Markov-switching State Space Models for Uncovering Musical Interpretation

Daniel J. McDonald*
Department of Statistics, Indiana University

Michael McBride
Department of Statistics, Indiana University

Yupeng Gu
Department of Informatics, Indiana University

Christopher Raphael
Department of Computer Science, Indiana University

April 30, 2019

Abstract

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. 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 research, we use data from the CHARM Mazurka Project—forty-six professional recordings of Chopin's Mazurka Op. 63 No. 3 by consumate artists—with the goal of elucidating musically interpretable performance decisions. Using information on the inter-onset intervals of the note attacks in the recordings, we apply functional data analysis techniques enriched with prior information gained from music theory to discover relevant features and perform hierarchical clustering. The resulting clusters suggest methods for informing music instruction, discovering listening preferences, and analyzing performances.

*The authors gratefully acknowledge support from the National Science Foundation (grants DMS-1407439 and DMS-1753171).

Keywords: keyword1; keyword2;

Contents

1 Introduction

Note: See here for the style guide (detailed) and here for the author instructions.

Statistical analysis of the musical contentent 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).

> You see, everybody thinks he knows what classical music is: just any music that isn't jazz, like a Stan Kenton arrangement or a popular song, like "I Can't Give You Anything but Love Baby," or folk music, like an African war dance, or "Twinkle, Twinkle Little Star." But that isn't what classical music means at all.

Bernstein goes on to discuss an important distinction between what we often call 'classical music' and other types of music which is highly relevant to the current study.

> The real difference is that when 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—even the exact number of instruments or voices; 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

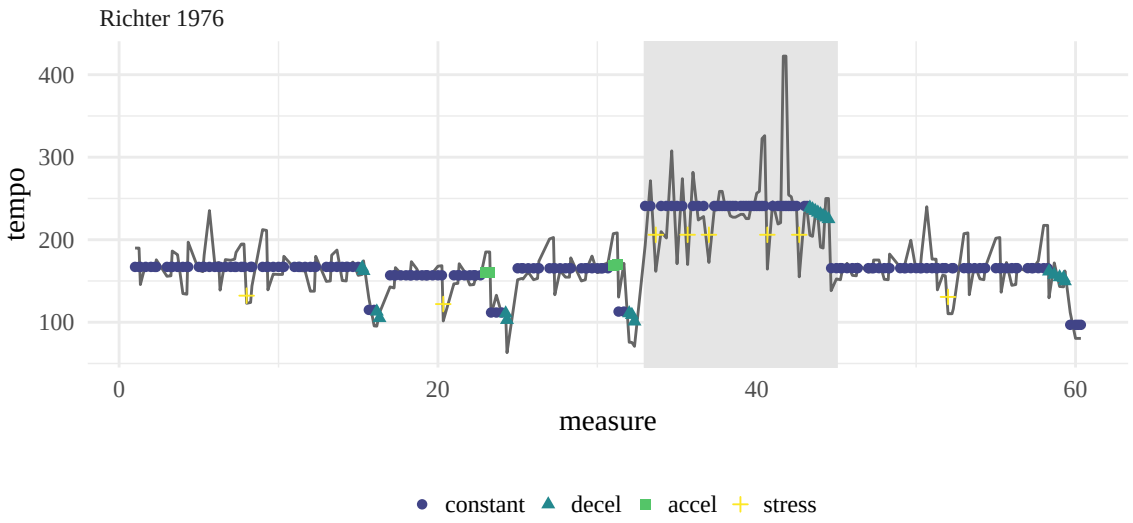

Figure 1: Note-by-note tempos for a recording of Chopin's Mazurka Op. 68 No. 3 by Sviatislav Richter. The solid line are the observed tempos, while the dots represent inferred tempo states from our model.

> have to know about what the composer wanted. But that's just what makes the performer's job so exciting—to try and find out from what the composer did write down as exactly as possible what he meant. Now of course, performers are all only human, and so they always figure it out a little differently from one another.

What separates classical music from other types of music is that the music itself is written down but performed millions of times in a variety of interpretations. 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 serve only to relate 'pieces' but not 'performances'. We need new methods in order 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[1] in beats-per-

[1]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".

minute (b.p.m.) of a recording made by Sviatislav Richter of Chopin's Mazurka Op. 68 No. 3. The solid line shows the actual tempo at which he plays each note, while the colored 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. 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.

1.1 Related work

The vast majority of work at the intersection of statistics or machine learning 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 learning 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 et al. (2018a) use variational autoencoders to learn 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 et al., 2012; Collins, 2016; Flossmann et al., 2013; Sturm et al., 2019). It is actively investigated, especially using deep learning (Hadjeres et al., 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 et al., 2018b).

The score alignment problem matches live or recorded performances to the musical score, a necessary processing step for any type of 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 seek to modulate playback in response to a live soloist who both makes interpretive timing decisions and mistakes. Off-line alignment (Earis, 2007) can be used for simply analyzing the recordings, as we do here, or for generating features that describe the performance (Thickstun et al., 2017), possibly for later analysis in recommender systems (McFee and Lanckriet, 2011; van den Oord et al., 2013). For an overview of these and related goals in music information retrieval, see Schedl et al. (2014).

Note: What do we do?

- We want to model tempo and dynamic decisions.
- We want a musician to understand what the parameters mean.

2 Materials and methods

2.1 Data and preprocessing

In this paper, we examine note-by-note tempos for 46 recordings of Chopin's Mazurka Op. 68, No.3. The data are part of a large collection of the complete Chopin Mazurkas and other recordings collected and analyzed by the Center for the History and Analysis of Recorded Music (CHARM) in the United Kingdom (CHA, 2009). The recordings were processed using the note-onset detection algorithm developed in (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. For the sake of reproducibility, we have included this data in our R package.

2.2 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 a particular 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

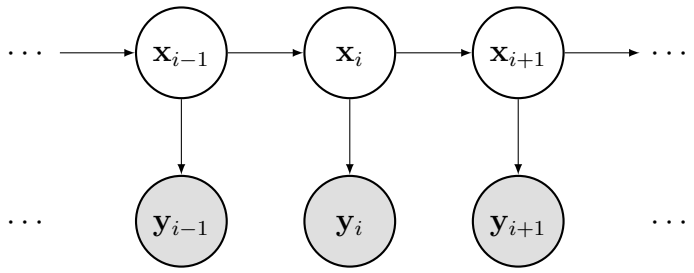

Figure 2: State-space model. Filled objects are observed, circles indicate that both hidden and observed states are continuous.

Markov property—induces a temporal model for the observations. The most general form of a state-space model is then characterized by the observation equation (the conditional probability of observations given the states), the state 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 independent and identically distributed as well as mutually independent. Both y_i and x_i can be (generally) 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. Figure 2 shows a directed acyclic graph for the dependence structure in the typical state-space model.

In the case where f_θ and g_θ are linear with ϵ_i and η_i normally distributed, (1) specializes to

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

where the matrices d, T, c, Z, R, Q, and G are allowed to depend on θ and can potentially vary (deterministically) with i. In this case, the Kalman filter, Algorithm 1 (see e.g. Harvey, 1990; Kalman, 1960), can be used to derive closed form solutions for the conditional distributions of the states and to calculate the likelihood of θ given data.

