Evolutionary Biology
https://doi.org/10.1007/s11692-020-09494-x

RESEARCH ARTICLE

Cross-validated Between Group PCA Scatterplots: A Solution to Spurious Group Separation?

Andrea Cardini[1,2] · **P. David Polly**[3]

Received: 20 December 2019 / Accepted: 26 January 2020

Abstract

Between group PCA (bgPCA) has been developed to summarize group differences in high dimensional spaces like in geometric morphometrics and microarray data where the number of variables is often larger than sample size. However, it has been very recently shown that this technique inflates apparent differences as seen in scatterplots and, in extreme cases, can even create differences where there are none, an effect that becomes more exaggerated as dimensionality increases. In this study, we explore whether leave-one-out cross-validated scatterplots, in which cross-validated scores are used to construct the final ordination instead of the conventional ones, can mitigate the issue. Using simulated data with both isotropic variation or covariance, and increasing the number of variables, we show that cross-validated bgPCs reduce but do not completely remove the distortion of mean differences. However, although scatterplots might still depict inaccurate relationships between group means and must therefore be interpreted with great caution, cross-validation largely solves the issue of spurious separation. Thus, cross-validated bgPCA offers a big improvement for faithfully summarizing overlap or separation among groups in high dimensional spaces and its results will be largely consistent with distance-based permutation tests of significance for group mean differences in the full data space.

Keywords Classification · Covariance · Geometric morphometrics · Group differences · Multivariate analysis · Sampling error

Introduction

Recently, it has been shown that bgPCA, a multivariate alternative to canonical variates analysis (CVA) for exploring between-group differences in high dimensional data sets (Culhane et al. 2002), can produce scatterplots that spuriously separate groups when individuals are projected into the space defined by axes derived from differences between group means (Cardini et al. 2019; Bookstein 2019). Spurious differences appear when small random variation on the p variables is projected into a g-1 bgPCA space (with g being the number of groups), especially when $g << p$. Thus, it has been suggested either to avoid bgPCA altogether (Bookstein 2019) or to use it with caution (Cardini et al. 2019). In particular, Cardini et al. (2019), who first discovered it, provided several clear examples of the problem with this method, but also discussed various scenarios in which the issues may be less serious (data with strong covariance; large samples; small p thanks to a careful selection of variables in relation to the study aims) and (p. 313) strongly emphasized that "scatterplots are not the only tool for assessing group differences" and, thus, argued that "results from a bgPCA should be complemented by tests of significance".

However, bgPCA can also be employed as a classification method (Culhane et al. 2002; Schlager 2017) and classification accuracy is not affected by the same dimensionality issue as long as the classification is cross-validated (x-validated) using, for instance, a leave-one-out approach (Cardini et al. 2019). This type of x-validation simply consists of

✉ Andrea Cardini
alcardini@gmail.com

P. David Polly
pdpolly@indiana.edu

[1] Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi, 103-41125 Modena, Italy

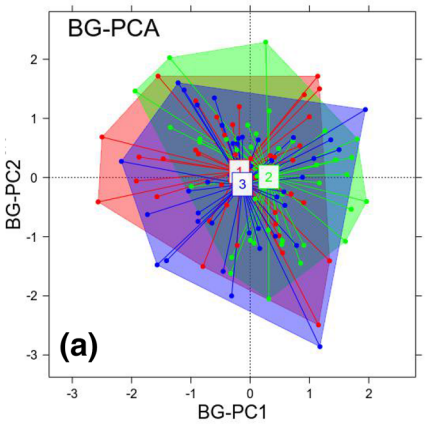
[2] Centre for Forensic Anthropology, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

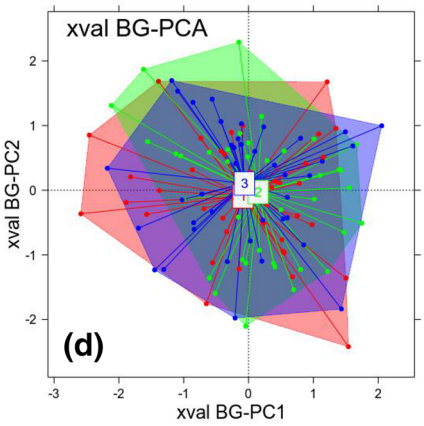
[3] Departments of Earth and Atmospheric Sciences, Biology, and Anthropology, Indiana University, 1001 E. 10th Street, 47405 Bloomington, IN, USA

