

Ten quick tips for effective dimensionality reduction

Nguyen, L.H. and Holmes, S.,

- 1: Choose an appropriate method
- 2: Preprocess continuous and count input data
- 3: Handle categorical input data appropriately
- 4: Use embedding methods for reducing similarity and dissimilarity input data
- 5: Consciously decide on the number of dimensions to retain
- 6: Apply the correct aspect ratio for your visualizations
- 7: Understand the meaning of the new dimensions
- 8: Find the hidden signal
- 9: Take advantage of multidomain data
- 10: Check the robustness of your results and quantify uncertainties

1. Choose an appropriate method

Method	Input Data	Method Class	Nonlinear	Complexity
PCA	continuous data	unsupervised		$\mathcal{O}(\max(n^2p, np^2))$
CA	categorical data	unsupervised		$\mathcal{O}(\max(n^2p, np^2))$
MCA	categorical data	unsupervised		$\mathcal{O}(\max(n^2p, np^2))$
PCoA (cMDS)	distance matrix	unsupervised		$\mathcal{O}(n^2p)$
NMDS	distance matrix	unsupervised		$\mathcal{O}(n^2h)$
Isomap	continuous*	unsupervised	✓	$\mathcal{O}(n^2(p + \log n))$
Diffusion Map	continuous*	unsupervised	✓	$\mathcal{O}(n^2p)$
Kernel PCA	continuous*	unsupervised	✓	$\mathcal{O}(n^2p)$
t-SNE	continuous/distance	unsupervised	✓	$\mathcal{O}(n^2p + n^2h)$
Barnes–Hut t-SNE	continuous/distance	unsupervised	✓	$\mathcal{O}(nh \log n)$
LDA	continuous (X and Y)	supervised		$\mathcal{O}(np^2 + p^3)$
PLS (NIPALS)	continuous (X and Y)	supervised		$\mathcal{O}(npd)$
NCA	distance matrix	supervised	✓	$\mathcal{O}(n^2h)$
Bottleneck NN	continuous/categorical	supervised	✓	$\mathcal{O}(nph)$
STATIS	continuous	multidomain		$\mathcal{O}(n^2P, nP^2)$
DiSTATIS	distance matrix	multidomain		$\mathcal{O}(n^2P, nP^2)$

Method	R function	Python function
PCA	stats::prcomp	sklearn.decomposition.PCA
CATPCA	gifi::princals	
CA	FactoMineR::CA	
MCA	FactoMineR::MCA	
PCoA (cMDS)	stats::cmdscale	sklearn.manifold.MDS
NMDS	ecodist::nmms	sklearn.manifold.MDS
Isomap	vegan::isomap	sklearn.manifold.Isomap
Diffusion Map	diffusionMap::diffuse	
(Barnes–Hut) t-SNE	Rtsne::Rtsne	sklearn.manifold.TSNE
LDA	MASS::lda	sklearn.discriminant_analysis.LinearDiscriminantAnalysis
PLS (NIPALS)	mixOmics::pls	sklearn.cross_decomposition.PLSRegression
DiSTATIS	DistatisR::distatis	
Procrustes	vegan::procrustes	scipy.spatial.procrustes

2. Preprocess continuous and count input data

- Normalizujte dáta. Ak napríklad používate PCA, ktoré sa snaží zachovať čo najviac variancie, tak ak máte atribút s veľkými hodnotami a veľkou varianciou, tak dostane väčšiu váhu vo výslednej reprezentácii.

