MARKOV-SWITCHING STATE SPACE MODELS FOR UNCOVERING MUSICAL INTERPRETATION

By Daniel J. McDonald[1], Michael McBride[2,*], Yupeng Gu[3,†], and Christopher Raphael[3,‡]

[1]*Department of Statistics, University of British Columbia, daniel@stat.ubc.ca*

[2]*Department of Statistics, Indiana University, *michmcbr@iu.edu*

[3]*School of Informatics, Computing, and Engineering, Indiana University, †yupeng.gu@gmail.com; ‡craphael@indiana.edu*

For concertgoers, musical interpretation is the most important factor in determining whether or not we enjoy a classical performance. Every performance includes mistakes—intonation issues, a lost note, an unpleasant sound—but these are all easily forgotten (or unnoticed) when a performer engages her audience, imbuing a piece with novel emotional content beyond the vague instructions inscribed on the printed page. In this research, we use data from the CHARM Mazurka Project—forty-six professional recordings of Chopin's Mazurka Op. 68 No. 3 by consummate artists—with the goal of elucidating musically interpretable performance decisions. We focus specifically on each performer's use musical tempo by examining the inter-onset intervals of the note attacks in the recording. To explain these tempo decisions, we develop a switching state space model and estimate it by maximum likelihood combined with prior information gained from music theory and performance practice. We use the estimated parameters to quantitatively describe individual performance decisions and compare recordings. These comparisons suggest methods for informing music instruction, discovering listening preferences, and analyzing performances.

1. Introduction.

Statistical analysis of the musical content of recordings has become more and more important to academics and industry. Online music services like Pandora, Last.fm, Spotify, and others rely on recommendation systems to suggest potentially interesting or related songs to listeners. In 2011, the KDD Cup challenged academic computer scientists and statisticians to identify user tastes in music with the Yahoo! Million Song Dataset (see Dror et al. (2012) for details of the competition). Pandora, through its proprietary Music Genome Project, uses trained musicologists to assign new songs a vector of trait expressions (consisting of up to 500 "genes" depending on the genre) which can then be used to measure similarity with other songs. However, most of this work has focused on the analysis of more popular and more profitable genres of music—pop, rock, country—as opposed to classical music.

Western classical music is a subcategory whose boundaries are occasionally difficult to define. But the distinction is of great importance when it comes to the analysis which we undertake here. Leonard Bernstein, the great composer, conductor, and pianist, gave the following characterization in one of his famous "Young People's Concerts" broadcast by the Columbia Broadcasting Corporation in the 1950s and 1960s (Bernstein, 2005).

> [. . . W]hen a composer writes a piece of what's usually called classical music, he puts down the exact notes that he wants, the exact instruments or voices that he wants to play or sing those notes [. . .] and he also writes down as many directions as he can think of. [. . .] Of course, no performance can be perfectly exact, because there aren't enough words in the world to tell the performers everything they have to know about what the composer wanted.

Keywords and phrases: classification and clustering, Kalman filter, hidden Markov model.

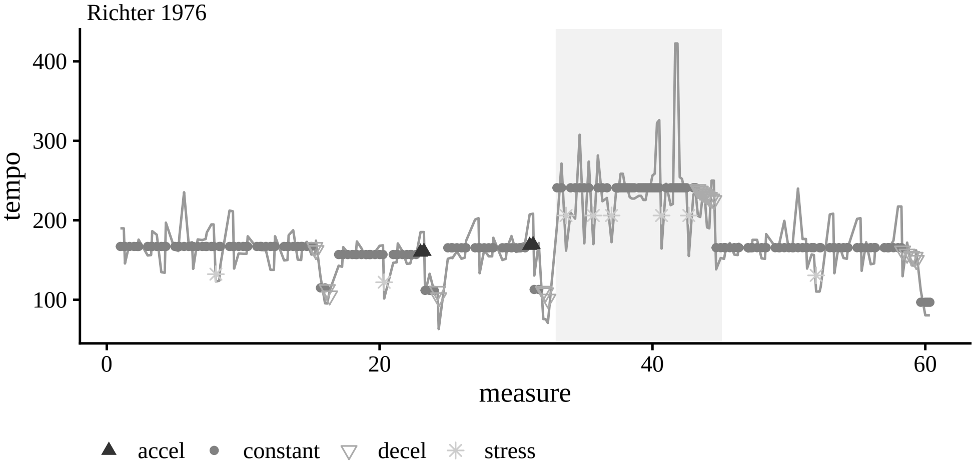

FIG 1. *Note-by-note tempos for a recording of Chopin's Mazurka Op. 68 No. 3 by Sviatoslav Richter. The solid line are the observed tempos, while the dots represent inferred tempo states from our model.*

What separates classical music from other types of music is that the music itself is written down very precisely but performed millions of times in a variety of interpretations.[1] There is no "gold standard" recording to which everyone can refer but rather a document created for reference. Therefore, the musical genome technique mentioned above will only relate "pieces" but not "performances". We need new methods to decide whether we prefer Leonard Bernstein's recording of Beethoven's Fifth Symphony or Herbert von Karajan's and to articulate why.

In this paper, we develop a statistical model for some of the decisions that a musician must make for classical music interpretations. We focus on how the musician modulates *tempo*, or speed, over the course of a recording. Figure 1 shows the tempo[2] in beats-per-minute (b.p.m.) of a recording made by Sviatoslav Richter of Chopin's Mazurka Op. 68 No. 3. The solid line shows the actual tempo at which he plays each note, while the points correspond to our model's inferences for his actual intentions. Some of this intent is prescribed by Chopin in his music, but the extent to which Richter observes Chopin's indications makes his recording different from those of other pianists. It is these differences that we hope to capture and understand.

1.1. *Related work.* The vast majority of work at the intersection of statistics and classical music analysis has focused on a handful of tasks, most notably structure analysis, music generation, and score alignment.

Analysis of musical structure and its relationships with interpretation forms the basis of music theory, and hence constitutes the core of standard conservatory curricula, along with history and performance. Automatically discovering musical structures from performances without expert input has become more relevant recently. Ren et al. (2010) use Dirichlet process models to identify similar sections of individual classical music performances. Roberts

[1] Of course jazz, rock, pop and other genres are often written down for future performances, but without the same precision. Compare for example a lead sheet John Coltrane's "Giant Steps" to his original 1960 recording.

[2] Technically, by "tempo", we mean the ratio of musical time to clock time as 0.25 beats $/$ 0.1 seconds $=$ 150 beats per minute. A musician would likely think of "tempo" more broadly as something like a "typical speed regime" akin to the "constant tempo" state we use in our decision model below. We will not generally distinguish between these two interpretations and use the more succinct "tempo".

et al. (2018) use variational autoencoders to discover long-term structure with an explicit goal toward improved automatic music composition.

Computer music generation and composition has a long history (Ariza, 2005; Boulanger-Lewandowski, Bengio and Vincent, 2012; Collins, 2016; Flossmann, Grachten and Widmer, 2013; Sturm et al., 2019). It is actively investigated, especially using deep learning (Hadjeres, Pachet and Nielsen, 2017), and has become commercially relevant for advertising and video games through companies like Aiva (`aiva.ai`), and Melodrive (`melodrive.com`). Google has developed the Magenta project to enable open-source music composition (Roberts, Hawthorne and Simon, 2018).

The score alignment problem matches live or recorded performances to the musical score, a necessary processing step for any automated analysis. On-line alignment processes audio waveforms in real-time and is sometimes called score following (Arzt and Widmer, 2015; Cont, 2010; Cont et al., 2007; Dannenberg and Raphael, 2006). Audio matched to the score can then be used as an input for automated musical accompaniment (Dannenberg, 1985; Raphael, 2010; Vercoe, 1984). Given recorded accompaniment, these systems modulate playback in response to a live soloist who both makes interpretive timing decisions and mistakes. Off-line alignment (Earis, 2007) can be used to analyze the recordings, as we do here, or for generating descriptive features of the performance (Thickstun, Harchaoui and Kakade, 2017), possibly for later analysis in recommender systems (McFee and Lanckriet, 2011; van den Oord, Dieleman and Schrauwen, 2013). For an overview of related goals in music information retrieval, see Schedl et al. (2014).

More closely connected to the work here is the literature on expressive synthesis (Arcos and Mantaras, 2001; Bresin, Friberg and Sundberg, 2002; Flossman S. Grachten and Widmer, 2012; Grindlay and Helmbold, 2006; Maezawa, 2019; Widmer, Flossmann and Grachten, 2009). In this domain, one seeks to create a musically satisfying performance of a score, often given a training set of score-performance pairs. This problem is highly challenging since musical interpretation relies on latent aspects of the music, such as structure, stress, grouping, closure, surprise, affect, and others not explicitly appearing in the score. Music theorists often describe the score as the *surface* of the music, thus recognizing that there is much that lies beneath this surface. Though we do not treat expressive synthesis, our work shares the need for explicitly representing expressive performances.

Most recent work in expressive synthesis is based in machine learning, both in terms of methodology and the agnostic spirit of the modeling. Here one parametrizes the performance in terms of variables that will be estimated from the score, such as rate of change of tempo or dynamics. For instance, Flossman S. Grachten and Widmer (2012) represent the joint configuration of local performance parameters and score parameters as a conditional Gaussian distribution, learned from training, and then used to estimate the expression on a new score. We also use conditional Gaussians as a modeling element, though we try to model the *process* nature of the music, rather than viewing each note independently. The work on expressive synthesis of Maezawa (2019) uses an autoregression to model the expressive parameters of a note, estimating the autoregressive parameters using deep learning on score attributes. The use of autoregression accomodates the smoothness normally found in expressive performance, while still not being overly prescriptive about the nature of the musical evolution.

Common to these approaches are the generic assumptions relating the musical score and the expressive performance parameters, hoping to push the hardest work — understanding what is important — onto the learning algorithm. In this spirit one seeks to discover what relevant sub-surface attributes of the score can be correlated with performance decisions. Perhaps the most extreme example of this musically agnostic approach is Grindlay and Helmbold (2006), who use a minimally-primed hidden Markov model to generate expression without giving any explicit meaning to the states.

We emphasize an important modeling difference between these approaches and what we propose. We explicitly model the performance in terms of a switching Kalman filter, thus making rather strong *a priori* assumptions derived from our musical sensibilities. Stowell and Chew (2012) similarly use an explicit parametrization of tempo evolution by automatically partitioning the music into segments represented by local quadratics. Our appoach considers a broader family of parametrizations but shares the basic approach of *building in* musical knowledge to the model.

1.2. *Our contributions.* In this paper, we develop a switching Kalman filter model for the tempo decisions a performer makes in recorded classical music. We present an algorithm for performing likelihood inference, estimate our model using a large collection of recordings of the same composition, and demonstrate how the model is able to recover performer intentions, and how they relate to standard musical analysis. We use the low-dimensional representations to compare and contrast the recordings, and discuss how this analysis facilitates more informed musical comparisons of the recordings. Such an analysis may help listeners to choose other performers whose tendencies are similar (or dramatically different!) from those they already enjoy, suggest new recordings to purchase, or motivate future concert attendance behaviors. Our analysis can also aid automatic performance generation by reusing a musician's estimated parameters on a reproducing instrument or potentially inform music education. In combination with other software, a music teacher could create visuals for a student's performance, such as those presented in this paper, and directly discuss areas for improvement.