While Algorithm 1 returns the likelihood for θ, χ_i and P_i represent the mean and variance of the conditional distribution of the unobserved component given only the observations

Algorithm 1 Kalman filter: estimate x_i conditional on $\{y_j\}_{j=1}^{i}$, for all $i = 1, \ldots, n$ and calculate the log likelihood for θ

Input: Y, x_0, P_0, d, T, R, c, Z, and G
$\ell(\theta) \leftarrow 0$ ▷ Initialize the log-likelihood
for $i = 1$ to n **do**
  $H = RQR^\top$ ▷ Effective state variance
  $\chi_i \leftarrow d + Tx_{i-1|i-1}, \quad P_i \leftarrow H + TP_{i-1|i-1}T^\top$ ▷ Predict current state
  $\widetilde{y}_i \leftarrow c + Z\chi_i, \quad F_i \leftarrow G + ZP_iZ^\top$ ▷ Predict current observation
  $v_i \leftarrow y_i - \widetilde{y}_i \quad K_i \leftarrow P_iZ^\top F^{-1}$ ▷ Forecast error and Kalman gain
  $x_{i|i} \leftarrow \chi_i + K_iv_i, \quad P_{i|i} \leftarrow P_i - P_iZ^\top K_i$ ▷ Update
  $\ell(\theta) = \ell(\theta) - v_i^\top F^{-1}v_i - \log(|F_i|)$
end for
return $\widetilde{Y} = \{\widetilde{y}_i\}_{i=1}^n$, $\chi = \{\chi_i\}_{i=1}^n$, $\widetilde{X} = \{x_{i|i}\}_{i=1}^n$, $P = \{P_i\}_{i=1}^n$, $\widetilde{P} = \{P_{i|i}\}_{i=1}^n$, $\ell(\theta)$

Algorithm 2 Kalman smoother (Rauch-Tung-Striebel): estimate $\widehat{X}$ conditional on Y

Input: χ, $\widetilde{X}$, P, $\widetilde{P}$, T, c, Z.
$t = n$,
$\widehat{x}_n \leftarrow \widetilde{x}_n$,
while $t > 1$ **do**
  $\widehat{y}_i \leftarrow c + Z\widehat{x}_i$, ▷ Predict observation vector
  $e \leftarrow \widehat{x}_i - \chi_i, \quad V \leftarrow P_i^{-1}$,
  $t \leftarrow i - 1$, ▷ Increment
  $\widehat{x}_i = \widetilde{x}_i + \widetilde{P}_iTVe$
end while
return $\widehat{Y} = \{\widehat{y}_i\}_{i=1}^n, \widehat{X} = \{\widehat{x}_i\}_{i=1}^n$

$\{y_j\}_{j=1}^{i}$: $\chi_i = \mathbb{E}[x_i \mid y_1 \ldots, y_i]$ and $P_i = \mathbb{V}[x_i \mid y_1, \ldots, y_i]$. To incorporate all future observations into these estimates, the Kalman smoother is required.

There are many different smoother algorithms tailored for different applications. Algorithm 2, due to Rauch et al. (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 $\widehat{x}_i = \mathbb{E}[x_i \mid y_1, \ldots, y_n]$ for the sake of computational speed. This version is more appropriate for inference in the type of switching models we discuss below.

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 (e.g., Andrieu et al., 2010; Koyama et al., 2010). However, these models are less useful for change-point detection or other forms of 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 3. Here, we assume S 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 3 gives the conditional independence properties we will use in our model for musical interpretation, but this represents just one of many possibilities. Switching state-space models have a long history with many applications 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). An excellent overview of the history, typography, and algorithmic developments can be found in (Ghahramani and Hinton, 2000). In (2), the parameter matrices were not time varying. In our switching model, we allow the switch states s_i, s_{i-1}, along with the parameter vector θ, to determine the specific dynamics at time i:

$$\begin{aligned} x_1 &\sim N(x_0,\ P_0), \\ x_{i+1} &= d(s_i, s_{i-1}) + T(s_i, s_{i-1})x_i + R(s_i, s_{i-1})\eta_i, \quad \eta_i \sim N(0, Q(s_i, s_{i-1})), \\ y_i &= c(s_i) + Z(s_i)x_i + \epsilon_i, \qquad\qquad\qquad\qquad \epsilon_i \sim N(0, G(s_i)). \end{aligned} \tag{3}$$

In other words, the hidden Markov (switch) state determines which parameter matrices govern the evolution of the system.

2.3 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 4 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

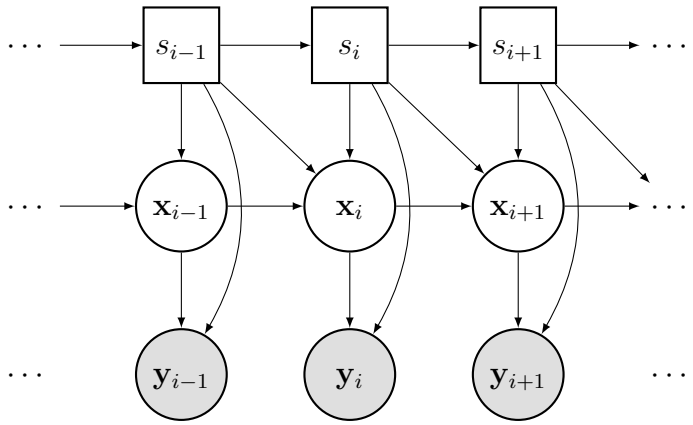

Figure 3: Switching state space model. Filled objects are observed, rectangles are discrete, and circles are continuous.

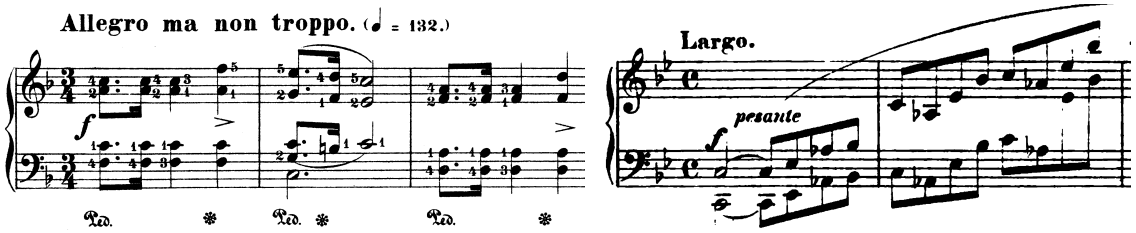

Figure 4: The beginning of two Chopin piano compositions: the Mazurka we analyze is on the left while the Ballade No. 1, Op. 23 is on the right.

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 being never plays precisely like a metronome, although they may try quite hard to do so. The observed ratio of musical time to clock time is therefore best viewed as stochastic, the sum of an intentional, constant tempo, plus noise representing

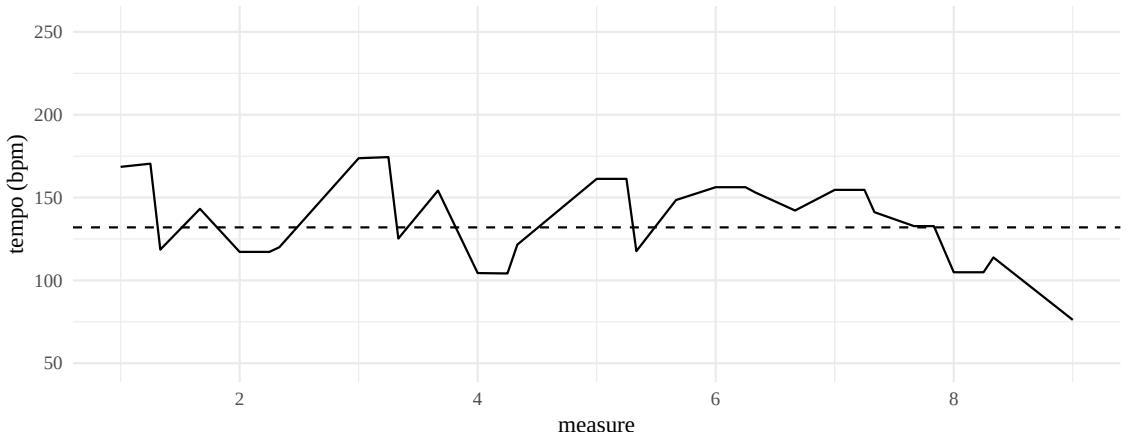

Figure 5: The solid line shows the observed note-by-note tempo for the beginning of the Mazurka as performed by Arthur Rubinstein in 1961. The dashed line indicates 132 b.p.m.

inaccuracy or, perhaps more charitably, unintentional variation which the listener fails to perceive as "wrong". For instance, the example in Figure 5 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.

