 m (top), through 1.1 x 1.1 km (middle) to 2.1 x 2.1 km (bottom). For the full image (i.e. a single window) the correlation is zero.

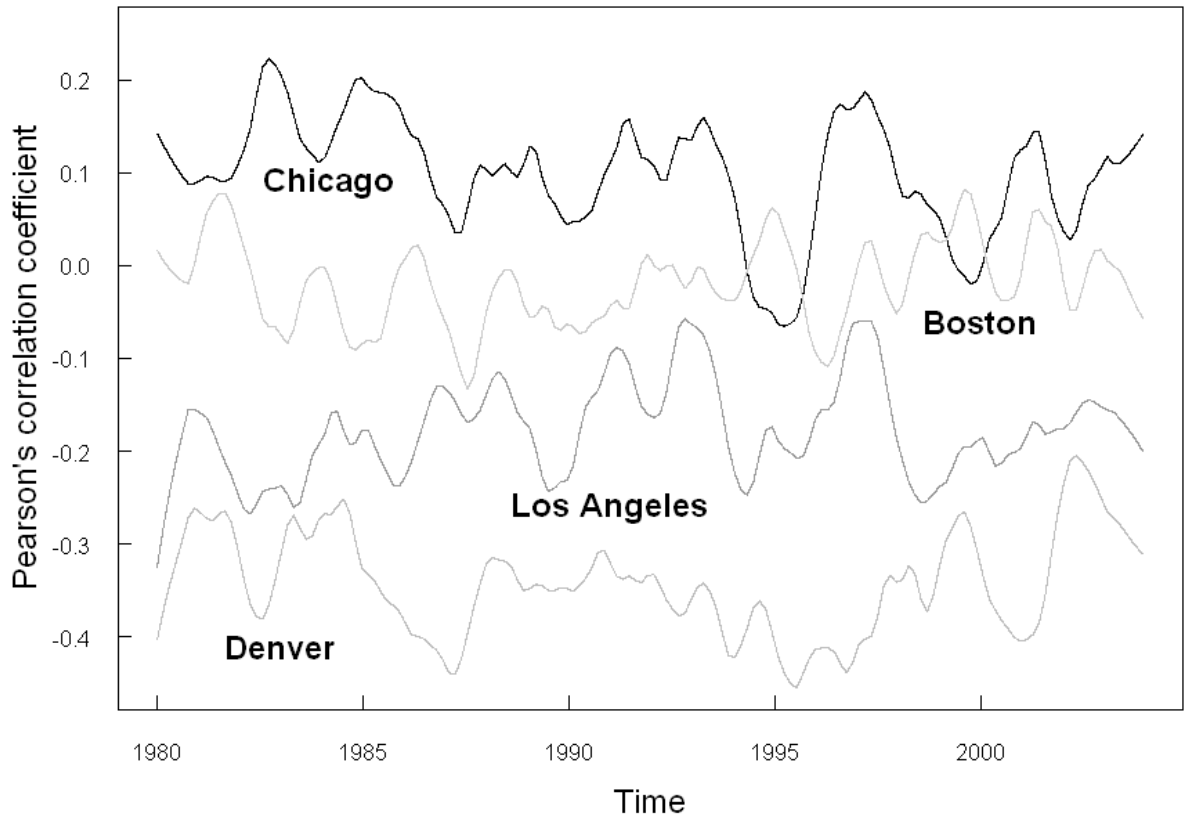


Fig. 3. Smoothed time-series of the correlation between mean daily temperature and precipitation for four US-American cities. Systematic seasonal variation was removed by Loess decomposition (contributing about twice as much as the long-term trend depicted here). Moving window width is 30 days (Data courtesy to Peter E. Thornton, Oak Ridge National Laboratories: http://www.daymet.org).

