STA 642 Mini-Project: Music Generation using DLMs and DDNMs

Anna Yanchenko

4/28/17

1 Introduction

Time is a crucial component of any musical piece. Composers use time to set the tempo, rhythm and form of a piece and all of these factors greatly impact the experience of listeners of music. Musical pieces are inherently dynamic over time and there has been much work in analyzing music from a time series perspective. Beran and Mazzola (1999) used statistical methods to analyze the harmony, melody and rhythm of classical music. Beran (2004) used harmonic regression models with local stationarity, periodograms and spectral methods to analyze musical sound waves. Temperley (2007) explored generative models of rhythm using probabilistic graphical models that evolved over time, as well as Hidden Markov Models (HMMs) for polyphonic key-finding. Coviello et al. (2011) explicitly modeled temporal qualities of music through dynamic texture models for automatic annotation of music. Dirst and Weigend (1993) used various time series models to explore completing Bach's last, unfinished fugue and Boulanger-Lewandowski et al. (2012) utilized dynamic models for the automatic transcription of polyphonic music. Finally, Fukino et al. (2016) used recurrence plots to analyze the underlying dynamics of various musical pieces.

Recent work in musical modeling has focused on the algorithmic composition of music. Recurrent Neural Networks (RNNs) are the current state of the art in computer-generated music, for example Hadjeres and Pachet (2016), Developers (2017) and Magenta (2016). For my Masters Thesis, I explored the utility of HMMs in algorithmic composition of classical music. However, most pieces generated by either RNNs or HMMs suffer from a lack of global structure, resulting in the extension of RNNs to Long-Short Term Memory units to model music (Johnson (2015), Eck and Schmidhuber (2002)). The first goal of this mini-project was thus to explore the use of Dynamic Linear Models (DLMs) to capture longer-term structure in classical, single instrument pieces, leading to generated pieces of music with more global structure than those generated by HMMs or RNNs.

Additionally, while my thesis work had focused on generating pieces of music with only one instrument, I would like to move towards generating orchestral pieces of music with multiple instruments. In addition to the univariate challenges of generating harmonically consonant pieces with melodic structure, polyphonic (multi-voice) pieces have the additional modeling challenge of capturing the dynamic dependencies between different voices. In typical orchestral pieces with many different instruments, it is computationally prohibitive to model each voice as fully connected to all of the other voices. Thus, the second goal of this mini-project was to explore the utility of Dynamic Dependence Network Models (DDNMs) in modeling the dependencies between different voices in polyphonic music, allowing for a reduced parameter space and a parallel implementation.

2 Methods

2.1 Data

For this mini-project, I considered modeling the Baroque era piece Canon in D by Johann Pachelbel ("Pachelbel's Canon"). Pachelbel's Canon is a simple piece both melodically and harmonically, with chords built on the perfect fourth interval, which I found to result in largely consonant pieces in my thesis work. The piece was originally downloaded in Musical Instrument Digital Interface (MIDI) format from Krueger (2016), then converted to CSV using open source software (Walker (2008)). MIDI is a data communications protocol that allows for the exchange of information between software and musical equipment and symbolically represents the data needed to generate the musical sounds encoded in the data (Rothstein (1992)). For this mini-project, I only modeled the note pitches over time. In MIDI format, the note pitches were encoded as integers between 0 and 127, where 60 represented middle C and each integer change corresponded to a pitch change of a half-step. I assumed that the notes were evenly spaced in time, which was a slight approximation to the actual data.

I considered two different arrangements of Pachelbel's Canon, one for a single voice (and thus a univariate time series, Figure 1) and another arrangement for four voices; soprano, alto, tenor and bass (Figure 2). For the multi-voice arrangement, I imputed missing values when one voice would rest for an extended period of time with the last observed note pitch for this voice, meaning that instead of resting, this voice sustained the last observed pitch. While both the univariate and multivariate series were discrete, I assumed that the values were continuous. Additionally, the assumption of normality for both the univariate and multivariate series was a clear approximation, as all series tended to be quite heavy - tailed (Figure 3). Each series, especially the univariate series, was subject to non-normal innovations and sudden changes in volatility. The ACF and PACF plots for the univariate series (Figure 4) indicated structure out to several lags and the ACF and PACF plots for each voice in the multivariate series showed similar trends. Thus, the goal for the generated pieces was to observe structure in the ACF and PACF plots out to several lags, indicating melodic structure over several measures in the generated pieces.