Similar to (Gu and Raphael, 2012), we propose a Markov model for S on four states for four different performance behaviors. Our model has transition with transition probability diagram given by Figure 6. 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 only allow certain transitions for musical reasons and for estimability. The marked transition probabilities are sufficient to infer the remainder. The fourth state, stress, corresponds to *tenuto*, a common feature of musical performance. Such

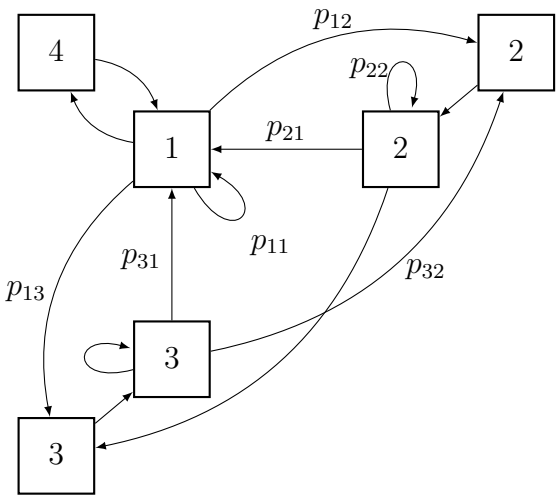

Figure 6: Transition diagram. The four states are: constant tempo (1), deceleration (2), acceleration (3), and emphasis (4).

stresses may be marked with a line over the note in question, but are more often a feature of the performer taste, corresponding to a longer-than-written duration of 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 there before leaving. Thus, the entire transition diagram is fully determined. This process can be viewed equivalently as a second order Markov chain. We discuss some potential improvements at the end of Section 3.

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, so, throughout this piece, a quarter-note (♩) has $l_i = 1/3$, an eighth note (♪) has $l_i = 1/6$, etc. This is because each measure contains three quarter notes. 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. So for any performance, we want to be able 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

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

2.4 Estimation and computational issues

To understand the performance decisions of individual musicians, we wish to simultaneously learn 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 4^n possible paths through the switch variables, so evaluating the likelihood to maximize over θ via Algorithm 1 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 in (Fearnhead and Clifford, 2003). Whiteley et al. (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 3 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 and 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 to the full discrete distribution over paths with a subset of support points by

Switch states		Transition equation parameter matrices		
s_i	s_{i-1}	d	T	$RQR^\top$
1	1	0	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
2	1	$\begin{pmatrix} l_i\mu_{\text{acc}} \\ \mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} l_i^2 & l_i \\ l_i & 1 \end{pmatrix}$
3	1	$\begin{pmatrix} -l_i\mu_{\text{acc}} \\ -\mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} l_i^2 & l_i \\ l_i & 1 \end{pmatrix}$
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	$\begin{pmatrix} 1 & l_i \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
3	2	$\begin{pmatrix} -l_i\mu_{\text{acc}} \\ -\mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} l_i^2 & l_i \\ l_i & 1 \end{pmatrix}$
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	$\begin{pmatrix} 1 & l_i \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
2	3	$\begin{pmatrix} l_i\mu_{\text{acc}} \\ \mu_{\text{acc}} \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\sigma^2_{\text{acc}} \begin{pmatrix} l_i^2 & l_i \\ l_i & 1 \end{pmatrix}$
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	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

Switch states	Measurement equation 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

Table 1: Parameter matrices of the switching state space model.

Algorithm 3 Discrete particle filter

Input: Y, θ, π_1 probability vector over initial states (paths), B maximum beam width
for $i = 1$ **to** n **do**
    Set $b_i = |\{\pi_i > 0\}|$, the number of current paths
    Use Algorithm 1 to calculate the 1-step likelihood ℓ_i for each path and every potential state s_{i+1} resulting in $b_i|S|$ particles
    Set $\pi_{i+1} \leftarrow \pi_i \ell_i p_i$: multiply the path probability by the likelihood and the probability of transitioning. Normalize π.
    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 π_n.

minimizing the mean squared error.

Algorithm 3 returns B paths along with their weights through the discrete state S 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 3 and then using the best path with Algorithm 1.

2.5 Penalized maximum likelihood

Even without the latent discrete states, 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. However, for the present application, we have reasonably strong prior information for many of the parameters. The three mean parameters μ_{tempo}, μ_{acc} and μ_{stress} have sign restrictions in addition to strong information about 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 reasonably strong information 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

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

Table 2: Informative prior distributions for the music model

conforming with musical intuition. The specific choices are shown in Table 2. We chose to fix σ^2_{acc} and σ^2_{stress} after numerical experiments suggested that they were poorly identified from the data.

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 parametric clusters determined by our model, contrast these with some alternative approaches to music modelling, and discuss some difficulties we encountered.

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 (cite Burkhardt). 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 (cite

Kallberg), 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 simply 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 repeated twice. The second time 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 *fermatta* in measure 24, indicating an arbitrary elongation while the piece concludes with a two-measure long *ritardando*. Throughout, frequent markings prescribe emphasis of the third beat of each measure. This emphasis is also in keeping with the mazurka style, an intentional thwarting of the listener's expectation of first-beat emphasis.

Figure 7 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." This harmonic progression, combined with the rhythmic repetition suggests a further division of this section, 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.

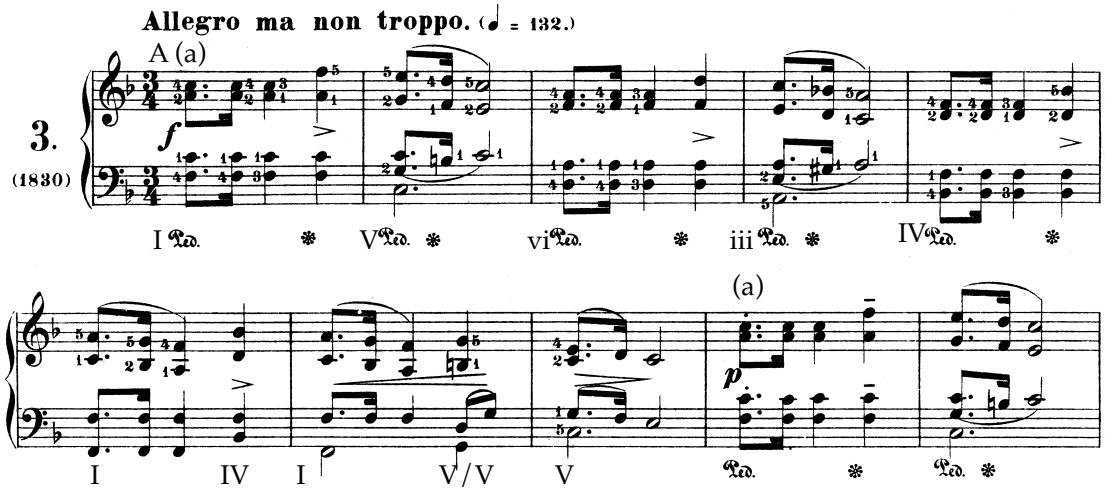

Figure 7: The first ten measures of Chopin's Mazurka Op. 68, No. 3. The harmonic progression is indicated below the staff in Roman numerals. Sections are marked above the staff, e.g. A (a). Analysis by the authors.

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 their own, and may even alter those decisions from performance to performance.

