We thank all reviewers for careful review and comments. We first address the questions that reviewers have in common:

1. Why ratio trace instead of trace ratio: Since WDA can be viewed as an extension to the classical Fisher linear discriminant analysis (FDA), we refer to the trace ratio and ratio trace formulations in the context of FDA. Statistically, these two formulations are both defined and are both served as criterion to maximize inter-class distance while minimizing intra-class distance (see [Fukunaga, 2013] page 446-447, eqn. (10.5) and (10.8)). When the reduced dimension $p = 1$, both the numerator and the denominator are scalars and these two formulations are equivalent. When $p > 1$, the ratio trace formulation iteratively finds p orthonormal vectors v_i to maximize $\frac{v_i^T A v_i}{v_i^T B v_i}$, while the trace ratio maximizes $\frac{\sum_{i=1}^p v_i^T A v_i}{\sum_{i=1}^p v_i^T B v_i}$. Both formulations are widely in the literature. For example, in our reference, [8, 41] used the ratio trace formulation of FDA while [14, 23] used the trace ratio formulation. Algebraically, these two formulations are not equivalent, and one is not upper/lower bounded by the other, so it is hard to quantify the difference. In an optimization sense, the ratio trace can be viewed as a relaxation to the trace ratio objective: the trace ratio problem is equivalent to the trace difference problem, which can be relaxed to the ratio trace problem.

2. What are the intuition behind assumptions A1, A2, A3, and why do they hold in practice: Since Theorem 1 and 2 essentially characterize how does the eigenspace vary when the matrix pair undergoes a small perturbation, the intuition behind assumptions A1 and A2 comes from matrix perturbation theory. When the matrices A and B are less sensitive to perturbation, the algorithm is easier to converge. This "sensitivity" is quantified in the Lipschitz constants ξ_a, ξ_b in A1. A2 and A3 are relaxations to a stronger assumption that assumes that there is an arctan gap between the p^{th} and $p + 1^{th}$ eigenvalue for $(A(P), B(P))$ constructed from any P, which guarantees that a discriminative subspace exists and is unique. For the toy example used to generate Figure 1 and Table 1 (described in Section 4.1), we know that the true discriminative subspace exists and has dimension 2, and we numerically checked that assumptions A1-A3 hold. For real datasets, whether the assumptions hold depends on the inherent structure of the data and specific choice of parameters, and the theory can provide some guidance in choosing the parameters λ, p and initialization P_0. For example, we can start with a small λ since it's easier to converge and adaptively increase λ. We can also initialize with the subspace found by FDA (append orthogonal columns if needed) because it is closer to the true subspace if λ is small.

Reviewer 1: 1. What if the C_w term is singular: There are two ways for circumventing this problem: first is to add a diagonal regularization term ϵI on C_w as we did in Line 213-219. By doing so we improve the numerical stability in computing the inverse. Another approach is to project away the null space as we discussed in Line 216-218.
2. Matrix 2-norm was used in Line 132. 3. Assumptions A1-A3: see the general comment 2 above.
4. Yes, s_k is defined to be $\| \sin \Theta(P_k, P_{k-1}) \|$. The definition was moved to the supplementary material.
5. Section 3.2 analyzed the convergence of WDA-eig under SCF framework. The usual WDA uses a gradient-based approach and has a different convergence criteria, so direct comparison between convergence curves may not be intuitive.
6. A small λ indeed implies a larger regularization to the Wasserstein distance. Here, however, we are perturbing the covariance matrices in FDA and not perturbing the true Wasserstein distance. When $\lambda = 0$ only the regularization term remains and WDA is FDA. When λ is small, WDA focuses more on global information and is more similar to FDA.

Reviewer 3: 1-3. Ratio trace iteratively finds directions that maximize the ratio of the inter-class and intra-class distance, while trace ratio maximizes total sum of inter-class distance while minimizing total sum of intra-class distance. In the context of classical FDA, these are two definitions that are commonly used in the literature. To our knowledge, there is no theoretical arguments showing one is strictly better than the other. See the general comment 1.
4. Global convergence: If the assumption A2 is made even stronger, we can show that the solution to the WDA problem exists and is unique, and the algorithm always converges to a global optimal (see the general comment 2). However, here we are not trying to prove that the algorithm converges to a global optimal, but rather, the algorithm converges to some point globally in a numerical sense, see Line 144-145.
5. Local convergence: If λ is small, A1 always holds. From the perspective of numerical computation, A3 always holds. Furthermore, if the data are well-separated in the true discriminative subspace, η in A3 should be large which yields faster convergence. We empirically observed linear convergence on toy example as well as on real datasets.
6. Writing styles: Yes, we agree that the mathematical concepts could be delayed to later sections.

Reviewer 4: Thanks for the suggestions. For the clustering algorithms, we emphasize that our work focuses on **finding a subspace** for high-dimensional data and can be combined with many other metrics to perform clustering. We have in fact performed other experiments where WDA is combined with other clustering methods such as spectral clustering and WDA is able to find a better subspace for clustering (KL divergence might need some tweak here since it is not symmetric and therefore is not strictly a metric). In the paper we intentionally selected WDA combined with K-Means as an example, because subspace clustering with K-means is studied extensively in the literature, but we can certainly include more in the future version. For data we used in our experiments, we just followed the standard experiments and datasets in the subspace clustering literature. Using graph-structured data that is more suitable for Wasserstein distance is an excellent insight and definitely worth pursuing in our future work.