Response to Reviewers' Comments

Manuscript No.	:	SIM-21-0956
Title	:	Spike-and-Slab Generalized Additive Models and Scalable Algorithms for High-Dimensional Data

First, we would like to thank you for handling this paper and thank the AE and both reviewers for their valuable comments. We have carefully addressed all the comments. Point-by-point responses are given below, where each comment is quoted in *italics*.

Response to Associate Editor

Comments to the authors:

Please follow carefully the comments, criticisms and suggestions of the three referees.

Response: We summarize the major changes of our manuscript as a response to address the comments from the referees:

- According to Reviewer #2's and Reviewer #4's comments, we modified the prior distribution of the non-linear effects inclusion indicator such that the strong hierarchy for bi-level selection is enforced.
- We moved the EM-IWLS fitting algorithm to the supporting information for three reasons. First of all, moving the EM-IWLS to the supporting information can help the audience focus on the model formulation and EM-Coordinate Descent algorithm, avoiding confusion and distraction. Secondly, many useful comments raised by the reviewers related to the EM-IWLS algorithm, e.g. constructing hypothesis testings and confidence

band, requires a more dedicated discussion. We are preparing another manuscript discussing those. Lastly, relocating the EM-IWLS section would not hurt the integrity of the current manuscript that emphasize the scalability and bi-level selection.

- According to Reviewer #1's and Reviewer #4's comments, we added linear lasso model and spikeSlabGAM model among the models to be compared.
- According to Reviewer #1's and Reviewer #4's comments, we added a new set of simulation scenarios where the functions of predictors are linear.
- According to Reviewer #4's comment, we summarized the variable selection performance for all models of comparison and bi-level selection performance for the models that is capable, i.e. the proposed model and spikeSlabGAM.
- According to Reviewer #4's comment, we recorded the computation performance of spikeSlabGAM.
- All existing figure and tables were updated to reflect the methodology change and simulation design change. Figures were added according to Reviewer #4's comment. [TODO: think what figures are needed]
- We fixed the additional typos and grammar errors according to Review #2. The revision of the manuscript is marked in red. Deletions are not highlighted.

Response to Reviewer #1

We sincerely thank you for the valuable comments. We have carefully addressed all the comments. A point-by-point response is given below, where each of your comments is quoted in *italics*. **For a summary of major changes, please see our response to the Associate Editor.**

Comments to the Authors

In this paper, the authors proposed a novel Bayesian hierarchical generalized additive model (BHAM) for high dimensional data analysis. Specifically, they incorporated smoothing penalties in the model via re-parameterization of smoothing function to avoid overly shrinking basis function coefficients. Incorporating the smoothing penalty allows the separation of the linear space of a smoothing function from the nonlinear space. Then they added a new two-part spike-and-slab spline prior on the smoothing functions for bi-level selection such that the linear and nonlinear spaces of smoothing functions can be selected separately. Two scalable optimization algorithms, EM-Coordinate Descent (EM-CD) algorithm and EM-Iterative weighted least square (EM-IWLS) algorithm, are developed and implemented.

The proposed framework, BHAM, provides computational convenience using independent mixture double exponential distribution during model fitting by using optimization algorithms instead of intensive sampling algorithms and addresses the incapability of bi-level selection.

Here are some comments.

1. In the real data study, it is very important to establish the correlation between response variables and explanatory variables with appropriate models. The authors should add the comparison of linear models to illustrate that whether using non-linear was better. In addition, except expertise knowledge, are there metrics to judge whether a model is suitable for building model with non-linear part?

Response: Thank you for the comment. We expanded our simulation study from two perspectives: 1) we added linear lasso models and spikeSlabGAM models, where linear lasso models serve as the reference to benchmark the performance of linear models; 2) we added new scenarios where the the

functional form of each predictors are linear, to demonstrate the robustness of the proposed, regarding prediction and variable selection performances, when the non-linear assumption no longer holds. Please see the changes and conclusion in *Section 3: SIMULATION STUDY*. To conclude, we observed [TODO: REPLACE HERE WITH CONCLUSION]. Hence, the proposed models perform better than linear models when functional are non-linear and as good as linear model when functions are linear. The expanded simulation study also demonstrates that the proposed model is flexible to model both linear and non-linear effects and broadly applicable in high-dimensional modeling without much constraints, which deems to be an additional strength.

2. In Weight Loss Maintenance Cohort, R2 was very small by the two GAM methods. I may think whether the data pretreatment and noise reduction process is appropriate.