In Section 2 we discuss our dataset, a collection of professional recordings of Chopin's Mazurka Op. 68 No. 3. We also present our model for tempo decisions, discuss its statistical estimation, and detail its utility for understanding interpretations as a musician would. Section 3 presents a music theory interpretation of the Mazurka. We discuss how different performers approach this piece through the lens of our model. We also examine groups of performances based on our model and interpret the musical meaning of these groupings. Finally, we contrast our approach with some alternative non-parametric smoothers, discusses their deficiencies relative our switching model, and examine some issues with our proposal.

2. Materials and methods.

In this paper, we examine note-by-note tempos for 46 recordings of Chopin's Mazurka Op. 68, No. 3. The data is part of a large collection of the complete Chopin Mazurkas and other recordings assembled and analyzed by the Center for the History and Analysis of Recorded Music (CHARM) in the United Kingdom (CHARM, 2009). The recordings were processed using the note-onset detection algorithm developed by Earis (2007) and are available for download (Earis, 2009). We use the data for "all rhythmic events", which includes the time of each note attack as well as it's relative loudness.

2.1. *Switching state-space models.* State-space models define the probability distribution of a continuous time series Y by reference to some imagined, continuous hidden state, X. In particular, the observation at time i is assumed to be independent of past and future observations conditional on the state at time i. Coupling with temporal dependence for X—most frequently obeying the Markov property—induces a temporal model for the observations. The most general form of a state-space model is then characterized by the measurement equation (the conditional probability of observations given the states), the transition equation (specifying the nature of Markovian dynamics), and an initial distribution for the state:

$$y_i = f_\theta(x_i, \epsilon_i), \quad x_{i+1} = g_\theta(x_i, \eta_i), \quad x_1 \sim F, \tag{1}$$

where ϵ_i are η_i are marginally and mutually independent and F is an arbitrary but specified distribution. Both y_i and x_i can be vector-valued, though in our application, y_i will be univariate. The vector $\{y_i\}_{i=1}^n$ is observed, and the goal is to make inferences for the unobserved states $\{x_i\}_{i=1}^n$ as well as any unknown parameters θ characterizing f_θ, g_θ, and the distributions of ϵ_i and η_i.

If f_θ and g_θ are linear, and ϵ_i, η_i, and F assumed to have Gaussian distributions, Equation (1) becomes

$$\begin{aligned} x_{i+1} &= d + Tx_i + \eta_i, \quad \eta_i \sim N(0,\, Q), \quad x_1 \sim N(x_0,\, P_0), \\ y_i &= c + Zx_i + \epsilon_i, \quad \epsilon_i \sim N(0,\, G), \end{aligned} \tag{2}$$

where the vectors c, d and matrices T, Z, Q, and G are allowed to depend on θ, can potentially vary with i, or can depend on previous values of x and y. In this case, the Kalman filter (see for example, Harvey, 1990; Kalman, 1960), provides closed form solutions for the conditional distributions of the states and gives the likelihood of θ given data. For completeness, we have included the Kalman filter inference algorithm in the Supplementary Material.

Although the Kalman filter returns the likelihood for θ, and is therefore all we need for parameter estimation, inference for the mean and variance of X is conditional only on the preceding observations $\{y_j\}_{j=1}^i$: $\mathrm{X}_i = E\left[x_i \mid y_1 \ldots, y_i\right]$ and $P_i = Var\left[x_i \mid y_1, \ldots, y_i\right]$. To incorporate all future observations into these estimates, and produce the inferred performance tempos shown in, for example, Figure 1, the Kalman smoother is required. Many different smoother algorithms have been tailored for different applications. The smoother we use, due to Rauch, Striebel and Tung (1965), is often referred to as the classical fixed-interval smoother (Anderson and Moore, 1979). It produces only the unconditional expectations of the hidden state $\widehat{x}_i = E\left[x_i \mid y_1, \ldots, y_n\right]$, which is all that is necessary for our analysis. This algorithm is again given in the Supplementary Material. We note that the smoother estimate of $Var\left[x_i \mid y_1, \ldots, y_n\right]$ is not necessary for any of the analysis discussed in this paper.

Linear Gaussian state-space models can be made quite flexible by expanding the state vector or allowing the parameter matrices to vary with time. Furthermore, this general form encompasses many standard time series models: ARIMA models, ARCH and GARCH models, stochastic volatility models, exponential smoothers, and more (see Durbin and Koopman, 2001, for many other examples). Nonlinear, non-Gaussian versions have been extensively studied (Durbin and Koopman, 1997; Fuh, 2006; Kitagawa, 1987, 1996) and algorithms for filtering, smoothing, and parameter estimation have been derived (for example, Andrieu, Doucet and Holenstein, 2010; Koyama et al., 2010). However, these models are less useful for change-point detection or other discontinuous behavior when the times of discontinuity are unknown.

To remedy this deficiency, one can use a switching state-space model as shown in Figure 2. Here, we assume $\{s_i\}_{i=1}^n$ is a hidden, discrete process with Markovian dynamics. Then, the value of the hidden state at time i, $s_i = k$ say, can determine the evolution of the continuous model at time i. The graphical model in Figure 2 gives the conditional independence properties we will use in our model for musical interpretation, representing just one of many possiblities. Switching state-space models have a long history with applications ranging from economics (Hamilton, 2011; Kim, 1994; Kim and Nelson, 1998) to speech processing (Fox et al., 2011) to animal movement (Block et al., 2011; Patterson et al., 2008). Ghahramani and Hinton (2000) provide an excellent overview of the history, typography, and algorithmic developments. In Equation (2), the parameter matrices were not time varying. We allow the switch states s_i, s_{i-1}, along with the parameter vector θ, to determine the specific dynamics at time i:

$$\begin{aligned} x_1 &\sim N(x_0,\, P_0), \\ x_{i+1} &= d(s_i, s_{i-1}) + T(s_i, s_{i-1})x_i + \eta_i, \quad && \eta_i \sim N(0, Q(s_i, s_{i-1})), \\ y_i &= c(s_i) + Z(s_i)x_i + \epsilon_i, && \epsilon_i \sim N(0, G(s_i)). \end{aligned}$$

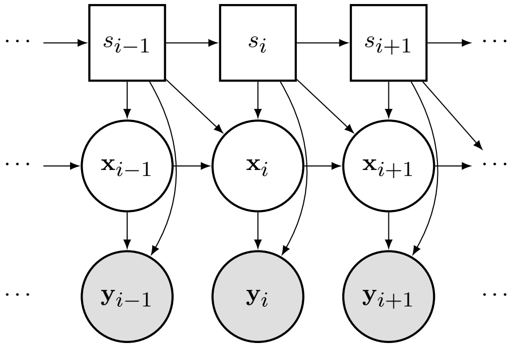

FIG 2. *Switching state space model. Filled objects are observed, rectangles are discrete, and circles are continuous.*

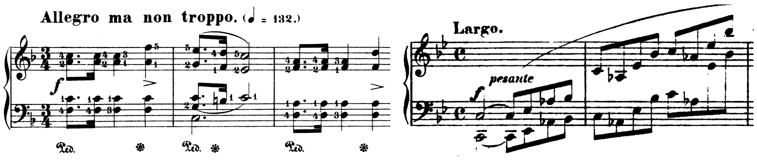

FIG 3. *The beginning of two Chopin piano compositions: the Mazurka we analyze is on the left while the Ballade No. 1, Op. 23 is on the right.*

In other words, the hidden Markov (switch) state determines the collection of d, c, T, Z, G and Q that govern the evolution of the system. 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.

2.2. *A model for tempo decisions.* In musical scores, *tempi* (the Italian plural of tempo) may be marked at various points throughout a piece of music. The beginning can be either explicit, with a metronome marking to indicate the number of beats per minute (b.p.m.), and/or with some words (e.g., *Adagio*, *Presto*, *Langsam*, Sprightly) which indicate an approximate speed. Figure 3 shows the beginning of two Chopin piano compositions: the Mazurka we analyze and the Ballade No. 1, Op. 23. The initial tempo of the Mazurka is given with a metronome marking as well as the Italian phrase *Allegro ma non troppo* ("cheerful, but not too much"). The beginning of the Ballade is marked *Largo*, which translates literally as "broad" or "wide", and modified by the stylistic indication *pesante* ("heavy"). Obviously, the metronome markings are much more exact, though even these are often viewed as suggestions rather than commandments. The metronome markings in most of Beethoven's compositions, for example, are notoriously fast, and some scholars believe that his metronome (one of the first ever made) was inaccurate (Forsén et al., 2013). Often, compositions will have numerous such markings later in the piece of music, but these are only some of the ways that tempo is indicated. Composers will also indicate periods of speeding-up (*accelerando*) or slowing-down (*ritardando*).

Absent instructions from the composer, performers generally maintain (or try to maintain) a steady tempo, and this assumption plays a major role in our model of tempo decisions. Of course, a normal human never plays precisely like a metronome, although she may try quite hard to do so. The observed ratio of musical time to clock time can therefore be thought of as stochastic, the sum of an intentional, constant tempo, plus noise representing inaccuracy or, perhaps more charitably, unintentional variation which the listener fails to perceive as

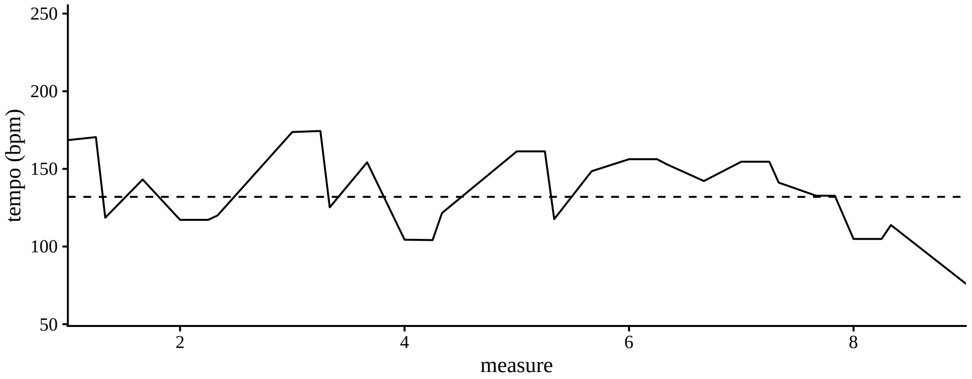

FIG 4. *The solid line shows the observed note-by-note tempo for the beginning of the Mazurka as performed by Arthur Rubinstein in 1961. The dashed line indicates 132 b.p.m.*

"wrong".[3] For instance, the example in Figure 4 shows the beginning of the piece as performed by Arthur Rubinstein in a 1961 recording. The solid line shows the actual, performed tempo, while the dashed horizontal line is placed at the indicated tempo of 132 b.p.m. The figure has three important lessons: (1) observed speed varies around intended tempo; (2) 132 b.p.m. is not necessarily the tempo a performer will choose despite the indication; and (3) performers have other tempo intentions which are not marked, like the pronounced slow-down in measures 7–8.