Published online: 03 February 2020

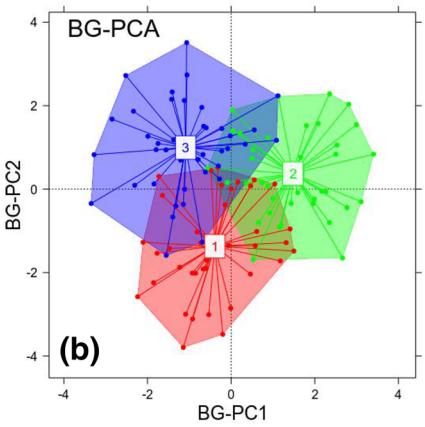
N=120

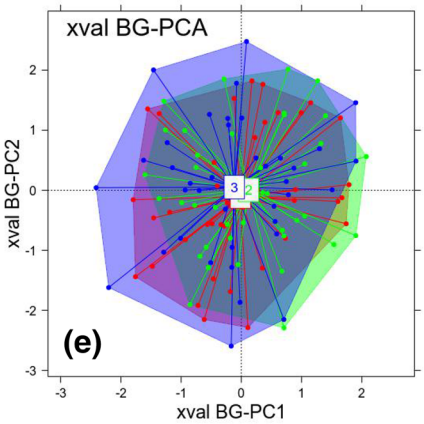
p=4



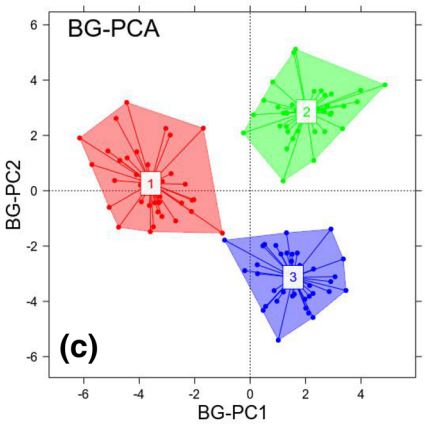


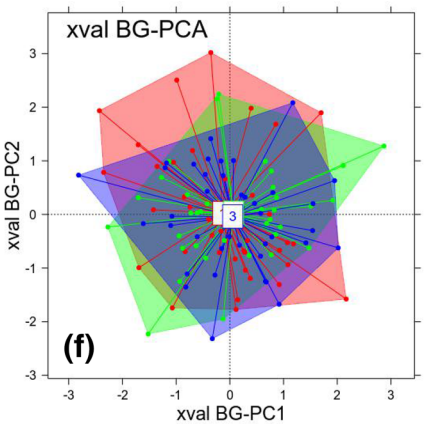
p=120





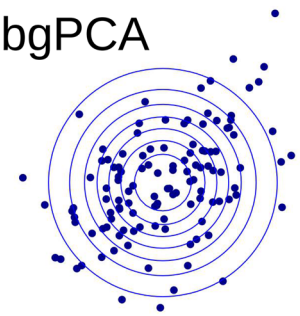
p=720





bgPCA

x-validated bgPCA



multivariate normal uncorrelated (ISO): NO REAL GROUPS (ex0)

◂**Fig. 1** bgPCAs scatterplots of isotropic random data with no real groups (ISO-ex0). In this figure (as well as in Figs. 2, 3 and 4) **a–c** are the conventional bgPCA scatterplots and **d–f** their leave-one-out cross-validated (xval) version. All bgPCA graphics were created with Adegraphics (Siberchicot et al. 2017)

excluding one observation at a time and using the remaining N-1 individuals to build the functions used to classify the omitted case. The same method is commonly used in discriminant function analyses (DAs) to avoid circularity that would otherwise occur by deriving functions from the same individuals that are being classified by those functions and therefore inflating classification accuracy (Kovarovic et al. 2011). Building on the x-validation strategy, in this short communication we explore whether it can mitigate the problem of spurious group differences in bgPCA scatterplots.

Methods

Simulated Data

Keeping the design simple in this first exploratory study, we investigated the case of g = 3 with constant total sample size (N = 120) and balanced samples (i.e., three groups, each having a sample size of 40). We varied p from four to 60, 120, 360 and, when possible (see below) 720. For each value of p we generated random numbers with either no differences between group means (this experiment is referred to as example 0, which is abbreviated ex0) or with systematically controlled differences (examples 1, 2 and 3, abbreviated as ex1, ex2 and ex3, respectively). The effects of the number of variables and distance between groups were then assessed by running the simulated data through bgPCA. For each combination of p and group difference, the simulation was repeated 100 times.

For the simulations in which groups were given different means, we constrained the three means to lie on a straight line in the multidimensional space with equal spacing between them. In the first simulations differences were introduced only in three of the p variables (ex1), but in the others all p variables were given a different mean (ex2-ex3, with ex3 having differences approximately twice bigger than ex2). Random normally distributed data were generated in R (R Core Team 2018) using the *mvrnorm()* function of the package MASS (Venables and Ripley 2002). As in Cardini et al. (2019), we used both isotropic data (independent variables with a standard deviation of one) and data with covariance among the variables, as described in more details below. For brevity, we will call ISO the isotropic model, and VCV the one using empirical covariances.

To generate the VCV simulated data we employed covariance matrices estimated from a set of traditional linear morphometric measurements obtained from a sample of 171 adult male vervet monkey skulls, which are part of a larger published dataset (Cardini and Elton 2017). The linear measurements were calculated from Cartesian coordinates of anatomical landmarks using PAST (Hammer et al. 2001). Where pairs of variables were highly redundant ($r > 0.85$), the one with the highest average correlation with other variables was removed before selecting random subsets of p variables for estimating covariance matrices for the sampling experiments. This reduced the total number of variables available to estimate covariances to 384, thus limiting the maximum number of covarying variables that we could simulate (the maximum number for the uncorrelated simulations was 720). The selection of less highly correlated variables may seem unnecessary, but it makes the model more consistent with actual morphometric practice (Marcus 1990) where an author is unlikely to include multiple variables that provide almost identical information (e.g., many slightly different measurements of cranial lengths, as one might get from a matrix of pairwise distances between all landmarks, as in the vervet cranial sample used in our study).

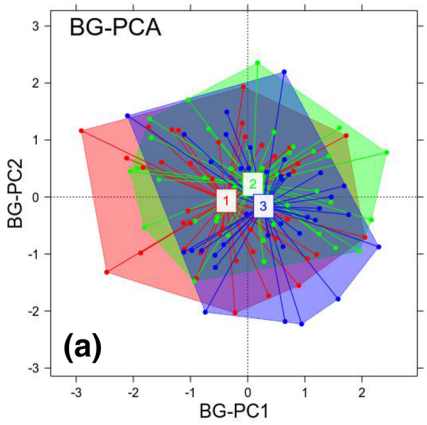
Estimates of Distortions of Group Relationships

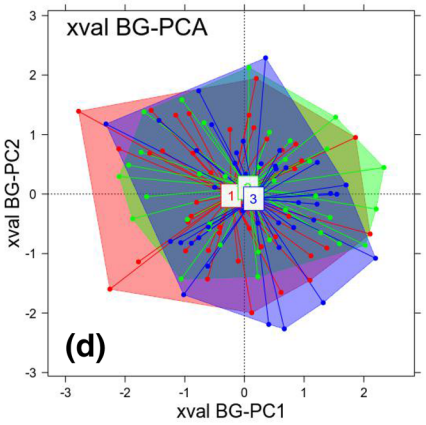
bgPCAs were performed using the package Morpho in R (Schlager 2017). For the datasets with no differences (ex0 using ISO or VCV), we tested the significance of group mean differences using a permutational ANOVA (1000 permutations) based on the Euclidean distance between the means (*adonis()* function in the Vegan package for R (Oksanen et al. 2013)). In each run of a simulation, the test was performed first in the full dataspace (as it is customarily done) and then using only bgPCs without and with x-validation. The test in full data space provides unbiased estimates of the significance of the between-group differences. The same test was then performed in the reduced dimensional bgPCA space (with and without x-validation) as a measure of how much the between-group differences are inflated by the method. If bgPCA did not introduce any distortion, the P values in the bgPCA space would be the same as in the full space (i.e., with a nominal threshold α for type I errors of 0.05, which implies at least 95% of non-significant tests when there are no real differences). Like with CVA, one would not normally use bgPCA space to test for between-group differences, because it is constructed to emphasize those differences relative to within-group variation, and the test is only used here for the specific aim of quantifying the degree of inflation of group structure in the bgPCA space.

