An Eigenvalue Test for spatial Principal Component Analysis

Running title: A new statistical test for sPCA

Word count: 2982

Montano V[1*] and Jombart T[2]

[1] School of Biology, University of St Andrews, Bute Building, St Andrews KY16 9TS, UK

[2] MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK

[*]Corresponding author: mirainoshojo@gmail.com

Abstract

- The spatial Principal Component Analysis (sPCA, Jombart 2008) is designed to investigate non-random spatial distributions of genetic variation. Unfortunately, the associated tests used for assessing the existence of spatial patterns (*global and local test*; Jombart et al. 2008) lack statistical power and may fail to reveal existing spatial patterns.
- Here, we present a non-parametric test for the significance of specific patterns recovered by sPCA.
- We compared the performance of this new test to the original *global* and *local* tests using datasets simulated under classical population genetic models. Results show that our test outperforms the original *global* and *local* tests, exhibiting improved statistical power while retaining similar, and reliable type I errors. Moreover, by allowing to test various sets of axes, it can be used to guide the selection of retained sPCA components. As such, it represents a valuable complement to the original analysis, and should prove useful for the investigation of spatial genetic patterns.

Keywords; eigenvalues; sPCA; spatial genetic patterns; Monte-Carlo

INTRODUCTION

The principal component analysis (PCA; Pearson 1901; Hotelling 1933) is one of the most common multivariate approaches in population genetic (Jombart et al 2009). Although PCA is not explicitly accounting for spatial information, it has often been used for investigating spatial genetic patterns (Novembre and Stephens 2008). As a complement to PCA, the spatial principal component analysis (sPCA; Jombart et al. 2008) has been introduced to explicitly include spatial information in the analysis of genetic variation, and gain more power for investigating spatial genetic structures.

sPCA finds synthetic variables, the principal components (PCs), which maximise both the genetic variance and the spatial autocorrelation as measured by Moran's *I* (Moran 1950). As such, PCs can reveal two types of patterns: '*global*' structures, which correspond to positive autocorrelation typically observed in the presence of patches or clines, and '*local*' structures, which correspond to negative autocorrelation, whereby neighboring individuals are more genetically distinct that expected at random (Jombart et al.. 2008). The *global* and *local* tests have been developed for detecting the presence of global and local patterns, respectively (Jombart et al. 2008). Unfortunately, while these tests have robust type I error, they also typically lack power, and can therefore fail to identify existing spatial genetic patterns (Jombart et al.. 2008). Moreover, they can only be used to diagnose the presence or absence of spatial patterns, and are unable to test the significance of specific structures revealed by sPCA axes.

In this paper, we introduce an alternative statistical test which addresses these issues. This approach relies on computing the cumulative sum of a defined set of sPCA eigenvalues as a test statistic, and uses a Monte-Carlo procedure to generate null

distributions of the test statistics and approximate p-values. After describing our approach, we compare its performances to the global and local tests using simulating datasets, investigating several standard spatial population genetics models. Our approach is implemented as the function *spca_randtest* in the package *adegenet* (Jombart 2008; Jombart and Ahmed 2011) for the R software (R Core Team 2017).

METHODS

Test statistic

As in most multivariate analyses of genetic markers, our approach analyses a table of centred allele counts or frequencies, in which rows represent individuals or populations, and columns correspond to alleles of various loci (Jombart et al 2008; Jombart et al 2009; Jombart et al 2010). We note X the resulting matrix, and n the number of individuals analysed. In addition, the sPCA introduces spatial data in the form of a n by n matrix of spatial weights L, in which the i^{th} row contains weights reflecting the spatial proximity of all individuals to individual i. The PCs of sPCA are then found by the eigen-analysis of the symmetric matrix (Jombart et al. 2008):

$$1/(2n)\, X^{T}(L^{T} + L)X \quad (1)$$

We note λ the corresponding non-zero eigenvalues. We differentiate the r positive eigenvalues λ^{+}, corresponding to global structures, and the 's' negative eigenvalues λ^{-}, corresponding to local structures, so that $\lambda = \{\lambda^{+},\lambda^{-}\}$. Without loss of generality, we assume both sets of eigenvalues are ordered by decreasing absolute value, so that $\lambda_1^{+} > \lambda_2^{+} > \ldots > \lambda_r^{+}$ and $|\lambda_1^{-}| > |\lambda_2^{-}| > \ldots > |\lambda_s^{-}|$.

