(A) Data in the Life: Authorship Attribution of Lennon-McCartney Songs

Mark E. Glickman*
Department of Statistics
Harvard University
glickman@fas.harvard.edu

Jason I. Brown
Department of Mathematics and Statistics
Dalhousie University
jason.brown@dal.ca

Ryan B. Song
School of Engineering and Applied Sciences
Harvard University
ryan.b.song@gmail.com

Abstract

The songwriting duo of John Lennon and Paul McCartney, the two founding members of the Beatles, composed some of the most popular and memorable songs of the last century. Despite having authored songs under the joint credit agreement of Lennon-McCartney, it is well-documented that most of their songs or portions of songs were primarily written by exactly one of the two. Furthermore, the authorship of some Lennon-McCartney songs is in dispute, with the recollections of authorship based on previous interviews with Lennon and McCartney in conflict. For Lennon-McCartney songs of known and unknown authorship written and recorded over the period 1962-66, we extracted musical features from each song or song portion. These features consist of the occurrence of melodic notes, chords, melodic note pairs, chord change pairs, and four-note melody contours. We developed a prediction model based on variable

*Address for correspondence: Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA. E-mail address: glickman@fas.harvard.edu. Jason Brown is supported by NSERC grant RGPIN 170450-2013. The authors would like to thank Xiao-Li Meng, David C. Hoaglin, the Co-Editor who oversaw our peer review, and the three anonymous referees for their helpful comments.

screening followed by logistic regression with elastic net regularization. Out-of-sample classification accuracy for songs with known authorship was 76%, with a c-statistic from an ROC analysis of 83.7%. We applied our model to the prediction of songs and song portions with unknown or disputed authorship.

Key words: authorship, elastic net, logistic regression, music, regularization, stylometry, variable screening

1 Introduction

The Beatles are arguably one of the most influential music groups of all time, having sold over 600 million albums worldwide. Beyond the initial mania that accompanied their introduction to the UK and Europe in 1962-63, and subsequently to the United States in early 1964, the Beatles' musical and cultural impact still has lasting influence. The group has been the focus of academic research to an extent that rivals most classical composers. Heuger (2018) has been maintaining a bibliography that contains over 500 entries devoted to academic research on the Beatles. Some recent examples of scientific study of Beatles music include Cathé (2016) who applied harmonic vectors theory to Beatles songs, Wagner (2003) who analyzed the presence of blues motifs in Beatles music, and Brown (2004) who used Fourier analysis to determine the true arrangement and instrumentation of the opening chord of "A Hard Day's Night."

The songwriting duo of John Lennon and Paul McCartney took the writing credits for most recorded Beatles songs. The two agreed prior to the Beatles' formation that all songs written by the two of them, either together or individually, would be credited to the partnership "Lennon-McCartney." After the Beatles broke up in 1970 and the Lennon-McCartney partnership dissolved, Lennon and McCartney attempted to clarify their contributions to their jointly credited songs. Most often, individual songs were acknowledged to be written entirely by either McCartney or Lennon, though in some cases one would write most of a song and

the other would contribute small portions or sections of the song. Compton (1988) provided a fairly complete accounting of the actual authorships of Lennon-McCartney songs to the extent they are known through interviews with each of Lennon and McCartney. According to this listing, several songs are of disputed musical authorship. Some examples include the entire songs "Ask Me Why," "Do You Want to Know a Secret?," "Wait," and "In My Life." Womack (2007) provided an interesting account of the discrepancy in Lennon and McCartney's recollection of the authorship of "In My Life" in particular: Lennon wrote the lyrics, McCartney asserted that he wrote all of the music, and Lennon claimed that McCartney's only contribution was helping with the middle eight melody. Given that further direct questioning about these songs is unlikely to reveal their true author, it is an open question whether musical features of Lennon-McCartney songs could provide statistical evidence of song authorship between Lennon versus McCartney.

The idea of using statistical models to predict authorship is one that has been around for over half a century. In one of the first successful attempts at modeling word frequencies, Mosteller and Wallace (1963, 1984) used Bayesian classification models to infer that James Madison wrote all of the 12 disputed Federalist papers. Other recent works related to authorship attribution include Efron and Thisted (1976) and Thisted and Efron (1987), who address questions related to Shakespeare's writing, and Airoldi, Anderson, Fienberg, and Skinner (2006), who examine authorship attribution of Ronald Reagan's radio addresses. Typical text analysis relies on constructing word histograms, and then modeling authorship as a function of word frequencies. Basic background on the analysis and modeling of word frequencies can be found in Manning and Schütze (1999), and these models applied to text authorship attribution can be found in Clement and Sharp (2003) and Malyutov (2005).

This paper is concerned with using harmonic and melodic information from the corpus of Lennon-McCartney songs from the first part of the Beatles' career to infer authorship of songs by John Lennon and Paul McCartney. It is not unreasonable to assume that

Lennon and McCartney songs are distinguishable through musical features. For example, both McCormick (1998) and Hartzog (2016) observed that Lennon songs have melodies that tend not to vary substantially in pitch (illustrative examples include "I Am the Walrus" and "Across the Universe"), whereas McCartney songs tend to have melodies with larger pitch changes (e.g., "Hey Jude" and "Oh Darling"). However, such anecdotal observations may not sufficiently characterize distinctions between Lennon and McCartney – a more scientific approach is necessary. Our analyses attempt to capture distinguishing musical features through a statistical approach.

Previous work applying quantitative methods to distinguish Lennon and McCartney songs is limited. Whissell (1996) performed a stylometric analysis of Beatles songs based on lyric content via text analyses to characterize the emotional differences between Lennon and McCartney over time. An unpublished paper by McDougal (2013) performed a traditional text analysis using word count methods to compare Lennon and McCartney's lyric usage, and supplemented the text analysis with auditory-derived information from the program The Echo Nest (`the.echonest.com`). More generally, a variety of statistical methods for inferring authorship from musical information have been published. Cilibrasi, Vitányi, and De Wolf (2004) and Naccache, Borgi, and Ghédira (2008) used Musical Instrument Digital Interface (MIDI) encoding of songs, which contains information on the pitch values, intervals, note durations, and instruments to perform distance-based clustering. Dubnov, Assayag, Lartillot, and Bejerano (2003) developed methods to segment music using incremental parsing applied to MIDI files in order to learn stylistic aspects of music representation. Conklin (2006) also introduced representing melody as a sequence of segments, and modeled musical style through this representation. A different approach was taken by George and Shamir (2014), who converted song data into two-dimensional spectrograms, and used these representations as a means to cluster songs.

Our approach to musical authorship attribution is most closely related to methods applied to

genome expression studies and other areas in which the number of predictors is considerably larger than the sample size. In a musical context, we reduce each song to a vector of binary variables indicating the occurrences of specified local musical features. We derive the features based on the entire set of chords that can be played (harmonic content) and the entire set of notes that can be sung by the lead singer (melodic content). From the point of view of melodic sequences of notes or harmonic sequences of chords behaving like text in a document, individual notes and individual chords can be understood as 1-gram representations. The occurrences of individual chords and individual notes form an essential part of a reduction in a song's musical content. To increase the richness of the representation, we also consider 2-gram representations of chord and melodic sequences. That is, we record the occurrence of pairs of consecutive notes and pairs of consecutive chords as individual binary variables. Rather than considering larger n-gram sequences (with $n > 2$) as a unit of analysis, we extract local contour information of melodic sequences indicating the local shape of the melody line to be a fifth set of variables to represent local features within a song. Using occurrences of pitches in the sung melodies, chords, pitch transitions, harmonic transitions, and contour information of Lennon-McCartney songs with known authorship permits modeling of song authorship as a function of musical content.

We developed our modeling approach as a two-step algorithm. First, we kept only musical features that had a sufficiently strong bivariate association with authorship, an application of sure independence screening (Fan, 2007; Fan & Lv, 2008). With the features that remained, we then modeled the authorship attribution as a logistic regression, but estimated the model parameters using elastic net regularization (Friedman, Hastie, & Tibshirani, 2010; Zou & Hastie, 2005), an approach that flexibly constrains the average log-likelihood by a convex combination of a ridge penalty (Le Cessie & Van Houwelingen, 1992) and a lasso penalty (Tibshirani, 1996, 2011). Many other approaches to regularization are possible. For example, Kempfert and Wong (2018), who predicted the authorship of Hadyn versus Mozart string quartets based on musical features, selected their model through subset selection on the

Bayesian information criterion (BIC) statistic.