Estimating intended *tempi* would be reasonably simple, perhaps, if the locations of the tempo changes were known. In such a case, the average of tempi between changes may be a good estimate as could the slope of known speed-ups or slow-downs. However, performers take liberties with these decisions, exactly the liberties we would like to discover. This suggests employing a switching model with a small number of discrete states.

We propose a Markov model for S on four states for four different performance behaviors with transition probability diagram given by Figure 5. The 4 switch states correspond to 4 different behaviors for the performer: (1) constant tempo, (2) speeding up, (3) slowing down, and (4) single note stress. As shown in the diagram, we allow only certain transitions for musical reasons and for estimability. The marked transition probabilities are sufficient to infer the remainder. One can imagine that a performer will remain mainly in the state 1 with departures to states 2 and 3 either due to markings by the composer, or, absent these, for interpretive reasons referred to collectively as *rubato*, which translates literally as "stolen time". The fourth state, stress, corresponds to *tenuto*, a common feature of musical performance. These stresses may be marked with a line over the note in question, but are more often a feature of performer taste, corresponding to a longer-than-written duration for a particular note. Such emphases occur for a variety of musical purposes—emphasis of the beat in running notes, the top of a phrase, a "landing point" where a phrase ends, etc.—but are always within the frame of constant tempo. Thus we allow stress to occur only after and before notes in state 1. Furthermore, we cannot allow state 2 or state 3 to return immediately to state 1, or else "stress" could happen through these pathways. We impose related constraints for a transition from state 2 to state 3 and vice versa. Essentially, transitions into these states must remain

[3] Some may argue with this explanation. As one anonymous reviewer pointed out, describing deviations from constant tempo as "unintentional noise" fails to account for the possibility that the performer is consciously or unconsciously controlling these small deviations, and their success as artists can be partially attributed to their preternatural abilities to exert such control. In the end, our model is smoothing such deviations away for the sake of providing a low-dimensional explanation of performance behavior. See Section 2.5 for a discussion of how much might be lost by this smoothing.

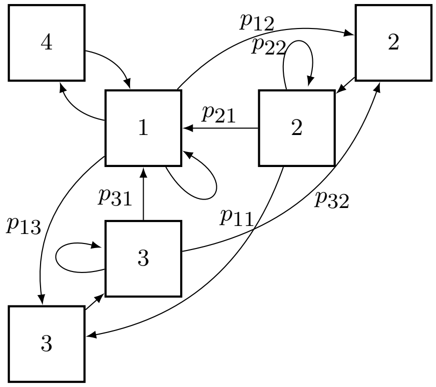

FIG 5. *Transition diagram. The four states are: constant tempo (1), deceleration (2), acceleration (3), and emphasis (4).*

there before leaving. Thus, the entire transition diagram is fully determined. This process can be viewed equivalently as a second-order Markov chain.

Generally, we feel that this model should be broadly applicable across composers and time periods as well as instrumentation. That is, it should work equally well for compositions by Mozart, Bach, Beethoven, or Stravinsky. While some composers, especially those that are more modern, have a more varied use of time signature and rhythm than the music we examine here, these generally require even more stringent adherence to "steady tempo". Some of Chopin's other compositions present more severe departures from steady tempo (the Nocturnes, for example), but we intend our model to be able to capture these features through states 3 or 4. We return to this issue in Section 3.6.

Our data gives y_i as the observed tempo (in b.p.m.) of the note (or chord) of the i^{th} note onset in Chopin's Mazurka Op. 68 No. 3. The hidden continuous variable (X_i) is taken to be a two component vector with the first component being the prevailing tempo and the second the amount of acceleration. The amount, or existence, of acceleration is determined by the current and previous switch states. We use ℓ_i to denote the musical duration of a particular note as given by the written score. Because, in this piece, each measure contains three quarter-notes, (♩), a quarter-note has $\ell_i = 1/3$, an eighth note (♪) has $\ell_i = 1/6$, etc. In more complicated music with changing time signatures or instances where the notation doesn't necessarily correspond with the time signature, more care would be required. The observed tempo is already normalized to account for variable note durations, but the intentional tempo and its variance should be proportional to ℓ_i. When the performer is in state 1 (or transits in and out of state 4), we take the prevailing tempo as constant with no acceleration: $X_{i+1} = X_i$. Corresponding to these configurations, the parameter matrices are given in Table 1 (transition equation) and Table 2 (measurement equation). So for any performance, we wish to estimate the following parameters: σ^2_{tempo}, σ^2_{acc}, σ^2_{stress}, σ^2_{ϵ}, the probabilities of the transition matrix (there are 7), and means μ_{tempo}, μ_{acc}, and μ_{stress}. Lastly, we have the initial state distribution

$$x_1 \sim N\left(\begin{pmatrix}\mu_1\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & 0\\0 & 0\end{pmatrix}\right) \quad \text{where } s_1 = 1.$$

To clarify this model, we explicate two different behaviors: discrete sequence $1 \to 4 \to 1$ (emphasis within constant tempo) and discrete sequence $1 \to 1 \to 2$ (constant tempo to slowing down). In the first case, the state space system has the following configurations

TABLE 1
Parameter matrices of the transition equation for the switching state space model.

Switch states		parameter matrices		
s_i	s_{i-1}	d	T	Q
1	1	0	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
2	1	$\begin{pmatrix} \ell_i \mu_{\text{acc}} \\ \mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} \ell_i^2 & \ell_i \\ \ell_i & 1 \end{pmatrix}$
3	1	$\begin{pmatrix} -\ell_i \mu_{\text{acc}} \\ -\mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} \ell_i^2 & \ell_i \\ \ell_i & 1 \end{pmatrix}$
4	1	$\begin{pmatrix} 0 \\ \mu_{\text{stress}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{\text{stress}} \end{pmatrix}$
2	2	0	$\begin{pmatrix} 1 & \ell_i \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
3	2	$\begin{pmatrix} -\ell_i \mu_{\text{acc}} \\ -\mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} \ell_i^2 & \ell_i \\ \ell_i & 1 \end{pmatrix}$
1	2	$\begin{pmatrix} \mu_{\text{tempo}} \\ 0 \end{pmatrix}$	0	$\begin{pmatrix} \sigma^2_{\text{tempo}} & 0 \\ 0 & 0 \end{pmatrix}$
3	3	0	$\begin{pmatrix} 1 & \ell_i \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
2	3	$\begin{pmatrix} \ell_i \mu_{\text{acc}} \\ \mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} \ell_i^2 & \ell_i \\ \ell_i & 1 \end{pmatrix}$
1	3	$\begin{pmatrix} \mu_{\text{tempo}} \\ 0 \end{pmatrix}$	0	$\begin{pmatrix} \sigma^2_{\text{tempo}} & 0 \\ 0 & 0 \end{pmatrix}$
1	4	0	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

TABLE 2
Parameter matrices of the measurement equation for the switching state space model.

Switch states	parameter matrices		
s_i	c	Z	G
4	0	$\begin{pmatrix} 1 & 1 \end{pmatrix}$	σ^2_ϵ
else	0	$\begin{pmatrix} 1 & 0 \end{pmatrix}$	σ^2_ϵ

$1 \to 4$

$$x_2 = \begin{pmatrix} 0 \\ \mu_{\text{stress}} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x_1 + \mathrm{N}\left(0,\ \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{\text{stress}} \end{pmatrix}\right)$$

$$y_2 = (1 \quad 1) x_2 + \mathrm{N}(0,\ \sigma^2_\epsilon)$$

$4 \to 1$

$$x_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x_2$$

$$y_3 = (1 \quad 0) x_3 + \mathrm{N}(0,\ \sigma^2_\epsilon),$$

Algorithm 1 Discrete particle filter

```
Input: $Y$, $\theta$, $\pi_1$ probability vector over initial states (paths), $B$ beam width
for $i = 1$ to $n$ do
    Set $b_i = |\{\pi_i > 0\}|$, the number of current paths
    Use the Kalman filter to calculate the 1-step likelihood $\mathcal{L}_i$ for each path and every potential state $s_{i+1}$ resulting in $b_i|S|$ particles
    Set $\pi_{i+1} \leftarrow \pi_i \mathcal{L}_i p_i$: multiply the path probability by the likelihood and the probability of transitioning. Normalize $\pi$.
    Set $b_{i+1} = |\{\pi_{i+1} > 0\}|$ . If $b_{i+1} > B$, resample the weights to get $B$ non-zero weights and renormalize
end for
Return $B$ paths $\{S_b\}_{b=1}^{B}$ along with their weights $\pi_n$.
```

while in the second

$$
\begin{array}{ll}
1 \to 1 & 1 \to 2 \\
x_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x_1 & x_3 = \begin{pmatrix} \ell_i \mu_{\text{acc}} \\ \mu_{\text{acc}} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} x_1 + \mathrm{N}\left(0,\ \sigma^2_{\text{acc}} \begin{pmatrix} \ell_i^2 & \ell_i \\ \ell_i & 1 \end{pmatrix}\right) \\
y_2 = (1 \quad 0) x_2 + \mathrm{N}(0,\ \sigma^2_\epsilon) & y_3 = (1 \quad 0) x_3 + \mathrm{N}(0,\ \sigma^2_\epsilon).
\end{array}
$$

Recall that in any case y_i is a scalar and $x_i \in \mathbb{R}^2$.

2.3. *Estimation and computational issues.* To understand the performance decisions of individual musicians, we wish to simultaneously estimate θ, S, and X. Because the switch states S and the continuous states X are both hidden, this becomes an NP-hard problem. In particular, there are approximately 4^n possible paths through the switch variables, so evaluating the likelihood to maximize over θ via the Kalman filter at each path is intractable. Ghahramani and Hinton (2000) give a variational approximation to estimate θ without also estimating S, but, as our goal is to learn both, we use the particle filtering approximation described by Fearnhead and Clifford (2003). Whiteley, Andrieu and Doucet (2010) refer to this algorithm as the Discrete Particle Filter, and it can be seen as an instance of the "Beam Search" optimization technique (Bisiani, 1992). The details are given in Algorithm 1 but the intuition is as follows: (1) for the first few time points, evaluate one step of the Kalman filter for each possible subsequent discrete state and store all these values; (2) calculate weights for each path by updating previous weights with the likelihood multiplied by the transition probability; (3) continue through time until the number of stored values exceeds some threshold storage limit; (4) from that point forward, subselect the "best" paths using a sampling scheme. These paths can be selected greedily, retaining only the highest values to that point, though we use the resampling procedure of Fearnhead and Clifford (2003) which is designed to approximate the full discrete distribution over paths with a subset of support points by minimizing the mean squared error.