We also estimated the percentage of variance explained by group differences (R^2) using the same R function as for the permutation test. This was done for all sets of simulations (without or with real differences in both ISO or VCV data). As with the estimation of P values, we calculated R^2

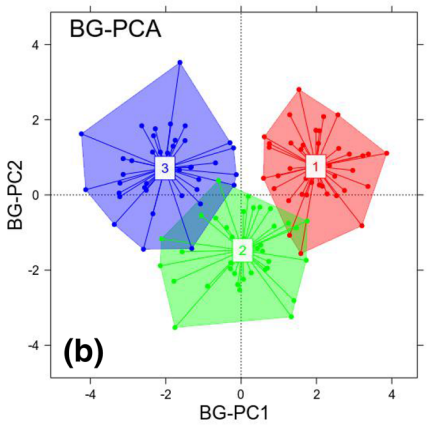
N=120

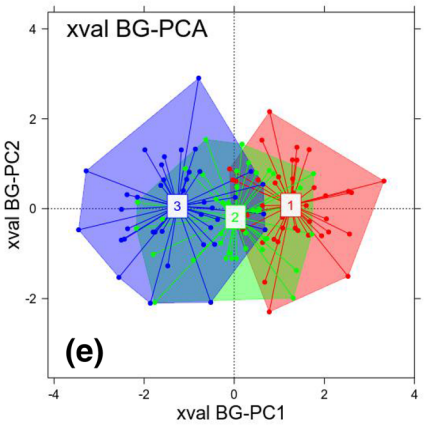
p=4



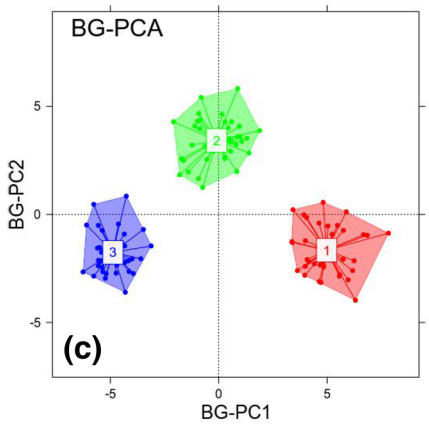


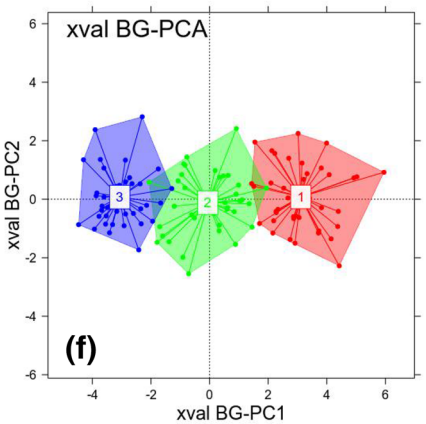
p=120





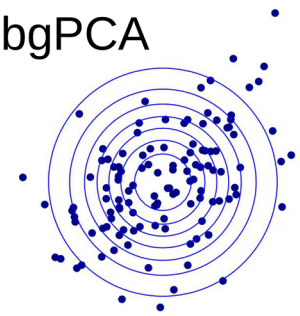
p=720





bgPCA

x-validated bgPCA



multivariate normal uncorrelated (ISO): DIFFERENCES in all p variables with means 'on a line' (ex2: R^2~3%)

◂**Fig. 2** bgPCA scatterplots of isotropic data with small group differences on all p variables (ISO-ex2; full space median $R^2 \sim 3\%$)

first in the full space (correct estimate) and then in the two reduced spaces (bgPCA and x-validated bgPCA). If bgPCA is immune from distortion, R^2 should be the same in full and bgPCA spaces. R^2 is a valid test statistic but is a biased estimator of effect size because R^2 increases along with group number or with decrease in sample size. A useful unbiased alternative is the multivariate Z-score proposed by Collyer et al. (2015). However, R^2 is intuitive and customarily used in studies of group differences (including our previous contribution on bgPCA – Cardini et al. 2019). More importantly, the bias does not affect comparability of results in this study, because N is constant in all simulations and datasets. Thus, R^2 will not be zero even when there are no real group differences, but this provides a useful warning for the readers that, because of sampling error, one should always expect to find differences and the stronger the error the larger they might be (which underscores again the importance of statistical testing in the context of studies of group variation).

Finally, we evaluated how much the relative positions of the group means are altered in bgPCA spaces (without or with x-validation) compared to the theoretical model used to generate the data (for instance, no differences-ex0- or three different means on a line-see above). For simulated data with no group differences, the three means should lie in precisely the same position (with some difference due to sampling error) and for those with simulated differences the means should lie equally spaced on a straight line. We used Procrustes superimposition (with no size standardization) to measure the difference in the positions of the three group means, which are equivalent to a triangle of three multidimensional points, to measure bgPCA's distortion by computing the Euclidean distances between the bgPCA and original positions. This is akin to measuring differences in form space (size and shape) of the two pairs of triangles (theoretical *vs* observed in bgPCA space, and theoretical *vs* observed in x-validated bgPCA space). Note that Procrustes fails if the three points lie in precisely the same location. To compensate for this, we added a tiny mean difference to the groups whose magnitude was less than 1/400 of the standard deviations used to generate the data to the means in the simulations. Because the added differences were so small relative to the variance in the simulated datasets, the three means were virtually identical. For brevity, we will abbreviate these Euclidean distances between triangles formed by the means (theoretical *vs* observed) using the acronym EDM.