This paper proceeds as follows. We describe the background of the song data collection and formation in Section 2. This is followed in Section 3 by the development of a model for authorship attribution based on a variable screening procedure followed by elastic net logistic regression. The application of the modeling approach is described in Section 4 where we summarize the fit of the model to the corpus of Lennon-McCartney songs of known authorship, and apply the model results for predicting songs of disputed authorship. The paper concludes in Section 5 with a discussion of the utility of our approach to wider musical settings. We provide relevant background on musical notes, scales, note intervals, chords, and song structure in Appendix A.

2 Song Data

The data used in our analyses consist of melodic and harmonic information based on Lennon-McCartney songs that were written between 1962 and 1966. This period of Beatles music is during the years they toured and occurred before the band's activities centered on studio productions when their songwriting approach likely changed significantly. The songs we included in our analyses were from the original UK-released albums *Please Please Me, With the Beatles, A Hard Day's Night, Beatles for Sale, Help!, Rubber Soul,* and *Revolver*, as well as all the singles from the same era that were not present in any of these albums. The essential reference for both the melodic and harmonic content of the songs was Fujita, Hagino, Kubo, and Sato (1993), although the Isophonics online database of chords for The Beatles songs (`http://isophonics.net/content/reference-annotations`) provided additional points of reference for each song.

The authorship of each Lennon-McCartney song, or whether the authorship credit was in

dispute, has been documented in Compton (1988), though for some songs we have found other documentation of song authorship. Aside from recording whether entire songs were written by Lennon versus McCartney, Compton also notes that in many cases songs had multiple sections with possibly different authors. For example, the song "We Can Work It Out" is credited to McCartney as the author, though the bridge section starting with the lyric "Life is very short..." was written by Lennon. In our analyses, we treat these sections as two different units of analysis with different authors. Furthermore, several songs that were acknowledged as full collaborations between Lennon and McCartney were excluded from the corpus of known authorship from which we develop our prediction algorithm. The song "The Word" is such an example of a full collaboration. It is plausible that some of the disputed songs were actually collaborations, but the current information about the songs did not permit these joint attributions. The total number of Lennon-McCartney songs or portions of songs with an undisputed individual author (Lennon or McCartney) was 70. Eight songs or portions of songs in this period were of disputed authorship.

Our process was to manually code each song's harmonic (chord) and melodic progressions. The song content that serves as the input to our modeling strategy is a set of representations of simple melodic and harmonic patterns within each song in the form of category indicators. That is, we let each song be represented by a vector of binary variables within the song, where each variable is the presence/absence of a musical feature that could occur in the song. We describe these representations in more detail below. The process to obtain these category indicators involved converting each song's melodic and harmonic content into a usable form. Melody lines were partitioned into phrases which were typically book-ended by rests (silence). An alternative approach would have been to model counts of musical features within songs, which is much more in line with authorship attribution analysis for text documents. A crucial difficulty with this approach is how to address repeated phrases (e.g., verses, choruses) within a song. As an extreme example that is not part of our sample, consider the later-Beatles period McCartney-written song "Hey Jude." The "na na na"

fadeout, which lasts roughly four minutes on the recording, is repeated 19 times (Everett, 1999). Keeping these repeated occurrences would likely over-represent the musical ideas suggested by the phrase. We explored models in which feature counts were incorporated, including versions where the counts were capped at an upper limit (i.e., winsorizing the larger counts), and versions involving the transformation of counts to the log scale, but these approaches resulted in worse predictability than our final model. The use of whether a musical feature was present in a song produced better discriminatory power in authorship predictions.

The key of every song was standardized relative to the tonic for songs in a major key, and to the relative major (up a minor third) for songs in a minor key. If a key change occurred in the middle of the song, the harmonic and melodic information from that point onward would be standardized to the modulated key.

We constructed five different sets of musical features within each song as follows based on processed melodic and harmonic data for the collection of songs. The first set of features was chord types. Seven diatonic chords, that is, I, ii, iii, IV, V, vi, vii, which are conventionally the building blocks for most popular Western music, were their own categories. The true diatonic chord on the seventh note of the scale is a diminished chord, which was only used once, in "You Won't See Me," while the minor vii was used more often. We therefore took the liberty of using the minor vii instead as our "diatonic" chord on the seventh. Because diminished and augmented chords were used rarely in general, we collapsed all occurrences of non-diatonic major chords along with augmented chords into a single category, and non-diatonic minor chords along with diminished chords into a single category. This resulted in a total of 9 categories. We explored other category divisions, including fewer instances of collapsed categories, but the sparsity of the data across the non-diatonic, augmented, and diminished chords resulted in less reliable predictability. Additionally, we decided to group all seventh and extended chords (e.g., ninth chord, eleventh chord) with their unaltered triad

counterparts.

The second set of features consisted of melodic notes. The octave in which a melodic note was sung was ignored in the construction, so that the number of note categories totaled 12 (the number of pitch classes on the chromatic scale).

The third set of features comprised chord transitions, that is, pairs of consecutive chords. As with individual chord categories, considering all combinations of chord transitions would have resulted in an unnecessarily large number of sparsely counted categories. We collapsed the chord categories as follows. Each transition among the tonic, sub-dominant (major fourth), and dominant (major fifth) was its own category. Every other transition from a diatonic chord to another diatonic chord, regardless of the order of the two chords, was its own category. For example, transitions from ii to V were grouped with transitions from V to ii. Transitions that involved the tonic and any non-diatonic chord were grouped into one category, and transitions that involved the dominant and any non-diatonic chord were also all grouped into one category. Chord transitions starting with any non-diatonic chord, and ending with a diatonic chord (other than the tonic or dominant) was its own category, and chord transitions ending with any non-diatonic chord, and starting with a diatonic chord (other than the tonic or dominant) was its own category. Finally, all chord transitions between two non-diatonic chords fell under one category. The total number of chord transition categories totaled 24 with these raw category collapsings. Empty categories from the canon of songs were ignored.

The fourth set of features involved melodic note transitions as pairs of notes. In contrast to the single melodic note categories, we considered the octave of the second note in the pair. Thus, each melodic note in a pair could be in a three-octave range. In addition, we considered the start and end rest of a phrase to be considered a note in constructing note transition categories. Thus a single note at the start or at the end of a phrase was each treated as

a note transition. Each start of a phrase on any diatonic note was its own category, and each end of a phrase on any diatonic note was its own category. All notes on the diatonic scale transitioning from or to the tonic was its own category. Any transition from a pitch on the diatonic or pentatonic scale (which includes the flat 3 and flat 7) to another pitch on the diatonic or pentatonic scale, including the same pitch, was its own category, regardless of octave. Upward movements by 2, 3, 4, or 5 notes on the diatonic scale were individual categories, and the corresponding downward movements were their own categories.

We performed a greater amount of collapsing of categories of melodic transitions when at least one note in the transition was not on the diatonic scale. All transitions between the two same non-diatonic notes (excluding the flat 3 and flat 7) were collapsed into the same category. All melodic phrases starting on a non-diatonic note were collapsed into the same category, and all melodic phrases ending on a non-diatonic note were collapsed into the same category. A semitone upward or downward movement from a diatonic note to a non-diatonic note formed two distinct categories, as did a semitone upward or downward movement from a non-diatonic note to a diatonic note. All upward movements of at least two semitones involving a non-diatonic note were collapsed into the same category, and all downward movements of at least two semitones were collapsed into the same category. The total number of nonempty categories of melodic transitions under this collapsing scheme was 65. It is worth noting that we had also considered an alternative set of melodic transition variables. These were based to a large extent on grouping upward and downward movements by the size of the interval, but without regard to the musical function of the transition. We feel that the main groupings described above are arguably more musically justifiable because they are more directly connected to the pitches within transition pairs rather than pitch distances.

The last set of features captured local contours in the melodic line of a song. Every consecutive 4-note subset within a melodic phrase (between its start and end) was partitioned into one of 27 different categories according to the direction of each consecutive pair of notes.

For each of the three pairs of consecutive notes in a 4-note melodic sequence, the transitions could be up, down or same if the melodic notes moved up, down, or stayed the same. Because each consecutive pair across the 4-note sequences allowed three possibilities, the representation consisted of $3 \times 3 \times 3 = 27$ categories. Longer contours (consecutive note subsets of 5 or more notes) would provide greater contour detail, but the number of implied categories would create difficulties in model fitting especially with the relatively low number of songs to use for model-building. The contour representation is an attempt to characterize local features in the melodic line beyond 2-gram representations but without the same level of detail.