Simply put, each eigenvalue quantifies the magnitude of the spatial genetic patterns in the corresponding PC: larger absolute values indicate stronger global (respectively local) structures. We note $V^{+} = \{v_1^{+}, \ldots, v_r^{+}\}$ and $V^{-} = \{v_1^{-}, \ldots, v_s^{-}\}$ the sets of corresponding PCs.The most natural choice of test statistic to assess whether a given PC contains significant structure would seem to be the corresponding eigenvalue. This would, however, not account the dependence on previous PCs: v_j^{+} (respectively v_j^{-}) can only be significant if all previous PCs $\{v_1^{+}, \ldots, v_{j-1}^{+}\}$ are also significant. To account for this, we define the test statistic for v_j^{+} as:

$$f_i^{+} = \Sigma_{i = 1, \ldots, j}\, \lambda_i^{+}$$

and as:

$$f_i^- = \Sigma_{i = 1, \ldots, j} |\lambda_i^-|$$

for v_j^-.

Permutation procedure

f_i^+ and f_i^- become larger in the presence of strong global or local structures in the first i^{th} global / local PCs. Therefore, they can be used as test statistics against the null hypotheses of absence of global or local structures in these PCs. The expected distribution of f_i^+ and f_i^- in the absence of spatial structure is not known analytically. Fortunately, it can be approximated using a Monte-Carlo procedure, in which individual genetic profiles are permuted randomly along the connection network, computing f_i^+ and f_i^- for each permutation. Note that the original values of the test statistic are also included in these distributions, as the initial spatial configuration is by definition a possible random outcome. The *p*-values are then computed as the relative frequencies of permuted statistics equal to or greater than the initial value of f_i^+ or f_i^-.

To guide the selection of global and local PCs to retain, this testing procedure can be used with increasing numbers of retained axes. Because each test is conditional on the previous tests, incremental Bonferroni correction is used to avoid the inflation of type I error, so that the significance level for the i^{th} PC will be α / i, where α is the target type I error. The entire testing procedure is implemented in the function `spca_randtest` in the package *adegenet* (Jombart 2008; Jombart and Ahmed 2011) for R (R Core Team 2016).

Simulation study

To assess the performance of our test, we simulated genetic data under three migration models: island (IS) and stepping stone (SS), using the software GenomePop 2.7 (Carvajal-Rodríguez 2008), and isolation by distance (IBD), using *IBDSimV2.0* (Leblois 2009). We simulated the IS and SS models with 4 populations, each with 25 individuals, and a single population under IBD with 100 individuals. 200 unlinked SNPs diploid loci were simulated. Populations evolved under constant effective population size $\theta = 20$, and interchanged migrants at three different symmetric and homogeneous rates (0.005, 0.01, and 0.1). We performed 100 independent runs for each of the three migration rates, for a total of 300 simulated dataset per migration model.

To quantify rates of errors type I for the *spca_randtest, global* and *local tests*, we extracted 100 random coordinates from 10 square 2D grids, using the function *spsample* from the *spdep* package (Bivand et al. 2013). In order to evaluate the rate of false negatives for global patterns, we manually generated 10 sets of 100 pairs of coordinates simulating gradients and/or patches from 2D grids. To test for the rate of false negatives for local patterns, we perform a principal component analysis on 10 random datasets simulated under the SS model with 0.005 migration rate. We used the coordinates of the individuals on the first principal component and set the second coordinate to zero for all individuals (1D). With the coordinates so produced, we used the function *chooseCN* in adegenet to obtain 10 neighbouring graphs where the most genetically distinct individuals (falling in the upper quartile of the pairwise genetic distances) are considered as neighbors, while the others are non-neighbors.