Response: We thank the reviewer for the comment. We didn't perform the data processing step ourselves. We used the publicly available data that has been pre-processed. The data processing step is described in [TODO: add citation]. Meanwhile, we respectively argue that the small R2 is reasonable, as in many applications, genomics data can only accounts small amount of data variation. We think this example would be the case, as environmental factors (not accounted in these models) would be more substantial when explaining weight loss.

3. In the paper, several priors were mentioned. How to choose the prior in different situations? What is the advantage of the spike-and-slab double exponential prior?

Response: Thank you for the comment. The purpose of mentioning the other several priors in the previous manuscript is to describe the generalization of the proposed algorithms. In the updated manuscript, we move the subsection describing the other priors to the supporting information so that the readers can focus on understanding the proposed scalably model. We included an additional sentence in the discussion section. [TODO: list the sentence here].

Compared to other priors, the spike-and-slab double exponential prior provides three advantages. First of all, the spike-and-slab double exponential prior provides a locally adaptive shrinkage when estimating the coefficients. Hence, the smoothing functions can be estimated more accurately. Secondly, the

spike-and-slab exponential prior encourages a sparse solution, making variable selection straight forward. Thirdly, the spike-and-slab double exponential prior motivates a scalable algorithm, the EM-CD algorithm, for model fitting, and hence is more feasible for high-dimensional data analysis.

We revised our method section to highlight these advantages, see Pxxx.

4. How to detect the necessary before using the proposed model.

Response: We don't think any preliminary checkings or testings are necessary before implementing the proposed model, as long as some general rules of thumb for high-dimensional modeling are followed. Such rules of thumb includes [TODO: Insert citaiton for ultra high-dimension thing, as the number of predictors is a function of the sample size.]. As a supporting evidence, our added simulation demonstrates the proposed model performs well, in reference to the linear lasso model, even when signals are linear and sparse (see response to the Comment 1).

We acknowledge the reviewer's comment that careful deliberation are necessary when conducting analysis, which includes domain knowledge. However, the discussion of modeling strategy, e.g. starting with more flexible models or linear models, or using statistical testings to examine assumptions, is out of the scope of the current manuscript. The current manuscript focuses on introducing a novel model that allows flexibility when modeling complex signals.

Response to Reviewer #2

We sincerely thank you for the valuable comments. We have carefully addressed all the comments. A point-by-point response is given below, where each of your comments is quoted in *italics*. **For a summary of major changes, please see our response to the Associate Editor.**

Comments to the Authors

This paper introduces a new way to impose group sparse regularization in high-dimensional generalized additive models (GAMs). Specifically, this work proposes a re-parametrization, which allows separating the predictor space into linear/nonlinear ones. A spike-and-slab LASSO prior is then imposed on the coefficients for the re-parametrized predictors. The authors further uses EM algorithm to fit this model, where the M-step can be solved with either coordinate-descent or iteratively reweighted least squares. Simulation results and metabolomics data analysis are presented.

The authors claim that the proposed model has the advantage of

1. *does not induce excess shrinkage while still estimating a smooth function,*
2. *allow us to deal with linear/non-linear terms separately and understand if nonlinear terms are necessary*
3. *more scalable than previous algorithms which require MCMC for model fitting*

While targeting an important question, the novelty of the solution provided by the authors seem a little bit limited. In addition, the key and most significant step is the re-parametrization, yet there are a few points which I do not completely understand and might need further clarification from the authors:

(1) It seems that after re-parametrization, the authors do not further consider the smoothing penalty (Equation (1)) and only include the spike-and-slab LASSO prior. I wonder how does this affect the smoothness of the fitted function and how does it resolves the issue of excess shrinkage.

Response: We thank the reviewer for the comment. The purpose of the reparameterization step is to factor the smoothing penalty matrix in the design matrix, such that the smoothing penalty is no long a function of the

smoothing penalty matrix. In other words, the smoothing penalty changes from $\lambda\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$ to $\lambda\boldsymbol{\beta}^{*T}\boldsymbol{\beta}^*$, where the $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ are the coefficients of the smoothing function before and after the reparameterization respectively. The spike-and-slab LASSO prior estimates whether to include the smoothing function and its appropriate smoothing parameter λ with local adaptivity. This contrasts to previous methods that don't consider smoothing penalty at all or impose a uniform sparsity penalty upon uniform smoothing penalty, where both approaches tend to induce excessive shrinkage for the smoothing function.

In the updated manuscript, we clarify the utility of the reparmeterization step and emphasize how the smoothness of the additive function are properiately modeled. Please see the revision of the method section, highlighted in red.

(2) I was wondering why the authors choose spike-and-slab LASSO prior in particular. It is claimed in the paper that this prior yields a fast coordinate-descent algorithm, yet coordinate-descent seems also available to other priors as well. I think more discussion on the properties of SSL should be helpful here.