The five sets of musical features together total 137 binary variables for each song. Our modeling approach, which relied mainly on cross-validating regularized logistic regression, could result in prediction instability when a feature is shared by very few or very many songs. We therefore removed 16 features in which five or fewer songs contained the feature, or where 66 or more songs (out of 70) contained the feature. The features shared by 66 or more songs included the tonic chord; melodic notes that included the tonic, second, third and fifth; and the 4-note contour (up, down, down). The features shared by five or fewer songs consisted of the minor seventh chord, chord transition from iii to V, upward and downward melodic transitions by 5 notes on the diatonic scale, repeated flat 3 notes, other repeated non-diatonic notes, upward melodic transition from flat 7 to flat 3, melodic transition between flat 3 and fifth, and melodic transition from flat 7 to fourth. With these exclusions, our analyses used a total of 121 musical features.

We display the most common musical characteristics by category, after the exclusions, in Table 1. Major 4th and major 5th chords are the most common among the 70 songs (after the tonic), and the melodic notes of a 4th and 6th are also common. These notes and chords are understood to be the building blocks of popular Western music. The chord transition from major 5th to tonic is also a common chord change in popular music, is well-represented

in early Lennon-McCartney songs, and is often utilized as a harmonic phrase resolution. The most common melody note transitions stay on the diatonic scale, which again is in keeping with Western songwriting. Finally, the two contours listed in Table 1 are both simple shapes in the melodic line.

Representation	Features
Chords	Major 4th (64), Major 5th (63)
Melodic notes	4th (62), 6th (63)
Chord transitions	Major 5th to Tonic (61)
Note transitions	Downward transition of one note on the diatonic scale (62), Upward transition of one note on the diatonic scale (60)
Contours	(down, down, down) (61), (down, down, up) (62)

Table 1. Musical features among the 121 that occurred in 60 or more of the 70 songs with known authorship, after eliminating features occurring in 65 or more songs. Numbers in parentheses indicate the number of songs with the listed feature.

3 A model for songwriter attribution

Our approach to modeling authorship involved a two-step process. First, we selected a subset of the 121 musical features that each had a sufficiently strong bivariate association with authorship. Second, conditional on the selected features, we modeled authorship using logistic regression regularized via elastic net penalization (Zou & Hastie, 2005) with tuning parameters optimized by cross-validation. The latter process was implemented in the R package `glmnet` (Friedman et al., 2010). We describe each step in more detail below.

For song i, $i = 1, \ldots, n$, where n is the number of songs with known authorship in the training data, let

$$y_i = \begin{cases} 0 & \text{if song } i \text{ was written by John Lennon} \\ 1 & \text{if song } i \text{ was written by Paul McCartney.} \end{cases} \tag{1}$$

We let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denote the vector of binary authorship indicators. For $j = 1, \ldots, J$, where J is the total number of dichotomized musical features, let for each $i = 1, \ldots, n$,

$$x_{ij} = \begin{cases} 0 & \text{if feature } j \text{ is not observed in song } i \\ 1 & \text{if feature } j \text{ is observed in song } i. \end{cases} \tag{2}$$

We let $\boldsymbol{X}$ denote the $n \times J$ matrix with elements x_{ij}, and let $\boldsymbol{X}_j$ denote the j-th column of $\boldsymbol{X}$.

The first step of our procedure is to determine a subset of the index set $\{1, 2, \ldots, J\}$ in which $\boldsymbol{X}_j$ is sufficiently associated with authorship. This can be accomplished by computing odds ratios of the j-th binary feature with authorship and retaining features with an odds ratio (or its reciprocal) above a specified threshold. Equivalently, the selection can be performed by retaining features in which tests for significant odds ratios have p-values below a specified level. This pre-processing of features, known as sure independence screening (SIS), has been developed and explored by Fan (2007), Fan and Lv (2008), and Fan and Song (2010). SIS is more typically employed in settings with a massive number of predictors, but in our setting provides a crude but effective way of reducing the number of features in our final model. Our final model evaluations exhibit better out-of-sample accuracy including SIS as a pre-processing step to modeling than omitting this step, as we describe in Section 4.

To implement SIS in our setting, we computed a p-value of a Pearson chi-squared test for each $j = 1, \ldots, J$, for the significance of the odds ratio in a 2×2 contingency table constructed from $\boldsymbol{y}$ and $\boldsymbol{X}_j$. When the elements of any of the contingency tables has low counts, the odds ratio estimate is unstable. The reference distribution for such settings is poorly approximated by a chi-squared distribution, so we instead simulated test statistics 10,000 times from the null distribution according to Hope (1968) to obtain more reliable p-values. This procedure is implemented in the `chisq.test` function in base R. The p-value for each test was then compared to a pre-specified significance level to determine inclusion

for modeling. See below for a detailed discussion about the specified significance level.

Suppose as a result of the variable screening we retained K variables, renumbered $1, \ldots, K$. The second step of the procedure involves a logistic regression model of the form

$$p_i = \Pr(y_i = 1 | \boldsymbol{x}_i, \beta_0, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-(\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}))} \tag{3}$$

where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})'$, and with model parameters β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. Given the possibly large number of musical features compared to the number of songs in our data set, we fit our logistic regression model through elastic net regularization. Letting

$$\ell(\beta_0, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}^*) = \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \tag{4}$$

be the log-likelihood of the model parameters, where $\boldsymbol{X}^*$ is the $n \times K$ matrix of x_{ij} retained from variable screening, elastic net regularization seeks to find estimates of β_0 and $\boldsymbol{\beta}$, conditional on α and λ, that minimize

$$f_{EN}(\beta_0, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}^*, \alpha, \lambda) = -\frac{1}{n} \ell(\beta_0, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}^*) + \lambda \left[(1 - \alpha) \frac{\|\boldsymbol{\beta}\|_2^2}{2} + \alpha \|\boldsymbol{\beta}\|_1 \right] \tag{5}$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{J} \beta_j^2$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{J} |\beta_j|$, and $\lambda \geq 0$ and $0 \leq \alpha \leq 1$ are tuning parameters. When $\alpha = 0$, regularization is of the form of a ridge (L_2) penalty, and when $\alpha = 1$ the logistic regression is fit with a Lasso (L_1) penalty.

Optimization of the elastic net logistic regression parameters proceeds as follows. We consider the equally-spaced grid of values for α in $\{0.0, 0.1, \ldots, 1.0\}$. For each candidate value of α, we consider 100 candidate values of λ. The choice of these candidate values is described in Friedman et al. (2010). For these $11 \times 100 = 1100$ candidate pairs (α, λ), we perform 5-fold cross-validation using the negative log-likelihoods evaluated at the withheld fold. Each fold is constructed by sampling songs stratified by author so that approximately 20% of Lennon

and 20% of McCartney songs are contained in each fold. This approach preserves the balance in authorship within fold relative to the overall sample. We choose the minimizing pair of α and λ, and then minimize the target function in (5) over the coefficients β_0 and $\boldsymbol{\beta}$. Zou and Hastie (2005) argued for considering the selection of λ based on a 1 standard error rule commonly used in regularization procedures, but we found in our application that choosing the minimum value resulted in better predictability.

A natural extension to regularized logistic regression is to include interactions among the predictors. Among the difficulties of including all interaction terms in a regularized regression is that the likely higher degree of sparsity among the interactions compared with the individual features makes it difficult to identify the important interactions. Futhermore, high correlations among the variables can negatively impact selection. Work aimed at discovering important interactions in a more principled manner has been explored. Ruczinski, Kooperberg, and LeBlanc (2003, 2004) developed logic regression, a procedure that finds Boolean combinations of binary predictors in an approach similar to Bayesian CART (Chipman, George, & McCulloch, 1998). Logic regression prevents overfitting through the reduction of model complexity in growing the number of Boolean combinations that are formed. Procedures such as those by Bien, Taylor, and Tibshirani (2013) and Lim and Hastie (2015) involve building interactions only when the main effect terms are selected, and this is carried out by taking advantage of the group-lasso (Yuan & Lin, 2006). We explored these extensions to our approach, based on having already eliminated the rarely-occurring or frequently-occurring musical features, but found that out-of-sample predictability was worse than using only the additive effects of our features. An argument could be made that including interactions would better account for sets of musical features that are highly correlated. However, the extra flexibility associated with including interactions resulted in greater variance in the predictions that degraded our model's performance.