We tested 100 simulations each for all the 30 sets of geographic coordinates (random, positive and negative), for each of the three migration rates (0.005, 0.01 and 0.1), for each

of the three migration models (IS, SS, IBD; total of 9,000 tests per migration model). We repeated all tests using a subset of 40 SNPs per individual, for a total of 18,000 tests in the absence of spatial structures, and and 36,000 tests in the presence of global or local structures.

RESULTS

Statistical power of the spca_randtest

We compared the performances of the *spca_randtest* with the *global* and *local* tests in three settings: in the absence of spatial structure, and in the presence of global, and local structures. The results obtained in the absence of spatial structure show that all tests have reliable type I errors (Table 1 and 2). The *spca_randtest* exhibited consistently better performances for detecting existing structures in the data than both *global* and *local tests* (Table 1 and 2, Figure 1). Although our simulated local spatial patterns turned out more difficult to detect than global patterns, the *spca_randtest* is twice to five times more effective than the *local test* (Table 1 and 2). Generally, the underlying migration model, the migration rate and the number of loci affect the ability of all tests to detect non-random spatial patterns. Both *spca_randtest* and *global* and *local tests* have in fact a lower sensitivity in presence of island migratory schemes, while results for stepping stone and isolation by distance models are more satisfying (Table 1 and 2). Increasing migration rates lead to a higher rates of false negatives for all tests, which can be overcome using more loci (Table 1 and 2).

Significant eigenvalues are assessed using a hierarchical Bonferroni correction which accounts for non-independence of eigenvalues and multiple testing (Figure 2). Strong patterns (e.g. IBD) tend to produce a higher number of significant components than weak patterns (e.g. island models with high migration rates), which are otherwise captured by fewer to no components.

CONCLUSIONS

We introduced a new statistical test associated to the sPCA to evaluate the statistical significance of global and local spatial patterns. Using simulated data, we show that this new approach outperforms previously implemented tests, having greater statistical power (lower type II errors) whilst retaining consistent type I errors. Our simulations also suggest that demographic settings and migratory models can substantially impact the ability to detect spatial patterns. The impact of specific factors such as the effective population size or the number of individuals sampled per population remain to be investigated.

Acknowledgements

The authors declare no conflict of interest

Author contributions

Test development: VM and TJ. Data analysis: VM. Wrote the manuscript: VM and TJ.

Literature

1. Bivand RS, Pebesma E, Gómez-Rubio V (2013) *Applied Spatial Data Analysis with R*. Springer, New York, 378pp.
2. Balkenhol N, Gugerli F, Cushman SA, Waits LP, Coulon A, Arntzen JW, Holderegger R, Wagner HH (2009) Identifying future research needs in landscape genetics: where to from here? *Landscape Ecology*, **24**, 455.
3. Carvajal-Rodríguez A (2008) GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics*, **9**, 223.
4. Cushman SA, Landguth EL (2010) Spurious correlations and inference in landscape genetics. *Molecular Ecology*, **19**, 3592–3602.
5. Hotelling H (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24,** 417.
6. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.
7. Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**, 92–103.
8. Jombart T, Pontier D, Dufour AB (2009). Genetic markers in the playground of multivariate analysis. *Heredity* **102,** 330–341.
9. Jombart T, Devillard S, Balloux F (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics* **11,** 94.
10. Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27,** 3070–3071.
11. Moran PAP (1950) Notes on Continuous Stochastic Phenomena. *Biometrika*, **37**, 17–23.

12. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

13. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, **2,** 559–572.

14. Peres-Neto PR, Jackson DA, Somers KM (2005) How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis,* **49**, 974 – 997.

15. R Core Team (2017) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

16. Schiffers K , Travis JMJ (2014) ALADYN - a spatially explicit, allelic model for simulating adaptive dynamics. *Ecography*, **37**, 1288–1291.