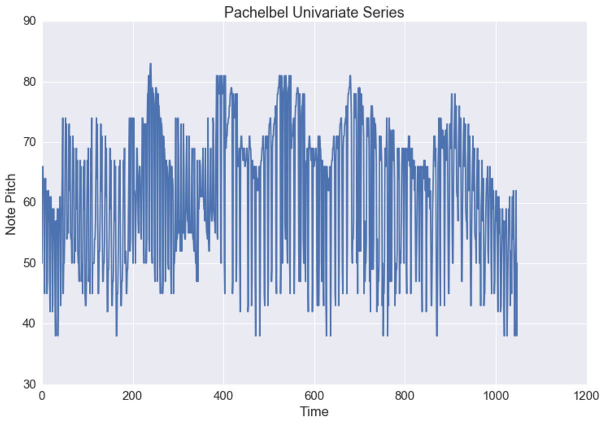

Figure 1: Univariate time series plot for Pachelbel's Canon.

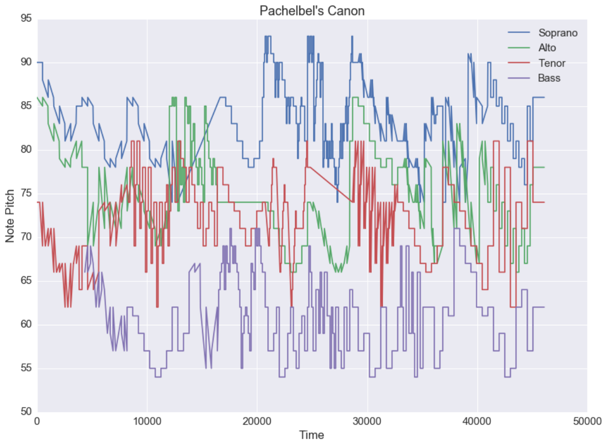

Figure 2: Multivariate time series plot for Pachelbel's Canon arranged for four voices.

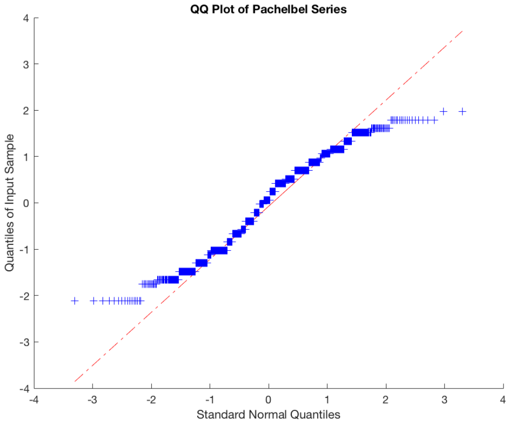

Figure 3: QQ plot for the univariate Pachelbel's Canon series.

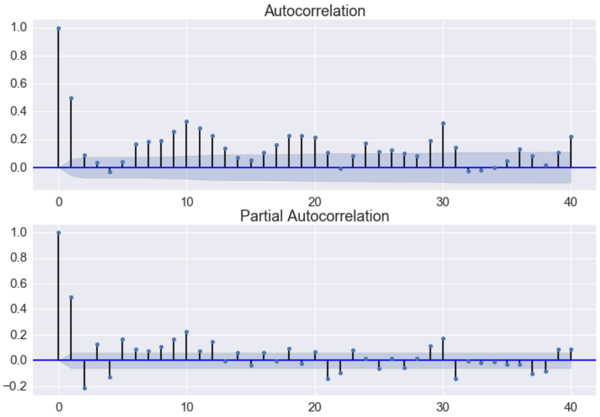

Figure 4: ACF and PACF plots for the univariate Pachelbel's Canon series.