Rather than specifying a single significance level threshold for variable screening followed

by regularized logistic regression, our selection procedure considered five different significance level thresholds: 1.0 (no variable screening), 0.75, 0.50, 0.25, and 0.10. We discuss in Section 5 the rationale for only four additional thresholds. We performed leave-one-out cross-validation in the following manner to choose the best threshold. Let $\boldsymbol{X}_{(i)}$ and $\boldsymbol{y}_{(i)}$ denote the predictor matrix and response vector with observation i deleted. First, for a fixed threshold $t \in \{1.0, 0.75, 0.50, 0.25, 0.10\}$, we performed variable screening on $\boldsymbol{X}_{(i)}$ followed by fitting elastic net logistic regression of $\boldsymbol{y}_{(i)}$ based on the retained features (with 5-fold cross-validation within the $n-1$ songs to obtain the elastic net parameter estimates). The out-of-sample predicted probability $\hat{p}_i^{(t)}$ for observation i and threshold t is then computed given $\boldsymbol{x}_i$ from the fitted logistic regression. The negative log-likelihood for threshold t is computed as

$$\mathrm{LL}^{(t)} = -\sum_{i=1}^{n} \left(y_i \log \hat{p}_i^{(t)} + (1 - y_i) \log(1 - \hat{p}_i^{(t)}) \right). \tag{6}$$

The threshold $t = t_{opt}$ with the minimum $\mathrm{LL}^{(t)}$ is the one chosen by this procedure. Once t_{opt} is determined, variable screening is performed using this threshold based on all n observations followed by performing regularized logistic regression on the remaining features.

4 Model implementation and results

We applied our approach to authorship attribution developed in Section 3 to the corpus of 70 Lennon-McCartney songs based on the musical features described in Section 2. We first describe model summaries applied to the 70 Lennon-McCartney songs in the training data. These summaries are based on a leave-one-out predictive analysis. We then fit our model to the full 70 songs, and use the results to make predictions on the songs and song portions that are of disputed authorship or are known to be collaborative.

4.1 Predictive validity and leave-one-out model summaries

A common approach to predictive validity in machine learning is to divide a data set into modeling, validation, and calibration subsets. Typically a model is constructed and validated iteratively on the first two subsets of the sample, and predictive properties of the approach are summarized on the withheld calibration set. See Draper (2013) for a good overview of this approach, which the author terms "calibration cross-validation." Given the small number of observations (songs) in our sample, our predictive accuracy would suffer by withholding a substantial calibration set, so instead we summarized our algorithm's quality of calibration through leave-one-out cross-validation. Specifically, we withheld one song at a time, and with the remaining 69 songs we performed the procedure described in Section 3. That is, with 69 songs at a time, we first optimized the choice of the p-value threshold for SIS through leave-one-out cross-validation (with a 68-versus-1 split to compute the out-of-sample negative log-likelihood), then with the variables selected based on the optimized p-value threshold we fit a logistic regression via elastic net regularization on the 69 songs (using 5-fold cross-validation to estimate the tuning parameters). Finally, based on the logistic regression fit, the probability estimate of the withheld song was computed. This process was performed for all 70 songs to obtain out-of-sample predictions for each song with known authorship.

Figure 1 displays histograms of the out-of-sample probabilities McCartney wrote each of the 70 songs or song portions with known authorship. The songs or fragments were divided into the 39 that Lennon wrote, and the 31 that McCartney wrote. Generally, the higher probability estimates tend to correspond to McCartney-authored songs, and lower probabilities correspond to Lennon songs. Using 0.5 as a threshold for classification, the model correctly classifies 76.9% of Lennon songs, and 74.2% of McCartney songs, with an overall correct classification rate of 75.7%. We display the leave-one-out probability predictions for the 39 songs known to be written by Lennon in Table 2, and for the 31 songs known to be written by McCartney in Table 3.

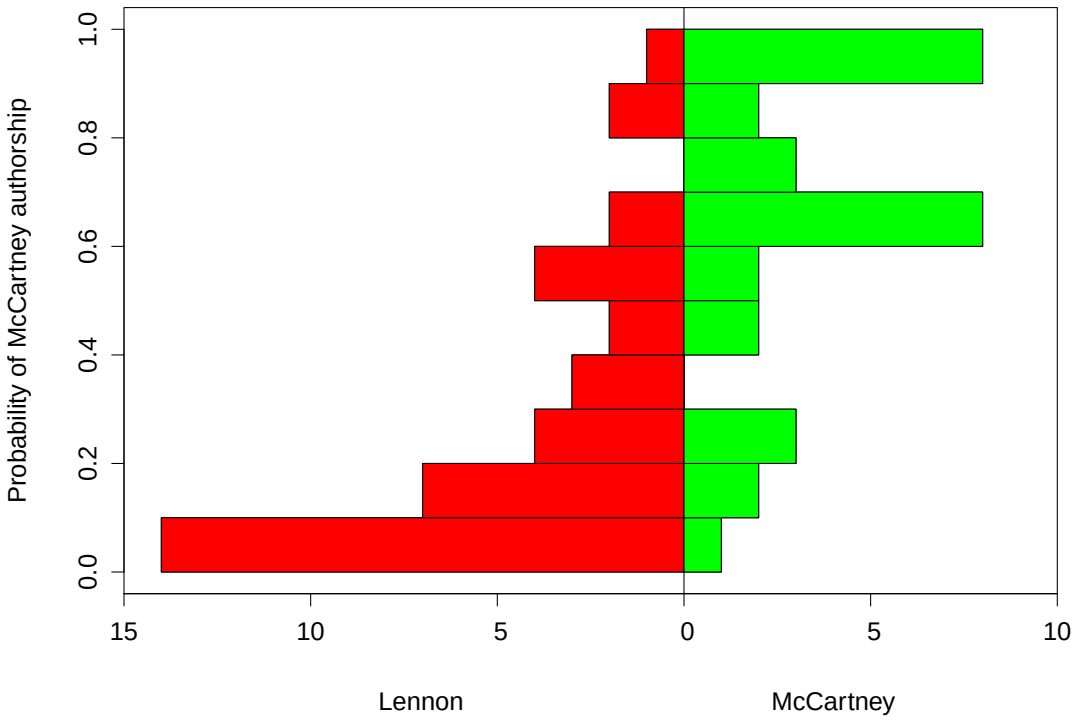

Figure 1. Back-to-back histograms of the out-of-sample prediction probabilities of songs of known authorship. Bars to the left represent 39 songs or song portions known to be written by Lennon, and bars to the right represent 31 songs or song portions known to be written by McCartney.

In addition to the simple classification results, we performed a receiver operating characteristic curve (ROC) analysis on the out-of-sample probability predictions for the 70 songs and fragments. The results of the analysis, which were performed using the `pROC` library in R (Robin et al., 2011), are summarized in Figure 2. The c-statistic (or area under the ROC curve, AUC) is 0.837, which indicates a strong level of predictive discrimination.

For each of the 70 applications of optimized variable screening followed by regularized logistic regression based on 69 songs at a time, we recorded the optimal variable screening p-value threshold. We discovered that among the p-value thresholds in the candidate set, the significance level of 0.25 was selected in 69 of the 70 applications of variable screening, and the

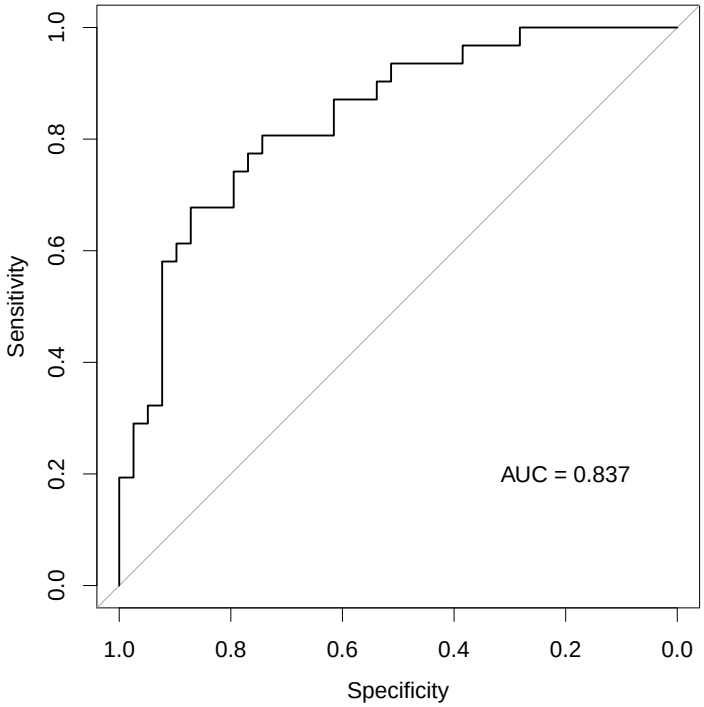

Figure 2. ROC plot for out-of-sample song probability predictions based on 70 songs or song fragments with known authorship.

significance level of 1.0 was selected for one application (corresponding to leaving out the song "You Won't See Me" by McCartney). Figure 3 shows boxplots across the 70 analyses of the leave-one-out predictive negative log-likelihoods for each p-value threshold. As seen in the figure, the negative log-likelihoods achieve their lowest values when discarding features that have a p-value for an odds ratio larger than 0.25. The second-best choice among these candidate thresholds was not to remove any variable prior to elastic net. Removing variables based on a threshold of 0.10 resulted in noticeably worse performance than any of the other choices.