Although observed means in simulated data will always be different from theoretical ones because of sampling error, if there was no distortion when p variables are 'squeezed' into g-1 bgPCs, EDM should not change as p increases. For both the R^2 as well as the EDM, we used the median of the 100 runs in each simulation as a summary.

Results

The first two figures show several example bgPCA scatterplots generated from isotropic data of three different dimensionalities (p = 4, 120, and 720) where there were no real group differences (Fig. 1) and with small differences (Fig. 2; $R^2 \sim 3\%$). Without x-validation and no real differences (Fig. 1a-c), the apparent group separation in bgPCA space spuriously increases along with the number of variables so that separation appears complete when p = 720 (Fig. 1c). In contrast, x-validation bgPC mitigates the spurious separation regardless of p (Fig. 1d-f). When the simulated data had small group differences in each of the p variables, bgPCA produced increasingly large group separations that also increased with p (Fig. 2a–c). This should not happen because adding new variables, each with the same amount of simulated difference (and thus also the same amount of variance unexplained by groups), does not change the overall multivariate difference in relation to the total variance in the p-dimensional space (i.e., the R^2). In this case, however, x-validation did not completely mitigate the effect, although it did reduce the bias somewhat (Fig. 2d-f). The distortion of the pattern of mean differences is clearly more pronounced if scores are not x-validated, especially with high p = 120 and 720, as indicated by the sharp deviation of the mean of group 2 from a straight line. Thus, the x-validated results not only mitigated the spurious differences between groups in bgPCA space, but they somewhat helped maintain their spatial relationships.

Results for data with covariance (VCV, Figs. 3, 4) are similar to those of the isotropic data (ISO). Spurious group separation increases with p in the non-x-validated bgPCA-x-validated bgPCA space (Fig. 3a–c) and the distortion of the positioning of group means is also stronger in non-x-validated bgPCAs (Fig. 4a–c). However, as shown by Cardini et al. (2019), spurious differences are much less pronounced when variables covary, as it is as if the real dimensionality of the data was smaller than p. Likely, the smaller effect of p on group separation in VCV datasets also depends on the strong pattern of covariance, which stretches bgPC1 so much that group means look closer and overlap larger. Interestingly, however, if covariance helps to reduce spurious separation, it also increases the distortion of the pattern of group mean differences (see also below). Yet, the distortion is again less pronounced after x-validation (Fig. 4d, e).

Figure 5 shows the percentage of runs from ex0 (no real differences) in which the test produced statistically significant differences between groups. As expected, spurious

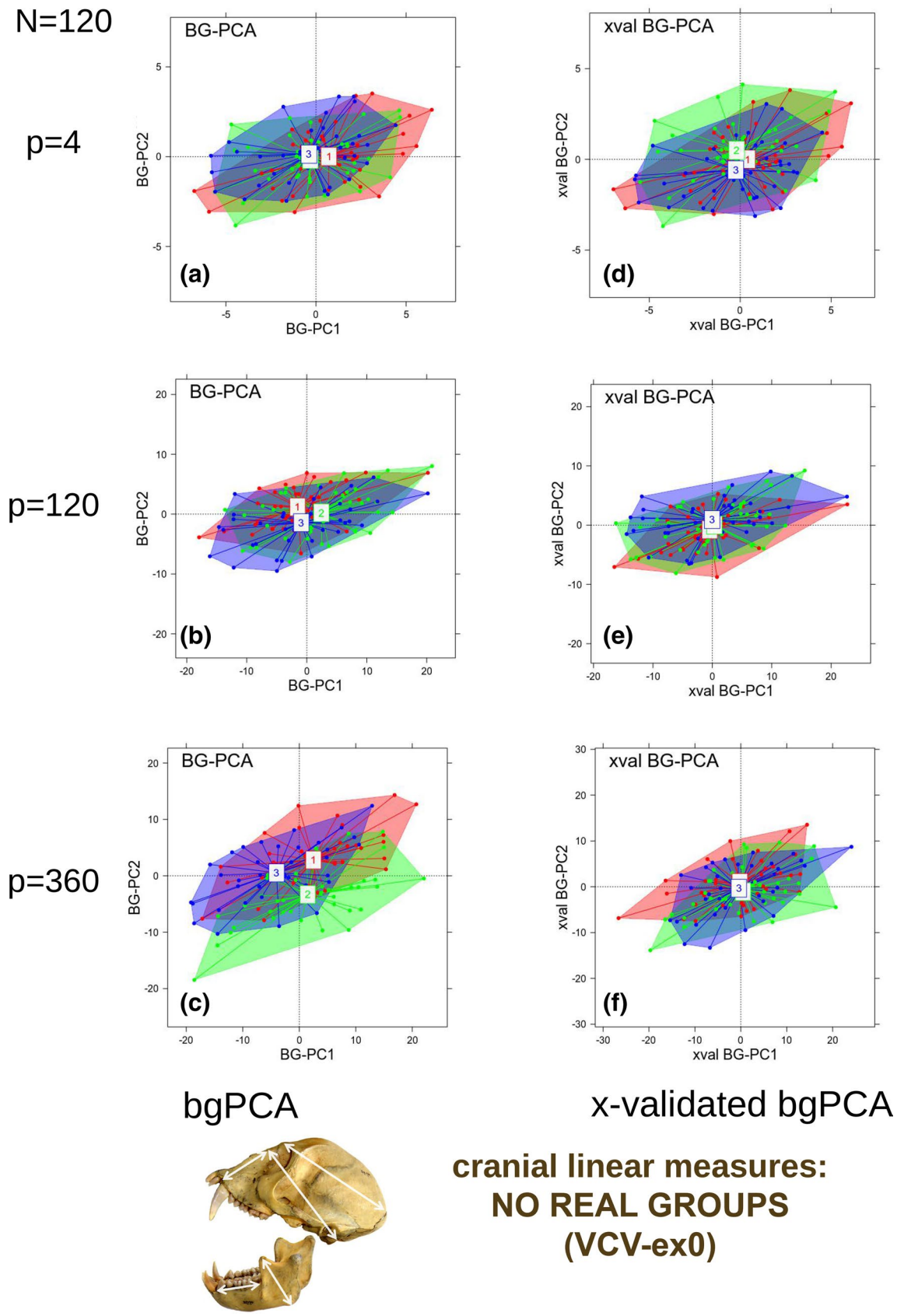


Fig. 3 bgPCA scatterplots of covarying data with no real group differences (VCV-ex0)