2.2 Models

2.2.1 Dynamic Linear Models

I first considered various dynamic linear models (DLMs) to model the univariate series. Following (Prado and West (2010)), I considered DLMs with a linear trend term only (Model 1), a quadratic trend term only (Model 2), an AR(1) process with a linear trend term (Model 3), an AR(1) process with a quadratic trend term (Model 4), an AR(2) process with a linear trend term (Model 5), an AR(2) process with a quadratic trend term (Model 6), a linear trend term modeling the differenced series (Model 7), an AR(1) process with a linear trend term for the differenced series (Model 8) and finally an AR(2) process with a linear trend term for the differenced series (Model 9). From the ACF plot of the univariate series (Figure 4), it appeared that there was autoregressive structure out to at least lag-2 and the plot of the series over time (Figure 1) suggested a quadratic trend term or to difference the data.

The marginal likelihoods and cumulative model probabilities for each of these nine models are plotted in Figure 5. Model 5 (an AR(2) process with a linear trend term) and Model 6 (an AR(2) process with a quadratic trend term) alternatively had the highest cumulative model probabilities over the course of the piece. While the pieces generated by each of these nine models did have some structure over time, primarily seen in the ACF and PACF plots with structure out to a few lags, none of these DLMs were able to capture the high degree of volatility and non-stationarity in the series (partially due to the non-normal innovations) or to generate pieces that sounded distinguishable from each other, which suggested considering TVAR models next.

2.2.2 Time Varying Autoregression

From both a modeling and a musical perspective, the univariate series for Pachelbel's Canon was non-stationary and volatile over time. TVAR models (Prado and West (2010)) are able to model this non-stationarity and are also able to handle the periodicity seen in the ACF plot of the original piece. By running a grid-search over several different parameter ranges, a TVAR(11) model with discount factors $\delta = 0.970$ and $\beta = 0.905$ maximized the marginal likelihood. Plots of the

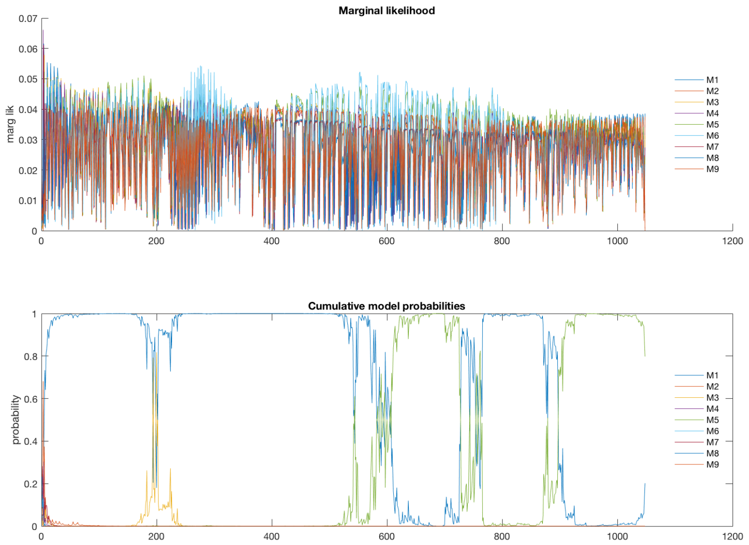

Figure 5: Marginal likelihoods and cumulative model probabilities for nine different DLM models: a linear trend term only (M1), a quadratic trend term only (M2), an AR(1) process with a linear trend term (M3), an AR(1) process with a quadratic trend term (M4), an AR(2) process with a linear trend term (M5), an AR(2) process with a quadratic trend term (M6), a linear trend term modeling the differenced series (M7), an AR(1) process with a linear trend term for the differenced series (M8) and finally an AR(2) process with a linear trend term for the differenced series (M9).

TVAR coefficients over time (Figure 6) and the decomposition of the series into latent components (Figure 7) suggested several different processes in the original piece that evolved in importance and magnitude over the course of the piece.

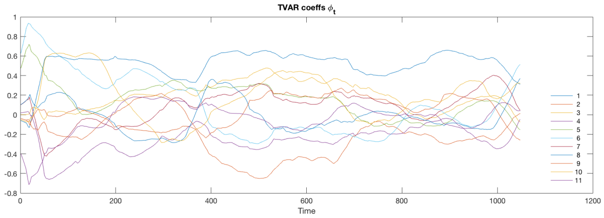

Figure 6: Plots of the eleven TVAR coefficients over time.