Response: Compared to other priors, the spike-and-slab double exponential prior provides three advantages. First of all, the spike-and-slab double exponential prior provides a locally adaptive shrinkage when estimating the coefficients. Hence, the smoothing functions can be estimated more accurately. Secondly, the spike-and-slab exponential prior encourages a sparse solution, making variable selection straight forward. Thirdly, the spike-and-slab double exponential prior motivates a scalable algorithm, the EM-CD algorithm, for model fitting, and hence is more feasible for high-dimensional data analysis.

We revised our method section to highlight these advantages, see Pxxx.

(3) Since the authors have already separated linear and nonlinear effects, and we might expect linear effects to enter the model first, I wonder why the authors does not incorporate this explicitly into their model by, say, changing the prior for gamma's to impose different sparsity.

Response: We thank the reviewer for the comment. We changed the prior distribution of the non-linear effects inclusion indicator to be $\gamma^*|\gamma, \theta \sim Bin(1, \gamma\theta)$ (See Equation ([TODO: add the equation number])), in comparison to

$\gamma^*|\theta \sim Bin(1, \theta)$ previously. This prior formulation is motivated by effect forcing discussed in (Chipman, 2004). Now, the inclusion of the nonlinear effects is strictly dependent on the inclusion of the linear effects, and hence, reflects the hierarchy of bi-level selection. The modification of the prior doesn't complicate the model fitting algorithm. The γ can be analytically integrated out of the density function (See Equation ([TODO: add the equation number]) and appendix) and requires minimum change of the proposed algorithms. In other words, the proposed model remain feasible for high-dimensional data analysis and retains the claimed computational advantages.

Besides, some writing of this paper should better be polished and some typos corrected. For example:

1. *In page 3, line 31, what does "in the prior density function" mean?*
2. *Same page, line 43 to line 45, the notation here is a little bit confusing.*
3. *Page 5, line 17-18, how shall we understand this statement and any explanations?*
4. *Page 5 line 37 and page 6 line 21-22, the implication seems to be 'SSL is not a continuous prior', which is not true.*
5. *Page 6, line 16, missing a period at the end.*
6. *Page 6, line 30, there are duplicate "the'".*
7. *Page 6, line 56, 'converge' should be 'converges'.*
8. *Page 9, line 24-25, there seems to be duplicated "variance-covariance matrix".*
9. *page 11, line 8, there are some problems with the table reference.*
10. *In section 4, what is definition for the "deviance"?*

Response:

Response to Reviewer #4

We sincerely thank you for the valuable comments. We have carefully addressed all the comments. A point-by-point response is given below, where each of your comments is quoted in *italics*. **For a summary of major changes, please see our response to the Associate Editor.**

Comments to the authors

1. Summary

The authors introduce the Bayesian hierarchical generalized additive model (BHAM) for high-dimensional data analysis with generalized additive models (GAMs). The BHAM framework allows for variable selection and estimation in GAMs. However, in contrast to previous works on Bayesian GAMs, bilevel selection is possible under BHAM, and BHAM allows practitioners to more easily distinguish linear effects from nonlinear effects. Two deterministic algorithms, EM-Coordinate Descent (EM-CD) and EM-Iterative Weighted Least Squares (EM-IWLS), are introduced to fit BHAM. The EM-IWLS algorithm allows for inference, which is its main advantage over EM-CD. The algorithms appear to be faster than GAMs based on group regularization since fitting a model with a lasso penalty is typically faster than fitting one with a group lasso penalty.

The paper is coherent and introduces a potentially promising methodology that addresses some of the shortcomings of existing Bayesian GAM approaches. However, in order to be suitable for publication in Statistics in Medicine, I have a few concerns that I think the authors should address. Namely, the authors need to do a better job illustrating the benefits of the BHAM approach over existing methods (such as the sparse Bayesian GAM method of (Bai et al. 2020; Bai 2021) or the spikeslabGAM of (Scheipl, Fahrmeir, and Kneib 2012)). I detail my main concerns below.

2. Main Comments and Questions

1. The incorporation of the smoothing penalty matrix S_j in BHAM allows for separation of the linear space of a smoothing function from the nonlinear

space. This is one of the touted benefits of BHAM - this separation allows for bilevel selection rather than "all-in-all-out" selection, so that the practitioner can distinguish linear effects from nonlinear ones. However, one concern is that the method might select the nonlinear effects but fail to select the linear component. The authors briefly allude to this in the Conclusion (p. 13 of the manuscript). The authors say that they "[include] the linear component in the model when non-linear component is selected."