Legends

Figure 1. Graphical representation of the results reported in Tables 1 and 2. Only test with threshold 0.05 are plotted.

Figure 2. Distributions of significant eigenvalues detected in the presence of global (blue bars) and local (green bars) spatial patterns after hierarchical Bonferroni correction, for 100 significantly positive and 100 significantly negative patterns. Black bars correspond to eigenvalues which are significant without Bonferroni correction. Bars' height indicates the frequency of observing a significant eigenvalue in a certain position (from most positive to most negative) over the 100 tested patterns.

Table 1. Significant results for g*lobal test* (g test), l*ocal tests* (ll test), and *spca_randtest* (r test +/-) for random, global and local patterns using 200 loci per individual. IS, SS, IBD indicate the migration models (see Methods); different migration rates are coded by number: 1 = 0.005, 2 = 0.01 and 3 = 0.1. Results show the proportion of significant tests over 1,000 replicates, based on 1,000 permutations with thresholds .05 and .01.

200 SNPs		**Random Patterns**				**Global Patterns**				**Local Patterns**			
Models	*p*-value*	g test	r test (+)	l test	r test (-)	g test	r test (+)	l test	rt est (-)	g test	r test (+)	l test	r test (-)
IS-1	.05	0.054	0.059	*0.041*	*0.047*	**0.947**	**0.985**	*0.029*	*0.001*	*0.047*	0.071	0.061	**0.284**
	.01	0.011	*0.007*	*0.009*	0.010	**0.822**	**0.948**	*0.005*	*0.001*	*0.008*	0.010	0.015	0.113
IS-2	.05	*0.040*	*0.041*	0.058	0.056	**0.227**	**0.564**	*0.044*	*0.018*	0.056	0.059	0.050	0.123
	.01	*0.007*	*0.009*	*0.009*	0.013	0.067	**0.302**	*0.005*	*0.002*	0.011	*0.007*	0.012	0.026
IS-3	.05	0.051	*0.040*	0.053	*0.041*	0.055	*0.049*	*0.045*	*0.047*	*0.049*	*0.047*	*0.044*	0.059
	.01	0.010	0.014	0.013	*0.008*	0.010	0.013	*0.007*	0.013	*0.002*	*0.014*	*0.008*	0.019
SS-1	.05	0.053	0.058	0.053	0.050	**0.986**	**0.996**	*0.022*	*0.000*	0.063	0.064	0.124	**0.582**
	.01	*0.007*	0.011	0.010	0.010	**0.960**	**0.988**	*0.002*	*0.000*	0.017	0.010	*0.041*	**0.398**
SS-2	.05	*0.044*	0.058	0.058	0.063	**0.798**	**0.909**	*0.047*	*0.004*	*0.034*	*0.044*	0.059	**0.316**
	.01	0.011	0.011	0.013	0.016	**0.676**	**0.771**	*0.010*	*0.000*	*0.004*	*0.005*	0.014	0.147
SS-3	.05	*0.047*	*0.046*	0.057	*0.049*	0.054	0.128	*0.040*	*0.042*	*0.044*	0.054	*0.049*	0.071
	.01	0.014	*0.007*	0.011	0.013	0.014	0.036	*0.006*	0.010	*0.003*	*0.009*	*0.006*	*0.009*
IBD-1	.05	*0.044*	0.050	0.053	*0.048*	**0.962**	**0.999**	*0.021*	*0.000*	*0.025*	0.087	**0.438**	**0.809**
	.01	*0.008*	0.012	*0.009*	0.010	**0.926**	**0.997**	*0.003*	*0.000*	*0.009*	0.023	0.192	**0.694**
IBD-2	.05	0.052	*0.045*	0.061	*0.038*	**0.967**	**0.998**	*0.023*	*0.000*	*0.046*	0.076	**0.451**	**0.794**
	.01	*0.009*	*0.008*	0.011	*0.009*	**0.932**	**0.997**	*0.004*	*0.000*	*0.009*	0.018	**0.208**	**0.672**
IBD-3	.05	0.052	*0.046*	0.053	0.050	**0.977**	**0.999**	*0.015*	*0.000*	0.050	0.083	**0.441**	**0.824**

	.01	0.013	*0.009*	0.011	0.012	**0.939**	**0.999**	*0.005*	*0.000*	*0.009*	0.023	**0.225**	**0.684**

218 ***p-values** are in italic when non significant and in bold when the fraction of true positive is above 20%

Table 2. Results for the same simulations reported in Table 1 using a subset of 40 loci per individual.