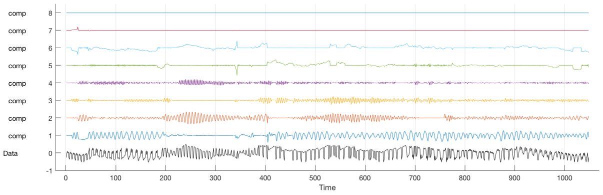

Figure 7: Decomposition of the univariate series into latent components for the TVAR(11) model.

2.2.3 Dynamic Dependence Network Models

Dynamic Dependence Network Models enable the modeling of multivariate time series where contemporaneous values of some of the series explicitly appear as predictors for other series (Zhao et al. (2016)). DDNMs allow for the specification of parent sets for the different univariate series and forward filtering and backwards sampling can be used to estimate dynamic regression coefficients between values of the different series at the same point in time. An important part of model specification is the determination of the parent set for each univariate series.

From a music theory perspective and in general counterpoint practice from the Baroque era, the bass voice in polyphonic music was the foundation on which the higher voices were composed. As a result, I considered a DDNM where the alto, tenor and bass voices were parents of the soprano voice, the tenor and bass voices were parents of the alto voice, the bass voice was a parent of the tenor voice and the bass line had no parents. The model structure attempted to mirror the dependence between voices and the order of composition of the actual composer.

Based on a TVAR analysis of each voice individually, each marginal likelihood was maximized when the TVAR model included only a lag 1 term. Additionally, when several DDNMs with the same parental structure as above but differing in the lags of the predictors included were analyzed in terms of the eight-step-ahead MSE of the predicted values for the last few measures of the original piece, the model with the lowest MSE was a DDNM which included the lag 1 term of each other term for every voice (that is, the note pitch for the soprano, alto, tenor and bass voices at time $t-1$ was a predictor for each of the voices at time t). Then, using the notation of (Zhao et al. (2016)), the DDNM could be written as:

$$\begin{aligned}
y_{j,t} &= \boldsymbol{x}_{j,t}\boldsymbol{\phi}_{j,t} + \boldsymbol{y'}_{pa(j),t}\boldsymbol{\gamma}_{j,t} + \nu_{j,t}, \quad j = 1:4 \\
y_{1,t} &= [y_{1,t-1}, y_{2,t-1}, y_{3,t-1}, y_{4,t-1}]\boldsymbol{\phi}_{1,t} + [y_{2,t}, y_{3,t}, y_{4,t}]\boldsymbol{\gamma}_{1,t} + \nu_{1,t} \\
y_{2,t} &= [y_{1,t-1}, y_{2,t-1}, y_{3,t-1}, y_{4,t-1}]\boldsymbol{\phi}_{2,t} + [y_{3,t}, y_{4,t}]\boldsymbol{\gamma}_{2,t} + \nu_{2,t} \\
y_{3,t} &= [y_{1,t-1}, y_{2,t-1}, y_{3,t-1}, y_{4,t-1}]\boldsymbol{\phi}_{3,t} + [y_{4,t}]\boldsymbol{\gamma}_{3,t} + \nu_{3,t} \\
y_{4,t} &= [y_{1,t-1}, y_{2,t-1}, y_{3,t-1}, y_{4,t-1}]\boldsymbol{\phi}_{4,t} + \nu_{4,t}
\end{aligned}$$

where 1 = soprano, 2 = alto, 3 = tenor and 4 = bass.

2.2.4 Backwards Sampling

For all of the models considered above, in addition to examining and interpreting the various model parameters, the primary interest of this mini-project was how well each model could generate new pieces of music, in particular, looking at how much long-term structure was evident in the generated pieces. To this end, after using forward filtering to fit each of the models considered above, I ran backwards sampling to obtain samples from the full posterior to generate new pieces. Since the note pitch values were discrete and only certain note pitches occured in the original piece, I then binned the generated note pitches to the nearest pitch that occurred in the original piece. Without this binning, the generated pieces were too chromatic. Since the data was discrete but all of the models assumed continuous data, the generated pieces tended to move in integer steps quite a lot, resulting in very chromatic pieces that did not match the nature of the original piece, since each integer corresponded to a half-step in the MIDI representation of the data. As a result, I binned the note pitches in the generated pieces so that the generated pieces were more representative of the style and tonality of the original Pachelbel's Canon.

