We thank the reviewers for their helpful feedback.

Motivation for variational inference and concern on dataset sizes

We did not clearly describe the computational challenge when using the (inverse) Wishart process. In particular, all reviewers state that our dataset sizes are manageable with exact inference; this is incorrect. Exact inference with the (inverse) Wishart process model is *intractable for even small dataset sizes*. As far as we know, an expression for the (exact) posterior distribution does not even appear anywhere in the literature. The only options for inference are to set up an MCMC procedure (as done in previous work) or to take the variational approach we now propose. We made this lack of clarity worse in the introduction (ll. 30–31) and in Sec. 3 (ll. 81–83) where we discuss "exact posterior inference on the underlying Gaussian processes." This was unclear wording: the aforementioned MCMC routines require an exact GP posterior inference computation (costing $O(N^3)$ on each of the $O(D^2)$ GPs) to be performed *during each step of their iterative algorithms*! So we clarify that our variational approach dramatically reduces this *per iteration* computational cost (while also having an easier implementation via black-box, gradient-based inference and not losing competitive performance). We will certainly clarify this in the paper.

All reviewers considered the datasets in Sec. 5 to be small, however, note that there are no previous examples in the literature of inference with the Wishart process scaling to this size (of both N and D). We reiterate: our experiments are the largest scale we're aware of in any work using the (inverse) Wishart process. We would like to point out the inherently overparameterized nature of this problem: We are estimating a sequence of N covariance matrices, each of $O(D^2)$ size, from a *single example*. So note that even $D = 20$ is NOT "small". This will always be a difficult task even for small N and D, and the problem will always be overparameterized no matter how big your dataset gets. With that said, there is no reason to believe that our variational approach would not scale to larger dataset sizes.

Finally, we also note that one additional benefit of the variational approach is a natural way to implement *online inference*. We can also mention this in the paper, though this is certainly not a point we're trying to emphasize.

Motivation for predicting the covariance matrices

R1 & R2 requested motivation for why one would want to predict the covariances Σ_n. The prominent application requiring such predictions is in the construction of optimal financial trading portfolios that minimize risk (), where returns are maximized based on a separate model for Y_n (not addressed in this paper) and a model for the *covariance of the residuals* is used to penalize risky (i.e., volatile) assets, which is the part we are tackling in our paper. It is for this reason that we present in the context of financial applications. Moreover, these predictions of "volatility" or "risk" are desired throughout finance and beyond. R4 specifically requested motivation beyond finance: such models have been used to analyze the spread of disease incidence () and XXX (). We will make these clarifications and additions.

R2 similarly questions why (large) covariance matrices would be useful without sparsity. Note that the *full* covariance matrix is required to construct an optimal financial portfolio. Sparse approximations are often imposed for computational reasons, but you will *always* construct a suboptimal portfolio with a sparse approximation. This would certainly never be desired if you can handle the computational burden.

Originality/novelty

All reviewers expressed concerns that our work appears too close to previous techniques. We re-emphasize that our work demonstrates the first alternative to MCMC for inference in the Wishart process model (improving on its computational efficiency and ease of implementation), and the first experiments scaling inference to the size of the datasets in Sec. 5.

Other points

R1 points out the apparent contradiction of "low-variance" MC gradient estimates (through reparameterization) vs. the study in Sec. 4. We should have instead said: "while such gradient estimates typically have low variance, the particular form of the Wishart process likelihood introduces computational instability that renders these estimates useless."

R1 & R4 requested some discussion of wall clock runtime of the methods. We will add these to the paper.

We are pleased that R4 recognized the significance of our identification and resolution of the computational issues involved with applying black-box variational inference to the Wishart process case. These techniques are often applied indiscriminately in practice, and it is our hope this study (the likes of which are rarely explored in the literature) will provide a useful warning to practitioners.

We will follow R1's suggestion to move the details of Sec. 3 to the supplement (R2 & R4 also point out the density/clutter and notation, which we will address); this will also free up enough space to add all the proposed clarifications.