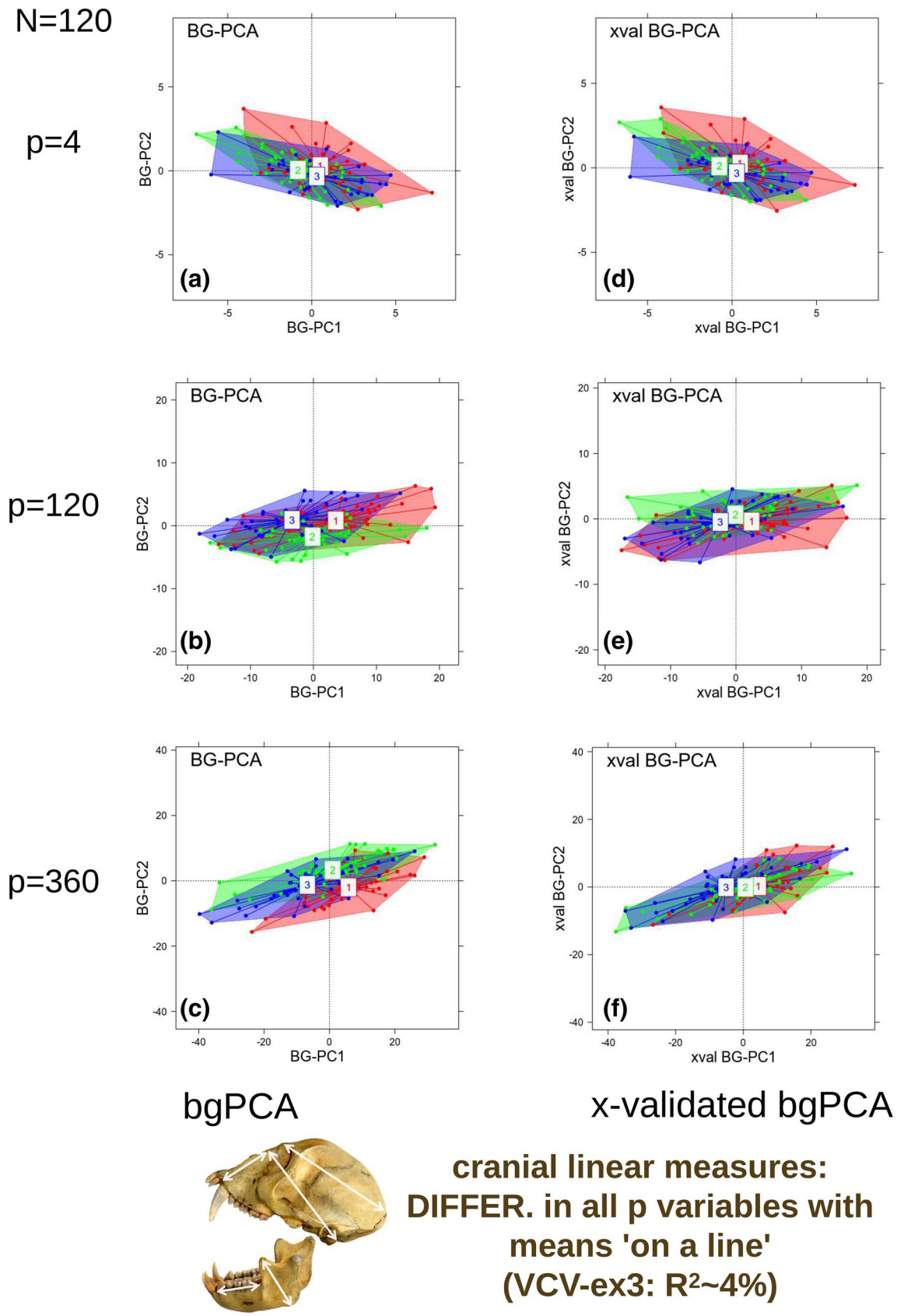


Fig. 4 bgPCA scatterplots of covarying data with larger group differences on all p variables (ISO-ex3; full space median $R^2 \sim 4\%$)

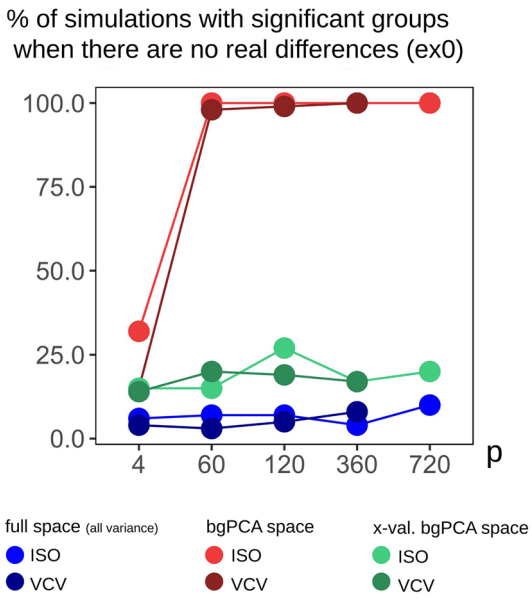


Fig. 5 Tests of group differences in data with no real differences (ex0). Profile plots of the percentage of simulations in which a permutation test for group mean differences (using the Euclidean distance between the means as a test statistics) was significant, as it should be not, given the absence of real groups. Tests were performed in the full space (as it is conventionally done), but also in the bgPCA space (i.e., using only bgPCs) without and with cross-validation. Note that normally one should not test differences using only bgPCs (which is circular reasoning, as it is, by definition, a space built ad hoc to capture differences) and this was done here only to provide an index that suggests how much differences are inflated in bgPCA space. With inflated differences, 5% or less of tests in bgPCA space should be significant, which is the expectation using $\alpha = 0.05$ and what is observed in the full data space. The same reasoning applies to R^2 in Fig. 6, that is also computed both in full space and in its bgPCA subspace