3 Results

3.1 TVAR(11) Model

The TVAR(11) model was fairly successful at generating pieces with longer term structure, as seen in the ACF and PACF plots for a single piece generated using backwards sampling of the TVAR(11) model (Figure 8). The ACF and PACF structure was reminiscent of the ACF and PACF structure

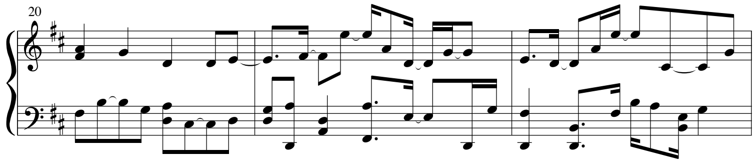

Figure 10: An example of a repeated melodic phrase in measures 21 and 22 of a TVAR(11) generated piece.

in the original piece (Figure 4), and in listening to the piece, there was more evidence of melodic progression than in the pieces generated by the DLMs. For example, the repetition of the same idea of an ascending line beginning with a dotted eighth note followed by an eighth note tied to a sixteenth note in measures 21 and 22 (Figure 10) indicated repetition of a measure of a short portion of melody. While the generated piece still had less global structure than the original Pachelbel's Canon, the TVAR(11) model was able to generate pieces that had melodic structure over a few measures at least, which was not the case with previously considered HMM generated pieces or the DLMs considered above. The generated piece was largely consonant as well, though the generated note pitches were binned to help ensure this. Furthermore, the time series of the generated pieces appeared very similar to the series of the original piece (Figure 9). The piece generated by the TVAR(11) is included as `TVAR11.mp3`.

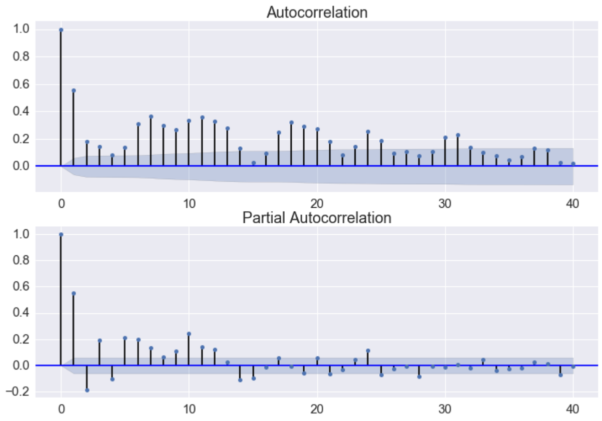

Figure 8: ACF and PACF plots of a piece generated by the TVAR(11) model.

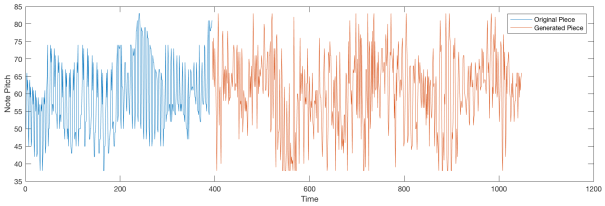

Figure 9: Time series plot of a portion of the original piece and a portion of a generated piece.

3.2 DDNM

After performing forward filtering for the DDNM described above, 2000 Monte Carlo samples were drawn at each time point t to obtain posterior estimates of the parameters $\boldsymbol{\phi}_{j,t}$ and $\boldsymbol{\gamma}_{j,t}$. The dynamic regression coefficients $\boldsymbol{\gamma}_{j,t}$ were of primary interest for this mini-project and all of the $\boldsymbol{\gamma}_{j,t}$ estimates for each series were dynamic over time. The DDNM parameter estimates were able to capture fundamental "regime shifts" in the original piece. For example, in the parameter estimates of $\gamma_{soprano,alto}$ over time, there was a clear shift in the parameter estimates at time $t = 175$, where