Note: maybe a short section here claiming the parameter reduced/smoothed/parsed version is a valid representation of the original recording? could use some audio examples if possible or maybe just use the results in the 2012 ISMIR paper saying "based on a user study conducted in ISMIR 2012 paper, the "smoothed" versions are not clearly distinguishable from the original recording

3.2 Archetypal performances

Here we will carefully investigate how our model learns interpretive decisions for three rather different performances. Figure 8 shows the inferred state sequence for recordings made by Joyce Hatto in 1993 and Sviatislav Richter in 1976. The B section is shaded in gray to better

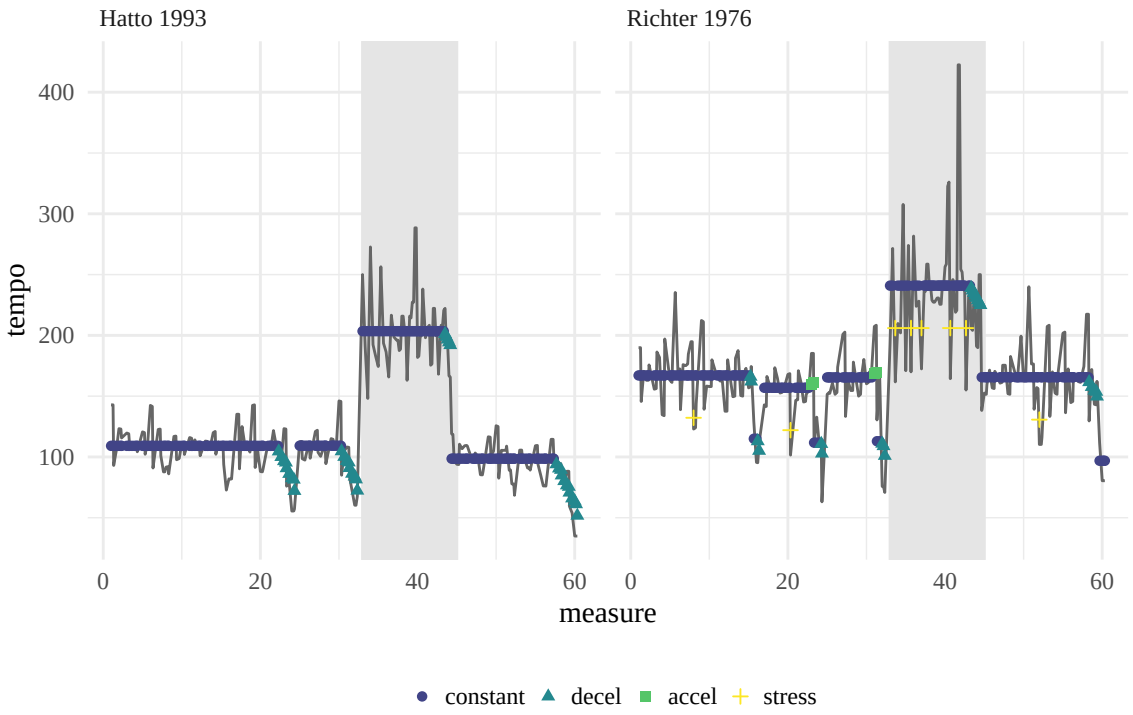

Figure 8: Inferred performer choices for two recordings.

illustrate the formal divisions discussed above.

In terms of our model, 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 better labelled as "constant", but these state transitions describe more severe decelerations than the model's linear assumption would allow. 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 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 *fermatta*, Richter's slowest note

Table 3: The estimated parameters for performances by Richter and Hatto.

	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}
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

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 3 shows the estimated parameters for these two performances. Richter has larger observation variance, σ^2_ϵ, 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 takeaways of Figure 8. It's easy to see the increased variability around the "constant tempo" in Richter's performance and the faster overall tempos in both the A and B sections.

While these two performances are quite different from each other, they also display some 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 9. Both in terms of the parametric model we propose, and if we simply compare the vectors of note-by-note tempos (discussed 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.

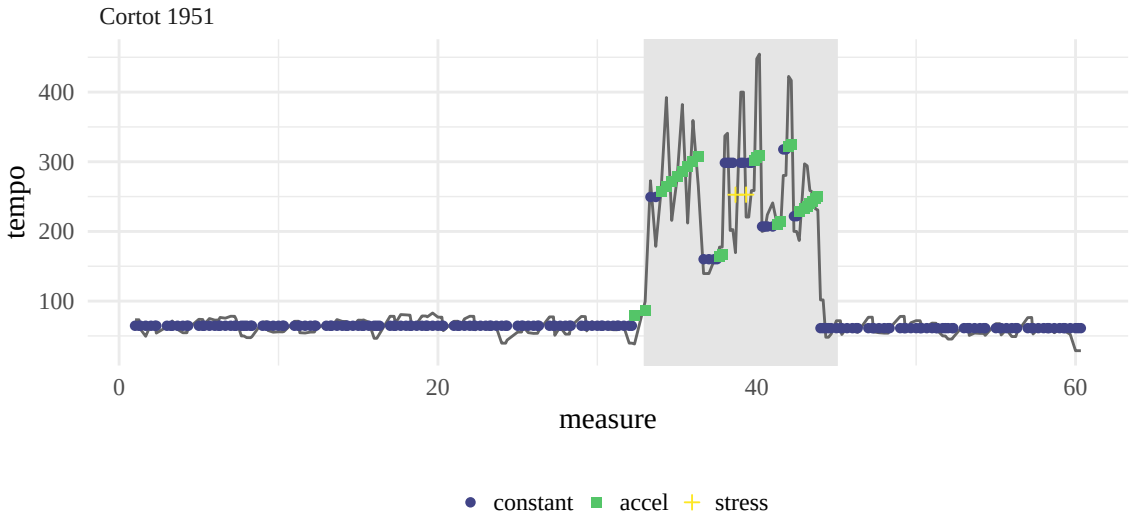

Figure 9: Inferred performance choices for Alfred Cortot's 1951 recording.

3.3 Clustering musical performances

To better understand how all the 46 performances relate to each other, we applied parametric clustering using the eleven-dimensional vector of estimated parameters. Because the estimated parameters are of different scales, have different domains, and can covary, we treat them differentially. To calculate distances between the mean and variance parameters, σ_ϵ^2, μ_{tempo}, μ_{acc}, μ_{stress}, and σ^2_{tempo} we simply use Euclidean distance on each individually. In the cases of the probabilities, we use weighted Euclidean distance in the prior precision. For example, for $p_{1\cdot}$, we calculate

$$d\left(p_{1\cdot},\ p'_{1\cdot}\right) = (p_{1\cdot} - p'_{1\cdot})^\top \Omega (p_{1\cdot} - p'_{1\cdot}),$$

where

$$\Omega_{ij}^{-1} := \Sigma_{ij} := \alpha_0^{-2}(\alpha_0 + 1)^{-1}\left(\alpha_i\alpha_0\delta_{ij} - \alpha_i\alpha_j\right),$$

is the covariance matrix of the Dirichlet distribution with $\alpha_0 = \sum_i \alpha_i$ and δ_{ij} the indicator that $i = j$. We then standardize each individual distance matrix to have a maximum distance of 1 and add them together. As such, the distance between any two performances can be no larger than 8.