40 SNPs		Random Patterns				Global Patterns				Local Patterns			
Models	p-value*	g test	r test (+)	l test	r test (-)	g test	r test (+)	l test	r test (-)	g test	r test (+)	l test	r test (-)
IS-1	.05	0.052	0.061	0.046	0.050	**0.591**	**0.807**	*0.033*	*0.004*	*0.036*	*0.000*	0.055	0.077
	.01	0.016	0.013	0.010	*0.007*	**0.393**	**0.592**	*0.005*	*0.000*	*0.004*	*0.000*	0.015	0.022
IS-2	.05	0.053	*0.047*	*0.038*	*0.042*	0.103	**0.226**	*0.046*	*0.020*	0.073	*0.000*	0.057	*0.038*
	.01	0.011	*0.009*	*0.006*	*0.006*	0.022	0.072	0.011	0.005	0.012	*0.000*	0.010	*0.006*
IS-3	.05	*0.047*	0.050	0.050	*0.045*	*0.048*	0.060	*0.044*	*0.042*	*0.036*	*0.000*	0.053	*0.026*
	.01	*0.009*	0.011	*0.008*	*0.007*	*0.009*	0.011	0.011	0.011	*0.002*	*0.000*	0.013	*0.001*
SS-1	.05	0.052	0.054	*0.039*	*0.049*	**0.898**	**0.949**	*0.017*	*0.000*	0.050	*0.001*	0.067	0.169
	.01	*0.009*	0.012	*0.005*	0.011	**0.826**	**0.865**	*0.006*	*0.000*	*0.007*	*0.000*	0.021	0.052
SS-2	.05	*0.046*	*0.045*	0.050	*0.046*	**0.528**	**0.588**	*0.044*	*0.009*	0.052	*0.000*	*0.048*	0.081
	.01	0.013	0.010	0.010	0.015	**0.377**	**0.370**	0.016	*0.000*	*0.005*	*0.000*	0.011	0.014
SS-3	.05	0.068	*0.040*	0.050	*0.048*	0.066	0.055	0.053	*0.033*	*0.026*	*0.000*	*0.047*	*0.023*
	.01	0.014	*0.005*	0.013	0.012	0.012	*0.009*	*0.005*	*0.006*	*0.006*	*0.000*	*0.008*	*0.000*
IBD-1	.05	*0.049*	0.053	0.052	0.057	**0.822**	**0.883**	*0.027*	*0.002*	*0.034*	0.055	0.124	**0.480**
	.01	*0.005*	*0.008*	0.013	0.013	**0.755**	**0.742**	*0.004*	*0.000*	*0.005*	*0.008*	0.032	**0.278**
IBD-2	.05	*0.043*	0.054	0.060	*0.049*	**0.835**	**0.880**	*0.028*	*0.001*	*0.043*	0.051	0.111	**0.458**
	.01	0.011	*0.007*	0.015	*0.009*	**0.755**	**0.732**	*0.005*	*0.000*	*0.008*	0.015	0.026	**0.259**
IBD-3	.05	*0.043*	*0.042*	0.051	0.050	**0.844**	**0.899**	*0.026*	*0.002*	*0.048*	0.058	0.115	**0.465**
	.01	0.012	0.013	0.012	0.010	**0.763**	**0.756**	*0.007*	*0.000*	*0.009*	0.010	0.023	**0.263**