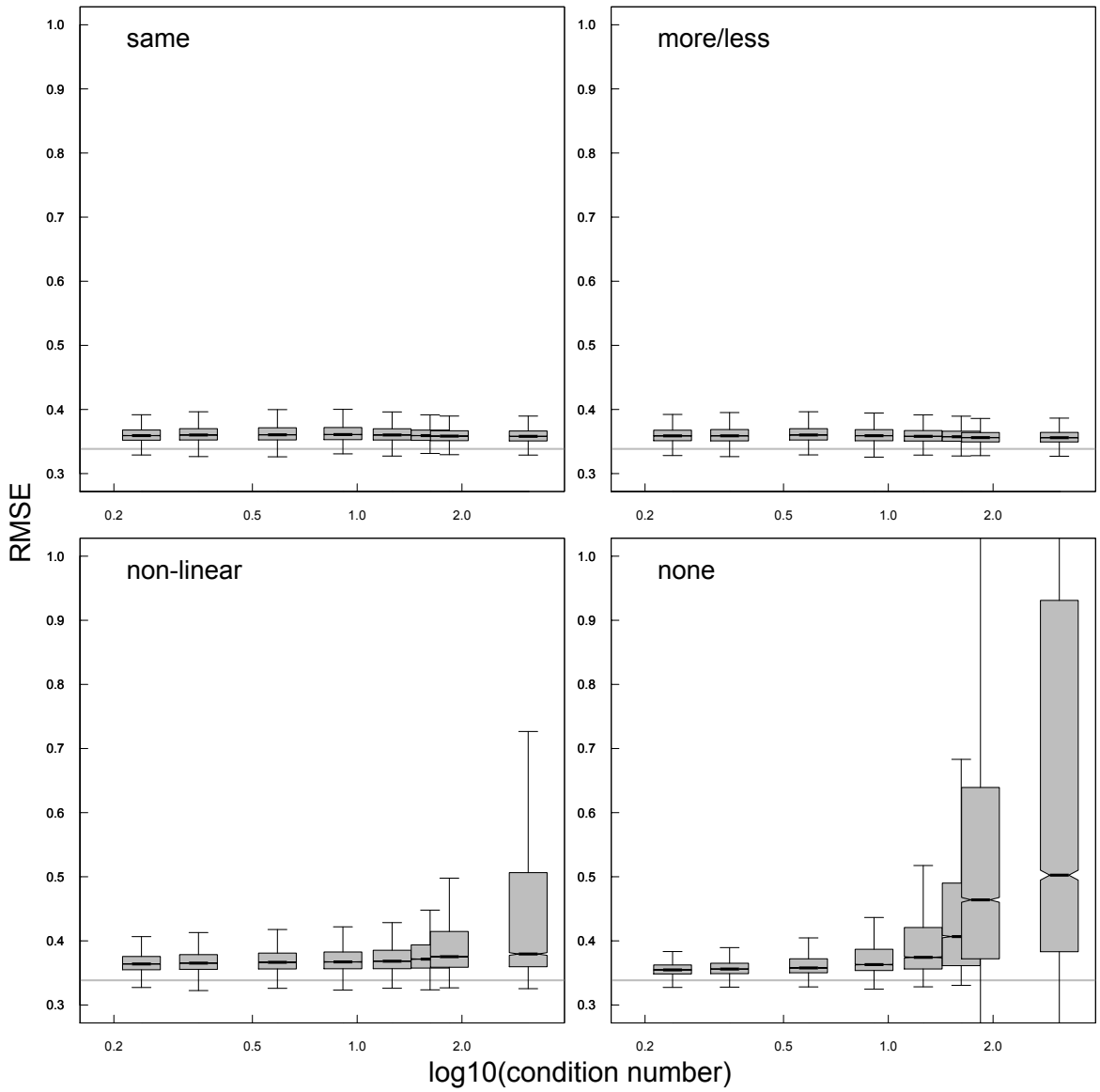
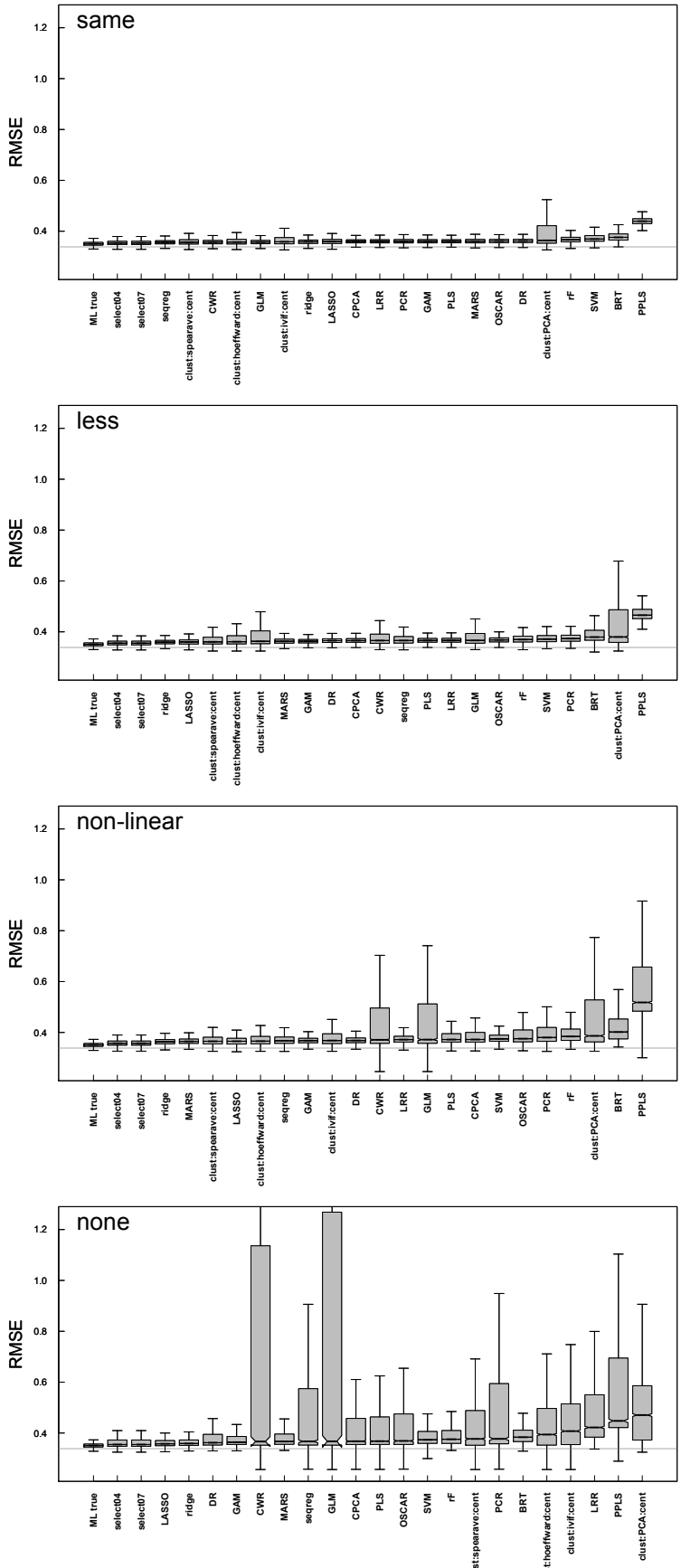
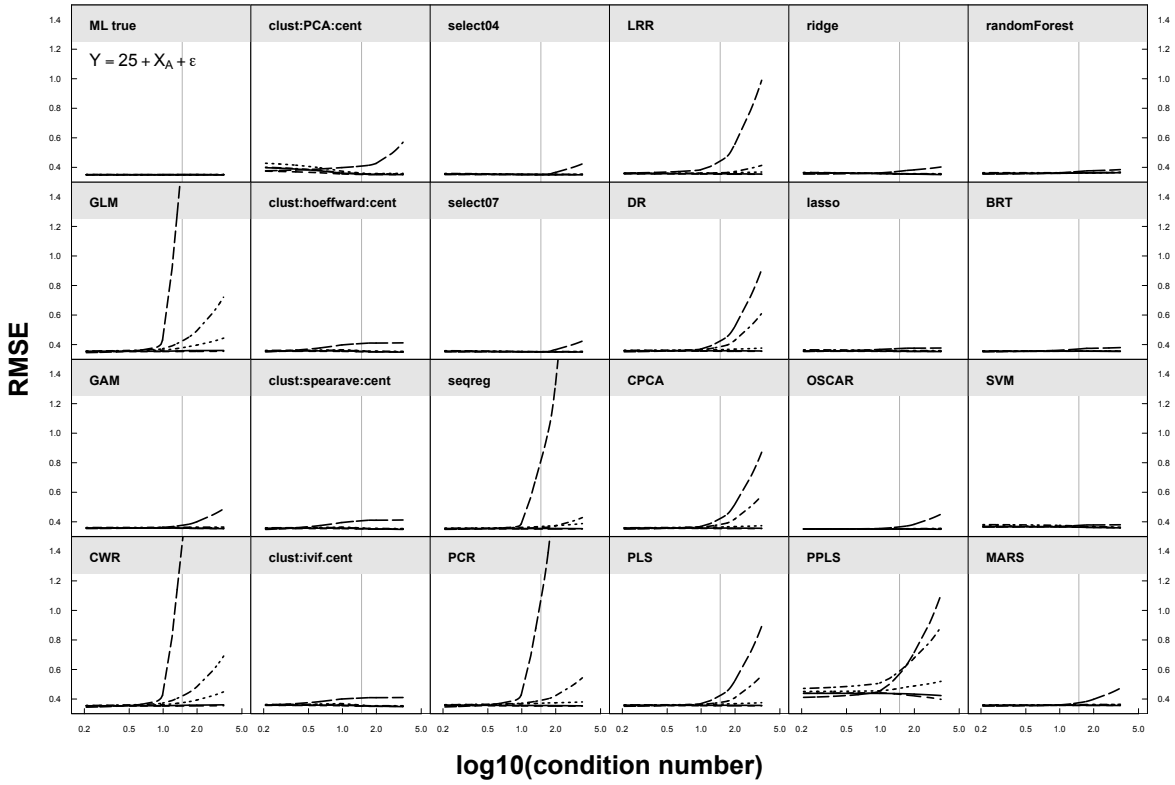


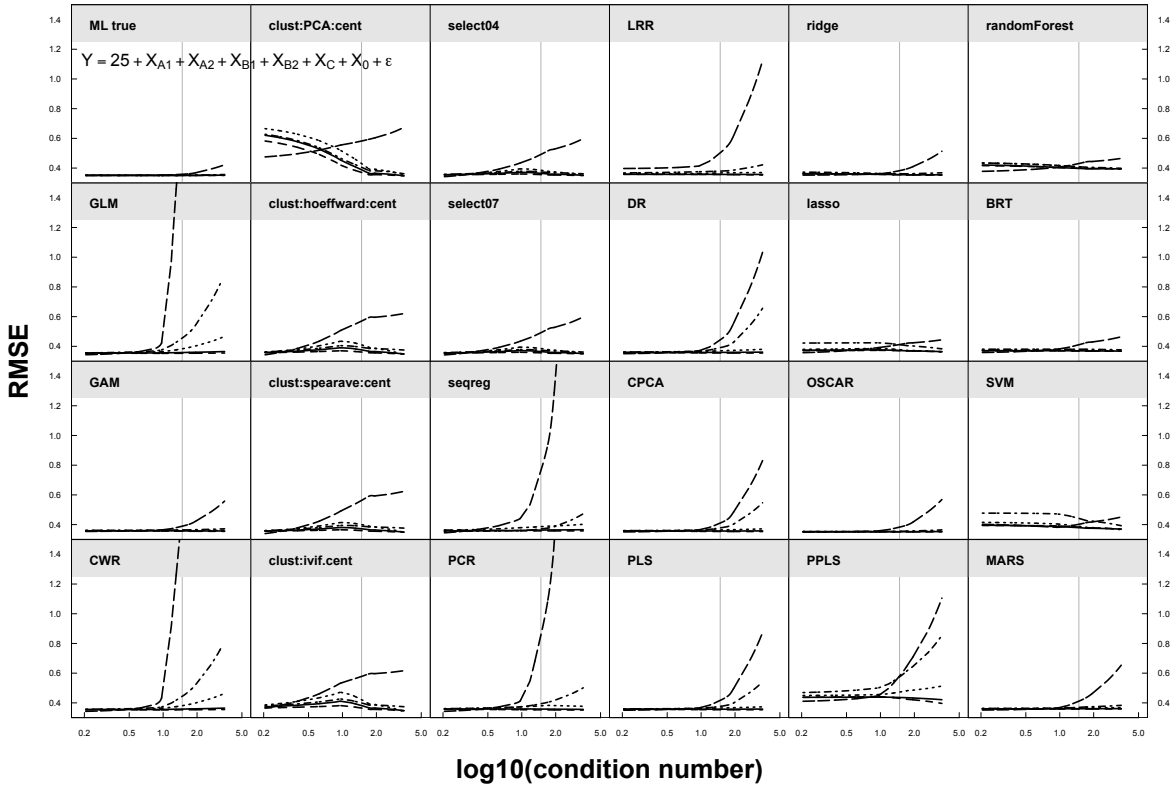
Fig. 4. Root Mean Square Errors across all simulations for the eight different levels of collinearity and using different collinearity structures for validation. Small linear changes, both increasing and decreasing absolute correlation (more/less), have little effect and are depicted together. Grey line indicates RMSE of the fit to the training data.

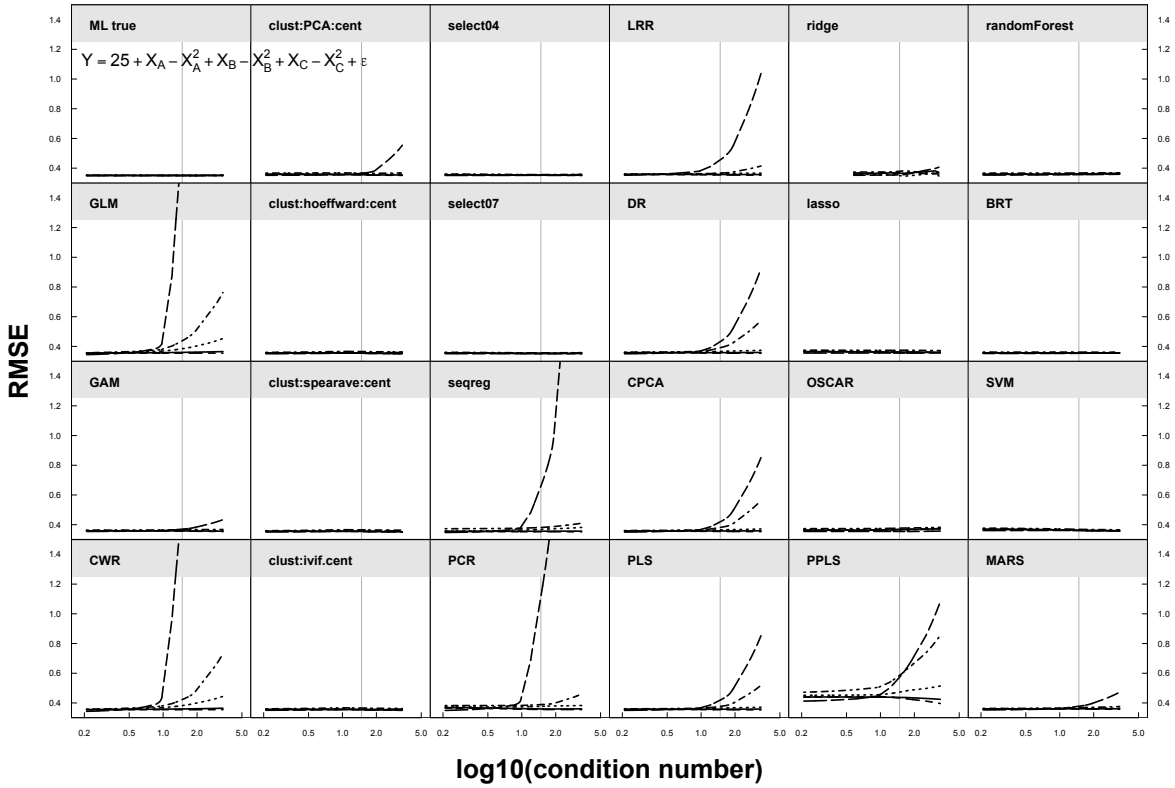