Algorithm 1 returns B paths through the discrete states along with their weights for a particular parameter value θ. One can view this as a (approximate) distribution over paths conditional on θ. Instead, we will simply take the path with the highest weight for inference via penalized maximum likelihood. Thus, the likelihood of a particular parameter vector θ is evaluated by computing the best path with Algorithm 1 and then using that best path with the Kalman filter.

2.4. *Penalized maximum likelihood.* Even without the latent discrete process, parameter estimation in state-space models is a difficult problem, often plagued by spurious local minima and non-identifiability. The addition of discrete states only exacerbates this issue.

TABLE 3
Informative prior distributions for the music model

Parameter		Distribution	Prior mean
σ_ϵ^2	$\sim$	$\textrm{Gamma}(40,\ 10)$	400 b.p.m.^{2}
$\mu_{\textrm{tempo}}$	$\sim$	$\textrm{Gamma}(\overline{Y}^2/100,\ 100/\overline{Y})$	$\overline{Y}$ b.p.m.
$-\mu_{\textrm{acc}}$	$\sim$	$\textrm{Gamma}(15,\ 2/3)$	10 b.p.m.
$-\mu_{\textrm{stress}}$	$\sim$	$\textrm{Gamma}(20,\ 2)$	40 b.p.m.
$\sigma_{\textrm{tempo}}^2$	$\sim$	$\textrm{Gamma}(40,\ 10)$	400 b.p.m.^{2}
$\sigma_{\textrm{acc}}^2$	$=$	1	1 b.p.m.^{2}
$\sigma_{\textrm{stress}}^2$	$=$	1	1 b.p.m.^{2}
$p_{1,\cdot}$	$\sim$	$\textrm{Dirichlet}(85,\ 5,\ 2,\ 8)$	
$p_{2,\cdot}$	$\sim$	$\textrm{Dirichlet}(4,\ 10,\ 1,\ 0)$	
$p_{3,\cdot}$	$\sim$	$\textrm{Dirichlet}(5,\ 3,\ 7,\ 0)$	

However, for the present application, we have reasonable informative prior information for many of the parameters. The three mean parameters $\mu_{\textrm{tempo}}$, $\mu_{\textrm{acc}}$ and $\mu_{\textrm{stress}}$ have sign restrictions in addition to reasonable constraints their magnitude: average tempo should be around the indicated 132 b.p.m., the average amount of acceleration should probably be less than the size of a stress. We also have can make musically informed choices about the probabilities of transitioning between states: self-transitions should be reasonably likely, long periods of speeding up are less likely than long periods of slowing down which are less likely than long periods in the constant tempo state. Because of this information, we use informative priors as penalties on all the parameters we estimate. This has the effect of introducing extra curvature to the optimization problem as well as conforming with musical intuition. The specific choices are shown in Table 3. We fix $\sigma_{\textrm{acc}}^2$ and $\sigma_{\textrm{stress}}^2$ to be 1 after numerical experiments suggested that they were poorly identified. Essentially, large values of these variances make the stress and deceleration states difficult to separate, so other values of similar magnitude make little difference. We defer justification and discussion of these choices to Section 3.6.

2.5. *Is this model reasonable.* It is reasonable to ask whether a simple model such as this can accurately represent performance practice without removing musically important information. In a statistical sense, this question is similar to the problem of tuning parameter selection in nonparametric estimation. Specifically, we do not want this model to "oversmooth" the performance, eliminating information necessary for listener appreciation. One way to examine such a question is to generate a performance using the smoothed tempos resulting from the fitted model and compare it aurally with the original recording. Gu and Raphael (2012) evaluate this question empirically: they surveyed nine graduate piano majors at a major conservatory on twelve different piano excerpts, both performed and synthesized. The pianists were not meaningfully able to distinguish between the two in the majority of experiments. We expect that a similar study with our model would yield better results. In the Supplementary Material, we allow the reader to decide for themselves: we have included a MIDI recording derived from Sviatoslav Richter's 1976 recording as well as one synthesized using our model. The only difference is the tempos of the individual beats.

The generative model in Gu and Raphael (2012) is also a switching model on four states like ours. It is quite a bit simpler, however, in terms of the transition matrix depicted in Figure 5. It is only first-order, so states 2 and 3 can be entered and left immediately. There's also no ability to go from state 3 to 2 or 2 to 1. This precludes the common feature of slowing down at the end of a phrase before returning to a new tempo. Furthermore, and importantly, their parameters are not estimated from data but chosen by eye. Comparing our model to theirs, we found that ours is more robust in that it is less likely to make spurious excursions to states 2 and 3 and more accurately uses state 4. It also has significantly lower RMSE on the data despite having only two additional parameters.

While an additive state space model is relatively easy to understand, some music theorists (Mead, 2007, for example) have argued that musicians make multiplicative tempo adjustments. That is, the ratio between the tempo of the current note and that of the previous note is important rather than their difference. Such a conception is fundamental to musical notation (quarter notes, eighth notes, etc) and frequently used to specify tempo changes within a piece of music. Unfortunately, switching linear models are challenging to estimate and nonlinear models are only more so. We examine a multiplicative model in the Supplementary Material. This model produces very reasonable interpretations of individual performances, but unfortunately, it is less useful for comparing performances.

3. Analysis of Chopin's Mazurka Op. 68 No. 3. We use the model and procedures developed above to estimate the parameters and performance choices for all 46 recordings of Chopin's Mazurka. Here we describe the inferences our model allows on some representative performances, describe performance groupings based on the estimated parameters, contrast our model with some alternative approaches to smoothing, and discuss some difficulties we encountered. All simulations and empirical calculations were performed with `R` (R Core Team, 2019) and C++ via `Rcpp` (Eddelbuettel, 2013). Figures and tables are generated using the `tidyverse` family of packages (Wickham, 2016, 2017). Most computations were implemented in parallel on a large memory computer cluster via the `batchtools` package (Lang, Bischl and Surmann, 2017).

3.1. *Musical analysis.* Throughout his life, Frédéric Chopin composed dozens of Mazurkas, of which 58 have been published. Inspired by a traditional Polish dance, these pieces gave Chopin an idiomatic style upon which to elaborate a wide variety of different compositional techniques, a practice German and Italian composers had employed frequently over the previous 3 centuries (Burkholder, Grout and Palisca, 2014). Repetition of themes, figures, or even small motives plays a central role in both the traditional dance and Chopin's compositions as do particular rhythmic gestures (Kallberg, 1996), especially the dotted-eighth sixteenth note pattern on the first beat of a measure.

Chopin's Op. 68 Mazurkas are a set of four similar works, published posthumously in 1855. The Op. 68 No. 3, which we analyze here, was composed in 1830, when Chopin was 20 years old. Around this time, Chopin, already a piano virtuoso and accomplished composer, left his native Warsaw and settled in Paris, where we would remain until his death in 1849.

This Mazurka has a rather simplistic ternary structure with two outer sections and a contrasting middle (ABA). The first A section is made up of four eight-bar phrases ($aaba$). The first phrase is echoed by the second phrase: they are nearly identical, with the two exceptions being that (1) the second is marked *piano* (soft) rather *forte* (strong) and (2) the second ends on the tonic (F major) rather than the dominant (C major). The fourth eight-bar phrase is an exact repetition of the second. The second A section is a repeat of the first two eight-bar phrases of the beginning. The intervening B section is 12 bars long, divided into three four-bar groups. The first four bars are simply a repeated interval of a perfect 5^{th} in the left hand. This *ostinato* will continue for the whole section. The remaining eight measures consist of a four-bar phrase in the right hand repeated twice. The second differs from the first only on the final two notes, preparing the recapitulation of the A section.

In terms of tempi, the B section is indicated to be faster, with the marking *Poco più vivo* (a little livelier). The B section ends with a *ritardando* into the following A section. The b section ends with a *fermata* in measure 24, indicating an arbitrary elongation while the piece

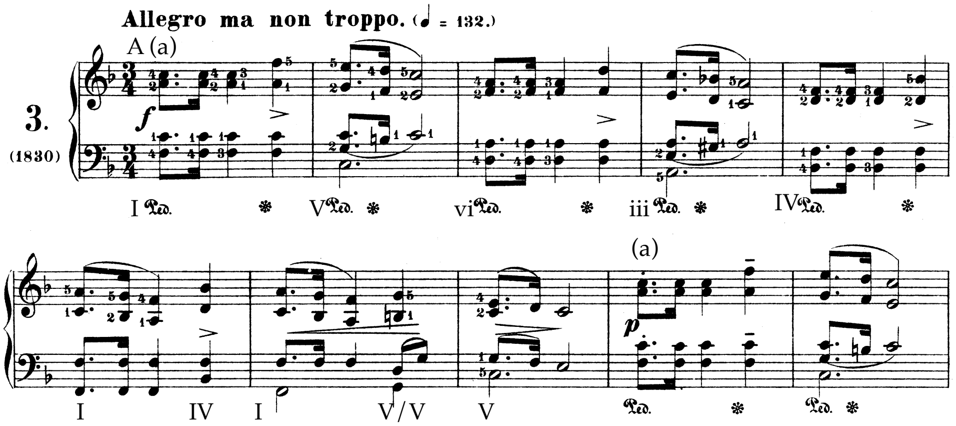

FIG 6. *The first ten measures of Chopin's Mazurka Op. 68, No. 3. The harmonic progression is indicated below the staff in Roman numerals. Sections are marked above the staff, e.g., A (a). Analysis by the authors. This image comes from the complete score published by Bote and Bock in 1880. This composition is in the public domain, and the score is freely available via the International Music Score Library Project.*

concludes with a two-measure long *ritardando*. Throughout, frequent markings prescribe emphasis of the third beat of each measure. This emphasis is in keeping with the mazurka style, an intentional thwarting of the listener's expectation of first-beat emphasis.

Figure 6 shows the first ten measures of the musical score with annotations for the sections discussed above and the harmonic progression in Roman numerals below the staff. The harmonies are standard, in fact, they are essentially the same as those of Pachelbel's *Canon*, familiar to many as "that song played at weddings." These harmonies, combined with the rhythmic repetition suggests a further division of this and all analogous sections into three small groupings: two two-measure phrases, followed by a four-measure phrase.

As a performer, these harmonic, rhythmic, and structural analyses aid in interpretation. The performer needs to decide how to emphasize or deemphasize these demarcations with slight or overt tempo or dynamic alterations. In a live performance, she could use physical motion to further suggest a particular interpretation. She can choose to emphasize long phrases, in this case, phrases of eight measures, or the shorter sub-phrases. Because of the repetition of similar phrases, she may choose to emphasize the long phrase on the first occurrence and shorter sub-phrases later on for variety, for example. While the musical structure suggests such possible interpretations, the performer must make these choices on her own, and may even alter those decisions from performance to performance.

3.2. *Archetypal performances.* Here we will carefully investigate the interpretive decisions implied by our estimated model for three rather different performances. Figure 7 shows the inferred state sequence for recordings made by Joyce Hatto in 1993 and Sviatoslav Richter in 1976. The B section is shaded in gray to better illustrate the formal divisions discussed above.

