AOAS2002-013, *Markov-switching state space models for uncovering musical interpretation*
Author response - Revision 1

Daniel J McDonald, Michael McBride, Yupeng Gu, Christopher Raphael

October 21, 2020

Major comments

We would like to thank the Editor and Referees for their very careful (and detailed!) comments. They have precipitated a number of changes which (we think) have improved the paper significantly.

Below, we address each of the major concerns and specifically mention some of the minor ones which were not simply grammatical or notational. In the latter cases, we have made the corrections discussed, hopefully without missing any. The original comments are reproduced (and occasionally shortened) in black while our responses are given in blue.

Note: We switched to the most recent `imsart` template file, so the page numbers have changed somewhat dramatically. Hopefully this doesn't cause too many issues. We have done our best point to the correct location.

Why not multiplicative?

> Referee 1: "The most serious concern has to do with viewing changes in tempo as additive rather than multiplicative. ... This would suggest that the authors' state-space model might be more appropriately applied to the log of tempo."
>
> Editor summary: "Referee 2 makes a similar comment about your approach to 'tempo'."

This is an excellent suggestion, but it has some unfortunate implications which we feel we should address at length, and somewhat informally. We had discussed trying such a model in the past, but our initial instinct was that using a nonlinear, multiplicative model, would be challenging to estimate in a switching scenario. The discrete particle filter is already a greedy approximation and the prior distributions (see below) are important to find a good (by which we mean musically interpretable) local minimum. Nonlinearity would seem to complicate the situation.

Given the suggestions of Reviewer 1 and the Editor, we went ahead and estimated it using the Extended Kalman Filter on one performance (Richter 1976) and were, frankly,

stunned at how well it performed (see Supplement Section 7) with very little tuning of the prior distributions. So we estimated the multiplicative model on all the performances, and it continued to fit well (in the sense that the interpretation plots look exceedingly reasonable). At this point, we considered rewriting the manuscript to focus on the multiplicative model rather than the additive one from our first draft.

However, it turns out that our first instinct is actually more accurate. While the model itself results in reasonable interpretations, the parameters of the model are no longer very meaningful. The data is only able to identify the observation variance σ_ϵ^2 and μ_{stress}. For the case of μ_{stress}, this is probably because, based on the estimated s_t sequence, some recordings use it and their parameters move from the prior while others don't, so the identification happens only for some recordings (the same holds for the linear model). Performing the same PCA analysis suggested by the Editor (see below), shows that the first PC concentrates entirely on σ_ϵ^2 while the second concentrates entirely on μ_{stress}. These two explain 99% of the variance across performances. Thus, while this model can explain *individual* performance interpretations well (when coupled with informative priors to drive parameter identification), it is not really able to distinguish *between* performances. At least not in a way that gives as much musical meaning as the linear model. For this reason, we have relegated the analysis to the supplement at this point. The supplement details the multiplicative model we use, shows some of the resulting interpretation fits, and describes the PCA estimates in more detail. We also included a shorter discussion in Section 2.5 of the manuscript.

One final note: we are not averse to investigating this model more thoroughly and potentially using it rather than the linear one in this manuscript. With some tweaks, it may well improve, but given the other tasks, we leave it here for now.

Prior sensitivity

> Referee 1: "I also had some concerns about the lack of checking the sensitivity to the prior distribution.. . . The decision to model variances as Gamma-distributed (as opposed to inverse-Gamma, or uniform as suggested by Gelman and others in similar settings) was a curious choice."
>
> Editor summary: "These "strong" statements almost scream for some sensitivity analysis - or else, *very* solid evidence that changing them would be non-sensical from a musical perspective."

We have done our best to examine these suggestions. While we would not describe our checks as "exhaustive" by any means, we hope that they are at least reasonably convincing. We are happy to expand them if desired.

Currently, we have examined (in one performance) using IG and Uniform distributions for the variance parameters as suggested. We also looked at uniform distributions on the transition probabilities and a prior which makes the observation variance, σ_ϵ^2 smaller. We added a paragraph to Section 3.6 of the main manuscript as well as presenting more details in the Supplement Section 8.