Could the authors elaborate on how they do this? I think this point should also be stressed and made more explicit earlier in the manuscript, possibly in Section 2 where the EM-CD and EM-IWLS algorithms are introduced. One potential downside of using the double exponential priors on individual basis coefficients, as opposed to the group spike-and-slab prior in SBGAM, is the fact that only individual coefficients are regularized. So it is conceivable that BHAM would end up including the nonlinear component but not the linear one, which is not sensible. How can strong hierarchy be enforced to ensure that the linear component is always included if the nonlinear one is selected (but not necessarily the other way around)? This should be detailed in the EM-CD and EM-IWLS algorithms, and the authors should explain how to enforce strong hierarchy.

For example, in bilevel selection with the sparse group lasso (Simon et al. 2013), Simon et al. (2012) state in their algorithm that they first check whether or not the group is selected, and only if the group is nonzero do they then proceed with regularizing the individual coordinates in the nonzero groups. Is a similar check being done for BHAM? This needs to be explained further and made explicit. If necessary, the authors of the present manuscript could include an additional subsection in Section 2 that explains how to enforce strong hierarchy (i.e. linear effect is always selected if the corresponding nonlinear effect for the jth covariate is selected).

Response: We thank the reviewer for the comment. We changed the prior distribution of the non-linear effects inclusion indicator to be $\gamma^*|\gamma, \theta \sim Bin(1, \gamma\theta)$ (See Equation ([TODO: add the equation number])), in comparison to $\gamma^*|\theta \sim Bin(1, \theta)$ previously. This prior formulation is motivated by effect forcing discussed in (Chipman, 2004). Now, the inclusion of the nonlinear effects is strictly dependent on the inclusion of the linear effects, and hence, reflects the hierarchy of bi-level selection. The modification of the prior doesn't

complicate the model fitting algorithm. The γ can be analytically integrated out of the density function (See Equation ([TODO: add the equation number]) and appendix) and requires minimum change of the proposed algorithms. In other words, the proposed model remain feasible for high-dimensional data analysis and retains the claimed computational advantages.

2. Although the authors claim that BHAM is beneficial because it can distinguish linear effects from nonlinear effects (therefore, we can tell if nonlinear effects are necessary), this touted benefit is not well-illustrated in the Simulation Study (Section 3) or the metabolomics data (Section 4). In particular, the simulation setting in Section 3.1 contains only one linear function. The authors should expand their simulation study to include a greater number of linear functions (in addition to a few nonlinear ones) to show how well BHAM is truly distinguishing the linear effects from the nonlinear effects.

It would also be useful to report things like FDR, FNR, Matthews correlation coef- ficient, etc. for the function selection to see how well the proposed BHAM model is performing with respect to selection of the nonzero effects, especially compared to other methods like SBGAM or spikeslabGAM. In addition, if the authors could re- port a performance metric of how well BHAM is distinguishing nonlinear from linear functions, e.g. possibly reporting the proportion of simulations where the linear func- tion (resp. nonlinear) was correctly identified as linear (resp. nonlinear), that would provide empirical support for BHAM over existing methods.

Response: We thank the review for the comments. We added a new subsection discussing the function selection result from the simulation studies. The suggested metrics, including aaa, bbb, ccc are described for all models of comparisons. Lastly,the bi-level selection results are discussed too.

3. The authors claim that BHAM is computationally beneficial because it does not rely on MCMC. The authors illustrate the computational benefit of the double exponential prior over the group spike-and-slab prior of (Bai et al. 2020; Bai 2021), but in my view, they should also compare it to the method spikeslabGAM of Scheipl, Fahrmeir, and Kneib (2012). There is an R package spikeslabGAM to implement this method. This method uses a Gaussian spike-and- slab prior on the basis coefficients and is implemented using MCMC. In order to assess the benefit of using double exponential priors in BHAM

vs. Gaussian spike-and-slab priors, as well as illustrating the computational efficiency of the EM-CD and EM- IWLS algorithms compared to MCMC, the authors should include comparisons to spikeslabGAM.

Response: We thank the review for the comment. In the update manuscript, we add the suggested mdoel, spikeslabGAM, in the simulation study, where all the performance metrics and computation time are recorded and compared. We see that the proposed method still preserves the computation advantage over spikeSlabGAM, particularly in the high-dimensional cases.