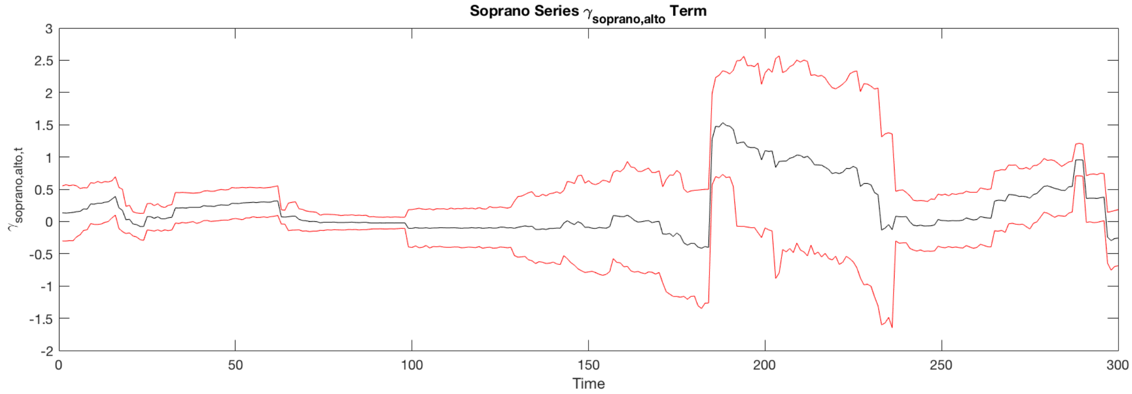

Figure 11: Plot of the $\gamma_{soprano,alto}$ dynamic regression coefficient over time. The black line is the Monte Carlo average of the parameter values and the red lines represent 90% credible intervals.

t refers to the number of the note pitch in the piece (Figure 11). At this point in the piece, the soprano voice transitioned to sixteenth notes, represented as many note pitch changes in the data, while the alto voice transitioned from half notes to quarter notes. The increase in moving notes was represented as a sharp change in the parameter estimate for the dynamic regression coefficient at this point in time, while the parameter estimate for $\gamma_{soprano,alto}$ was approximately 0 at other points in time. Similar trends were observed in the parameter estimates of $\gamma_{soprano,tenor}$.

The parameter estimates for $\gamma_{soprano,bass}$ were much more dynamic over time (Figure 12) and were almost always non-zero, representing the fact that apart from “regime changes” in melody between the upper voices, the bass voice was the most important in representing dynamic dependencies. Again, in the parameter estimates for $\gamma_{soprano,bass}$, the sharp spikes and changes in the parameter values corresponded to fundamental melodic changes in the original piece. For example, at time $t \approx 70$, there was a sign change in the parameter estimate, corresponding to when the soprano voice dropped out for several measures (the data values were imputed at this point) and there was another sign change at $t \approx 100$ when the soprano voice came back in and the data values were no longer imputed.

The dynamic regression coefficient estimates over time for $\gamma_{alto,bass}$ and $\gamma_{tenor,bass}$ were similar to those for $\gamma_{soprano,bass}$ and were more dynamic and non-zero over time than the estimates for $\gamma_{alto,tenor}$ and $\gamma_{soprano,alto}$. Thus, as expected from musical theory, the bass voice was the most important to explicitly include as a parent for the upper voices and the upper voices were very dependent on the bass voice over time. However, the parameter estimates also suggested that apart from “regime” melodic changes that occured between upper voices (which are often known a priori), the dynamic regression coefficients for contemporaneous values in the upper voices were often approximately zero. Thus, in future work, contemporaneous dependencies between the bass line and upper voices are very important to model, while dependencies between upper voices could perhaps be ignored, except for melodic shifts in the music.