Our estimated model suggests that these two performers are quite different from each other. Hatto maintains a constant tempo carefully, remaining in state 1 with the exception of four periods of deceleration. All four periods coincide with the most significant phrase endings: at the end of the A section at measure 32, the end of the B section at measure 48, at the end of the piece, and the minor transition from $b \to a$ in the first A section (measure 24). According to our inferred model, she never accelerates or uses the transitory stress state.

In contrast, Richter uses all four states from our model. The short blips of acceleration before the B section and before the $b \to a$ transition are slightly out of place, and are likely

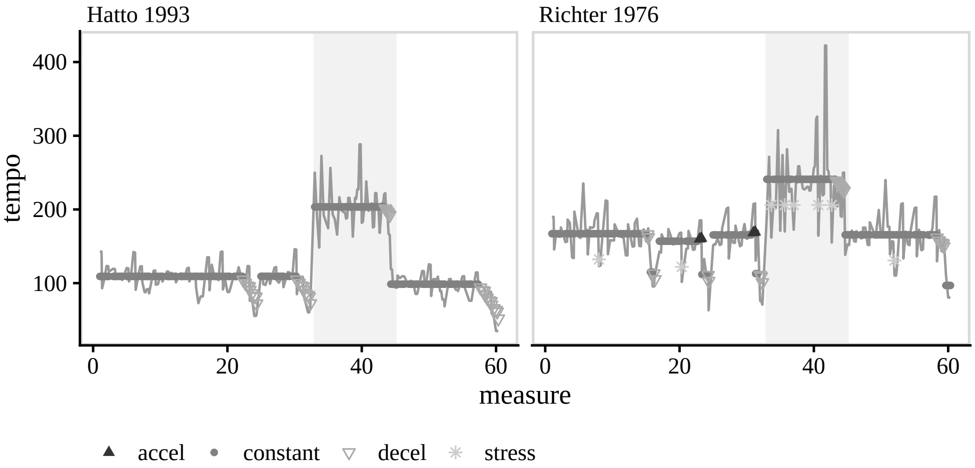

FIG 7. *Inferred performer choices for two recordings.*

TABLE 4
The estimated parameters for performances by Richter and Hatto.

	σ_ϵ^2	μ_{tempo}	μ_{acc}	μ_{stress}	σ_{tempo}^2	p_{11}	p_{12}	p_{22}	p_{31}	p_{13}	p_{21}	p_{32}
Richter 1976	426.70	136.33	-11.84	-34.82	439.38	0.85	0.05	0.74	0.44	0.02	0.25	0.17
Hatto 1993	405.57	130.36	-13.57	-27.93	408.99	0.94	0.03	0.82	0.36	0.01	0.16	0.19
Cortot 1951	403.71	182.84	-21.43	-45.67	460.82	0.92	0.02	0.71	0.34	0.03	0.23	0.09

better labelled as "constant", but these state transitions describe more severe decelerations than the model's linear assumption would allow (the multiplicative model in the Supplementary Material is much better). Richter uses stress frequently. Some may well be attributable to larger variance around constant tempo (picked up as frequent stress rather than larger σ_ϵ^2), but most correspond to interesting note emphases, for example the second beat of measure 20. This note is essentially a minor phrase ending, but it is also marked in the score with a *sforzando* (with sudden emphasis). It's the first of two such occurrences in the piece, the second coming four measures later on the *fermata*, Richter's slowest note in the entire piece. Richter likely chooses to make this prescribed emphasis with a sudden slow down in part because it takes place within the context of an already loud passage, precluding the use of extra volume. Table 4 shows the estimated parameters for these two performances.[4] Richter has larger observation variance, σ_ϵ^2, slightly faster average tempo, lower acceleration, and larger stress. He also has a larger tempo variance, meaning that returns to state 1 can start at relatively different tempos. On the other hand, Hatto is much more likely to remain in states 1 or 2. These inferences are largely consistent with the visual messages of Figure 7. The variability definitely increases around the constant tempo in Richter's performance and he uses faster overall tempos in both the A and B sections. While these two performances are quite different from each other, they also display similarities. Both take a faster tempo in the B section versus the A sections. Both performers slow down at the end of the piece, at the end of the B section, immediately preceding the B section, and at the $b \to a$ transition.

Alfred Cortot's 1951 performance is displayed in Figure 8. Both in terms of the parametric model we propose, and if we simply compare the vectors of note-by-note tempos (discussed

[4] In the Supplementary Material, we provide the estimates along with some measures of uncertainty for all 46 recordings.

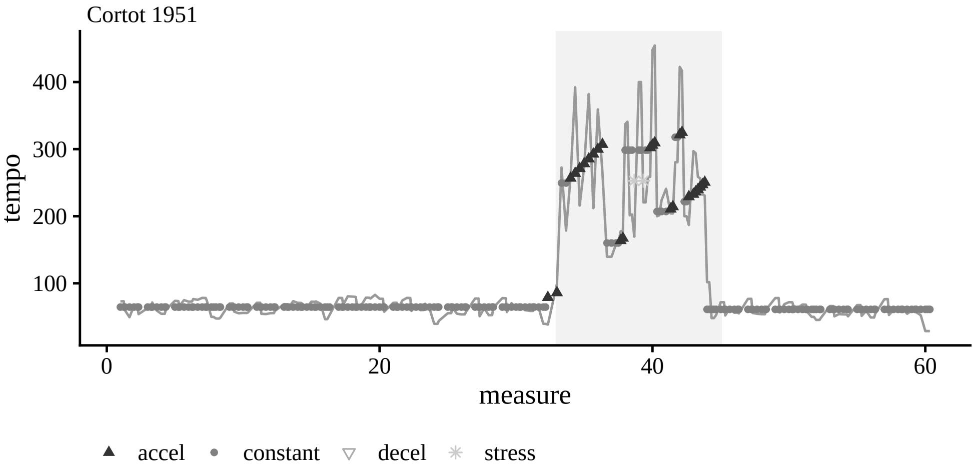

FIG 8. *Inferred performance choices for Alfred Cortot's 1951 recording.*

in more detail below), this performance is an outlier. Cortot never uses the deceleration state, and he remains in constant tempo for the entirety of both A sections. While the model describes his performance well, it also illustrates a deficiency of this approach: Cortot, more than any other performer, has large contrasts between the A and B sections. His A section is the slowest of all 46 recordings at around 64 b.p.m., half the marked tempo. The next slowest is Maryla Jonas's recording at around 84 b.p.m. Meanwhile, his B section is among the fastest of all the recordings and contains the fastest individual note. Additionally, there is stunningly little tempo variability in his A sections, but dramatic variation in the B section coupled with frequent uses of the acceleration and emphasis states. Taken together, Cortot's performance may be better described by estimating our model separately on the two sections.[5]

3.3. *Comparing performances.* To better understand how the 46 recordings relate to each other, we measured the distance between their vectors of paramater estimates. Because the parameters have different scales, have different domains, and can covary, we use Mahalanobis distance to scale by the inverse of the prior covariance. That is, the distance between performance i and performance j is given by

$$d_{i,j} = \left(\widehat{\theta}_i - \widehat{\theta}_j\right)^\top \Omega \left(\widehat{\theta}_i - \widehat{\theta}_j\right) \tag{3}$$

where $\widehat{\theta}$ corresponds to the estimates in, e.g. Table 4, and the prior precision, $\Omega = \Sigma^{-1}$, is calculated based on the distributions in Table 3. We use the prior covariance rather than the covariance of the estimated parameters because any parameters that are poorly identified by the model will have small estimated variance, and therefore dominate the distance calculation. Using the prior avoids this pathology and while still properly accounting for scale and structural dependence.

Figure 9 shows the distance matrix calculated from the estimated parameters for all 46 performances. The dendrogram helps to visualize similarities between the performances, but should be taken only as a heuristic. Across a variety of clustering procedures, methods for choosing the number of clusters (e.g., Dudoit and Fridlyand, 2002; Tibshirani, Walther and

[5] Cook (2013) suggests that this recording is not due to Cortot at all but part of a scandal at the Concert Artist label referred to as the "Hatto hoax", wherein her husband, owner of the label, released over 100 recordings made by others but listing her as the performer.

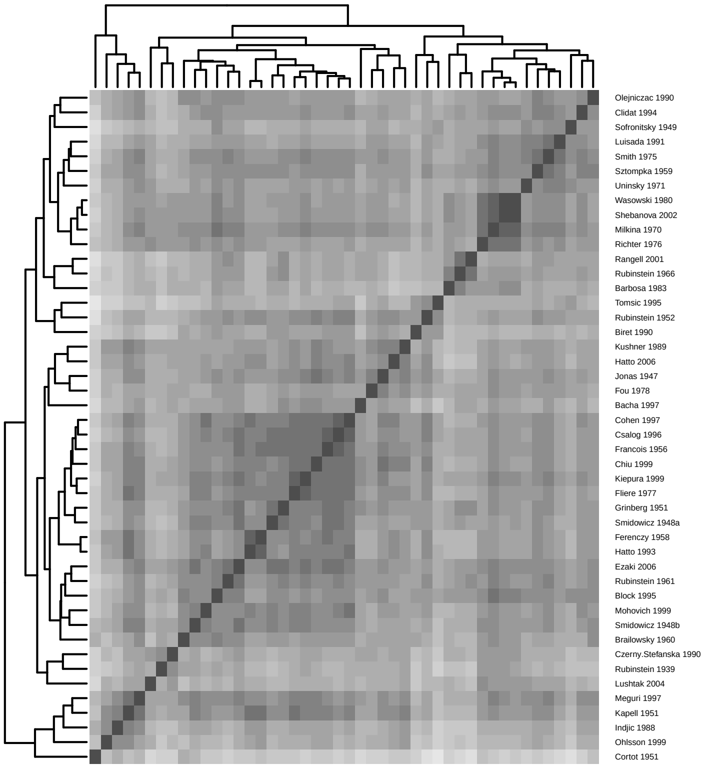

FIG 9. *Distance matrix using the estimated parameters for all 46 performances.*

Hastie, 2001) consistently suggest that there is only one cluster. The procedure developed by Tibshirani, Walther and Hastie (2001) chooses the number of clusters by finding the first maximum of the Gap statistic, subject to a measure of uncertainty. For the Chopin performances, the global maximum occurs at 7 clusters, which is both implausible and not robust to uncertainty. Nonetheless, for the purposes of organizing this section, we will consider those performances within a "group" to be more similar than across groups.

In order to inspect these performances visually, we follow the advice of an anonymous reviewer and perform principal components analysis on the matrix of estimated parameters. Only about 45% of the variance is explained by the first 2 components, and we would need 7 to explain 90%, but Figure 10 nonetheless corresponds somewhat closely to the groups suggested by the dendrogram in Figure 9. In the Supplementary Material, we plot all the inferred performance decisions by group and give the factor loadings for the first few principal components. In the remainder of this section, we describe typical behaviors of the performances within a few groups that have relatively small within-group variability.