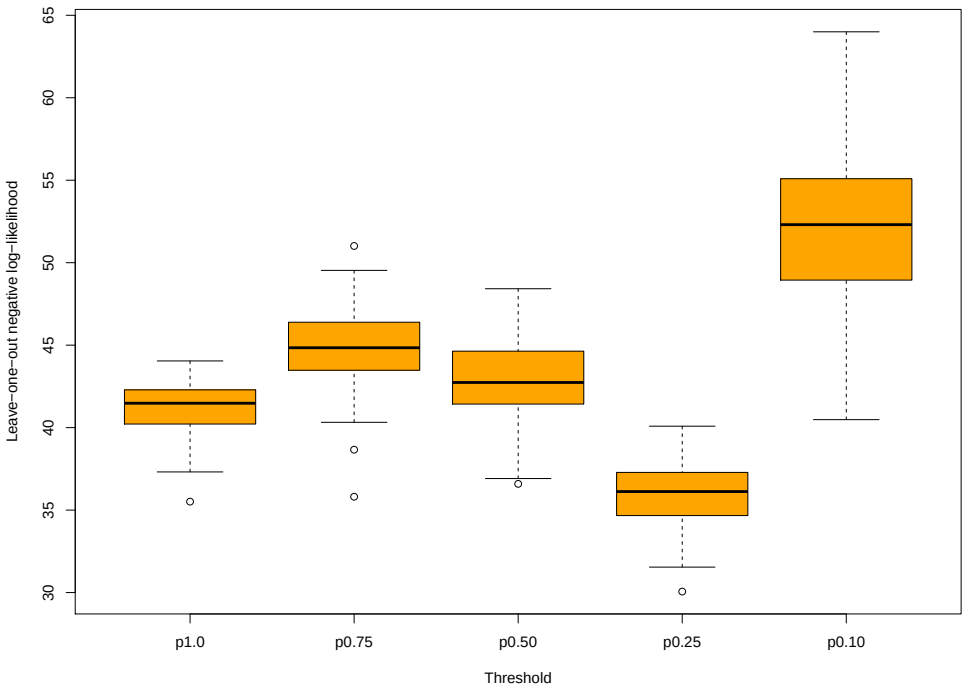

Figure 3. Boxplots of the leave-one-out negative log-likelihoods for each choice of p-value threshold (1.0, 0.75, 0.5, 0.25, 0.10). Each box consists of the distribution of negative log-likelihoods across the 70 leave-one-out analyses.

4.2 Probability predictions for disputed and collaborative songs

We applied our algorithm from Section 3 to the full set of 70 songs. The resulting logistic regression model was then used to make predictions on disputed songs and song portions, and on songs known to be collaborations between Lennon and McCartney. The optimal significance level threshold for the variable screening was 0.25 based on leave-one-out cross-validation. Conditional on selecting variables using the 0.25 p-value threshold, the tuning parameters in elastic net logistic regression were optimized at $\alpha = 0.3$ and $\lambda = 0.0359$. Thus, the final logistic regression model for predictions involved an average of L_1 and L_2 penalties, but more heavily weighted towards a ridge penalty. Of the 40 features that were selected through sure independence screening, 29 were non-zero in the final model as a result of elastic net logistic regression. The full set of 29 variables is listed in Table 4.

Distinguishing song features of Lennon and McCartney authorship can be learned from the coefficient estimates of the logistic regression. Positive coefficients are indicative of features used more associated with McCartney's songs, and negative coefficients are indicative of features more associated with Lennon's songs.

These results offer interesting interpretations of musical features that distinguish McCartney and Lennon songs. One clear theme that emerges is that McCartney tended to use more non-standard musical motifs than Lennon. For example, the harmonic transitions between I $\rightarrow$ vi and vi $\rightarrow$ I are moves that are natural and reasonably direct in popular music, and Lennon used these chord changes much more frequently than McCartney (coefficient of -0.315). These chord changes also create an ambiguity about whether the song is in the major or relative minor key. Lennon songs like "It's Only Love" start with two sets of alternations between I and vi. In contrast, the chord change between ii and IV (coefficient of 1.428) is less standard, and is used more frequently by McCartney, as offering a different "flavor" to the often used sub-dominant, and is used, for example, in McCartney's "I'm Looking Through You."

Another example is that Lennon's melodic note changes tended to remain much more often on the notes of the diatonic scale, whereas McCartney tended to use melodic note transitions that moved off the diatonic scale. This is exhibited in the negative coefficients for note transitions moving up or down one note on the diatonic scale, and the positive coefficient (1.135) for upward note transitions in which one was not on the diatonic scale. Lennon also more often started melodic phrases at the 3rd or ended phrases at a 5th, both of which are notes on the diatonic scale. In contrast, McCartney more often used a flat 3, and transitions from the flat 3 to the tonic in his sung melodies, both of which are notes often associated with a blues scale and not the diatonic scale. This observation is at odds with the often-held notion that Lennon composed songs in a more traditional "rock-and-roll" style. In general, these results suggest that the greater complexity in McCartney's music is a distinguishing

feature exhibited by the coefficients in Table 4 that are positive.

In addition to the coefficients, we report a measure of variable importance in the third column of Table 4. Our measure has close connections to an early approach developed in the context of random forests (Breiman, 2001). In particular, the importance of a variable can be assessed by randomly permuting its values across observations, and then computing an overall measure of model performance. The lower the performance measure after permuting the variable, the more important the variable. For our approach, randomly permuting the values of a musical feature across songs is effectively equivalent to having the feature removed because sure independence scanning should eliminate the feature in the first step of our prediction algorithm. Thus, our variable importance measure was computed as follows. First, we removed the musical feature whose importance we wanted to assess. We then applied our out-of-sample procedure from Section 4.1 and computed 70 leave-one-out predicted probabilities. We performed an ROC analysis on these probabilities and the known authorship of the 70 songs and summarized the c-statistics in the third column of Table 4. Lower values of the c-statistic indicate greater variable importance. The c-statistic without eliminating any features is 0.837, but some of the values in Table 4 can be higher given the random assignments in the stratified cross-validation procedure. Generally, higher absolute values of coefficient estimates correspond to lower c-statistics. Musical features with the lowest c-statistics, all less than 0.80, include the McCartney features (1) the occurrence of the 4th note on the diatonic scale, (2) the chord transition between ii and IV, (3) the note transition downward from the flat 3rd to the tonic, and (4) the note transition downward a half step from a non-diatonic note to a diatonic note. The only feature with a Lennon leaning and having a c-statistic less than 0.80 is the note transition up a half step from a non-diatonic note to a diatonic note. Compared with the McCartney feature of a downward half-step move, upward half-step moves may correspond to particular note transitions that are distinct from the downward moves.