The main takeaway is that (apart from the "smaller observation variance" setting) these different specifications don't have a dramatic effect: the fit to the data remains similar both quantitatively (as measured by RMSE and negative loglikelihood) and qualitatively

(as determined by examining the inferred performance graphics). We would also emphasize that the optimization technique is not overly influenced by starting values, because we use an overdispersed initial collection.

We also feel that our characterization of these priors as "strong" is perhaps the wrong word, so we have made minor changes to this wording where it appears. The prior modes and sometimes the local curvature are important for some parameters to avoid non-identifiability. And occasionally, as described in the manuscript, to enforce more musically meaningful switching behaviors. On the other hand, the prior tail shape is not particularly important here because we're estimating posterior modes rather than performing a full Bayesian analysis with accompanying credible intervals. So in reference to Gelman's standard suggestions, we're not so worried about properly quantifying uncertainty far from the mode.

Clustering

> Reviewer 1: "The material on clustering could be improved. For clustering according to Mahalanobis distances, I was not clear why the authors did not choose to standardize by the inverse of the posterior covariance matrix. Also, given the likely skewed distribution of the estimated variances across performers, it might have been more sensible to include the variance parameters in computing Mahalanobis distances on the log scale. It was also unclear how the specific number of clusters was determined. Methods exist (e.g., Tibshirani's gap statistic) to make this choice."
>
> Editor: "'The resulting clusters suggest methods for informing music instruction, discovering listening preferences, and analyzing performances.' I hope you will say how the clusters do this. Perhaps a music student might identify more with musicians in Cluster A and hence study them more, is that what you have in mind? "
>
> Editor report (minor): "If you were able to narrow the number down from 12 to, say, 3 linear combinations of the 12 (say, via PC?), could you plot them, and show visually how much of an outlier Cortot is?"

We have revised this section (3.3 in the manuscript) dramatically. The procedure we used was rather too finicky and likely obscured our goal. We have attempted to rectify this in a few ways. Here is the short overview, with more detailed explanations/justifications to follow.

1. We switched to the term "group" from "cluster" and added language that indicates that these are heuristic rather than statistically justifiable.
2. We focused on the PC suggestion in the body of the manuscript, and removed the density plots, which were not really statistically justified.
3. We used Tibshiriani's Gap statistic to choose the number of groups (with a caveat).
4. We continue to use Mahalanobis distance in the prior. We hope our reasoning below is convincing to Reviewer 1.

Our main goal with this section is to, as the editor suggests, be able to make statements about the similarity of performances such that a student or listener may be able to benefit (either through study or say buying new recordings). The clusters themselves are only really interesting if they are revealing about the similarities of the performances within them and if those performances have similar parameter estimates in the model.

We initially used 4 clusters, because the cut off seemed reasonable, and digestible, though this followed initially throwing some performances away. Instead, using the Gap Statistic, suggests 1 cluster. To be clear on what this statistic does: it is one of a number of methods that provides a statistic to maximize, but also measures uncertainty in that statistic. So, one would choose the number of clusters that maximizes the statistic, subject to that number of clusters being significant relative to smaller numbers of clusters. The choice of 1 cluster is robust across different clustering methods, screening steps, and other similar metrics for choosing the number of clusters. Thus, in light of Reviewer 1's comment, we don't think it's fair to claim to be doing "clustering" in the body of the manuscript. Now, ignoring the uncertainty, the Gap Statistic is actually maximized at 7 clusters (after throwing out Cortot only). So we have rewritten the section to describe those groups as being "similar". We discuss a few that are fairly obvious in the distance matrix (Figure 9), but we do not highlight them all. We use the grouping mainly to orient our discussion and to display all the performances in the Supplement.