significant differences occur in the full data space only about 5% of the time. This indicates that the type I error rate (rejection of the null hypothesis of no differences when there are indeed no differences) is generally appropriate (using a nominal α of 0.05 or 5%) in all datasets and simulations when tested in the full data space. In sharp contrast with this, when tests without x-validation are performed in bgPCA space, even with just four variables ($p = 4$), type I error rates are clearly higher than expected, as significance is found in approximately 25–30% of tests. With $p = 60$ or larger, virtually all tests in bgPCA space become significant, despite the absence of real differences and regardless of the model (ISO or VCV). However, if bgPC scores are x-validated, for both ISO and VCV data, significance is found only ca. 15% to about 25% of times. This is on average larger than the 5% expected using the conventional $\alpha = 0.05$, but clearly much better than in non-x-validated bgPCA space.

Figure 6 summarizes the proportion of variance explained by group membership (R^2) using the median of the 100 runs in each simulation. When no differences exist between groups (ex0, Fig. 6a), R^2 should remain close to zero, which it does in the full data space and the x-validated bgPCA space for both ISO and VCV models. The fact that R^2 remains low in the x-validated space is congruent with the lack of separation in the x-validated space in Figs. 1, 2. In sharp contrast, non-x-validated bgPCAs produce increasing R^2 in relation to p, with a very strong effect for the ISO data (up to more than 75% of explained bgPCA variance) and a definite but less pronounced one for the VCV data (up to ca. 10% of explained variance). This is again consistent with the scatterplots of Figs. 1, 2, which suggested a much less serious problem of spurious group separation when data covary and also indicated that x-validated scores further mitigate or even completely control for this artefact.

For simulated data with real between-group differences (Fig. 6b–d) the patterns are more complex, but they still consistently indicate that x-validated bgPCs more accurately represent between group differences than do the conventional bgPCs. In fact, in simulated data with co-varying variables, the x-valided bgPCA produces R^2 values that are close to the correct estimates obtained in the full data space. In Fig. 6b (VCV-ex1), the R^2 values in the full data space decrease as p increases because the simulated differences only involved three variables, which are progressively swamped with random noise as p gets larger. In contrast, the R^2 values in the full data space remain more or less constant, with a median of 3–7% and 2–4% respectively for ISO and VCV data, when all the variables were simulated with a smaller (ex2) or larger (ex3) between-group difference (Fig. 6c–d). In these simulations, ISO bgPCAs always lead to inflated R^2, although inflation is smaller after x-validation. In the same simulations, non-x-validated bgPCAs of VCV data always produce R^2 much smaller than those of ISO data and only slightly bigger than those from x-validated bgPCAs, which are, in turn, fairly close to the correct R^2 value obtained in full data space.

Collectively, these results confirm that the spurious between-group separations introduced by bgPCA is worse in isotropic, uncorrelated variables, as already shown by Cardini et al. (2019). While this model is unlikely to correspond to any real biological data where most variables covary, the mitigating effect of covariance simply means that the spurious separation is less in covarying data for the same p. With covariance, spurious separation is very modest but does increase with p. With x-validated scores, however, the inflation of differences is very small (Fig. 6b–d) and almost absent when there are no real differences (Fig. 6a). Thus, larger p will always result in larger spurious differences in bgPCA if x-validation is not used.