We applied the fit of our model to make predictions for eight songs or song portions with disputed authorship, and for 11 known to be collaborations. The prediction probabilities were derived by applying the fitted logistic regression to the songs of unknown and collaborative authorship. We accompanied the probability predictions with approximate 95% confidence intervals calculated in the following manner. For each song of disputed or collaborative authorship, we computed 70 probability predictions based on leaving out each one of the 70 songs in our training sample. An approximate 95% confidence interval is constructed from the 2.5%-ile and 97.5%-ile of the 70 probability predictions for each song. It is worth noting that these intervals are conservative because one fewer song is used than with the corresponding point prediction. The probability predictions and corresponding confidence intervals are displayed in Tables 5 and 6. We also display the distributions over the 70 predicted probabilities for each disputed song as density estimates in Figure 4. For the songs and fragments of disputed authorship, all of the probabilities are lower than 0.5 suggesting that each individually had a higher probability of being written by Lennon. The 95% confidence intervals are mostly less than 0.5, though "Wait" and the bridge of "In My Life" have confidence intervals that cross 0.5. The density plots in Figure 4 demonstrate the substantial uncertainty in the probability prediction for the bridge of "In My Life" and to a lesser extent for "Wait." In most instances, the conclusions based on our model seem to match up with the suspected authorship, as discussed by Compton (1988). According to Compton, the song "Ask Me Why," which Lennon sang, was likely written by Lennon. Similarly, "Do You Want to Know a Secret?" was one that Lennon recalled having written and then given to George Harrison to sing. In "A Hard Day's Night," the verse and chorus are known to have been written by Lennon (Rybaczewski, 2018; Wiener, 1986), but McCartney seemed to remember having collaborated, perhaps with the bridge, which he sang. While McCartney wrote most and possibly all of "Michelle," Lennon claimed in some interviews that he came up with the bridge on his own, but in other interviews asserted that the bridge was a collaboration with McCartney (Compton, 1988). "Wait" is also suspected to have been written

by Lennon according to Compton (1988), though in Miles (1998) McCartney remembered the song as mostly his. Lennon wrote "What Goes On" several years prior to the formation of the Beatles, and it is disputed whether McCartney (and Ringo Starr) helped write the bridge section when recording the song with the Beatles. We discuss "In My Life" in more detail below.

For the songs during the study period that were understood to be collaborative, it is unclear to what extent Lennon and McCartney shared songwriting efforts. Our model's probability predictions can be viewed as demonstrating similarities with the patterns inferred in songs and fragments with undisputed authorship. However, it is worth noting that our model was developed on a set of songs and song portions that were of single authorship, and that applying our model to songs of collaborative authorship may result in predictions that are not as trustworthy. As with the information in Table 5, most of the collaborative songs in Table 6 were inferred to be mostly matching the style of Lennon. While four songs were inferred to be written more in McCartney's style, two exceptions are worth noting. The songs "Baby's in Black" and "The Word," according to Compton (1988), were both entirely collaborative, with Lennon having claimed that "The Word" was mostly his work. It is curious, in particular, that "The Word" is inferred with near certainty of being McCartney-authored. One feature of the song is the predominance of the flat third. This McCartney-like motif may be responsible for the high probability that the song is inferred to be written by McCartney. The other two songs, "From Me to You" and "She Loves You," were also more likely to be McCartney-authored. Compton (1988) reported that the former was claimed to be entirely collaborative, and that the latter was initiated by McCartney even though the song was written collaboratively.

Two of the collaborations are worthy of comment. While Lennon and McCartney co-wrote "She Loves You," Lennon remembered that "it was Paul's idea" (Compton, 1988), and the probability indicates that the song is weighted towards McCartney. On the other hand, our

model's probability prediction for "I Want to Hold Your Hand," which was written "eyeball to eyeball" (Compton, 1988), is that the song is much more characteristic of Lennon's style. Indeed, in one of the Jann Wenner interviews (Wenner, 2009), Lennon opined about the beauty of the song's melody, and picked out that song along with his song "Help!" as the two Beatles' songs he might have wanted to re-record. However, perhaps the song might have been special to him as it had much more of his imprint.

Of all Lennon-McCartney songs, "In My Life" has probably garnered the greatest amount of speculation about its true author. *Rolling Stone* magazine considered it to be the 23rd greatest song of all time (*Rolling Stone*, 2011). Our model produces a probability of 18.9% that McCartney wrote the verse, and a 43.5% probability that McCartney wrote the bridge, with a large amount of uncertainty about the latter. Because it is known that Lennon wrote the lyrics, it would not be surprising that he also wrote the music. Lennon claimed (Compton, 1988) that McCartney helped with the bridge, but that was the extent of his contribution. Breaking apart the song into the verse and the bridge separately, it is apparent that the verse is much more consistent stylistically with Lennon's songwriting. Thus, a conclusion by our model is that the verse is consistent with Lennon's songwriting style, but the bridge less so. The bridge having a probability that McCartney wrote the song closer to 0.5 may be indicative of their collaborative nature, as suggested by Lennon, of this part of the song.

5 Discussion

The approach to authorship attribution for Lennon-McCartney songs we developed in this paper has connections to methods used in attribution analysis of text documents. One important difference is that typical text analysis models rely on the relative frequencies of occurrence of words or word combinations. In a musical context, where repeats of musical features are intrinsic to a song's construction, the relative frequencies of the occurrence of

the musical "words" may obscure their importance in characterizing an author's composition style. Another difference from typical text analysis problems is that songs include more than just one text stream. For our work, we specifically included songs' melodic note sequence and chord sequence as two streams in parallel. Our particular choice in the representation and analysis of Lennon-McCartney songs of the early Beatles period seemed to be sufficient in recovering a song's author with greater than 75% accuracy, and with a high level of discrimination (c-statistic of 0.837 from the ROC analysis).

Our model predictions, particularly for the songs with disputed authorship, seem to be supported generally with the stories that accompany the songs' origins. While it is tempting to interpret the results of our model as revelations of a song's true author, other interpretations are just as compelling. For example, a disputed song such as "In My Life" which according to our model has a high probability of the verse and bridge each being written by Lennon, may in fact have been written by McCartney who stated he composed the song in the style of Smokey Robinson and the Miracles (Turner, 1999), but actually wrote in the style of Lennon, whether consciously or subconsciously. Songs with high probabilities of being written by Lennon or McCartney are mainly indications that the songs have musical features that are consistent with the Lennon or McCartney songs used in the development of our model. To this end, one use of our model is to investigate whether certain sections of disputed or collaborative songs are suspected of being more consistent with particular composition styles. For example, the song of disputed authorship "Wait," which our model estimates a probability of 0.391 of being written by McCartney, is sung in harmony by Lennon and McCartney throughout the song except in the bridge section where McCartney sings alone. It is natural to ask whether that section may be more in the style of McCartney who may have had a freer hand in writing that portion of the song. Indeed, our model applied to just the bridge section resulted in a 0.646 probability of McCartney authorship, suggesting that the bridge is more in the style of McCartney than Lennon.

In typical text analyses, the choice of "stop" words, i.e., the ones used in analyses to distinguish authorship style, is often made subjectively or at least by convention. The analogous decision in a musical context is arguably much more difficult, as the complexity of choices is far greater. In our work, we needed to make many subjective decisions that influenced the construction of musical features. Such decisions included what constituted the beginning and ending of melodic phrases, whether a key change (modulation) should reset the tonic of the song, whether ad-libbed vocals should be considered part of the melodic line, how to include dual melody lines that were sung in harmony, and so on. Our guiding principle was to make choices that could be viewed as the most conservative in the sense of having the least impact on the information in the data. For example, we omitted melodic information from ad-libbed vocals, and made phrasings of melodic lines as long as possible, as shorter lines introduced extra "rests" as part of the melodic transitions. Also, when it was not clear in cases of dual melody lines which was the main melody, we included both melody lines.

It is worth noting that the model developed here was not our first attempt. We explored variations of the presented approach before arriving at our final model, including versions that permitted interactions, alternative variable selection procedures such as recursive feature elimination and stepwise variable selection, models for the musical features as a function of authorship that were inverted using Bayes rule, random forests, as well as several others. A danger in exploring too many models, especially with our small sample size and without a true test/holdout set, is the potential to overfit. This concern may not be apparent in the presentation of our analytic summaries, which was the culmination of a series of model investigations. The concern of overfitting limited some of our explorations. For example, after having modest success using elastic net logistic regression without any variable preprocessing, we inserted variable screening parameterized by a p-value threshold based only on four threshold values. Using a greater range of thresholds, especially after having learned that elastic net alone was a promising approach, and that we were tuning the model parameters based on the same leave-one-out validation data, would have had the potential to

produce overfitted predictions. We suspect that our final model, however, does not suffer from overfitting concerns in any appreciable way. First, the approach we present is actually fairly simple: the removal of musical features based on bivariate relationships with the response followed by regularized logistic regression. More complex procedures might raise questions about their generalizability. Second, we were cautious about optimizing the prediction algorithm and calibrating the predictability using out-of-sample criteria. For example, probability predictions involved leaving out data (one song at a time) to optimize the p-value threshold for variable screening, followed by leaving out portions of data (20% of the data that remained) to optimize the elastic net tuning parameters; and this entire procedure was performed leaving out one song at a time when making predictions for the songs of known authorship. This cascading application of cross-validation mitigates some of the natural concerns about possible overfitting.