We continue to use Mahlanobis distance in the prior precision (inverse covariance). In terms of our goal to make comparisons, we need the scales of the parameters to be similar for any type of clustering to work, so this seems sufficient. The prior inverse covariance still handles the structural dependence in the probabilities correctly. Another reason to avoid Whitening by the posterior precision, (we have also added this explanation to the text at the beginning of section 3.3), is that parameters which are nearly constant across performances (say if poorly identified) will have near zero posterior variance and therefore be dramatically overemphasized in the distance calculation. Using the prior avoids this pathology. As for transforming to handle the skewness, we feel that our approach should perform better. Suppose that our data was 1-dimensional from a 2 component normal mixture model with common variance 1, component means 0 and 5, and the probability of the large component is 0.2. This is "right-skewed" when the components are unknown. If we take a log transformation, this will make it harder to recover the two groups.

After doing things this way, we plot the data in the manuscript for the first few PCs and use the point type by the "clusters". This at least aids the eye for groupings, but doesn't emphasize too strongly the validity of 7 clusters (which is almost certainly invalid). In the manuscript, we emphasize that this grouping is only heuristic and not justified by the Gap statistic. This change also helps to cut down on color figures and allows us to remove the density plots, some of which had too few observations to be meaningful anyway.

Is any of this musically interesting?

> Reviewer 3: "it is not at all clear that the results produced are very interesting from a musical perspective. I consulted a musician colleague in considering this paper and he/she thought that there was nothing interesting produced by the output of this procedure."

We are disappointed to hear that the reviewer and colleague feel this way. We would argue that all 4 of the authors are musicians, some with degrees from well-known music programs. At the same time, we thought it may be useful if there were a more musical way of understanding our contributions. To this end, we are including two MIDI files in the supplementary material. The first is Richter's performance with the recording's tempos. The second replaces them with those produced by our model. We would argue that it is challenging to hear the difference (without being told) even for a trained musician. So from this perspective, the model is able to recover musically meaningful information.

Minor comments

We address here any minor comments that required more than simple grammatical or wording corrections.

Reviewer 1

- The abstract is clear, but does not specify that the focus of the paper is on modeling tempo changes within music. The abstract should be revised accord- ingly. We have revised the abstract to address this and the Editor's comment below. Also in light of the major comments above.

- Page 7: It would be helpful to explain up front that in the switching state- space model, the continuous hidden states are functions of both the current and previous switch states. Some rationale should be provided early on for this choice of a second-order Markov model, i.e., that the states depend on both velocity (tempo) and acceleration. We added the following sentence: "Allowing d, T, and Q, to depend on s_{i-1} in addition to s_i (the diagonal arrow in Figure 2) will allow us to incorporate acceleration as well as velocity into our model for tempo decisions."

- I think it would be helpful to discuss whether the switching model developed using four discrete states is more generally applicable beyond Chopin. The first three states seem reasonably generalizable, and perhaps single note stress is widely applicable in piano performance, but this is not entirely clear. We would imagine that "single note stress" is also widely applicable. It's a near necessity in performing Bach (though, see a recent NYTimes Review for a description of it's overuse in a famous recent recording). We added a paragraph to Section 2.2 (p8) and discuss it further in Section 3.6 (p20) to try to address this concern and the AE's similar statement below.

- Page 13: The Fernhead and Clifford citation should appear as "Fernhead and Clifford (2003)". Several other similar instances of incorrectly displayed citations should be fixed. We have hopefully caught all of these.

- What are the differences between the model in the current manuscript and the generative model in Gu and Raphael (2012)? This question arises in the context of section 2.5. We expanded on this section in 2.5 with additional detail. This also seemed like

a good place to point to the musical examples we have synthesized (see Reviewer 3's comment above)

Reviewer 2

- The only point with which I strongly disagree is their characterization of the tempo variability within a state. We agree that this is an important point. We have added a footnote describing the reviewers characterization (hopefully accurately) and pointing out that we are losing something by smoothing. We also point toward the discussion in section 2.5 where we describe the MIDI files we have synthesized. Although, there's even more lost by converting to MIDI!

Editor

- Abstract, l.-5: "we learn a switching state space model ..." Seems to me that you have started with assuming a switching state model, and what you really do in the MS is "model a musician's tempo via a Markov-switching state-space model, and estimate its parameters using prior information, and then cluster musicians via hierarchical clustering of model parameters." Is that correct? (See p4, para 3, l.3-4.) If so, then you can say so directly, more succinctly, and more clearly. (In my view, the verb "learn" has been grossly overused - and often improperly used - in the stat literature.) We have revised the abstract to remove some of the flowers and focus on the contributions. See also Reviewer 1's comment above.