3. Handle categorical input data appropriately

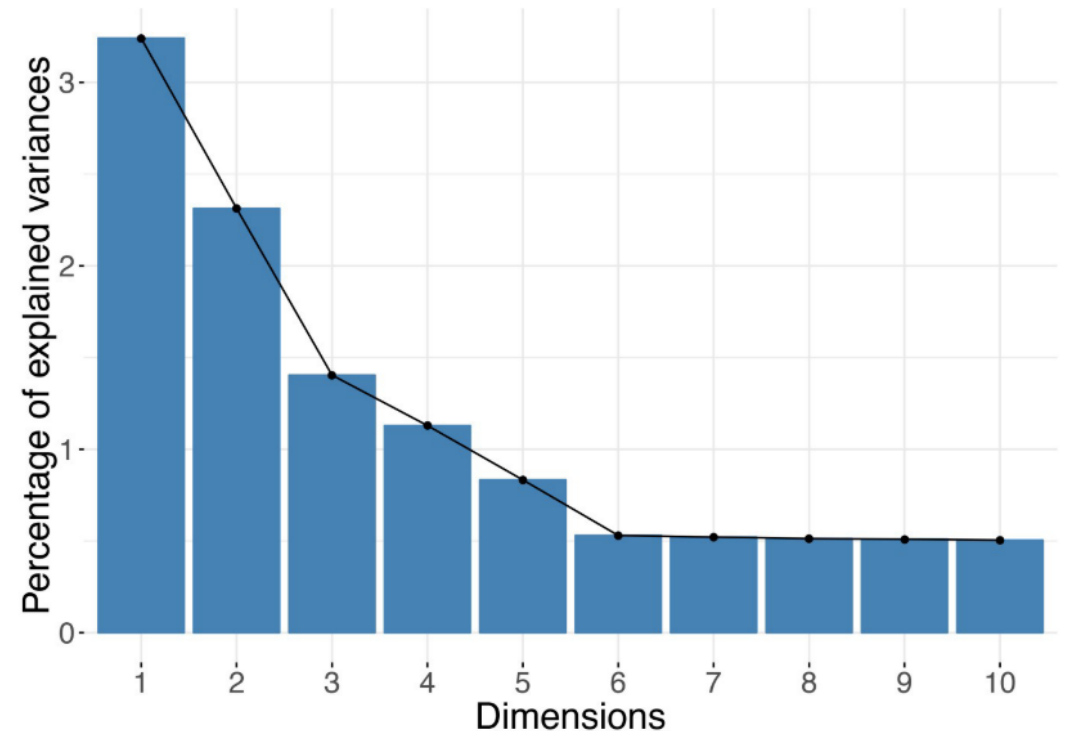
- Ak máme kategorický atribút, na ktorý aplikujeme One-hot encoding, tak môžeme dostať veľmi riedky dataset. Nie každá metóda redukcie dimenzionality sa na to hodí rovnako. Napríklad metódy založené na maticovej faktorizácii sú celkom vhodné. Obyčajné PCA ale nie. Vždy zvážte špecifickú metódu pre daný typ dát.

4. Use embedding methods for reducing similarity and dissimilarity input data

- Na redukcii dimenzionality nám niekedy môžu stačiť podobnosti pozorovaní (napríklad ak nevieme spočítať črty pozorovaní ale len **podobnosti/rozdielnosti**).
- Vtedy sa dajú použiť metódy transformujúce dáta do nejakého priestoru vektorovej reprezentácie (Např. t-SNE)
- Pozor tieto metódy sa snažia zachovať **lokálne interakcie** medzi pozorovaniami a **nie globálnu štruktúru**. Interpretovateľnosť takejto reprezentácie je teda výrazne odlišná. Napríklad ak použijete t-SNE na vizualizáciu, tak sa nemôžete pozerať na absolútnu pozíciu bodov (lebo nemá zmysel) ale len na vzdialenosť medzi bodmi a zoskupenia bodov (lebo odrážajú lokálne vlastnosti).

5. Consciously decide on the number of dimensions to retain

- Elbow method na výber počtu komponentov
- Pri vizualizácii v 2/3D netreba nutne používať len prvé komponenty