same
RMSE
1.2
1.0
0.8
0.6
0.4
ML true
select04
select07
seqreg
clust:spearave:cent
CWR
clust:hoeffward:cent
GLM
clust:ivif:cent
ridge
LASSO
CPCA
LRR
PCR
GAM
PLS
MARS
OSCAR
DR
clust:PCA:cent
rF
SVM
BRT
PPLS
less
RMSE
1.2
1.0
0.8
0.6
0.4
ML true
select04
select07
ridge
LASSO
clust:spearave:cent
clust:hoeffward:cent
clust:ivif:cent
MARS
GAM
DR
CPCA
CWR
seqreg
PLS
LRR
GLM
OSCAR
rF
SVM
PCR
BRT
clust:PCA:cent
PPLS
non-linear
RMSE
1.2
1.0
0.8
0.6
0.4
ML true
select04
select07
ridge
MARS
clust:spearave:cent
LASSO
clust:hoeffward:cent
seqreg
GAM
clust:ivif:cent
DR
CWR
LRR
GLM
PLS
CPCA
SVM
OSCAR
PCR
rF
clust:PCA:cent
BRT
PPLS
none
RMSE
1.2
1.0
0.8
0.6
0.4
ML true
select04
select07
LASSO
ridge
DR
GAM
CWR
MARS
seqreg
GLM
CPCA
PLS
OSCAR
SVM
rF
clust:spearave:cent
PCR
BRT
clust:hoeffward:cent
clust:ivif:cent
LRR
PPLS
clust:PCA:cent

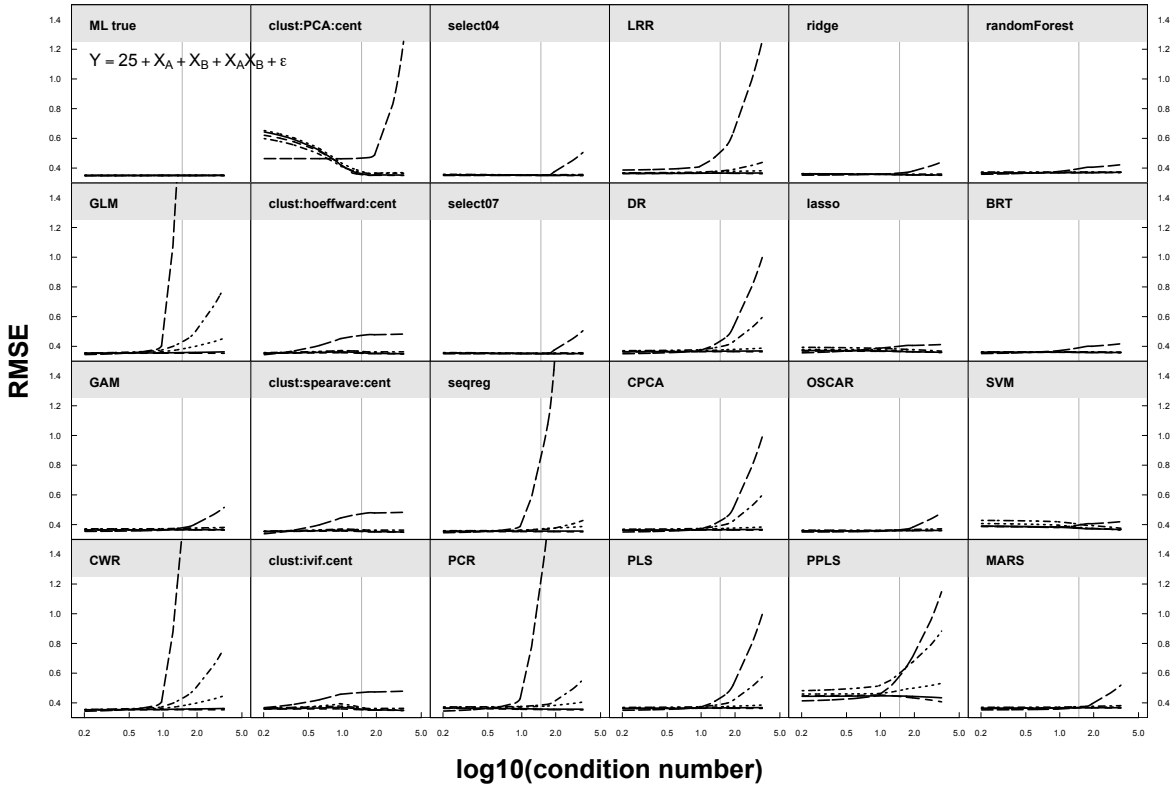
1130

Fig. 5. Root Mean Square Errors across all simulations for the different methods and using different collinearity structures for validation, sorted by median. Top: Same correlation structure, bottom: none. Grey