- "The resulting clusters suggest methods for informing music instruction, discovering listening preferences, and analyzing performances." I hope you will say how the clusters do this. Perhaps a music student might identify more with musicians in Cluster A and hence study them more, is that what you have in mind? We added some more details along these lines to the first paragraph in Section 1.2 (p4) and also a bit in the abstract (see Reviewer 1's comment above).

- p3, Fig 1: How desperately do you need color, here and in your subsequent figures? Color is very expensive to print, and AOAS, like many journals, ask authors to use it only when essential for accurate reader comprehension. You may need it in Fig 9 but perhaps gray scales and shadings will work for Fig 10-11. We appreciate your attention to this cost-saving measure. We have moved entirely to BW figures in the manuscript. Hopefully still legible. We eliminated the density plots (discussed above) and we think the distance matrix (Figure 9) is okay as is. The performance plots seem the most challenging to discern, but we tried to make the point-types as obvious as possible.

- p6, l.-2: Do you "lose" anything with this smoother that "produces only the unconditional expectations"? Is it more realistic or useful to focus on only the unconditional, rather than conditional expectations? (It makes sense to me but perhaps a Bayesian might see it otherwise?) We added a few sentences to the paragraph in the middle of, now, p5 to clarify. We don't actually need the variance computation at all, so this one is simply easier to implement. One could use any of the others if one desired.

- p9, para 2, l.-1: "small number of discrete states": "small" arises from the assumption on p8, l.-4: "performers generally maintain (or try to maintain) a steady tempo"? Or might one argue that this assumption might be reasonable for Chopin or 18-19thC composers? (Stravinsky doesn't seem very "steady" to me, but I'm not a musician!) We added some discussion of this to the 2nd paragraph on p20, though this is rather far from p9. See also Reviewer 1's comment above. We also expanded a bit on p8 to try to further distinguish "tempo" from "rhythm/time signature", and to include some of the following discussion. The Stravinsky example you mention is likely best described in terms of a much more varied use of rhythm and time signature than would exist in the 19th century. But at the same time, playing Stravinsky generally requires an even more stringent adherence to "steady tempo". Chopin's own music has rather severe departures from steady tempo in, for example, his Nocturnes, which regularly contain periods of improvisatory diversion. Somewhat related is the use of "impossible rhythm" (see Hook, 2011), which encompasses the Chopin improvisatory use but also includes other purposeful notations which could potentially break the model. These "bad" notations also mean notes whose notated durations don't correspond to the amount of time they would be played. One could potentially adjust via ℓ_t, but it would be challenging.

- p18, Table 4: Do we need SEs on these parameters? Which parameters are most meaningful in characterizing a performer? If you were able to narrow the number down from 12 to, say, 3 linear combinations of the 12 (say, via PC?), could you plot them, and show visually how much of an outlier Cortot is? We appreciate the suggestion of using PCA, and we have incorporated this idea in to the manuscript directly, in place of the previous density plots (see major comment above). We think this is certainly an improvement. As for confidence intervals, we went ahead and generated them using the observed Fisher information from the optimization routine. However, it's not entirely clear what these mean. For one, they ignore any uncertainty in the state sequence (see Figure 9 in the SM for some thoughts on the scale of this uncertainty). They also depend on identifiability, the priors, and the approximation to the posterior. Producing the MAP depends on the approximation accuracy at the MAP, but producing the Hessian needs that as well as the accuracy of about 700 additional function evaluations. And because we need the inverse, inaccuracies can explode. Our current feeling is that, short of performing a fully-Bayesian analysis, we would hesitate to attach much certainty to the metrics of uncertainty we provide. For this reason, we have put it in the supplement.

- p24, Sec 3.5: Any ideas how to modify the model to incorporate these limitations? (The ideas can be left as "Future work.") We added a few sentences to hint at fixes (e.g. using the multiplicative model suggested by reviewer 1, or modulating the priors as described in Section 3.6).