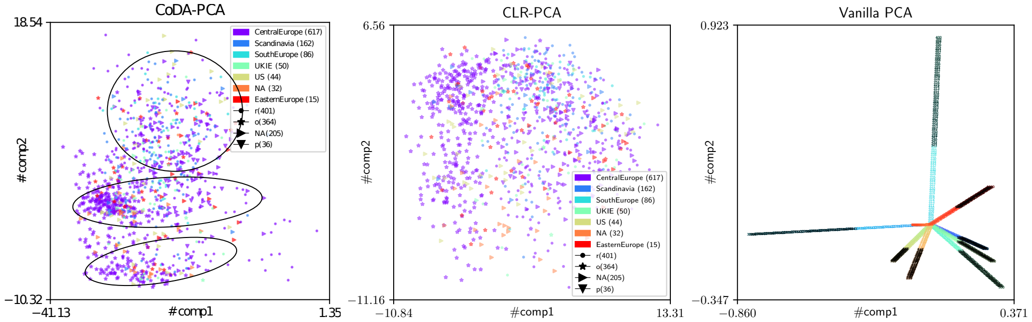

(a) Atlas: proposed approach (b) Atlas: classical approach (c) 10-arms

✉ **To Reviewer #1:** Thank you for recognising that *"a more principled method of compositional PCA would certainly be an interesting development"*. We will add a visualization but leave alr/ilr/clr+PCA to future work.

① **missing visualization:** This is mainly due to space limit, and we prefer to show quantitative comparisons. We do have some potentially meaningful visualizations on the studied datasets. For example, figure (a,b) shows the 2D plots on the Atlas dataset, where the colors mean the nationality and markers mean the DNA extraction method. CoDA-PCA in Fig. (a) uncovers a clustering structure (sketched with ellipses) which is not presented by CLR+PCA in Fig. (b).

② **alr/ilr/clr+PCA** are equivalent to each other, as alr/ilr/clr are different coordinate systems of the same affine space and can be linearly transformed back-and-forth. See figure (c) for our toy dataset under **(vanilla) PCA**, where the black dots indicates that the PCA reconstruction is outside of the simplex. PCA cannot be directly adapted to CoDA: the projection on the principal components, although still in the affine hull, may go beyond the convex hull of the vertices.

③ **nonlinear structure**: we will rephrase L52 and following: using Bregman divergences make explicit a dual affine coordinate space which is in fact the log coordinates of Aitchison. It is in this space that we have affine constraints, which are therefore non linear in the "primal", ambient space (we will quote Shun-Ichi Amari's latest book).

④ Our comment on L91 about the Poisson distribution are misleading. Both CoDA and our proposed approach is not limited to only count data and applies to also real valued data on the simplex. However we do not work out the corresponding exponential family distribution when the base measure is Lebesgue (as opposed to counts).

✉ **To Reviewer #3:** Thank you for agreeing that we are *"convincingly demonstrating an improvement over the current standard"*. We will focus to make the fundamental message clearer.

① **transformed vs original data:** The baselines (CLR-PCA; CLR-AE) work on CLR transformed data and learning is separated in two stages: CLR transformation→PCA/AE. The proposed methods ((S)CODA-PCA; CODA-AE) work on original data based on a unified cost function that fits better in the simplex geometry. We show in figure 2 that the proposed methods generally outperforms the baselines.

② It is straightforward to include an **"intercept" term** (e.g. changing the constraint in eq. (21) to $\mathsf{Y} = \mathsf{C}(\mathsf{U}^\top \mathsf{B} + \boldsymbol{b}\boldsymbol{e}^\top)$), which is already used by all the compared methods in our experiments. For simplicity we didn't show it in our formulations (will make clear of this in our revision).

③ **visualization of real data**: see our reply ① to Reviewer #1.

④ L65 (singular covariance): being clr, our representation faces the same clr properties **but** our methods avoid the computational pitfalls of these properties (Other comments will be carefully addressed in the revised version).

✉ **To Reviewer #4:** Thank you for seeing that our paper *"provides interesting connections of some theoretical ideas"*, and we will make links clearer between: exponential family PCA, scaled Bregman theorem, and compositional data.

① **extension to ilr**: see our reply ② to Reviewer #1. One of our contributions relies on a principled formulation, simplification and optimisation of clr-PCA loss (Section 4). Ilr coordinates project further the clr coordinates via matrices on the Stiefel manifold: a corresponding approach for ilr-PCA should exploit its Riemannian structure; it is therefore out of the scope of our paper but can be built on top of our Section 4 – we take it as an exciting direction.

② **unclear formulations** will be carefully checked and polished to make sure of consistency and self-containedness.

③ **misleading title**: We respectfully disagree, and see PCA as learning a representation as well (linear projection).

④ In machine learning, Eq.(5) can be attributed to [9] (see for example Section 4 of Roy, Gordon, Thrun, JAIR 2005). However, as rightfully observed, there is an older connection to the information geometry of Bregman divergences, see our reply ③ to Reviewer #1. We will make this more explicit with proper citations to Amari.

⑤ L132 the first term on RHS is the regular Bregman divergence induced by the "perspective" of φ ($\breve{\varphi}(\cdot)$ as in eq.(8)).