The first group (indicated as $\circ$ in Figure 10) corresponds to reasonably staid performances. This group is the largest and corresponds to the block from Cohen to Brailowsky in Figure 9. In this group, the emphasis state is rarely visited with the performer tending to stay in the constant tempo state with periods of slowing down at the ends of phrases. Acceleration is almost never used. Furthermore, these performances have relatively slow average tempos, and not much difference between the A and B sections. Joyce Hatto's recording in Figure 7 is typical of this group.

Recordings in the fourth group ($\oplus$ in Figure 10) are those in the upper right of Figure 9, from Olejniczac to Richter. These recordings tend to transition quickly between states, especially constant tempo and slowing down, accompanied by frequent transitory emphases. The probability of remaining in state 1 is the lowest while the probability of entering state 2

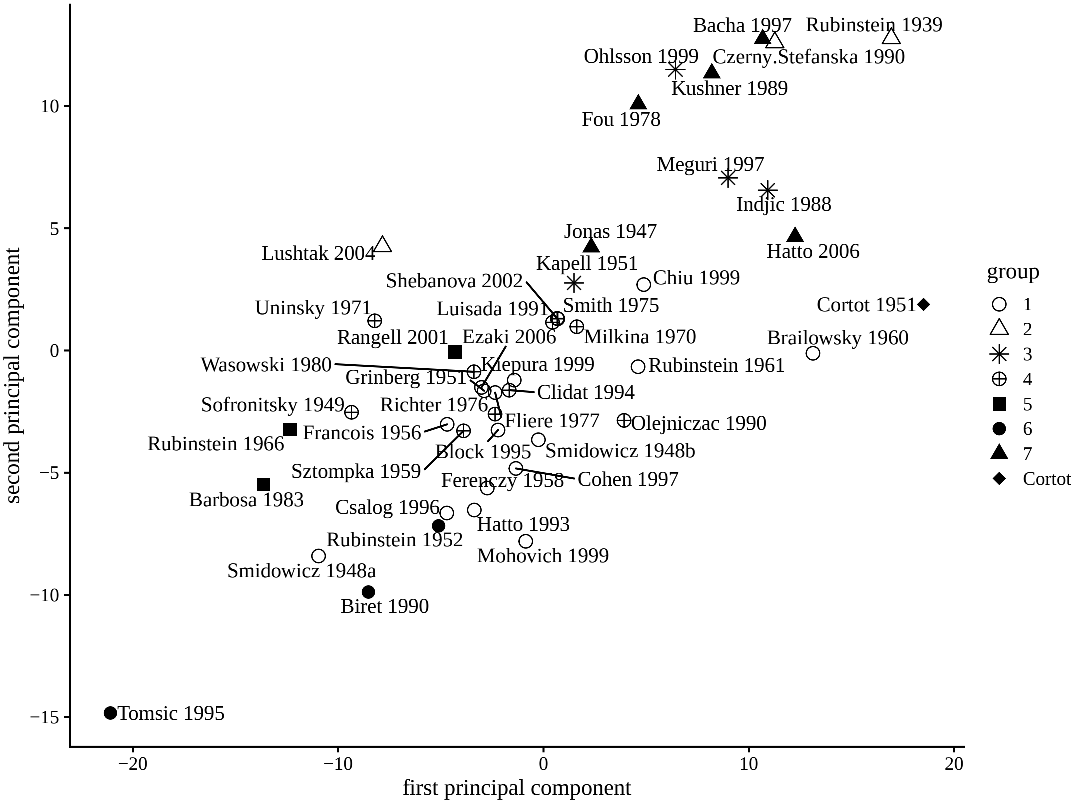

FIG 10. *The first two principal components of the matrix of estimated parameters. Similar performances are indicated by shape.*

from state 1 is the highest. The acceleration state is rarely visited. Four of the most similar performances are in this group, shown in Figure 11, along with Richter's 1976 recording.

The three performances in group six (●) are actually quite like others, but with small exceptions. Biret's 1990 performance is very much like those in group 1, but with a much larger contrast between tempos in the A and B sections. The recording by Rubinstein in 1952 is similar, though with a faster A section that has less contrast with the B section. Tomsic's 1995 performance is actually most similar to those in group three (✳), but played much faster and with a large σ_ϵ^2.

Comparing our groups to those we would find by applying the same procedure to the distances between note-by-note tempo vectors reveals a number of differences (see the Supplement for the distance matrix calculated in this way). The four similar recordings in Figure 11 would be spread across three different groups, for example, as would our group one. On the other hand, grouping by tempo vectors often (somewhat miraculously) groups recordings by the same pianist together: both recordings by Smidowicz (same as grouping by parameters), three of the four recordings by Rubinstein, and both recordings by Hatto. Both metrics see Cortot's recording as a strong outlier (the remote ◆ in Figure 10). In terms of Equation (3), Cortot's recording is 1.7 times farther from it's nearest neighbor than is the case for the next most dissimilar recording.

Figure 12 shows all four Rubinstein recordings. The 1939 recording is rather odd in that the measures 24–32 are so slow relative to the rest of the A section. The variability in the 1966 recording nearly obscures the contrast between the B section and the surrounding A sections. These two recordings are nonetheless grouped together by the tempo vectors. Our method

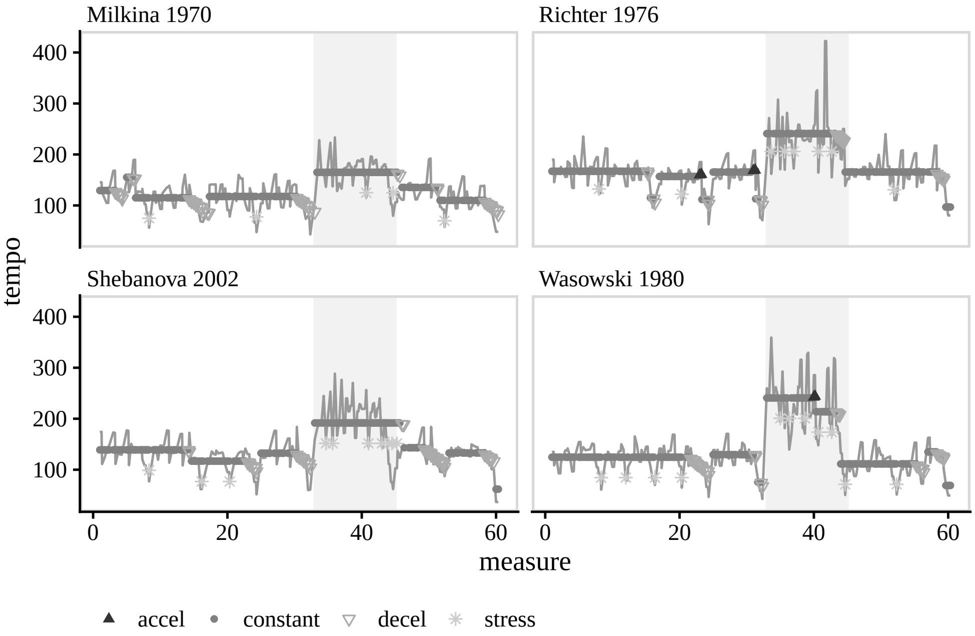

FIG 11. *Four similar performances, all in the fourth group (⊕).*

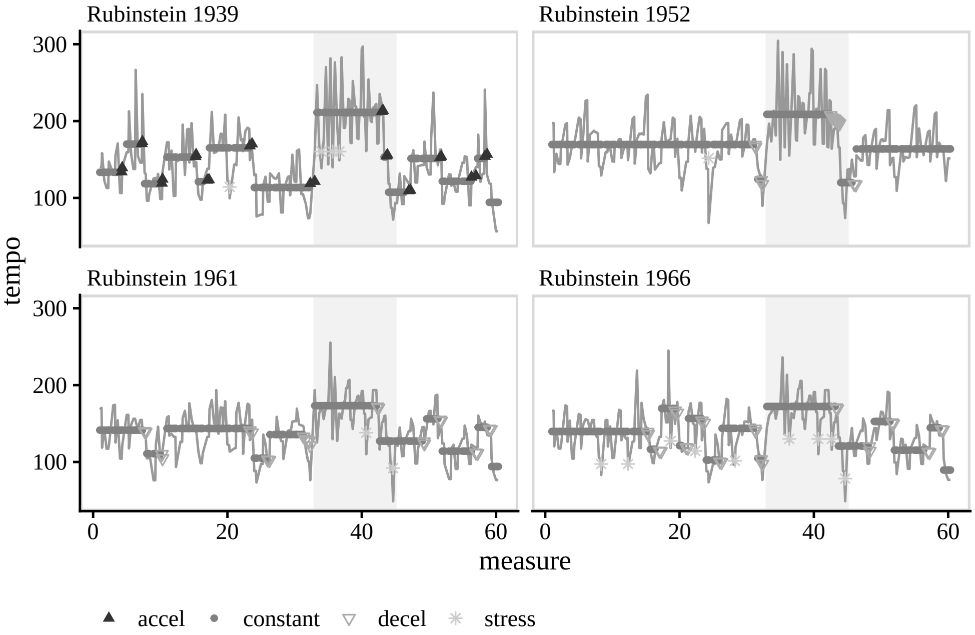

FIG 12. *The four recordings by Arthur Rubinstein. Our clustering puts the 1952 and 1961 recordings in clusters one and two while leaving the others out. Clustering by tempo vector separates 1952 from the other three.*

on the other hand, puts these four recordings in different groups. The estimated parameters for these four performances are shown in the bottom half of Table 5. The top half shows the parameters for the four similar performances in Figure 11. There is much larger variability across Rubinstein's recordings, as we would expect.

TABLE 5
The estimated parameters for the four similar performances in group four and those for all four by Arthur Rubinstein.