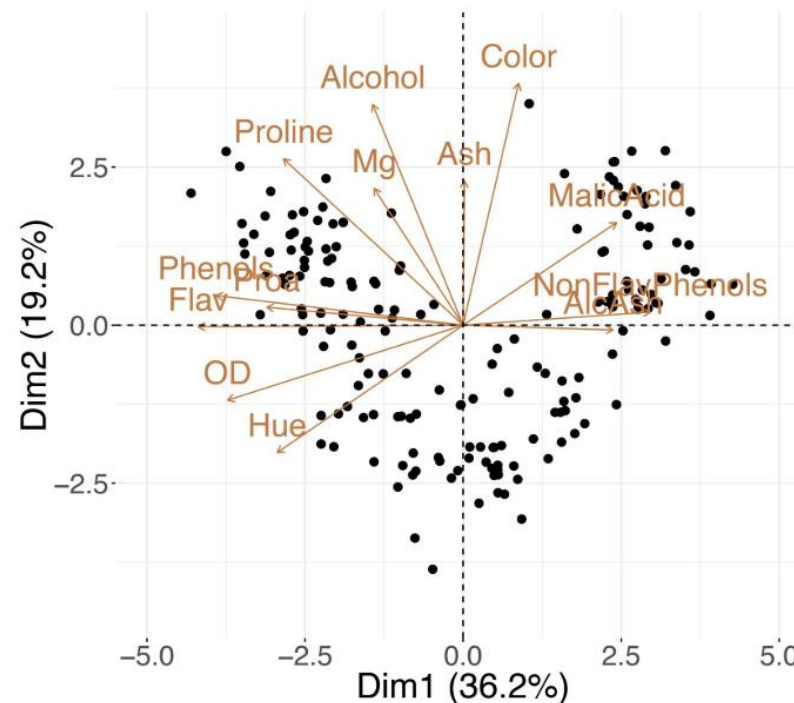


6. Apply the correct aspect ratio for your visualizations

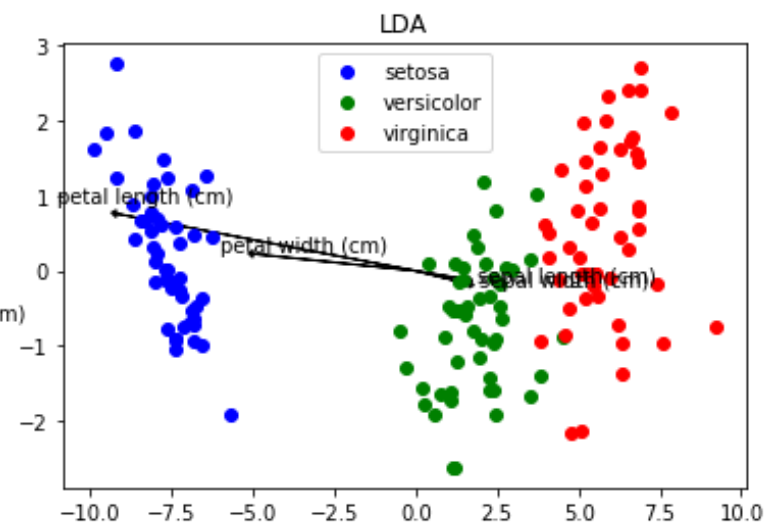
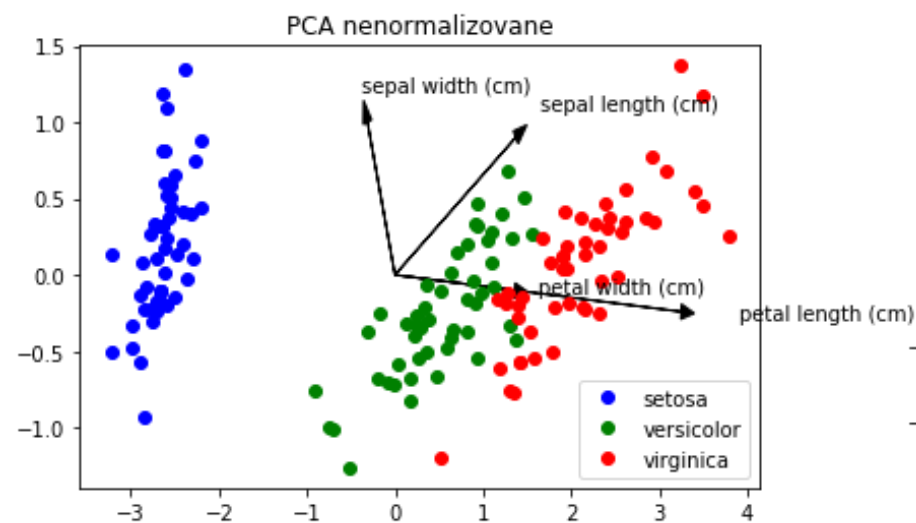
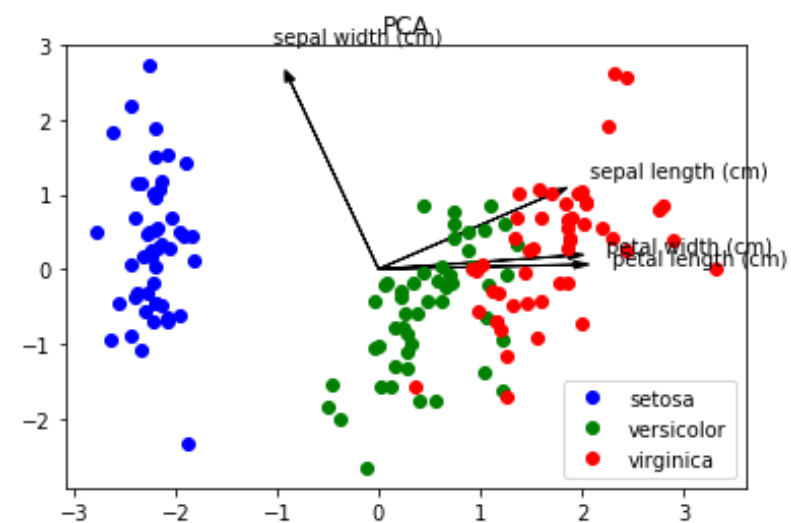
- Pomer škál osí je dôležitý
- Pri embedding metódach môže zlé rozmiestnenie v priestore indikovať nedostrénovaný model a bez správneho škálovania to nemusí byť vidieť.
- Pri PCA (a pri ďalších metódach) by mali škály odzrkadľovať podiel vysvetlenej informácie.