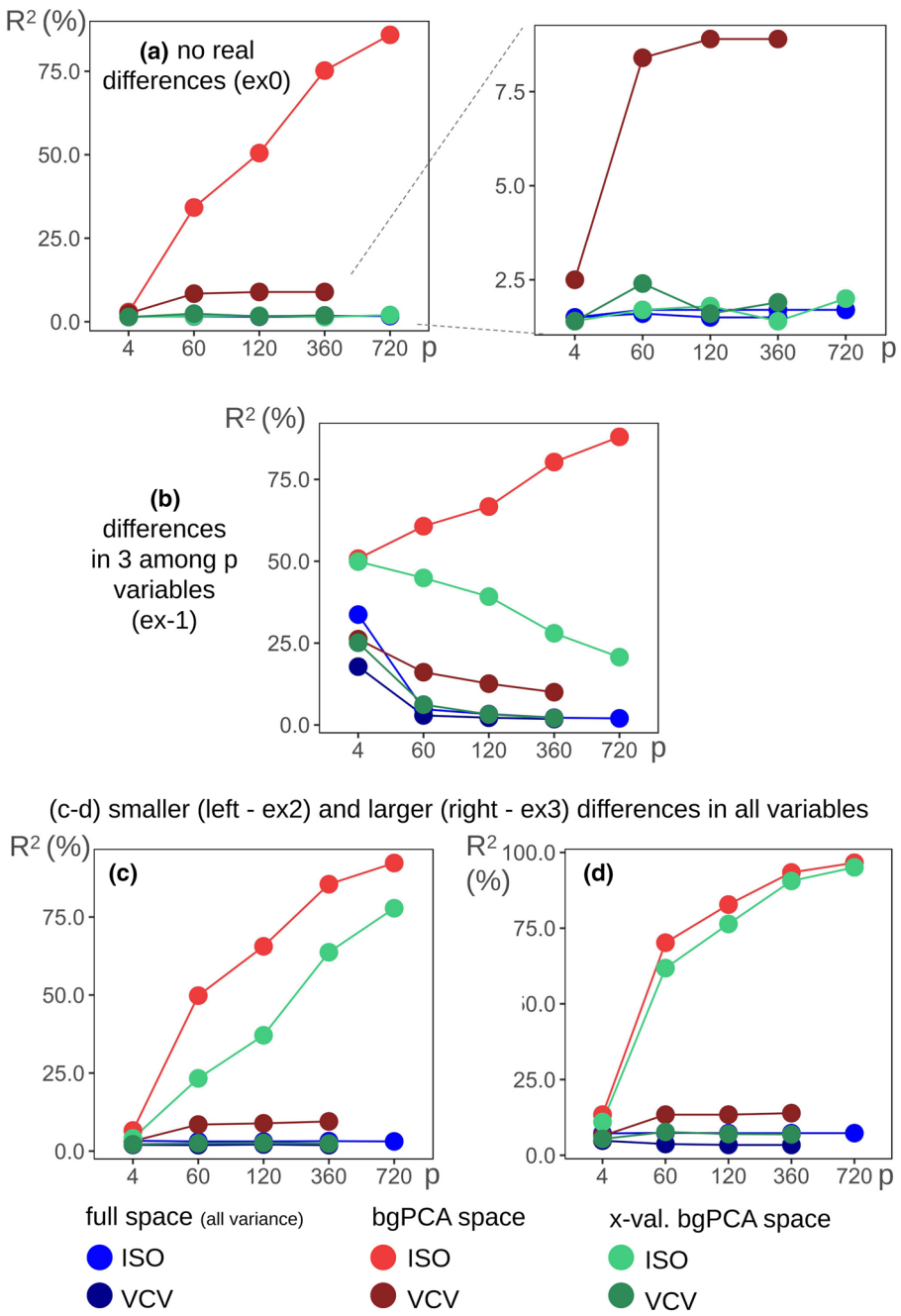


Fig. 6 Median percentage of variance explained by groups (multivariate R^2). As anticipated in Fig. 5, R^2 was computed not only in the full space, but also in the subspaces of the bgPCA (without or with cross-validation). The R^2 in the bgPCA space only serves as an index of the inflation of group differences in this subspace: the further from the full space R^2, the stronger the inflation of differences

Finally, Fig. 7 summarizes the estimates of the distortion of the group mean relationships using the median EDM of results from the 100 runs of each simulation. In this respect, these findings are congruent with the previous analyses, but this analysis nevertheless has an element of novelty. In agreement with all previous findings, x-validated bgPC scores outperform conventional non-x-validated analyses by showing less distorted patterns of between-group separation. However, the results in Fig. 7 show that the spatial arrangement of group means in the x-validated bgPCA space is more distorted with the VCV data than with ISO.

Conclusions

In conclusion, our study reconfirms that bgPCA potentially inflates group differences and can even introduce apparent differences where none exist. This effect is more pronounced as the number of variables increases. However, using x-validation to construct the bgPCA space can completely counteracts these spurious effects or largely mitigates them. Thus:

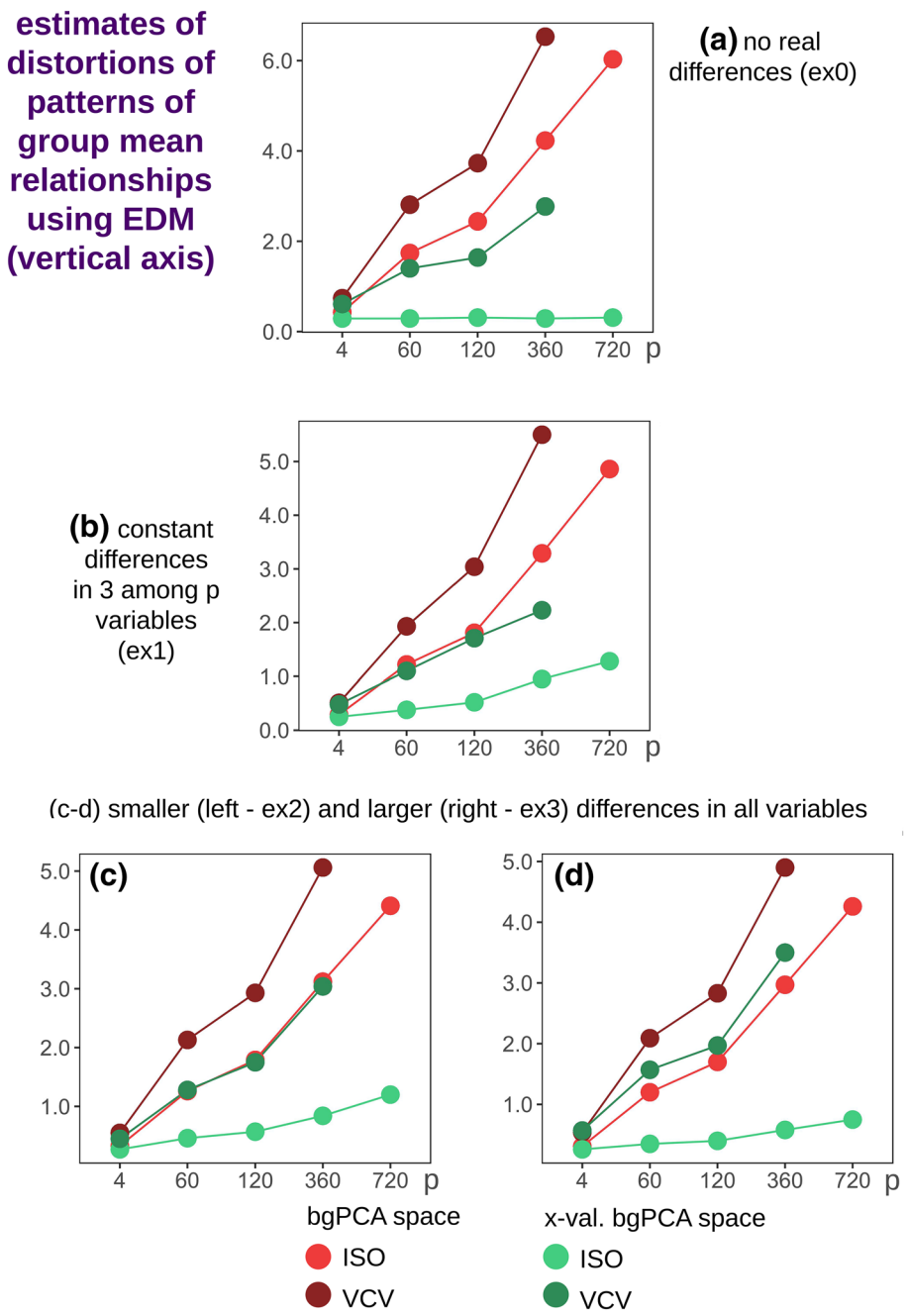


Fig. 7 Median EDM used to estimate the distortion of relationships between group means in bgPCA spaces (without or with x-validation) relative to the theoretical pattern of mean differences used in the model which generated the random data: with no distortion, EDM should be about zero