	σ^2_ϵ	μ_{tempo}	μ_{acc}	μ_{stress}	σ^2_{tempo}	p_{11}	p_{12}	p_{22}	p_{31}	p_{13}	p_{21}	p_{32}
Wasowski 1990	414.99	132.00	-10.00	-40.00	425.00	0.85	0.05	0.67	0.34	0.02	0.26	0.2
Shebanova 2002	439.98	132.00	-10.00	-40.00	400.02	0.85	0.05	0.67	0.33	0.02	0.27	0.20
Richter 1976	426.70	136.33	-11.84	-34.82	439.38	0.85	0.05	0.74	0.44	0.02	0.25	0.17
Milkina 1970	435.25	136.38	-9.68	-40.02	400.01	0.87	0.05	0.68	0.33	0.02	0.26	0.21
Rubinstein 1939	520.32	145.26	-7.89	-50.82	345.64	0.89	0.02	0.83	0.56	0.05	0.13	0.16
Rubinstein 1952	481.13	128.13	-7.76	-17.59	409.30	0.93	0.04	0.68	0.32	0.01	0.28	0.19
Rubinstein 1961	434.23	139.17	-8.34	-35.08	355.00	0.90	0.06	0.56	0.46	0.01	0.41	0.19
Rubinstein 1966	380.95	127.24	-8.80	-42.28	473.69	0.87	0.07	0.36	0.34	0.01	0.61	0.20

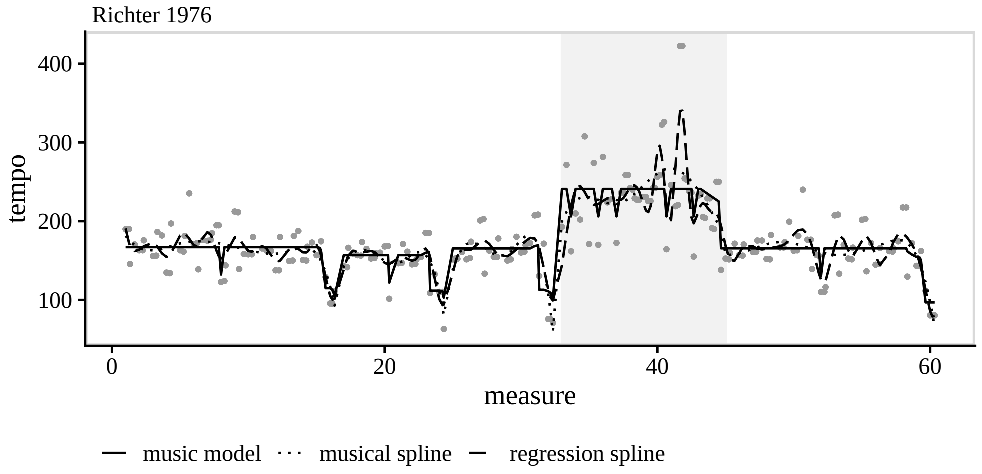

FIG 13. *Smoothing with splines and musical models*

3.4. *Alternative smoothers.* Our model is just one type of smoothing one could imagine using to find low-dimensional structure for the vector of note-by-note tempos. Alternative statistical techniques are common, and examining how they compare with our method helps to illuminate some of its benefits. The most obvious alternative is to use splines (Craven and Wahba, 1978; Wahba, 1990) though total-variation denoising or trend filtering (Kim et al., 2009; Tibshirani, 2014) are other reasonable alternatives. These statistical techniques perform smoothing by encouraging small changes in derivatives (splines) or bounded total variation (trend filtering). But musical performances do not conform to these assumptions because tempo interpretations rely on the juxtaposition of local smoothness with sudden changes and emphases to create listener interest. It is exactly the parts of a performance that are poorly described by statistical smoothers that render a performance interesting. Furthermore, many of these inflections are notated by the composer or are implicit in performance practice developed over centuries of musical expressivity. Consequently, smoothing that incorporates domain knowledge leads to better statistical and empirical results.

Figure 13 shows the note-by-note tempo of Richter's 1976 recording. Regression splines with equally spaced knots are shown with a dashed line. We use generalized cross validation (Golub, Heath and Wahba, 1979) to select the number of knots (one knot per measure). The dotted line shows a regression spline fewer knots, but whose locations were chosen manually to coincide with the musical phrase endings discussed in Section 3.1. Knots at phrase

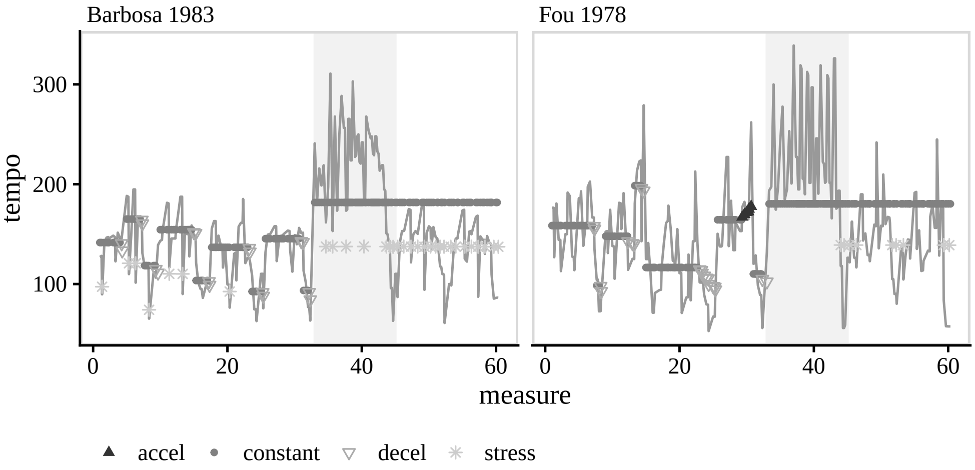

FIG 14. *Estimation errors on two performances.*

endings were duplicated up to four times to allow for discontinuities. The solid line shows the estimated smooth tempo from our model (the same as in Figure 7). The regression spline with equally spaced knots undersmooths in constant tempo areas in an attempt to capture sudden emphases and dramatic changes in others. The spline with informed knot choice does much better, picking up the periods of deceleration at the ends of phrases. Our model learns these behaviors on its own while also capturing individual emphases that are missed in the musical analysis but are idiosyncratic to Richter's playing. It is also more parsimonious to musical interpretation, inferring constant tempo periods rather than resulting in smoothly varying tempos in stable periods.

3.5. *Problems with the model and estimation.* While our model of musical decision making yields interesting insights into performance practice most of the time, it also suffers from some deficiencies. As discussed above in reference to Alfred Cortot's recording, the assumption that all parameters are stable over the entire piece may not always be accurate. The μ_{tempo} parameter especially, should be estimated separately in different sections. This problem will only be compounded in more complex music with many contrasting sections. A related issue is the current form for the slowing down and speeding up sections. Our model assumes that both occur linearly, with a constant decrease of μ_{acc} b.p.m. An ability to slow increasingly as one remains in the state may improve the model fit. The multiplicative model described in more detail in the Supplementary Material addresses this issue but requires further work.

There is nothing intrinsic to the model which forces states 2, 3, or 4 to always go in the correct direction. If for example, μ_{acc} is small in magnitude relative to σ^2_{acc}, a purposeful acceleration could be explained as time spent in state 2 but with large positive errors. For this piece, the priors help to avoid such occurrences, but this aspect of the Gaussian state-space model could be improved by enforcing non-Gaussian behavior. Of course, such constraints would complicate likelihood evaluation since the Kalman filter could no longer be used.

Relatedly, our model produced objectively incorrect inferences on two performances (Figure 14). Here, the estimated path failed to transition to state 1 at the recapitulation of the A section. In both cases, the resulting path stands out dramatically, remaining in the much faster constant tempo state from the B section with overly frequent emphases. Both of these performances are quite volatile, making estimation difficult. Altering the prior distributions along the lines suggested in the next section may help.

3.6. *Prior sensitivity and generalization.* As discussed in Section 2.4, the main reason for the prior distributions shown in Table 3 is that they help to identify the parameters. It is this identifiability issue that mainly guided our choices. For instance, if μ_{stress} is too similar to μ_{acc}, then a sequence like $1 \to 4 \to 1 \to 1$ will be hard to distinguish from $1 \to 2 \to 2 \to 1$. So, while it is important that those priors have low probability on common support, their shapes are less important for inference. Similar arguments hold for the other parameters.

With this intuition in mind, these priors should allow the model to work reasonably on similar piano pieces, even from different eras (Baroque, Classical, Modernist). That said, the specific values of prior means (for example, 10 for μ_{acc}) would be better expressed relative to the overall average speed, perhaps as $\overline{Y}/10$ b.p.m. or similar. In this way, really slow or fast pieces could be more easily accommodated. Since we only looked at one score, it is sufficient to simply fix some values. Using these 4 states should be enough for many types of music, even beyond piano recordings. However, music with more sections, say a piano sonata or a Prelude by Claude Debussy, would likely benefit from the inclusion of more discrete states. Adding another layer to Figure 2 that can handle formal divisions is one such option, while employing a Dirichlet process similar to that used by Ren et al. (2010) may also work.

To gauge prior sensitivity (subject to the constraints above), we also estimated the model under alternative specifications. In particular we examine Sviatoslav Richter's 1976 performance under four alternative priors: (1) replacing all Gamma distributions with inverse Gamma to examine the influence of tail shape; (2) making σ_ϵ^2 smaller; (3) setting $p(\sigma_\epsilon^2) = p(\sigma_{\text{tempo}}^2) \propto 1$; (4) replacing the informative transition probability priors with uniform distributions. We calculate the root-mean squared error (RMSE) and the negative log likelihood under different choices. The results are roughly similar both in terms of quantitative metrics and the qualitative inferences based on the figures. The most different setting occurs for choice (2) which decreases the quality of the fit, eliminates the use of the "emphasis" state, and has trouble dealing with the faster B section. A more comprehensive evaluation of these prior choices is included in the Supplement where we show the specific distributions and the inferred performance decisions (like in Figure 7) under each choice.

4. Discussion.

Musical interpretation is the most important factor in determining whether or not concertgoers enjoy a classical performance. Every performance includes mistakes—intonation issues, a lost note, an unpleasant sound—but these are all easily forgotten (or unnoticed) when a performer engages her audience, imbuing a piece with novel emotional content beyond the vague instructions inscribed on the printed page. While music teachers use imagery or heuristic guidelines to motivate interpretive decisions, combining these vague instructions to create a convincing performance remains the domain of the performer, subject to the whims of the moment, technical fluency, and taste.

In this paper, we develop a statistical model for tempo to elucidate performance decisions from classical music recordings. We present an algorithm for performing likelihood inference, estimate our model using a large collection of recordings of the same composition, and demonstrate how the model is able to recover performer intentions, and how they relate to standard musical analysis. While our methods perform well, our analysis reveals a number of avenues for future work and improvement. For the piano, apart from tempo decisions, the performer can also control dynamics differentially. Similar techniques to those employed here could be used to describe levels of loudness, and creating a model that combined both is desirable. Pianists have relatively few variables under their control for interpretation: tempo, dynamics, and pedalling. On the other hand, string players have many more. Bowing decisions, fingerings, vibrato, and broken chords are all important tools which are difficult to discern aurally from a recording, let alone describe with a simple statistical model. Significant work would be required to generalize our techniques to more detailed interpretative analysis. Examining more complex genres—sonatas, string quartets, symphonies—would also be interesting for future work.

Another avenue we wish to pursue in the future is to examine how our model's implications may be useful for teaching students. Can we estimate it quickly to provide immediate feedback to novice pianists? In this paper, we used a dataset in which the note-by-note tempos were annotated by experienced musicians. Combining our model with existing approaches to solving the note-score alignment problem (Dannenberg and Raphael, 2006; Lang and Freitas, 2005; Raphael, 2002), perhaps to their benefit would be the first step. Together, this could produce an immediate graphical representation that students and teachers could use to evaluate and improve their practice.