7. Understand the meaning of the new dimensions

- Redukcia dimenzionality nie je úplne čierna skrinka. Minimálne pri reprezentáciách používajúcich globálnu štruktúru sa dá interpretácia vytvorených dimenzií zobrazíť pomocou Biplotu.

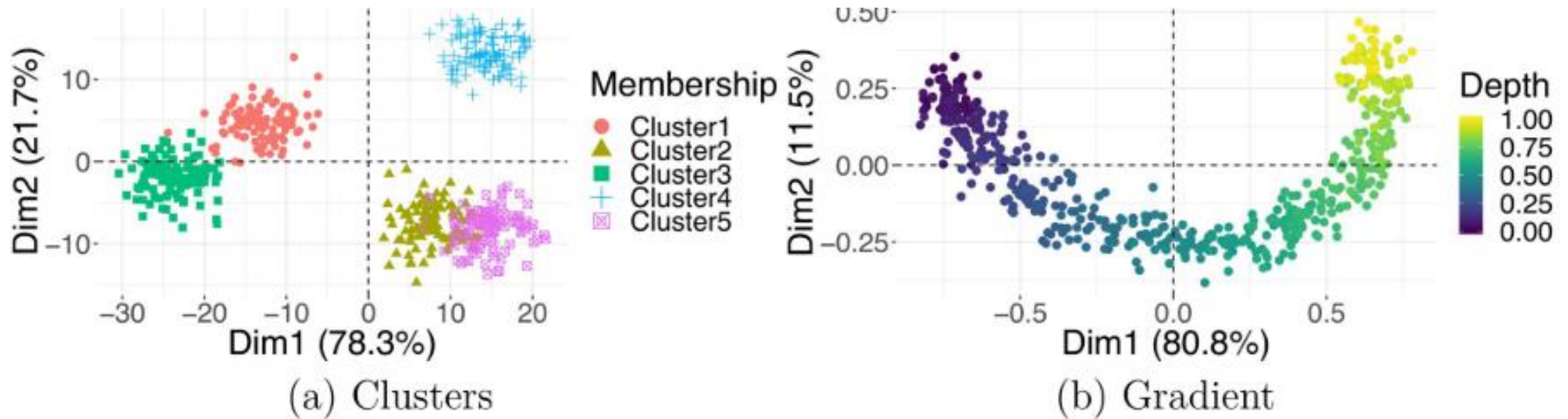


Biplot na rôznych redukciách datasetu IRIS



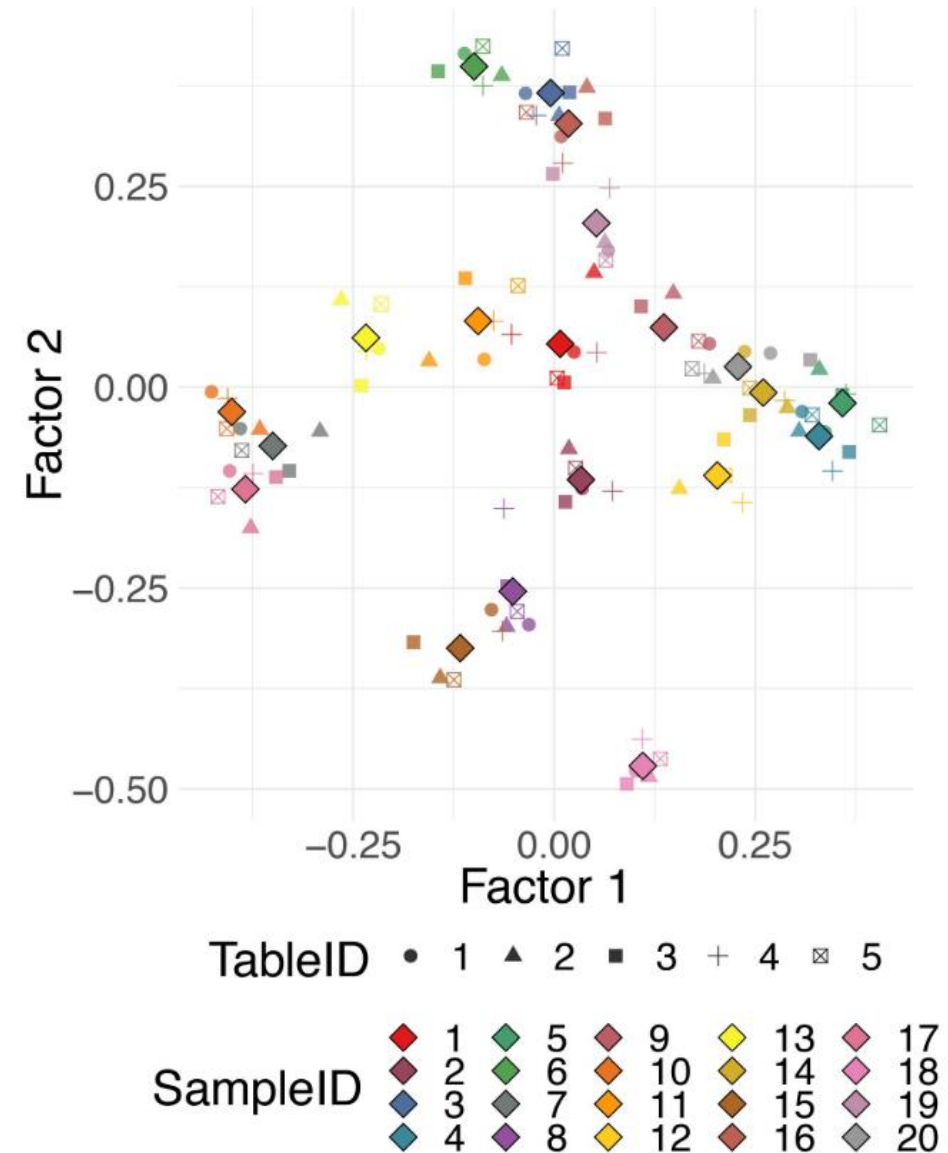
8. Find the hidden signal

- Overte či reprezentácia konzistentne odzrkadľuje signál ktorý chceme zachovať. Štruktúra, ktorú chceme zachovať zostala zachovaná. Podobné veci sú stále pohromade.



9. Take advantage of multidomain data

- Existujú metódy redukcie dimenzionality, ktoré dokážu spracovať nezávislé črty rovnakých pozorovaní. (STATIS, DiSTATIS)
- Na každú reprezentáciu sa dokonca dajú použiť rôzne metódy redukcie dimenzionality a potom sa dá hľadať mapovanie, ktoré ich spojí (Procrustes transformation)



10. Check the robustness of your results and quantify uncertainties

- Ako veľmi je výsledok redukcie závislý od malej zmeny parametrov alebo dát? Neovplyvnili mi príliš výsledok nejaké outliers?