(1) bgPCA inflates spurious group differences as p increases, as reported in previous studies (Cardini et al. 2019 and Bookstein 2019).
(2) The spurious effects of bgPCA are less evident in data sets where the variables covary, which is the case in most real biological datasets, because the bias increases more slowly as p increases than in data where the variables are uncorrelated.
(3) x-validation almost completely removes the spurious separation between groups from bgPCA scatterplots. Nevertheless, we recommend that users report an estimate of effect size such as R^2 or Z and the P value of a test for group mean differences, both calculated in the full multivariate dataspace. For instance, finding a small but significant percentage of variance accounted for by group differences will tell users that, even if groups may look well separated in the x-validated plot, differences are likely to be real but most of the variation in the total sample is not due to group structure. On the other hand, regardless of how large R^2 is and what the scatterplot may suggest, non-significance will indicate

that samples are not large enough for confidently infer patterns of group variation.

(4) In all instances, and especially for covarying variables, the relative positions of group means in bgPCA space become increasingly distorted as p increases, even with x-validation (although in this latter case the impact is less pronounced).

If the results we obtained using g = 3, balanced samples and the covariance structure of traditional morphometric measures from adult male vervet skulls generalize, it would seem that x-validated bgPC scores almost completely solve the issue of spurious and inflated group separation when the variance present in a set with a large number of variables is 'squeezed' in the small g-1 space of a bgPCA. Because increasing g (with the same p and within group sample size) mitigates the problem of spurious and inflated group differences in bgPCAs, it is likely that our results will be robust when applied to data with a larger number of groups, as, for instance, in most taxonomic studies (see Table 1 of Cardini et al. 2019). However, the effect of highly heterogeneous sample size should be explored and, when p is large, the pattern of group mean differences captured by a bgPCA will definitely become less reliable. This, as suggested also by other morphometric studies (Cardini et al. 2015; Cardini and Elton 2007), reinforces the general observation that accurate estimates of group means require large samples and samples must be especially large if many variables are analysed. As obtaining large numbers of variables becomes more common, thanks to new technologies and methodological advancement (e.g., semilandmark analyses in geometric morphometrics (Adams et al. 2013) and microarray data in genetics (Culhane et al. 2002), the importance of very large samples to achieve accuracy becomes even more crucial, an observation that raises profound questions especially for vertebrate palaeontologists and student of human evolution, that almost inevitably have to rely on small sample sizes of precious, but rare and frequently fragmented material.

Acknowledgements We are deeply grateful to Norm Campbell for his feedback on this study and more generally on the problems with bgPCA scatterplots. In fact, Norm independently came up with the same idea (i.e., using x-validated scores). We greatly thank him for sharing his thoughts and hope he will go on researching and developing solutions to this and related issues concerning group ordinations based on highly dimensional data. We are in debt also to Dean Adams and an anonymous reviewer for their constructive and helpful comments (as well as R code!), which contributed to greatly improve this paper.

References

D. Adams, F. Rohlf, D. Slice: A field comes of age: Geometric morphometrics in the 21st century. Hystrix, the Italian Journal of Mammalogy **24**(1), 7–14 (2013)

Bookstein, F. L. (2019). Pathologies of between-groups principal components analysis in geometric morphometrics. *Evolutionary Biology, 46*, 280–302.

Cardini, A., & Elton, S. (2007). Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology, 126*, 121–134.

Cardini, A., & Elton, S. (2017). Is there a" Wainer's rule"? Testing which sex varies most as an example analysis using GueSDat, the free Guenon Skull Database. *Hystrix, the Italian journal of mammalogy, 28*(2), 147–156.

A. Cardini, K. Seetah, G. Barker: How many specimens do I need? Sampling error in geometric morphometrics: Testing the sensitivity of means and variances in simple randomized selection experiments. Zoomorphology **134**(2), 149–163 (2015). https://doi.org/10.1007/s00435-015-0253-z

A. Cardini, P. O'Higgins, F.J. Rohlf: Seeing distinct groups where there are none: spurious patterns from between-group PCA. Evolutionary Biology **46**, 303–316 (2019)

Collyer, M. L., Sekora, D. J., & Adams, D. C. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity, 115*(4), 357–365. https://doi.org/10.1038/hdy.2014.75.

Culhane, A. C., Perrière, G., Considine, E. C., Cotter, T. G., & Higgins, D. G. (2002). Between-group analysis of microarray data. *Bioinformatics, 18*(12), 1600–1608. doi:https://doi.org/10.1093/bioinformatics/18.12.1600.

O. Hammer, D. Harper, P. Ryan: PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica **4**(1), 1–9 (2001)

Kovarovic, K., Aiello, L. C., Cardini, A., & Lockwood, C. A. (2011). Discriminant function analyses in archaeology: are classification rates too good to be true? *Journal of Archaeological Science, 38*, 3006–3018.

Marcus, L. F. (1990). Traditional morphometrics. In *Proceedings of the Michigan morphometrics workshop* (Vol. 2, p. 77–122). Special Publication.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). vegan: Community Ecology Package. http://CRAN.R-project.org/package=vegan.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schlager, S. (2017). Morpho and Rvcg – Shape Analysis in R. In G. Zheng, S. Li, & G. Szekely (Eds.), *Statistical Shape and Deformation Analysis* (pp. 217–256). Academic Press.

A. Siberchicot, A. Julien-Laferrière, A.B. Dufour, J. Thioulouse, S. Dray: Adegraphics: An S4 lattice-based package for the representation of multivariate data. The R Journal **9**(2), 198–212 (2017)

W.N. Venables, B.D. Ripley: Modern Applied Statistics with S. Springer, New York (2002)