Our particular modeling approach does permit extensions to address wider sets of songwriter attribution applications. Our model assumes only two authors, but this is easily extended to multiple songwriters in larger applications by modeling authorship in a multinomial logit model, for example. Another extension of our model can address changes in an author's style over time. Our application to Lennon-McCartney songs focused on a time period where the songwriters' musical styles were not changing in profound ways. To include larger spans of time where a songwriter's style may be changing, one possibility is to assume a stochastic process on the musical feature effects for each author, such as through an autoregressive process. Such an approach acknowledges that an author's style is likely to evolve gradually over time and with an uncertain trajectory. This approach would be straightforward to implement in a Bayesian setting, though implementing such a model in conjunction with variable screening would involve methodological challenges.

Several other limitations are worth mentioning. Our approach assumes that each song or (more relevantly) song portion contains sufficiently rich detail to capture musical information

for distinguishing authorship. Shorter song fragments would have a scarcity of features, and probability predictions are expected to be less reliable. Furthermore, if the goal of this work was to make the most accurate predictions of a song's author, then our approach could clearly be improved by incorporating readily available additional information. Lyric content, information on a song's structure, use of rhythm, song tempo, time signature, and the identity of a song's actual singer or singers are all likely to be highly predictive and distinguishing of a song's authorship. Our decision to ignore this extra information is consistent with the larger goal of being able to establish the stylistic fingerprint of a songwriter based solely on a corpus of songs' musical content, using Lennon-McCartney songs as a sandbox for understanding the potential for this approach. Ultimately, the reduction of a songwriter's musical content into low-dimensional representations, such as a vector of musical feature effects, is the first step towards establishing musical signatures for songwriters that can be used for further analysis. For example, with many songwriters' styles characterized in a reduced form, it is possible to establish influence networks to learn about the diffusion of the creative process in popular music. With recent improvements in technology to convert audio information into formats amenable to the type of analysis we developed in this paper (Casey et al., 2008; Fu, Lu, Ting, & Zhang, 2011), larger-scale analyses of songwriters' styles are a potential area of exploration.

A Musical Background

A justification for the musical features chosen requires an understanding of Western popular music. *Middle C*, often denoted as C4, has frequency 261.6Hz, and the well known equally-tempered 12-tone *chromatic scale* starting on note C4 is the sequence of notes

C4, C#4, D4, D#4, E4, F4, F#4, G4, G#4, A4, A#4, B4

where each successive note is derived from the previous one by multiplying the frequency by $2^{1/12}$. In the above sequence, notes preceding the "4" (i.e., C, C#, D, D#, E, F, F#, G, G#, A, A#, B) are the *pitches*, and the number 4 refers to the *octave* of the note. The continuation of the sequence above is the same set of pitches, but at the next higher octave, that is, C5, C#5, D5, and so on. The 12 notes can also be visualized in a piano diagram in Figure 5.

For the current discussion, we can represent a note as $Z[i, j]$, where $i \geq 1$ indexes the pitch of the note and $j \geq 1$ indexes the octave of the note. We set $Z[1, 4] =$ C4, and all other notes are relative to this anchoring choice. Given the circular ordering of pitches in the chromatic scale, $Z[i + 12, j] = Z[i, j + 1]$ for all i and j. Thus, a specific note has multiple representations using this notation. By convention, the octave of a note is the value j in which the representation $Z[i, j]$ has $i \leq 12$.

The notes $Z[i, j]$ and $Z[i + 1, j]$ are said to be a *semitone apart*, while the notes $Z[i, j]$ and $Z[i + 2, j]$ are said to be a *whole tone* apart. Notes $Z[i, j], Z[i, j + 1], Z[i, j + 2], \ldots$, are said to be in the same *pitch class*. Thus, D3, D4, D5, and so on, are in the same pitch class, but reside in different octaves. It is worth noting that while the *sharp* symbol # denotes raising a note a semitone, one can also use the *flat* suffix $\flat$ to lower a note a semitone. One can translate or *transpose* the chromatic scale to start on any note given its circular structure, and to the human ear all such chromatic scales played in sequence sound essentially the same. A chromatic scale can start on any note $Z[i, j]$ and consists of the 12 notes $(Z[i, j], Z[i + 1, j], \ldots, Z[i + 11, j])$.

The basis of Western music is the *diatonic scale*, which, starting on a given note $Z[i, j]$, called the *tonic* of the scale, consists of the subsequence of seven notes from the chromatic scale

$$(Z[i, j], Z[i + 2, j], Z[i + 4, j], Z[i + 5, j], Z[i + 7, j], Z[i + 9, j], Z[i + 11, j]).$$

For example, beginning on an A at any octave, the diatonic scale with tonic A is (A, B, C#, D, E, F#, G#). Chromatic notes that are not part of the diatonic scale are called *non-diatonic*. Thus the non-diatonic notes with respect to the diatonic scale starting on A include A#, C, D#, F, and G.

The diatonic scale permeates much of Western music, and most popular songs (or portions of songs) can be analyzed to be based on a diatonic scale starting at a specific note belonging to one of the 12 pitch classes; the lowest note of the diatonic scale is called the *major key*, or just the *key*, of the song, and the note itself is the tonic of that key. Songs are often to be found in a "minor" key, based on a *minor scale*. For our purposes, we associate, as is often done, the minor key with the major key three semitones up, as they share the same seven notes. This particular definition of a minor key is often called the *natural minor*, and is the *relative minor* of the associated major key. For example, the key of A minor (as a natural minor) consists of the notes (A, B, C, D, E, F, G), which are the same as those in the major key of C (C, D, E, F, G, A, B), so that A minor is the relative minor associated with C major. Because the major key and relative minor share the same notes on the diatonic scale, in our work we classify songs being in the major key as a proxy for the diatonic notes.

With a given key of a song, non-diatonic notes are usually specified by their relation to the tonic. So, for example, in the key of C, the flat third and flat seventh are E♭ and B♭ (and they could, equivalently, be called the raised second and raised sixth, as well). In fact, in pop/rock music, the flat third and flat seventh play a large role, as they appear in the five note *pentatonic* (or the *blues*) scale, which consists of the notes $(Z[i,j], Z[i+3,j], Z[i+5,j], Z[i+7,j], Z[i+10,j])$, where $Z[i,j]$ is the tonic of the pentatonic scale. Thus, the pentatonic scale starting on tonic C is (C, E♭, F, G, B♭).

A *note transition* or an *interval* is a pair of notes, where the *size* of the interval depends on the number of semitones between them. Some sample intervals include:

- *unison* is between two identical notes (e.g., C4 → C4).
- a *major second* consists of two notes where the second is two semitones (whole tone) up from the first (e.g., C4 → D4, F4 → G4).
- a *major third* consists of two notes where the second is four semitones (two whole tones) up from the first (e.g., C4 → E4, F4 → A4).
- a *perfect fourth* consists of two notes where the second is five semitones up from the first (e.g., D4 → G4).
- a *perfect fifth* consists of two notes where the second is seven semitones up from the first (e.g., A4 → E5).
- a *major sixth* consists of two notes where the second is nine semitones up from the first (e.g., D4 → B4).
- a *major seventh* consists of two notes where the second is 11 semitones up from the first (e.g., F4 → E5).
- an *octave* consists of two notes where the second is 12 semitones up from the first (e.g., C4 → C5).

The minor second, third, sixth, and seventh intervals arise by lowering the second note of the corresponding major interval by a semitone. For example, C → E♭ is a minor third. For intervals of a fourth and fifth, the term *diminished* applies when the top note of the corresponding interval is decreased by a semitone, and the term *augmented* applies when raising the top note a semitone. As an example, the interval C → G# is an augmented fifth in the key of C. In our choice of note transitions within pop songs, the diatonic notes (always relative to the key) have prime importance, with special emphasis on diatonic transitions to and from the tonic, transitions between small steps on the diatonic scale (which are fairly common in melody writing), and transitions along the pentatonic/blues scale.