4. The authors introduce the EM-IWLS method which can be used for inference of the regression coefficients. Does this method work for $p > n$? Or is it only limited to the case where $p \leq n$? Even if this algorithm is limited to only the $p \leq n$ case, the benefit of inference is not illustrated anywhere in the manuscript. Could the authors report how to use the standard errors from the error variance-covariance matrix returned from EM-IWLS to construct pointwise confidence intervals? There should be a section in the manuscript for Uncertainty Quantification which explains how to use the results from the EM-IWLS algorithm for the regression coefficients to construct pointwise confidence intervals for the univariate functions themselves.

In addition, the coverage probability of the 95% pointwise intervals constructed by BHAM in simulations should also be reported. Furthermore, the methods mgcv and spikeslabGAM also return interval estimates, so the authors would do well to compare the coverage probability of the confidence intervals constructed by BHAM vs. the confidence/credible intervals returned by mgcv and spikeslabGAM.

Response: We thank the reviewer for the comment. After careful deliberation, we decide to focusing an explaining the EM-CD method well. The inclusion of EM-IWLS would distract the reader from the main focus of manuscript, scalable model, good prediction and bi-level selection. We fully agree that the questions raised by the reviewer are very important and deserves careful discussion. Nevertheless, that would be out of the scope of the current work. We will dedicate another manuscript to detail many of the questions here, for example estimating effective degree of freedom, limiting distribution of the test statistics to the EM-IWLS method.

5. The authors only provide one figure (Figure 1) which illustrates the hierar-

chical model as a DAG. However, there are no other figures in the paper, e.g. plots of function estimates from BHAM. The authors may consider adding a plot of the function esti- mates vs. the ground truth in the simulation study (Section 3) for one of the nonlinear functions and one of the linear functions (these could be, for instance, the left panel and right panel of a figure). This will illustrate that BHAM can accurately recover both nonlinear functions and linear functions.

The authors may also give one or two plots from the real data analysis in Section 4 of some of the significant effects that they discovered. These plots should contain not just the point estimates but also the the pointwise confidence intervals from the EM-IWLS method. This will illustrate the benefit of EM-IWLS since it allows practitioners to assess the uncertainty, in addition to an estimate of the function itself.

Response: We thank the reviewer for the comment. We added more graphs displaying the effects of the significant effects.

3. Minor Comments

1. In Section 1, the authors say that nonparametric methods are an alternative to "black box" methods like random forest. This is a bit misleading, since the "black box" methods (random forest, deep learning, etc.) are als nonparametric methods. The authors might consider rephrasing it (e.g. they could say that semiparametric methods like GAMs allow for easier interpretation and selection of the significant effects than the fully nonparametric "black box" methods).

Response: We thank the reviewer for the comment. We agree that using "nonparametric methods" to differentiate from "black box" methods would be misleading. However, we respectively argue that our word choice, i.e. "nonparametric regression models," is not misleading. First of all The black box models normally are not treated as *regression* models. Moreover, the term "nonparametric regression models" has been consistently used in the generalized additive models literature, for example Hastie and Tibshirani (1987). Last but not least, to the authors' knowledge, there is a subtle distinction between nonparametric regression models and semiparametric regression models: semiparametric regression models assume some predictors have linear

effects, while nonparametric regression models assume unknown functions for all predictors. To give an example, see Du, Ma, and Liang (2010). The subtle definition distinction would not matter as much in this manuscript, but would be relevant to the future work, incorporating predictor structure in the additive model, e.g. gene network. We hope to establish the consistency in the word choice to avoid future confusion.

Nevertheless, we edited the sentence for clarity. [TODO: check if this is done]

Reference

Bai, Ray. 2021. "Spike-and-Slab Group Lasso for Consistent Estimation and Variable Selection in Non-Gaussian Generalized Additive Models." *arXiv:2007.07021v5.*

Bai, Ray, Gemma E. Moran, Joseph L. Antonelli, Yong Chen, and Mary R. Boland. 2020. "Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models." *Journal of the American Statistical Association*, June, 1–14. https://doi.org/10.1080/01621459.2020.1765784.

Du, Pang, Shuangge Ma, and Hua Liang. 2010. "Penalized Variable Selection Procedure for Cox Models with Semiparametric Relative Risk." *Annals of Statistics* 38 (4): 2092.

Hastie, Trevor, and Robert Tibshirani. 1987. "Generalized additive models: Some applications." *Journal of the American Statistical Association* 82 (398): 371–86. https://doi.org/10.1080/01621459.1987.10478440.

Scheipl, Fabian, Ludwig Fahrmeir, and Thomas Kneib. 2012. "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models." *Journal of the American Statistical Association* 107 (500): 1518–32. https://doi.org/10.1080/01621459.2012.737742.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2013. "A Sparse-Group Lasso." *Journal of Computational and Graphical Statistics* 22 (2): 231–45.