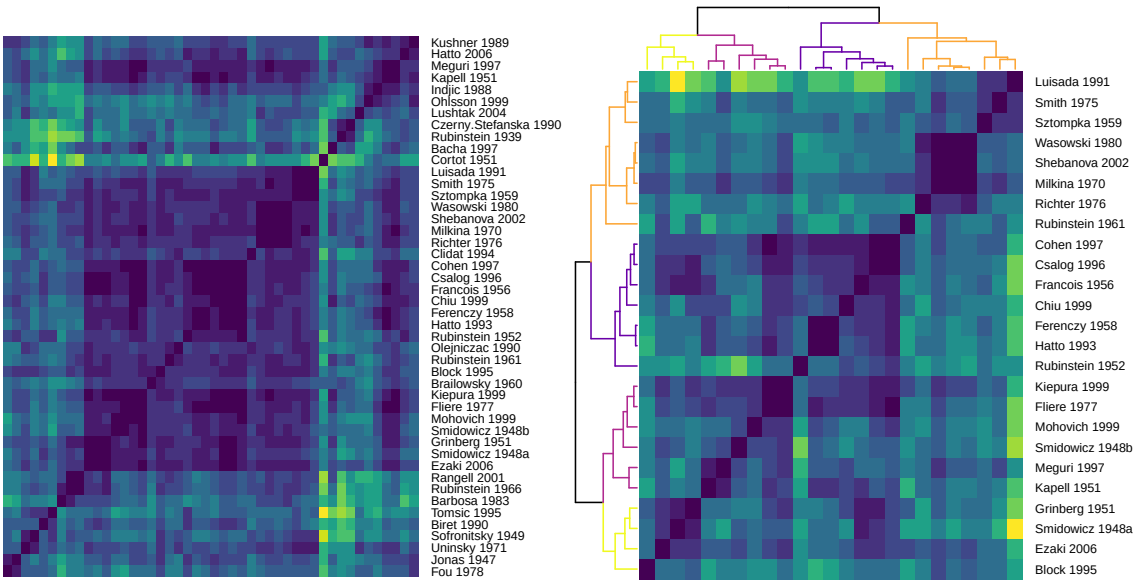

Figure 10: Distance matrix using estimated model parameters. Left: the matrix for all 46 parameters. Right: The same matrix with "outlying" performances removed and a dendrogram from hierarchical clustering.

Figure 10 shows the distance matrix calculated from the estimated parameters for all 46 performances (left) and the same matrix with "outlying" performances removed (right). To determine outlying performances, we calculated the distance to the third nearest performance. We then removed those performances that exceeded a threshold, meaning that the nearest similar performances were "far away". This screening left 25 performances. We used hierarchical clustering on these 25 performances, trying between two and five clusters. The remaining 21 were grouped together as "other". The right panel of Figure 10 displays this subset along with a dendrogram and four clusters.

The first cluster corresponds to performances which are reasonably staid. 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 never used. Such state preferences are clearly inferred by the model as shown in, e.g. the top row of Figure 11. Furthermore, these performances have relatively low average tempos, and not much difference between the A and B sections. Joyce Hatto's performance shown in Figure 8 is typical of this cluster.

Recordings in the second cluster tend to transition quickly between states, especially constant tempo and slowing down accompanied by frequent transitory emphases. The prob-

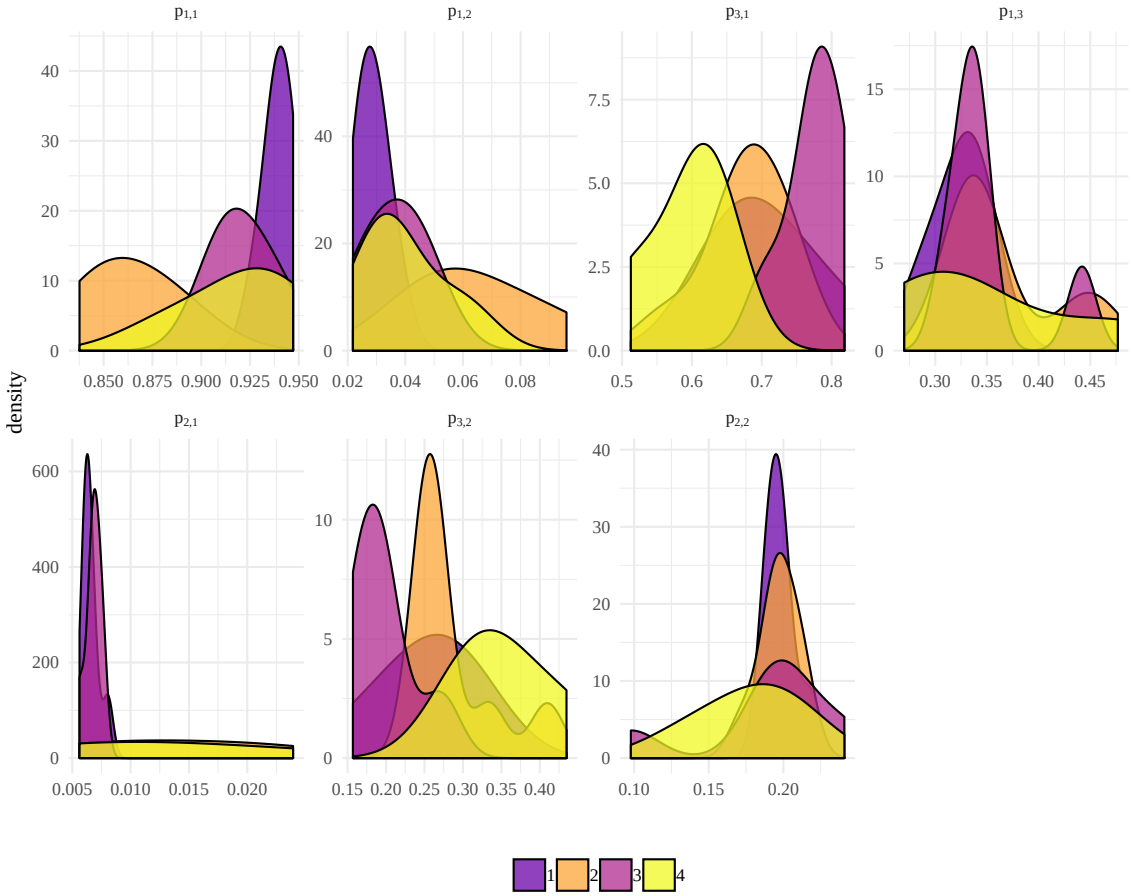

Figure 11: Cluster densities for Markov transition probabilities.

ability of remaining in state 1 is the lowest for this cluster while the probability of entering state 2 from state 1 is the highest. The acceleration state is visited only rarely. Four of the most similar performances are in this cluster, shown in Figure 13, along with Richter's 1976 recording.

Cluster three is somewhat like cluster one in that performers tend to stay in state 1 for long periods of time, but they transition more quickly from state 3 back to state 1. They also use state 4 frequently whereas cluster one did not. They also tend to have very large tempo contrasts between the A and B sections. Cluster four has both faster average tempos and more variability from one period of constant tempo to the next. State 4 is rare, with fast constant tempo changes that persist for small amounts of time tending to reflect note emphases.

Comparing our clusters to those we would find from simply clustering the distances between note-by-note tempo vectors reveals a number of differences. The four recordings in

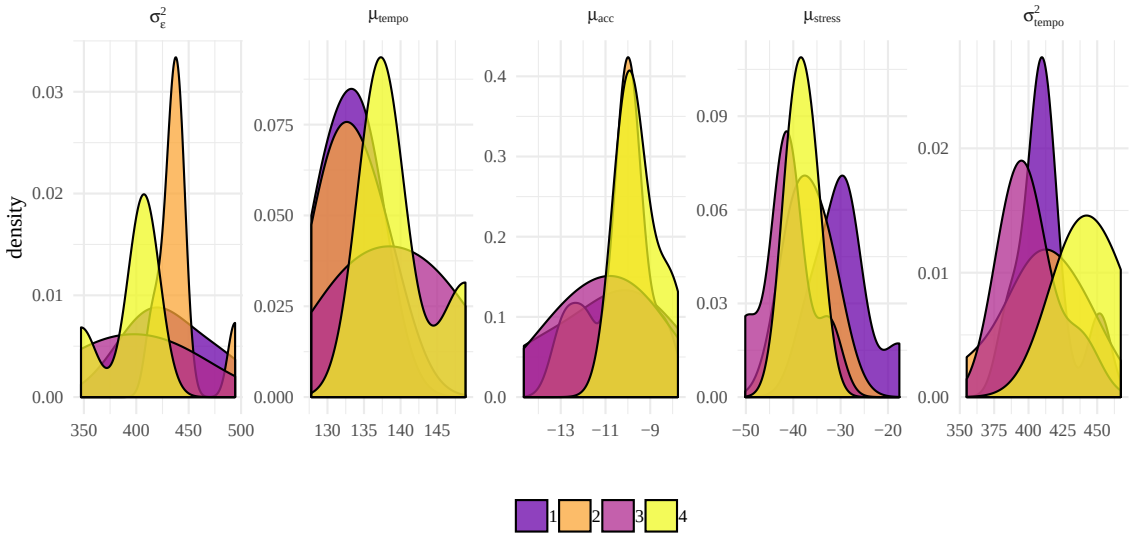

Figure 12: Cluster densities for mean and variance parameters.