SUPPLEMENTARY MATERIAL

Supplement: Additional figures, source code, and musical examples (https://github.com/dajmcdon/dpf). The supplementary material online contains the three components. First, the R-package "`dpf`" containing code to perform the methods described in the article. The package also contains all data sets used as examples in the article. We also provide additional `R` code necessary to reproduce all analyses and graphics. Second, a separate appendix with additional graphics for all clusters and analysis of all 46 recordings. Third, MIDI files for a real and synthesized performance by Richter.

REFERENCES

ANDERSON, B. D. O. and MOORE, J. B. (1979). *Optimal filtering*. Prentice-Hall, Englewood Cliffs, NJ.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **72** 1–33.

ARCOS, J. and MANTARAS, R. L. (2001). An interactive cbr approach for generating expressive music. *Journal of Applied Intelligence* **21** 115–129.

ARIZA, C. (2005). Navigating the Landscape of Computer Aided Algorithmic Composition Systems: a Definition, seven Descriptors, and a Lexicon of Systems and Research. In *Proceedings of International Computer Music Conference*.

ARZT, A. and WIDMER, G. (2015). Real-Time Music Tracking Using Multiple Performances as a Reference. In *International Society for Music Information Retrieval (ISMIR)* 357–363.

BERNSTEIN, L. (2005). *Young People's Concerts*. Amadeus Press, Pompton Plains, NJ.

BISIANI, R. (1992). Beam Search. In *Encyclopedia of Artificial Intelligence* 2nd ed. (S. Shapiro, ed.) John Wiley and Sons.

BLOCK, B. A., JONSEN, I. D., JORGENSEN, S. J., WINSHIP, A. J., SHAFFER, S. A., BOGRAD, S. J., HAZEN, E. L., FOLEY, D. G., BREED, G., HARRISON, A.-L. et al. (2011). Tracking apex marine predator movements in a dynamic ocean. *Nature* **475** 86.

BOULANGER-LEWANDOWSKI, N., BENGIO, Y. and VINCENT, P. (2012). Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription. In *Proceedings of the 29th International Conference on Machine Learning*.

BRESIN, R., FRIBERG, A. and SUNDBERG, J. (2002). Director musices: The KTH performance rules system. In *Proceedings of SIGMUS-46*.

BURKHOLDER, J. P., GROUT, D. J. and PALISCA, C. V. (2014). *A History of Western Music*, 9th ed. WW Norton & Company.

CHARM (2009). Centre for the History and Analysis of Recorded Music. Online; accessed 12 March 2019.

COLLINS, N. (2016). A Funny Thing Happened on the Way to the Formula: Algorithmic Composition for Musical Theater. *Computer Music Journal* **40** 41-57.

CONT, A. (2010). A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** 974–987.

CONT, A., SCHWARZ, D., SCHNELL, N. and RAPHAEL, C. (2007). Evaluation of real-time audio-to-score alignment. In *International Symposium on Music Information Retrieval (ISMIR)*.

COOK, N. (2013). *Beyond the score: Music as performance*. Oxford University Press.

CRAVEN, P. and WAHBA, G. (1978). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik* **31** 377–403.

DANNENBERG, R. (1985). An On-Line Algorithm for Real-Time Accompaniment. In *Proceedings of the 1984 International Computer Music Conference* 193–198. International Computer Music Association.

DANNENBERG, R. B. and RAPHAEL, C. (2006). Music score alignment and computer accompaniment. *Communications of the ACM* **49** 38–43.

DROR, G., KOENIGSTEIN, N., KOREN, Y. and WEIMER, M. (2012). The Yahoo! Music Dataset and KDD-Cup'11. In *KDD Cup* 8–18.

DUDOIT, S. and FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3** research0036.1.

DURBIN, J. and KOOPMAN, S. J. (1997). Monte Carlo Maximum Likelihood Extimation for Non-Gaussian State Space Models. *Biometrika* **84** 669–684.

DURBIN, J. and KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford Univ Press, Oxford.

EARIS, A. (2007). An Algorithm to Extract Expressive Timing and Dynamics from Piano Recordings. *Musicae Scientiae* **11** 155-182.

EARIS, A. (2009). Mazurka in F Major, Op. 68, No. 3. accessed 12 March 2019.

EDDELBUETTEL, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York.

FEARNHEAD, P. and CLIFFORD, P. (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 887–899.

FLOSSMAN S. GRACHTEN, M. and WIDMER, G. (2012). Expressive Performance Rendering with Probabilistic Models. In *Guide to Computing for Expressive Music Performance* (A. Kirke and E. Miranda, eds.) Springer.

FLOSSMANN, S., GRACHTEN, M. and WIDMER, G. (2013). Expressive performance rendering with probabilistic models. In *Guide to Computing for Expressive Music Performance* 75–98. Springer.

FORSÉN, S., GRAY, H. B., LINDGREN, L. O. and GRAY, S. B. (2013). Was Something Wrong with Beethoven's Metronome? *Notices of the AMS* **60**.

FOX, E. B., SUDDERTH, E. B., JORDAN, M. I. and WILLSKY, A. S. (2011). A STICKY HDP-HMM WITH APPLICATION TO SPEAKER DIARIZATION. *The Annals of Applied Statistics* **5** 1020–1056.

FUH, C.-D. (2006). Efficient Likelihood Estimation in State Space Models. *Annals of Statistics* **34** 2026–2068.

GHAHRAMANI, Z. and HINTON, G. E. (2000). Variational learning for switching state-space models. *Neural Computation* **12** 831–864.

GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

GRINDLAY, G. and HELMBOLD, D. (2006). Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning* **65** 361–387.

GU, Y. and RAPHAEL, C. (2012). Modeling Piano Interpretation Using Switching Kalman Filter. In *International Society for Music Information Retrieval (ISMIR)* 145–150.

HADJERES, G., PACHET, F. and NIELSEN, F. (2017). DeepBach: a Steerable Model for Bach Chorales Generation. In *Proceedings of the 34th International Conference on Machine Learning* (D. PRECUP and Y. W. TEH, eds.). *Proceedings of Machine Learning Research* **70** 1362–1371. PMLR, International Convention Centre, Sydney, Australia.

HAMILTON, J. D. (2011). Calling Recessions in Real Time. *International Journal of Forecasting* **27** 1006–126.

HARVEY, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

KALLBERG, J. (1996). *Chopin at the boundaries: Sex, history, and musical genre*. Harvard University Press.

KALMAN, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* **82** 35–45.

KIM, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics* **60** 1–22.

KIM, C. J. and NELSON, C. R. (1998). Business Cycle Turning Points, a New Coincident Index, and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime Switching. *Review of Economics and Statistics* **80** 188–201.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). ℓ_1 Trend Filtering. *SIAM Review* **51** 339-360.

KITAGAWA, G. (1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association* **82** 1032–1041.

KITAGAWA, G. (1996). Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics* 1–25.

KOYAMA, S., PÉREZ-BOLDE, L. C., SHALIZI, C. R. and KASS, R. E. (2010). Approximate Methods for State-Space Models. *Journal of the American Statistical Association* **105** 170–180.

LANG, M., BISCHL, B. and SURMANN, D. (2017). batchtools: Tools for R to work on batch systems. *The Journal of Open Source Software* **2** 135.

LANG, D. and FREITAS, N. D. (2005). Beat tracking the graphical model way. In *Advances in Neural Information Processing Systems* 745–752. MIT press, Cambridge, MA.

MAEZAWA, A. (2019). Deep Linear Autoregressive Model of Interpretable Prediction of Expressive Tempo. In *Proceedings of the 16th Sound and Music Computing Conference*.

MCFEE, B. and LANCKRIET, G. (2011). Learning multi-modal similarity. *Journal of Machine Learning Research* **12** 491–523.

MEAD, A. (2007). On Tempo Relations. *Perspectives of New Music* **45** 64–108.

PATTERSON, T. A., THOMAS, L., WILCOX, C., OVASKAINEN, O. and MATTHIOPOULOS, J. (2008). State–space models of individual animal movement. *Trends in ecology & evolution* **23** 87–94.

RAPHAEL, C. (2002). A hybrid graphical model for rhythmic parsing. *Artificial Intelligence* **137** 217–238.

RAPHAEL, C. (2010). Music Plus One and Machine Learning In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (J. FÜRNKRANZ and T. JOACHIMS, eds.) 21–28.

RAUCH, H. E., STRIEBEL, C. and TUNG, F. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA journal* **3** 1445–1450.

REN, L., DUNSON, D., LINDROTH, S. and CARIN, L. (2010). Dynamic Nonparametric Bayesian Models for Analysis of Music. *Journal of the American Statistical Association* **105** 458–472.

ROBERTS, A., HAWTHORNE, C. and SIMON, I. (2018). Magenta.js: A JavaScript API for Augmenting Creativity with Deep Learning. In *Joint Workshop on Machine Learning for Music (ICML)*.

ROBERTS, A., ENGEL, J., RAFFEL, C., HAWTHORNE, C. and ECK, D. (2018). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the 35th International Conference on Machine Learning* (J. DY and A. KRAUSE, eds.). *Proceedings of Machine Learning Research* **80** 4364–4373. PMLR, Stockholmsmässan, Stockholm Sweden.

SCHEDL, M., GÓMEZ, E., URBANO, J. et al. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval* **8** 127–261.

STOWELL, D. and CHEW, E. (2012). Bayesian MAP estimation of piecewise arcs in tempo time series. In *Proceedings of Computer Music Multidisciplinary Research*.

STURM, B. L., BEN-TAL, O., MONAGHAN, Ú., COLLINS, N., HERREMANS, D., CHEW, E., HADJERES, G., DERUTY, E. and PACHET, F. (2019). Machine learning research that matters for music creation: A case study. *Journal of New Music Research* **48** 36–55.

R CORE TEAM (2019). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

THICKSTUN, J., HARCHAOUI, Z. and KAKADE, S. M. (2017). Learning Features of Music from Scratch. In *International Conference on Learning Representations (ICLR)*.

TIBSHIRANI, R. J. (2014). Adaptive Piecewise Polynomial Estimation via Trend Filtering. *Annals of Statistics* **42** 285–323.

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63** 411–423.

VAN DEN OORD, A., DIELEMAN, S. and SCHRAUWEN, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 2643–2651. Curran Associates, Inc.

VERCOE, B. (1984). The synthetic performer in the context of live performance. In *Proceedings of the 1984 International Computer Music Conference* 199–200. International Computer Music Association.

WAHBA, G. (1990). *Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

WHITELEY, N., ANDRIEU, C. and DOUCET, A. (2010). Efficient Bayesian Inference for Switching State-Space Models using Discrete Particle Markov Chain Monte Carlo Methods Technical Report No. 10:04, Bristol University.

WICKHAM, H. (2016). *ggplot2: Elegant graphics for data analysis*, 2nd ed. Springer.

WICKHAM, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse' R package version 1.2.1.

WIDMER, G., FLOSSMANN, S. and GRACHTEN, M. (2009). YQX Plays Chopin. *AI Magazine* **30** 35.