#ToR1. **Q1: ["Weaknesses...-METHOD"]. A1**: Thanks for your valued and constructive comments. The main motivation of our manuscript is to propose robust additive models (with theoretical analysis and applications) for realizing nonlinear estimation and structure discovery simultaneously, even data contaminated with complex noise and without priori knowledge of variable structure. There are **key differences** between our MAM and [11,37,34]: *1)Hypothesis space (nonlinear VS linear).* The proposed MAM employs the spline-based additive hypothesis space, which is more flexible than the linear assumption for regression function [11,37,34], and leads to better performance for characterizing the nonlinear relationship in data (e.g., [13,15,21,29,41]). *2)Sparsity within group ($\nu \in [0,1]^P$ VS $\nu \equiv \mathbf{I}_P := (1,...,1)^T$).* The flexible selection of ν in our Outer Problem can unveil main effect variables across all tasks, which is useful to remove ambiguity for model identifiability. Indeed, [11] ignores the sparsity within group and [37,34] don't consider the structure discovery. *3) Optimization (bilevel non-convex optimization VS bilevel convex optimization).* The non-convexity of outer and inner problems is the main challenge for MAM's optimization, which can not be tackled directly by the optimization strategy for bilevel convex optimization problem[11]. A computing algorithm for bilevel non-convex scheme is proposed by developing an effective optimization method for the non-convex non-smooth inner problem. *4) Application*. As we known, MAM is the first attempt to the interpretable CMEs analysis by data-driven structure discovery strategy. In short, our work is new for structure prediction, and can't be established by easily combining [11] and [37]. We will add these remarks after Table 1 (line 69).
Q2: ["Weaknesses..-ALGORITHM"]. A2: The non-convexity of outer and inner problem is the main difficulty in the optimization. To tackle this problem, an optimization strategy is proposed based on building blocks (ProxSAGA [30], HQ optimization[24] and DFBB[36]). Theorem 1 in Supp C guarantees that our computing algorithm can obtain the local optimal solution. Indeed, it's always a challenge to give the complexity analysis on bilevel optimization (we leave this direction for future work). Following your suggestion, Remark 2 (line 126 in Supp) is added to illustrate the extra complexity (computation time and space) of our optimization algorithm (compared with BiGL [11]).
Q3: ["Correctness...-EVALUATION"]. A3: Thanks for your valued suggestions. Some competitors (BiGL[11], Lasso[35], RMR[37], SpAM[29]) with detailed analysis are added in Table 3 and Supp D. The AAE for each task is BiGL($11.09h$, $59.75km/s$, $63.51km/s$, $95.97km/s$, $4.77nT$), Lasso($12.16h$, $62.56km/s$, $59.81km/s$, $95.34km/s$, $4.38nT$), RMR($12.02h$, $58.65km/s$, $49.03km/s$, $98.13km/s$, $4.20nT$) and SpAM($10.02h$, $50.48km/s$, $54.97km/s$, $72.33km/s$, $3.88nT$), respectively. We can observe that MAM (in Table 3) can achieve smaller AAE than other competitors due to its ability of robust nonlinear approximation and structure discovery.
Q4: ["Clarity-It..."]. A4: We have polished the manuscript following your suggestions and will proofread it carefully.
Q5: ["Relation to .."]. A5: In Table 1, MAM has been related to [11,37,26,41] in terms of hypothesis space, evaluation criterion, robustness and sparsity. Additionally, we have enriched Table 1 by adding the (non)convexity of objective function, and added remark to discuss the extra complexity of MAM compared with methods[11,37].
Q6: ["Additional feedback"]. A6: **[-When...]:** mGAM with an oracle variable structure is the baseline of MAM. For easy reading, we have added Remark 3 (pg.5) to state the relation of mGAM and MAM. **[-There..]:** τ_j is the weight for each group, by controlling which we can emphasize some groups that are known a priori. **[-Line 53]:** "In theory,..." -> "In theory, we provide the convergence analysis of the proposed optimization algorithm". **[-Line 143-144]:** The first term is used for robust estimation. The second term guarantees the strong convexity of transformed inner objective (See Eq.4 in Supp). The third term incorporates the variable structure ϑ. **[-Line 173]:** We have corrected ε->μ, where $\varepsilon = n\sigma^2\mu$ is the penalty parameter in the transformed inner problem (line 26 in Supp). Note that ε is different from the noise ϵ. **[-It...]:** A description (line 152) is added to explain that the second set of gradients is computed by backpropagating through the iterations of a smooth algorithm. **[-What...]:** The information appearing in parenthesis is standard deviation. We have revised the title of Table 2. **[-In...]:** We have revised Equation 2.
#ToR3. **Q1: ["Weaknesses.."]. A1**:Thanks for the constructive comments. In Supp, Section E (pg.12) is added to discuss the generalization bound based on algorithmic stability. Firstly, we proved that the mode-induced metric satisfies Quadratic Growth condition [Charles 2018] under mild assumptions. Based on the definition of multi-task uniform stability and Theorem 2 [Zhang 2015], we quantify the stability of MAM and establish its generalization bound. In addition, a brief comparison between overlapping group lasso[16] and our MAM is provided in Supp.
Q2: ["Clarity.."]. A2: 1) We have added more key parameters in Table 4 and a flowchart of optimization before starting with Section B.1. 2) The description ("Even...") is added in line 196. 3) Some competitors are added in CMEs experiment (See A3 of #ToR1). 4)We have corrected these typos and will proofread the whole manuscript.
#ToR5. **Q1: ["Weaknesses.."]. A1**: A detailed description of the selection of group number L and $|V|$ is added in line 171(pg.6). In fact, we have discussed the impact of group number L in Fig 8, which indicates a satisfactory result can also be obtained even if L is set to be larger than oracle group number. Moreover, our method can realize data-driven variable selection without the inactive variables $|V|$ being set initially (e.g., the red pixel in Figs 2 and 3).
Q2: ["..discuss.."]. A2: In hyper-parameter selection problem (e.g., variable structure ϑ), bilevel framework has better generalization than traditional hyper-parameter selection method (e.g, Cross-validation), since the hyper-parameter is tuned over a continuous space by minimizing the outer problem [K.P. Bennett et al., IJCNN 2006]. According to your suggestions, a further discussion on the advantage of bilevel framework is provided in Introduction (line 42).