Figure 13 would be spread across three different clusters, for example, as would our cluster one. On the other hand, both metrics see Cortot's recording as a strong outlier, and clustering by tempo vectors often (somewhat miraculously) groups recordings by the same pianist together: both recordings by Smidowicz, three of the four recordings by Rubinstein, both recordings by Hatto. Figure 15 shows all four Rubinstein recordings. The 1939 recording is rather odd in that the fourth 8 measures of the opening A section is so slow relative to the rest. The variability in the 1966 recording nearly obscures the contrast between the B section and the surrounding A sections. These two recordings are nonetheless clustered together by the tempo vectors. Our method on the other hand, puts both in the "other" grouping. The estimated parameters for these four performances are shown in the bottom half of Table 4. The top half shows the parameters for the four similar performances in Figure 13. There is much larger variability across Rubinstein's recordings, as we would expect.

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 smoothing splines (Craven and Wahba, 1978; Wahba, 1990) though total-variation denoising or trend filtering (Kim et al., 2009;

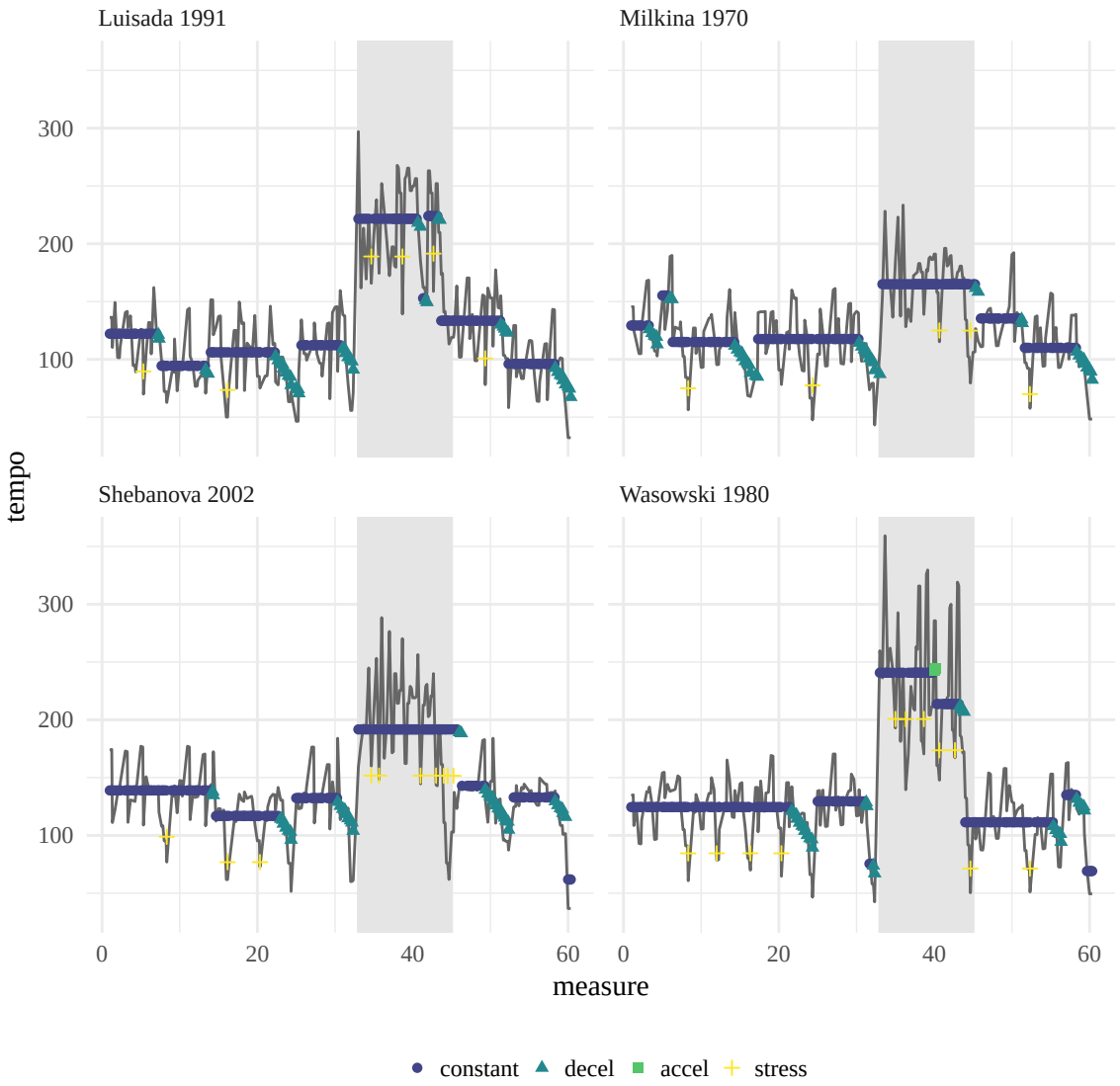

Figure 13: Four similar performances, all in the second cluster.

Tibshirani, 2014) are other reasonable alternatives. These statistical techniques perform smoothing by encouraging small changes in derivatives (smoothing splines) or bounded total variation (trend filtering). But musical performances do not conform to these assumptions because tempo and dynamic 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, as discussed above, 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

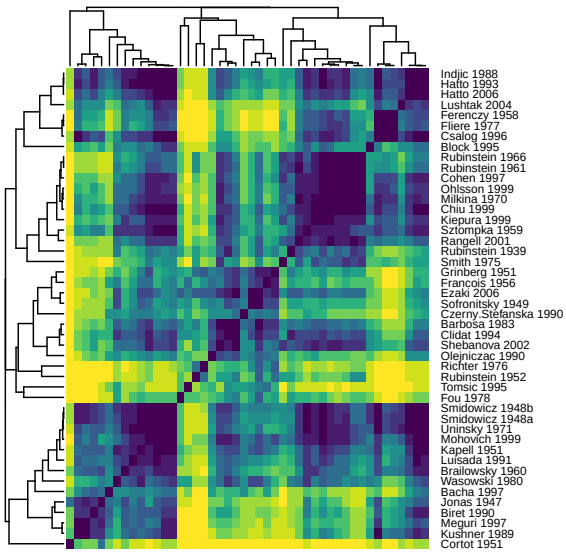

Figure 14: Distance matrix and dendrogram calculated using the note-by-note tempo vector for each recording.

statistical and empirical results.

Figure 16 shows the note-by-note tempo of Richter's 1976 recording. Splines with equally spaced knots are shown in yellow. We use generalized cross validation (Golub et al., 1979) to select the number of knots (one knot per measure). The red 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 endings were duplicated up to four times to allow for discontinuities. The blue line shows the estimated smooth tempo from our model (the same as in Figure 8), which automatically learns note emphases and discovers

Table 4: The estimated parameters for the four similar performances in Cluster Two 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 1980	414.99	132.00	-10.00	-40.00	425.00	0.85	0.05	0.67	0.34	0.02	0.26	0.20
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
Luisada 1991	494.33	127.80	-10.24	-32.56	411.63	0.84	0.10	0.71	0.35	0.01	0.26	0.19
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

phrases without purposeful knot duplication. 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, such as measures 1–16.