The DDNM was also successful at generating pieces with some longer-term structure; for example, the ACF plot for the generated bass line (Figure 13) showed autoregressive structure out to high lags and the ACF plots for the other voices were very similar. More importantly, the DDNM was able to model the dependencies between the voices successfully. The generated piece was largely

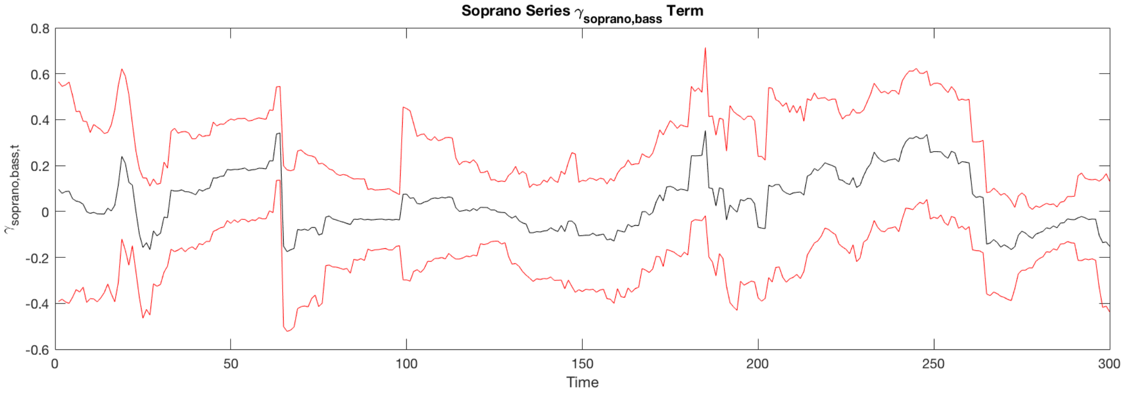

Figure 12: Plot of the $\gamma_{soprano,bass}$ dynamic regression coefficient over time. The black line is the Monte Carlo average of the parameter values and the red lines represent 90% credible intervals.

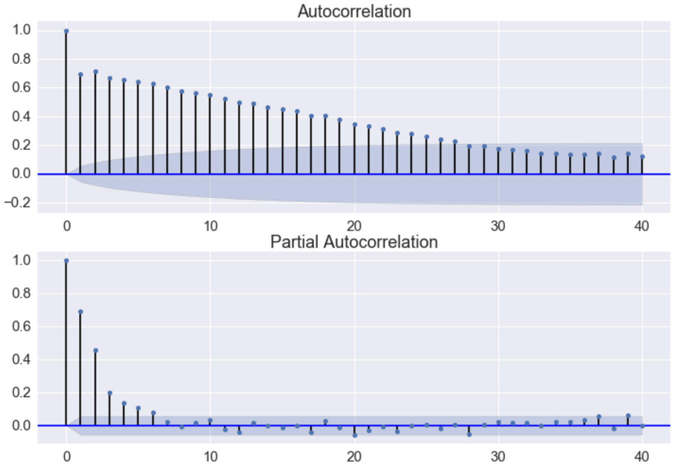

Figure 13: ACF and PACF plots of the bass line for a piece generated by the DDNM model.

consonant between the four voices, with minimal dissonance occurring between voices. Additionally, as in the original Pachelbel's Canon arranged for four voices, the moving notes tended to occur in only one voice at a time, so that when one voice had the melody, the other voices had a supporting role. Thus, the generated piece did not sound overly chaotic with all voices having moving notes at the same time. Additionally, the moving notes and melody transitioned between voices throughout the generated piece, just as in the original piece. While there were points in the generated piece where one voice was very repetitive, this was an artifact of the imputed values at certain points in the original piece for each voice. Overall, the DDNM was able to capture the dependencies between each voice well, resulting in a generated piece that was largely consonant between all four voices and was not overly chaotic, with each voice playing an appropriate role throughout the course of the piece. The generated piece is attached as `DDNM.mp3`. Please note that in the generated piece, the same instrument (a piano) is playing each voice, but there are four separate voices that were modeled and generated explicitly.

4 Conclusions and Future Work

Overall, the TVAR(11) model and the DDNM were successful at capturing longer term structure in Pachelbel's Canon, resulting in generated pieces that had melodic structure over at least a few measures, which was not the case in previous work using HMMs. The DDNM in particular was quite successful at capturing the dependencies between different voices in Pachelbel's Canon arranged for four voices, with the generated pieces largely consonant between voices and exhibiting a clear "trade-off" of the melodic line between voices at different points in the piece. While none of the generated pieces had as much melodic progression as the original piece, there was progress in structure occurring over a longer period of time. Furthermore, while I did not focus on evaluating other qualities of the generated pieces beyond the global structure, the generated pieces tended to be largely consonant and did not sound as similar to the original training piece as some of the previous HMM generated pieces.