lines refer to RMSE on training data. Note that sequence of models is different in each panel. Test data "more" was very similar to those of "less", hence only the latter is shown.


ML true
clust:PCA:cent
select04
LRR
ridge
randomForest
Y = 25 + X_A + ε
GLM
clust:hoeffward:cent
select07
DR
lasso
BRT
GAM
clust:spearave:cent
seqreg
CPCA
OSCAR
SVM
CWR
clust:ivif.cent
PCR
PLS
PPLS
MARS
RMSE
log10(condition number)


ML true
clust:PCA:cent
select04
LRR
ridge
randomForest
Y = 25 + X_A1 + X_A2 + X_B1 + X_B2 + X_C + X_0 + ε
GLM
clust:hoeffward:cent
select07
DR
lasso
BRT
GAM
clust:spearave:cent
seqreg
CPCA
OSCAR
SVM
CWR
clust:ivif.cent
PCR
PLS
PPLS
MARS
RMSE
log10(condition number)


ML true
$Y = 25 + X_A - X_A^2 + X_B - X_B^2 + X_C - X_C^2 + \varepsilon$
clust:PCA:cent
select04
LRR
ridge
randomForest
GLM
clust:hoeffward:cent
select07
DR
lasso
BRT
GAM
clust:spearave:cent
seqreg
CPCA
OSCAR
SVM
CWR
clust:ivif.cent
PCR
PLS
PPLS
MARS
RMSE
log10(condition number)


ML true
$Y = 25 + X_A + X_B + X_A X_B + \varepsilon$
clust:PCA:cent
select04
LRR
ridge
randomForest
GLM
clust:hoeffward:cent
select07