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.

There is nothing intrinsic to the model which forces states two, three, or four to always go in the correct direction. If for example, μ_{acc} is small in magnitude relative to σ^2_{acc}, an acceleration could be learned as time spent in state two but with large positive errors. For this piece, the penalties help to avoid such occurrences, but this aspect of the Gaussian state-space model could be improved by enforcing non-Gaussian behavior. Or 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 17 shows two performances where the estimated path failed to transition to a new constant tempo state 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 were clustered as "other". Both of these performances are quite volatile which makes estimation difficult.

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 an 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 had, string players have many more. Bowing decisions, fingerings, vibrato, broken chords are all important tools which are difficult to learn 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. On the other hand, focusing simply on tempo can be useful with solo performances or with larger ensembles. 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

R-package "dpf": R-package containing code to perform the methods described in the article. The package also contains all data sets used as examples in the article. (GNU zipped tar, also on Github)

Source code: Additional R code necessary to reproduce all analyses and graphics. (GNU zipped tar)

Appendix: Supplement with additional graphics for our clusters and analysis of all 46 recordings. (PDF)

References

Centre for the History and Analysis of Recorded Music, 2009. URL http://www.charm.rhul.ac.uk/about/about.html. Online; accessed 12 March 2019.

Brian D.O. Anderson and John B. Moore. *Optimal filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 72(2):1–33, 2010.

Christopher Ariza. 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*, 2005.

Andreas Arzt and Gerhard Widmer. Real-time music tracking using multiple performances as a reference. In *International Society for Music Information Retrieval (ISMIR)*, pages 357–363. Citeseer, 2015.

Leonard Bernstein. *Young People's Concerts*. Amadeus Press, Pompton Plains, NJ, 2005.

Roberto Bisiani. Beam search. In Stuart Shapiro, editor, *Encyclopedia of Artificial Intelligence*. John Wiley and Sons, 2nd edition, 1992.

Barbara A Block, Ian D Jonsen, Salvador J Jorgensen, Arliss J Winship, Scott A Shaffer, Steven J Bograd, Elliott Lee Hazen, David G Foley, GA Breed, A-L Harrison, et al. Tracking apex marine predator movements in a dynamic ocean. *Nature*, 475(7354):86, 2011.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

Nick Collins. A funny thing happened on the way to the formula: Algorithmic composition for musical theater. *Computer Music Journal*, 40(3):41–57, 2016.

Arshia Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987, 2010.

Arshia Cont, Diemo Schwarz, Norbert Schnell, and Christopher Raphael. Evaluation of real-time audio-to-score alignment. In *International Symposium on Music Information Retrieval (ISMIR)*, 2007.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

Roger Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the 1984 International Computer Music Conference*, pages 193–198. International Computer Music Association, 01 1985.

Roger B Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM*, 49(8):38–43, 2006.

Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The Yahoo! music dataset and KDD-Cup'11. In *KDD Cup*, pages 8–18, 2012.

J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods*. Oxford Univ Press, Oxford, 2001.

James Durbin and Siem Jan Koopman. Monte Carlo maximum likelihood extimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, September 1997.

Andrew Earis. An algorithm to extract expressive timing and dynamics from piano recordings. *Musicae Scientiae*, 11(2):155–182, 2007.

Andrew Earis. Mazurka in F Major, Op. 68, No. 3, 2009.

Paul Fearnhead and Peter Clifford. On-line inference for hidden markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4): 887–899, 2003.

Sebastian Flossmann, Maarten Grachten, and Gerhard Widmer. Expressive performance rendering with probabilistic models. In *Guide to Computing for Expressive Music Performance*, pages 75–98. Springer, 2013.

Sture Forsén, Harry B Gray, LK Olof Lindgren, and Shirley B Gray. Was something wrong with Beethoven's metronome? *Notices of the AMS*, 60(9), 2013.

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A): 1020–1056, 2011.

Cheng-Der Fuh. Efficient likelihood estimation in state space models. *Annals of Statistics*, 34(4):2026–2068, 2006.

Zoubin Ghahramani and Geoffrey E Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864, 2000.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Yupeng Gu and Christopher Raphael. Modeling piano interpretation using switching kalman filter. In *International Society for Music Information Retrieval (ISMIR)*, pages 145–150, 2012.

Gaëtan Hadjeres, François Pachet, and Frank Nielsen. DeepBach: a steerable model for Bach chorales generation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1362–1371, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

J.D. Hamilton. Calling recessions in real time. *International Journal of Forecasting*, 27: 1006–126, 2011.

Andrew C Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, 1990.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Chang-Jin Kim. Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1-2):1–22, 1994.

C.J. Kim and C.R. Nelson. 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(2):188–201, 1998.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

G. Kitagawa. Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, pages 1032–1041, 1987.

G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, pages 1–25, 1996.

Shinsuke Koyama, Lucia Castellanos Pérez-Bolde, Cosma Rohilla Shalizi, and Robert E. Kass. Approximate methods for state-space models. *Journal of the American Statistical Association*, 105(489):170–180, March 2010.

Dustin Lang and Nando D Freitas. Beat tracking the graphical model way. In *Advances in Neural Information Processing Systems*, pages 745–752, Cambridge, MA, 2005. MIT press.

Brian McFee and Gert Lanckriet. Learning multi-modal similarity. *Journal of Machine Learning Research*, 12:491–523, 2011.

Toby A Patterson, Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos. State–space models of individual animal movement. *Trends in ecology & evolution*, 23(2): 87–94, 2008.

C. Raphael. Music plus one and machine learning. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 21–28, Haifa, Israel, 2010.

Christopher Raphael. A hybrid graphical model for rhythmic parsing. *Artificial Intelligence*, 137(1):217–238, 2002.

Herbert E Rauch, CT Striebel, and F Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

Lu Ren, David Dunson, Scott Lindroth, and Lawrence Carin. Dynamic nonparametric Bayesian models for analysis of music. *Journal of the American Statistical Association*, 105:458–472, 2010.

Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4364–4373, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR.

Adam Roberts, Curtis Hawthorne, and Ian Simon. Magenta.js: A javascript api for augmenting creativity with deep learning. In *Joint Workshop on Machine Learning for Music (ICML)*, 2018b.

Markus Schedl, Emilia Gómez, Julián Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, 8 (2-3):127–261, 2014.

Bob L. Sturm, Oded Ben-Tal, Úna Monaghan, Nick Collins, Dorien Herremans, Elaine Chew, Gaëtan Hadjeres, Emmanuel Deruty, and François Pachet. Machine learning research that matters for music creation: A case study. *Journal of New Music Research*, 48(1):36–55, 2019.

John Thickstun, Zaid Harchaoui, and Sham M. Kakade. Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)*, 2017.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42:285–323, 2014.

Aaron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2643–2651. Curran Associates, Inc., 2013.

Barry Vercoe. The synthetic performer in the context of live performance. In *Proceedings of the 1984 International Computer Music Conference*, pages 199–200. International Computer Music Association, 1984.

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

Nick Whiteley, Christophe Andrieu, and Arnaud Doucet. Efficient bayesian inference for switching state-space models using discrete particle markov chain monte carlo methods. Technical Report 10:04, Bristol University, 2010.