The main modeling challenge that occurred in this mini-project was the fact that the data I modeled was not continuous, was heavy-tailed and exhibited non-normal innovations. In the future, considering models that either explicitly take into account the discrete nature of the data, or allow for non-normal innovations, such as innovations that follow a t-distribution with a low degree of freedom, might improve the ability of the models to generate pieces with even longer-term structure. Additionally, including an symbol for rests in a discrete modeling context would help resolve the issue of repetitive generated notes as a result of imputed values in some of the voices. This will be particularly important when more complex orchestral pieces are considered, where certain instruments rest for long stretches of time instead of just a few measures as in Pachelbel's Canon.

Additionally, Pachelbel's Canon is a very simple piece from a melodic and harmonic standpoint and it was straightforward to determine the appropriate parent sets in the DDNM based on musical knowledge alone. In more complicated orchestral pieces with many different instruments, it is not as straightforward as to what the parent sets for each instrument should be from a musical standpoint. Dynamic PCA will likely be helpful in determining parent sets for more complicated orchestral music. It will be interesting to explore in the future how well DDNMs are able to model more complex pieces. Also, I plan to consider other single instrument pieces that are more complex melodically and harmonically to explore how well TVAR models apply to this setting. In particular, the HMMs that I explored in my thesis work were unable to generate consonant pieces when trained on an original piece based on major or minor third chords, so it will be interesting to explore if TVAR models offer any improvement in the harmonic, as well as melodic, modeling of pieces. Finally, these generated pieces need to be evaluated by several listeners to determine if there has been improvement in longer-term melodic structure from a listening perspective, as compared to my original HMM thesis work.

References

Beran, J. (2004), *Statsitics in Musicology*, Chapman & Hall/CRC.

Beran, J. and Mazzola, G. (1999), "Analyzing Musical Structure and Performance - A Statistical Approach," *Statistical Science*, 14, 47–79.

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 (ICML-12)*, pp. 1159–1166.

Coviello, E., Chan, A. B., and Lanckriet, G. (2011), "Time Series Models for Semantic Music Annotation," *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 1343–1359.

Developers, G. (2017), "Magenta: Music and Art Generation (TensorFlow Dev Summit 2017)," `https://www.youtube.com/watch?v=vM5NaGoynjE`.

Dirst, M. and Weigend, A. S. (1993), *Time Series Prediction: Forecasting the Future and Understanding the Past*, chap. Baroque Forecasting: On Completing J. S. Bach's Last Fugue, Addison-Wesley.

Eck, D. and Schmidhuber, J. (2002), "A First Look at Music Composition using LSTM Recurrent Neural Networks," Tech. rep., Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale.

Fukino, M., Hirata, Y., and Aihara, K. (2016), "Music Visualized by Nonlinear Time Series Analysis," `https://sinews.siam.org/Details-Page/music-visualized-by-nonlinear-time-series-analysis`.

Hadjeres, G. and Pachet, F. (2016), "DeepBach: a Steerable Model for Bach chorales generation," *CoRR*, abs/1612.01010.

Johnson, D. (2015), "Composing Music with Recurrent Neural Networks," `http://www.hexahedria.com/2015/08/03/composing-music-with-recurrent-neural-networks/`.

Krueger, B. (2016), "Classical Piano MIDI Page," `http://www.piano-midi.de/midi_files.htm`.

Magenta (2016), "Magenta," `https://magenta.tensorflow.org/welcome-to-magenta`.

Prado, R. and West, M. (2010), *Time Series: Modeling, Computation and Inference*, Chapman & Hall/CRC.

Rothstein, J. (1992), *MIDI: A Comprehensive Introduction*, The Computer Music and Digital Audio Series, A-R Editions, Inc.

Temperley, D. (2007), *Music and Probability*, The MIT Press.

Walker, J. (2008), "MIDI-CSV," `http://www.fourmilab.ch/webtools/midicsv/#midicsv.5`.

Zhao, Z. Y., Xie, M., and West, M. (2016), "Dynamic Dependence Networks: Financial Time Series Forecasting and Portfolio Decisions," *Applied Stochastic Models in Business and Industry.*