DR
lasso
BRT
GAM
clust:spearave:cent
seqreg
CPCA
OSCAR
SVM
CWR
clust:ivif.cent
PCR
PLS
PPLS
MARS
RMSE
log10(condition number)

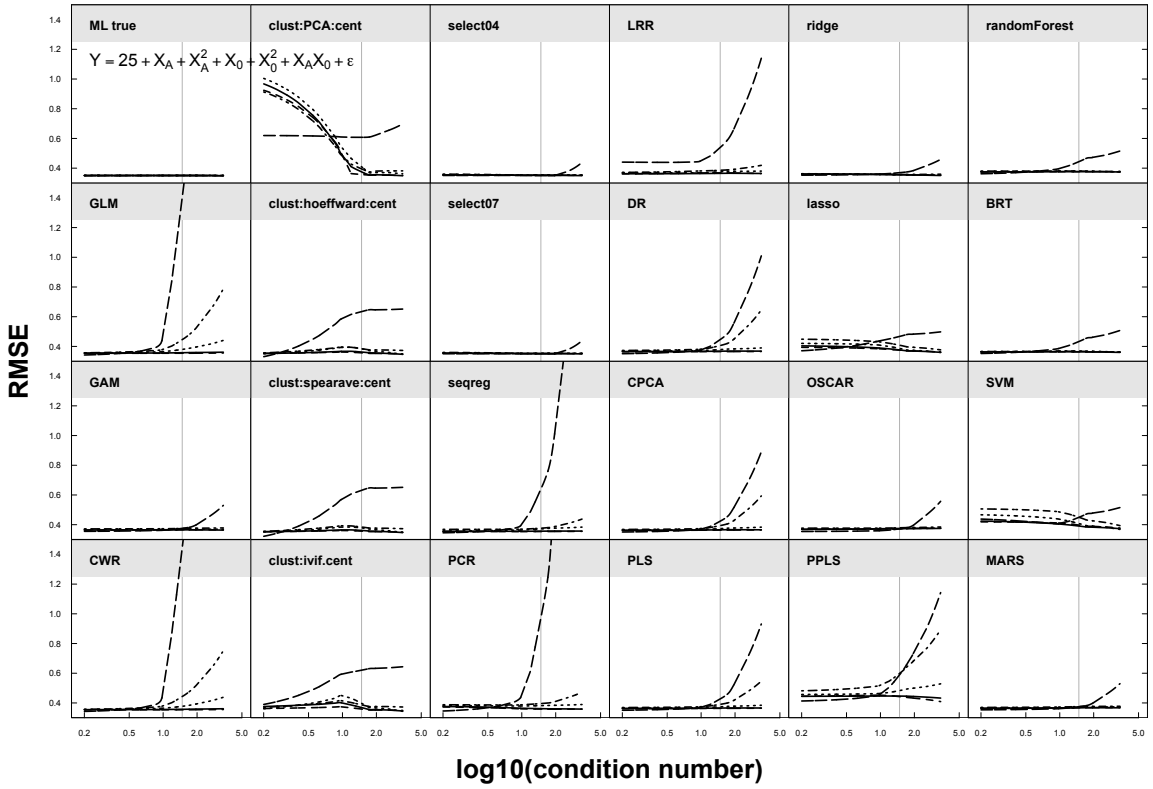


Fig. 6 Relative prediction accuracy on test data for an ideal model (ML true) and 23 collinearity methods as a function of collinearity in the data set. In each panel, solid/short-hatched/dotted/dash-dotted/long-hatched locally-weighted smoothers (lowess) depict model predictions on same/more/less/non-linear/no correlation data sets accordingly (not discernable in function 5 for select07 and select04 because they yield nearly identical values). *X*-axis is log(Condition Number), depicted logarithmically. That is, x-values are in fact double-log-ed CNs (one log for the fact that CN is a ratio, the second because we chose logarithmic scaling of collinearity decay rates when generating the data). Data are scaled relative to simulated truth: an R^2 of 1 indicates as perfect prediction as possible. Vertical line (at CN = 30) indicates the rule-of-thumb threshold for CN beyond data set collinearity is deemed problematic.