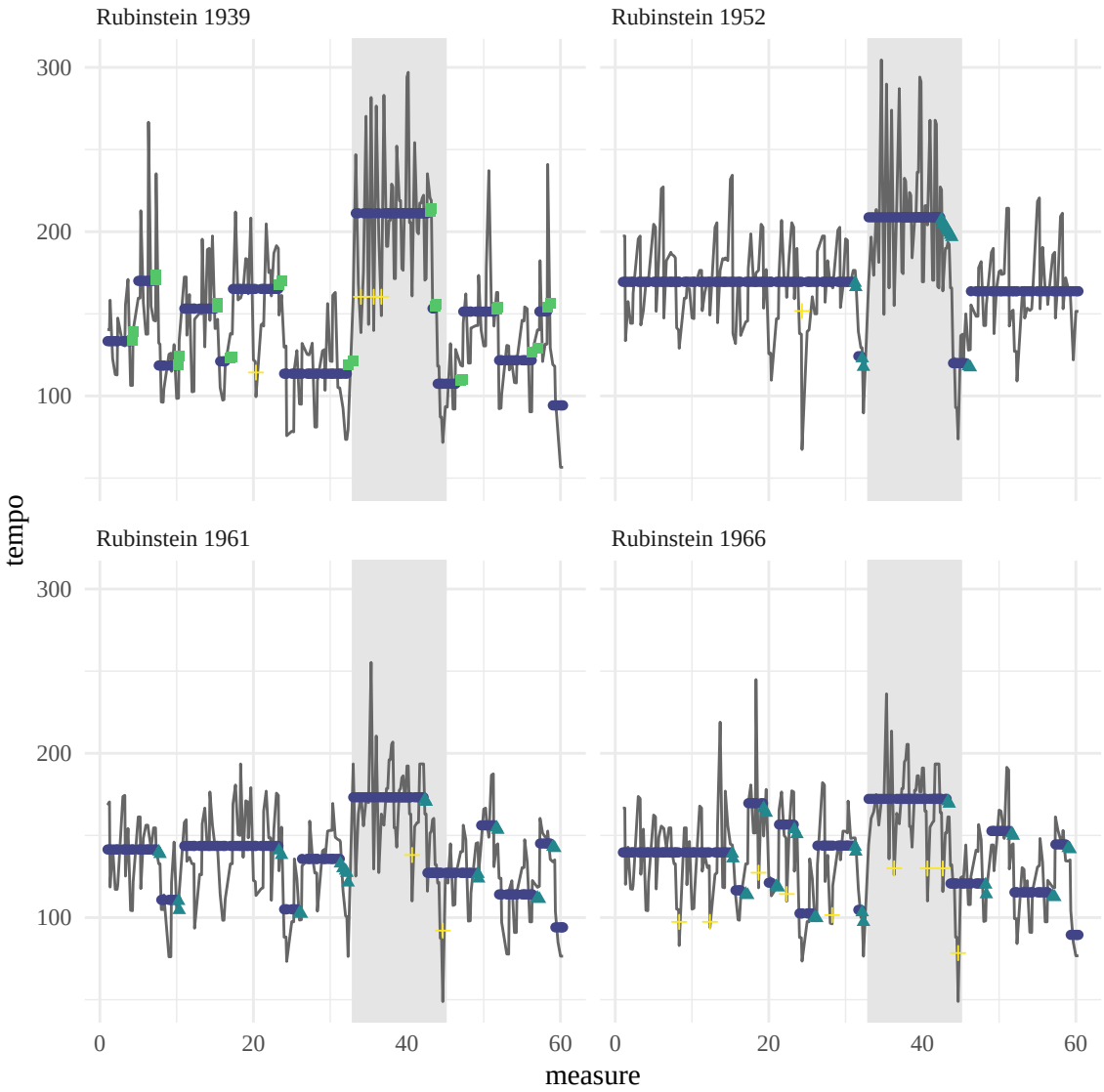

Figure 15: The four recordings by Arthur Rubinstein. Our clustering puts the 1952 and 1961 recordings in clusters one and two while leaving the others out. Clustering by tempo vector separates 1952 from the other three.

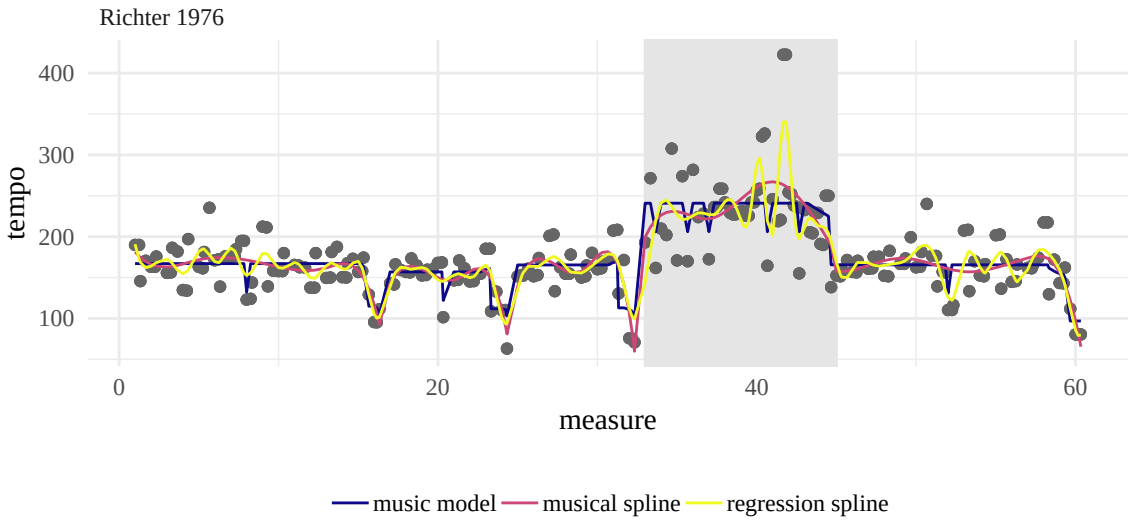

Figure 16: Smoothing with splines and musical models

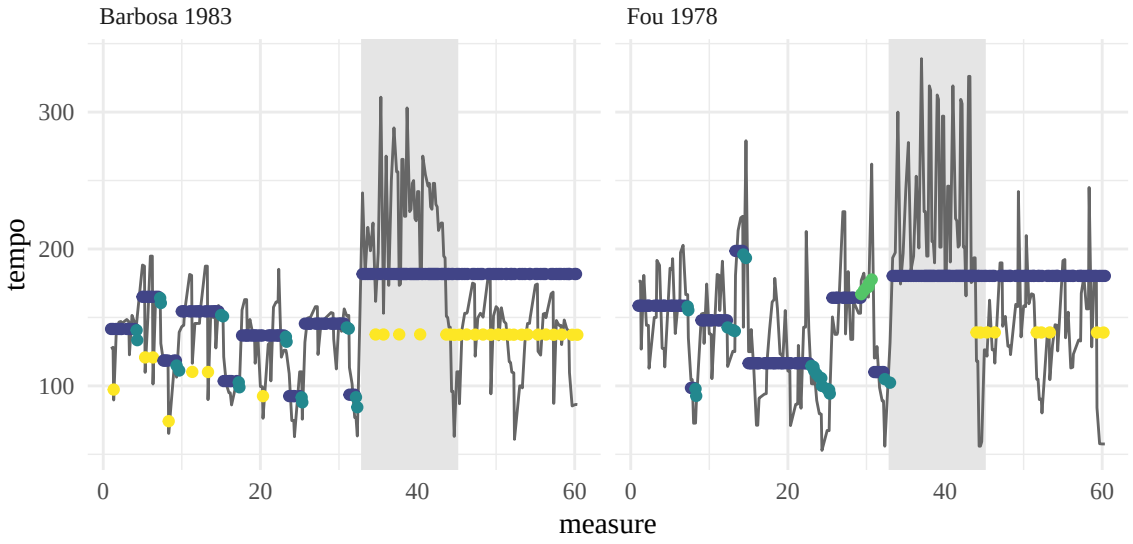

Figure 17: Estimation errors on two performances.