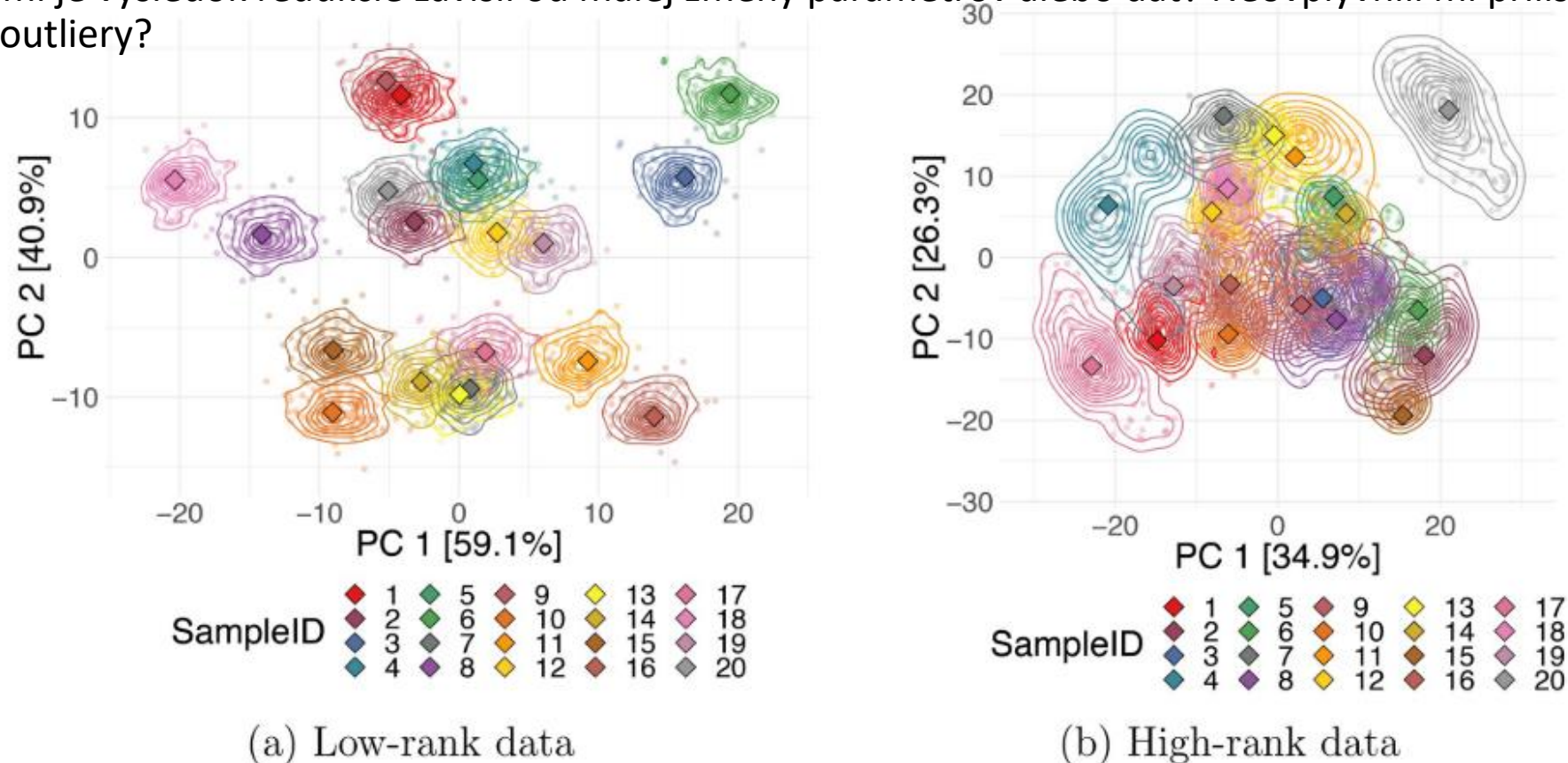


Fig 9. Data point uncertainties. Stability in the DR output coordinates for each data point. Projections of bootstrap samples for two 10D simulated datasets with rank 2 (a) and rank 5 (b) onto the first two PCs aligned using a Procrustes transformation. Smaller, circular markers correspond to each bootstrap trial, and larger, diamond markers are coordinates of the full dataset. DR, dimensionality reduction; PC, principal component.