**p-values* are in italic when non significant and in bold when the fraction of true positive is above 20%

221 **Figure 1**

222

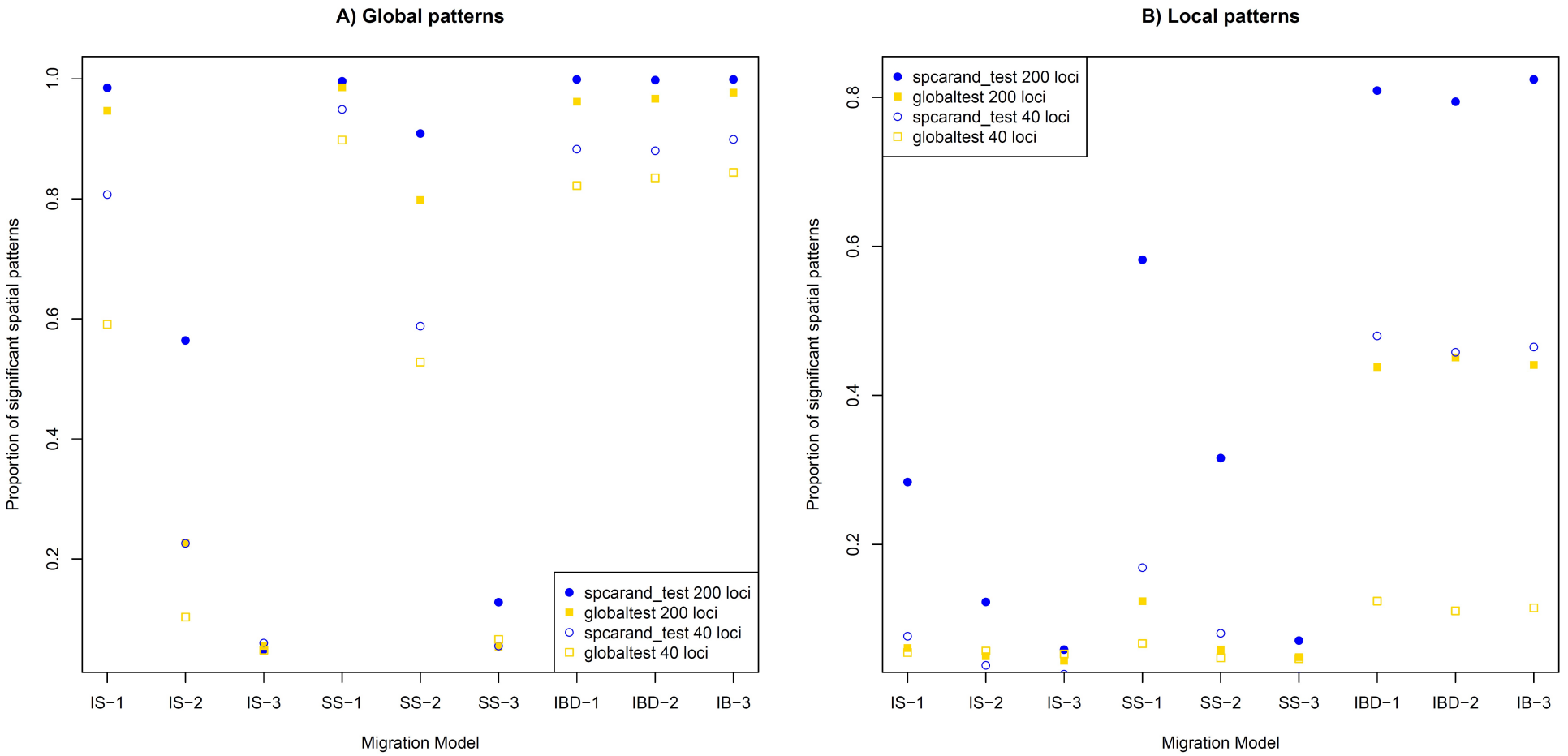
