I would like to thank the reviewers for their thorough and constructive comments, and I am confident that all of the feedback can be easily incorporated in a revision. The comments broadly fell into three areas, which I address below: 1) an application of the non-overlapping mixtures trick (NOMT) to real data; 2) a comparison of the NOMT to a broader class of methods for obtaining approximate posteriors; and 3) changes to the exposition.

Real Data Application: While the primary result of this work is theoretical, I agree with Reviewers 2 and 4 that including an application to real data would strengthen the manuscript. To that end, I performed an analysis of a publicly available single cell RNA-seq dataset of peripheral mononuclear blood cells (PMBCs) from 10X Genomics ("`pbmc_1k_v3`") using the NOMT for sparse PCA. Sparse PCA better separates known cell types with sparser loadings than classical PCA (using only about 12% as many genes). Furthermore, by having an explicitly sparse model, the posterior probability that a feature is non-zero can be interpreted as a Posterior Inclusion Probability (PIP). Genes with high PIP in the first few PCs include known markers of immune cell lineage.

Comparison to other methods: To further highlight the utility of the NOMT, I compared the VI scheme for the LDpred model (coming from GWAS) based on the NOMT to boosting black-box VI (BBBVI) [3] as implemented in `pyro` [1] and to adaptive random walk MCMC as implemented in NIMBLE [2]. I simulated 20 datasets as described in the manuscript at each level of noise: $\sigma_e^2 \in \{1.0, 0.5, 0.1, 0.05\}$. For BBBVI, I used independent Gaussians for each β_j and independent Bernoullis for each Z_j as the variational family, and did 5 iterations of boosting – allowing mixtures of 5 distributions from the variational family. NIMBLE was run for 1000 passes over the data which took about 12 hours per dataset. BBBVI took about 3 hours per dataset. The NOMT took less than 5 seconds per dataset. The results are presented below (mean across simulations $\pm$ two standard errors): – the NOMT performs almost identically to MCMC by the metrics considered, while BBBVI performs substantially worse.

	MSE $\times 1000$				Correlation			
Method	$\sigma_e^2 = 1.0$	$\sigma_e^2 = 0.5$	$\sigma_e^2 = 0.1$	$\sigma_e^2 = 0.05$	$\sigma_e^2 = 1.0$	$\sigma_e^2 = 0.5$	$\sigma_e^2 = 0.1$	$\sigma_e^2 = 0.05$
MCMC	9.1 ± 0.4	8.2 ± 0.3	2.7 ± 0.2	1.5 ± 0.1	0.16 ± 0.02	0.33 ± 0.03	0.84 ± 0.01	0.89 ± 0.01
BBBVI	9.8 ± 0.5	9.0 ± 0.4	9.1 ± 0.5	8.7 ± 0.5	0.08 ± 0.00	0.11 ± 0.00	0.23 ± 0.01	0.29 ± 0.01
NOMT	9.1 ± 0.4	8.2 ± 0.3	2.7 ± 0.2	1.5 ± 0.1	0.16 ± 0.02	0.34 ± 0.03	0.84 ± 0.01	0.89 ± 0.01

Changes to the exposition:

Reviewer 3 noted that the spike-and-slab model does not satisfy the non-overlapping support assumption of Theorem 1. For any non-atomic distribution such as the Gaussian component in the spike-and-slab we may remove a set of measure zero without changing the density. We may therefore define an exponential family with the same density but supported on $\mathbb{R} \setminus \{0\}$, which does not overlap the point mass at zero.

Reviewer 2 pointed out that there is an interesting asymmetry in Theorem 1 with respect to component K. This arises because the mixture weights are constrained to sum to one, so the K^{th} mixture weight is completely determined by the other mixture weights. It would be possible to have a "symmetric" version of the theorem, but it would describe a curved exponential family and many downstream results pertain only to exponential families that are not curved.

Reviewer 2 suggested using reconstruction error as a metric for the sparse PCA application. I have computed reconstructions errors for 5 more replicate simulations as described in the manuscript. I computed the reconstruction error for the "signal" in the matrix, that is the matrix before gaussian noise added to each element. The mean reconstruction error for the NOMT was 4,261; for the best naive VI scheme it was 7,191; for standard PCA it was 29,407; and for the "oracle" version of PCA it was 3,985.

Reviewers 1 and 3 pointed out additional methods for fitting approximate posteriors: structured VI and a number of black-box VI methods. While many of these methods are similar in spirit, they apply to different types of models than the present work. I will include a discussion of these similarities and differences in the revision.

Reviewer 1 asked if the supports of the mixture distributions must defined *a priori*. They could certainly be treated as hyper-parameters and optimized to maximize the ELBo in a variational analog of Empirical Bayes.

Reviewer 1 suggested changing the title as it is currently too broad. I agree, and will change the title to "Flexible mean field variational inference using mixtures of non-overlapping exponential families."

[1] Eli Bingham et al. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

[2] Perry de Valpine et al. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

[3] Francesco Locatello et al. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, pages 3401–3411, 2018.