Referee's comments: "Generalized Low-rank plus Sparse Tensor Estimation by Fast Riemannian Optimization"

This paper provides a generalized low-rank plus sparse tensor model to account for heterogeneous signals and outliers. A fast algorithm that combines Riemannian gradient descent and a pruning procedure is developed. Under some conditions on loss function and initialization, the final estimates enjoy statistical guarantees. Authors also present data analysis on international commodity data to support the proposed algorithm.

This paper is overall well written. The model is not new though; same statistical formulations have been proposed before [Gu et al., 2014, Zhou and Feng, 2017]. The main contribution of this paper lies in the algorithmic development and optimization accuracy analysis. In this regard, the paper may be appealing more to the optimization community but less to the statistical community. Nevertheless, I do like the connection presented in Section 5 and the applications therein. The visualization on international commodity data analysis is also impressive.

Major points:

1. Initialization. The major issue is the lack of initialization algorithm for $\hat{\boldsymbol{T}}$. Although tensor algorithm is nonconvex, the optimization is often believed to exhibit benign convexity. Under Assumptions 2-3 of the paper, the local optimization landscape near true parameter is simple (nearly convex or has unique global optimums). Therefore, the major challenge lies in the initialization. The proposed algorithm assumes good initialization of $\hat{\boldsymbol{T}}$ (which is impossible to check) and leaves the SNR unaddressed as "beyond the scope of the paper" in discussion.

   An initialization algorithm would be, in my view, not just useful but imperative. It would be great to provide a practical initializations with accuracy guarantees (in terms of signal-to-noise ratio; even if not optimal), and discussion to which level the statistical-algorithmic gap arises. There are some informal consensus that tensor optimization depends more on initialization but less on the subsequent refinement steps (e.g. first-order gradient descent, vs. second-order search, vs. alternating optimization). Could authors provide more guidance on this issue?

2. Comparison. The low-rank plus sparse tensor model was initially proposed in Gu et al. [2014], Zhou and Feng [2017]. I'd like to see the (1) the statistical benefits (e.g. signal-to-noise ratio, error rate), and (2) the algorithmic benefits, in light of earlier algorithms. Current simulations and real data are verification-type studies. It would be more convincing to show comparative analysis. For example, in the international commodity data analysis, authors may apply existing methods [Gu et al., 2014, Zhou and Feng, 2017] and assess the improvement quantitatively (in prediction error) or quantitatively (compared to clustering from standard tensor models).

3. Hyperparameter. The simulation in section 6 shows the sensitivity of estimation error to the tuning parameter γ. In addition, the main Theorem 4.1 suggests that the stepsize β should be in a proper range. More guidance on the selection of γ and β would be useful for practical applications. Algorithm 1 involves other hyperparameter, such as rank $\boldsymbol{r}$, sparsity parameter α, and μ_1. The suggested algorithm uses gradient pruning with $\gamma\alpha$ instead of α, where $\gamma > 1$ is a tuning parameter. I am curious why the gradient pruning with exact α is not allowed in Theorem 4.1. Some intuition about $\gamma > 1 + c_{4,m}^{-1} b_u^4 b_\ell^4$ would also be helpful.

4. The sup-norm error of $\hat{\boldsymbol{S}}_{l_{\max}}$ in Theorem 4.3 is great but the argument for recovering support of $\boldsymbol{S}^*$ seems not convincing. The requirement that non-zero entries in $\boldsymbol{S}^*$ having magnitude two times greater than the sup-norm error seems rather strict. Intuitively, the definition of sparse tensor has nothing to do with its entry magnitudes. Can authors provide some justification for this condition, or show the sup-norm is often negligibly small in practice?

Minor points:

1. Suggest to change "nonlinear" model to "generalized linear" model. GLM with a non-identity link is still a (generalized) linear model.
2. Figure 3 in the international commodity data analysis implies that the low-rank estimate is sensitive to the sparsity α. Authors draw the conclusion that economic similarity gradually dominates the geographical relations as α increases. However, it is hard to understand the relationship between the sparsity α and economic similarity. Perhaps more explanation is needed?
3. A typo in Line 43, page 15, $d^* = d_j d_j^{-1} = d_1 \cdots d_m$.

References

Quanquan Gu, Huan Gui, and Jiawei Han. Robust tensor decomposition with gross corruption. *Advances in Neural Information Processing Systems*, 27:1422–1430, 2014.

Pan Zhou and Jiashi Feng. Outlier-robust tensor pca. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2263–2271, 2017.