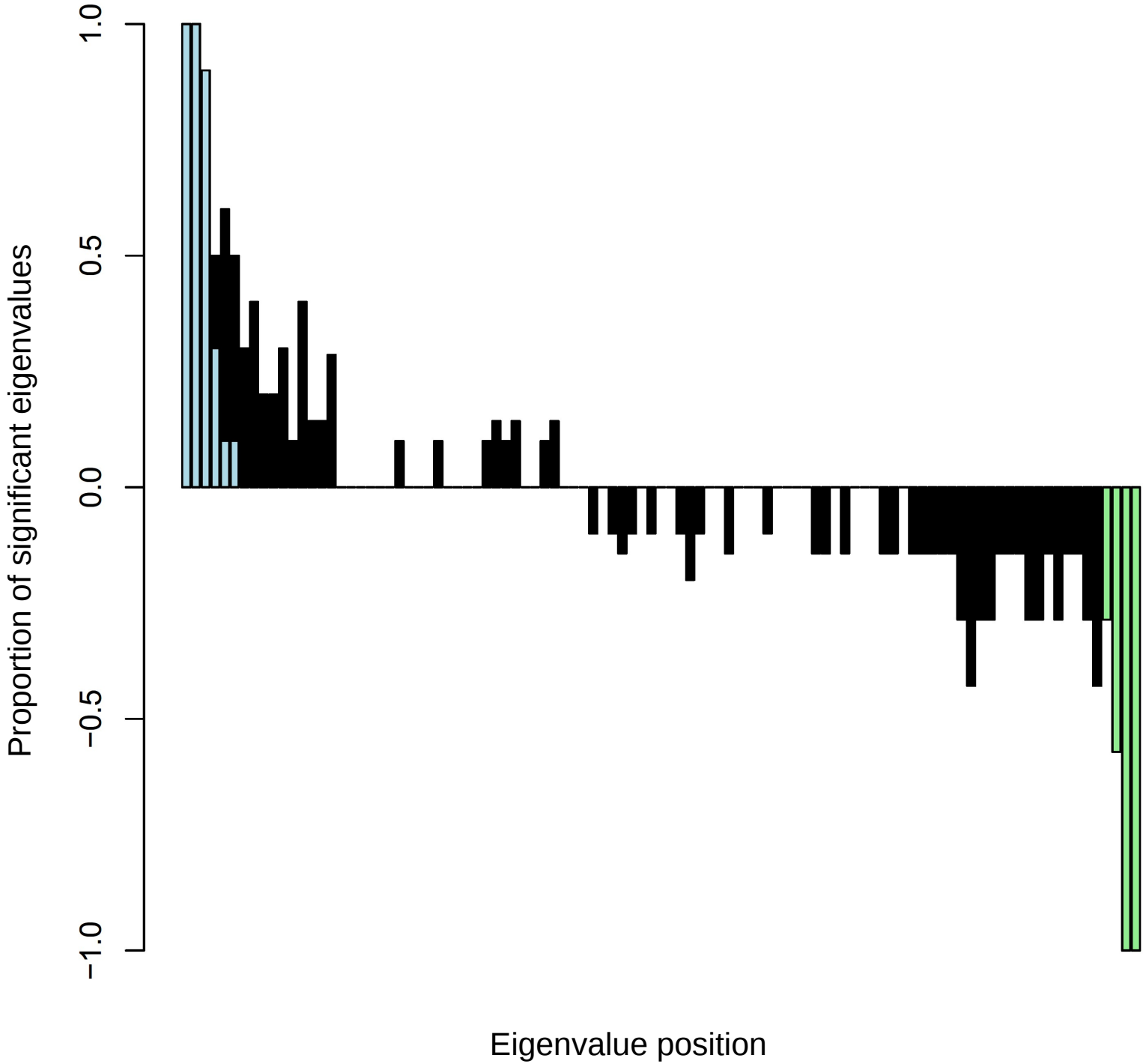


Figure 2


Proportion of significant eigenvalues
1.0
0.5
0.0
−0.5
−1.0
Eigenvalue position