Chords, for our purposes, consist of three notes played simultaneously (called a *triad*), and form the basis of most of the harmony in pop songs. The two most common types of chords are *major* chords and *minor* chords. A major chord is formed, using $Z[i, j]$ as the root of the chord, as $(Z[i, j], Z[i+4, j], Z[i+7, j])$. A minor chord, in contrast, is formed as $(Z[i, j], Z[i+3, j], Z[i+7, j])$. Less common are diminished chords, formed as $(Z[i, j], Z[i+3, j], Z[i+6, j])$, and augmented chords, formed as $(Z[i, j], Z[i + 4, j], Z[i + 8, j])$. Building chords from the diatonic scale consists of taking a starting note within the scale and successively layering on two extra notes above it, skipping a note each time. For example, in the key of C, the *diatonic chords* are:

- C major, the I major chord (the *tonic*), consisting of notes C, E, and G.
- D minor, the ii minor chord, consisting of notes D, F, and A.
- E minor, the iii minor chord, consisting of notes E, G, and B.
- F major, the IV major chord (the *subdominant*), consisting of notes F, A, and C.
- G major, the V major chord (the *dominant*), consisting of notes G, B, and D.
- A minor, the vi minor chord, consisting of notes A, C, and E.
- B diminished, the vii° diminished chord, consisting of notes B, D, and F.

All of these diatonic chords are "native" to the scale in which they reside; all other chords, with respect to the scale, are deemed to be *non-diatonic chords*. The diatonic chords are the most common ones in popular songs, although non-diatonic chords are often added for variety and creating emotional tension. In particular, in rock-and-roll music, the major chords on the flat third and the flat seventh (and sometimes the flat sixth) play a significant role in that genre.

In pop/rock music, the diatonic chords are all prevalent, especially the tonic (I), subdominant (IV), and dominant (V) chords, with the exception of the diminished seventh chord on the seventh note of the diatonic scale; this chord is rarely used. The minor chord on the seventh note occurs more often, and is sometimes considered a replacement as one of the diatonic chords.

Transitions between chords are a cornerstone of pop/rock music. *Chord progressions* are sequences of chords that often repeat throughout a song. Transitions between diatonic chords form the bulk of the chord transitions. Less common (but not infrequently, when grouped together) are transitions between non-diatonic chords and the tonic (I) or dominant (V).

Entire songs can be viewed in their most basic form as the superposition of chord progressions along with melodic lines. Songs are divided into sections within which chord progressions and melodies are identical or nearly identical. Two of the main sections that appear in most pop/rock songs are the verse and the chorus. The *verses* within a song typically have identical musical content, but usually contain different lyrics. The *chorus* of a song typically has greater musical and emotional intensity than the verse, and contains identical lyrics across repeats within the song. It is common for songs to have a third musical section inserted between an occurrence of the chorus and a subsequent verse, called the *bridge* section. This section musically functions as a connector between the chorus and verse, and may even undergo a *modulation*, that is, resetting the song to a different key, if only temporarily. Other types of sections may appear in typical pop/rock music (e.g., intro, pre-chorus, outro), but the verse, chorus, and bridge are nearly universal components of a song.

More details about the basics of melodic and harmonic structure of popular music can be found in Benward (2014) and Middleton (1990).

References

Airoldi, E. M., Anderson, A. G., Fienberg, S. E., & Skinner, K. K. (2006). Who wrote Ronald Reagan's radio addresses? *Bayesian Analysis*, *1*(2), 289–319.

Benward, B. (2014). *Music in theory and practice, volume 1.* McGraw-Hill Higher Education.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, *41*(3), 1111-1141.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, J. I. (2004). Mathematics, physics and A Hard Day's Night. *CMS Notes*, *36*(6), 4–8.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696.

Cathé, P. (2016). La nostalgie chez les Beatles: vers une application de la théorie des vecteurs harmoniques à la musique pop? *Volume!*, *12*(1), 181–191.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

Cilibrasi, R., Vitányi, P., & De Wolf, R. (2004). Algorithmic clustering of music based on string compression. *Computer Music Journal*, *28*(4), 49–67.

Clement, R., & Sharp, D. (2003). *N*-gram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, *18*(4), 423–447.

Compton, T. (1988). McCartney or Lennon?: Beatles myths and the composing of the Lennon-McCartney songs. *The Journal of Popular Culture*, *22*(2), 99–131.

Conklin, D. (2006). Melodic analysis with segment classes. *Machine Learning*, *65*(2), 349–360.

Draper, D. (2013). Bayesian model specification: Heuristics and examples. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 409–431). New York: Oxford University Press.

Dubnov, S., Assayag, G., Lartillot, O., & Bejerano, G. (2003). Using machine-learning methods for musical style modeling. *Computer*, *36*(10), 73–80.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, *63*(3), 435–447.

Everett, W. (1999). *The Beatles as musicians: Revolver through the anthology.* Oxford

University Press, USA.

Fan, J. (2007). Variable screening in high-dimensional feature space. In *Proceedings of the 4th international congress of chinese mathematicians* (Vol. 2, pp. 735–747).

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, *38*(6), 3567–3604.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1). Retrieved from `http://www.jstatsoft.org/v33/i01/`

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, *13*(2), 303–319.

Fujita, T., Hagino, Y., Kubo, H., & Sato, G. (1993). *The Beatles: Complete Scores.* Hal Leonard Publishing Corporation.

George, J., & Shamir, L. (2014). Computer analysis of similarities between albums in popular music. *Pattern Recognition Letters*, *45*, 78–84.

Hartzog, B. (2016, March). *The Beatles' songwriting.* Retrieved from `http://www.brianhartzog.com/beatles/beatles-songwriting.htm` (Accessed 07-June-2017)

Heuger, M. (2018). *Beabliography: Mostly academic writings about the Beatles.* Retrieved from `http://www.icce.rug.nl/~soundscapes/BEAB/index.shtml` (Accessed 11-July-2018)

Hope, A. C. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *30*(3), 582–598.

Kempfert, K. C., & Wong, S. W. (2018). Where does Haydn end and Mozart begin? Composer classification of string quartets. *arXiv preprint arXiv:1809.05075*.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*(1), 191–201.

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, *24*(3), 627–654.

Malyutov, M. B. (2005). Authorship attribution of texts: a review. *Electronic Notes in Discrete Mathematics*, *21*, 353–357.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.*

MIT Press.

McCormick, N. (1998, January 10). *Must it be Lennon or McCartney?* Retrieved from `http://www.telegraph.co.uk/culture/4711552/Must-it-be-Lennon-or-McCartney.html` (Accessed 07-June-2017)

McDougal, C. (2013, August). *Multi-dimensional computer-driven quantitative analysis of the music and lyrics of the Beatles* (Technical report). Northeastern University. Retrieved from `https://cedricmcdougal.com/4/papers/beatles.pdf`

Middleton, R. (1990). *Studying Popular Music.* McGraw-Hill Education (UK).

Miles, B. (1998). *Paul McCartney: Many Years from Now.* Macmillan.

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, *58*(302), 275–309.

Mosteller, F., & Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Springer.

Naccache, M., Borgi, A., & Ghédira, K. (2008). A learning-based model for musical data representation using histograms. In *International symposium on computer music modeling and retrieval* (pp. 207–215).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rolling Stone. (2011, April). *The Beatles, In My Life.* Retrieved from `https://www.rollingstone.com/music/music-lists/500-greatest-songs-of-all-time-151127/the-beatles-in-my-life-57758/` (Accessed 19-August-2018)

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, *12*(3), 475–511.

Ruczinski, I., Kooperberg, C., & LeBlanc, M. L. (2004). Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. *Journal of Multivariate Analysis*, *90*(1), 178–195.

Rybaczewski, D. (2018). *A Hard Day's Night.* Retrieved from `http://www.beatlesebooks.com/hard-days-night` (Accessed 19-August-2018)

Thisted, R., & Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, *74*(3), 445–455.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *58*(1), 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *73*(3), 273–282.

Turner, S. (1999). *A Hard Day's Write: The stories behind every Beatles song.* Carlton, Dubai.

Wagner, N. (2003). "Domestication" of blue notes in the Beatles' songs. *Music Theory Spectrum*, *25*(2), 353–365.

Wenner, J. (2009). *John Lennon Remembers - Jann Wenner Interview Part 5.* Retrieved from `http://tittenhurstlennon.blogspot.com/2009/07/jann-wenner-interview-part-5.html` (Accessed 14-January-2019)

Whissell, C. (1996). Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities*, *30*(3), 257–265.

Wiener, A. J. (1986). *The Beatles: A Recording History.* McFarland & Co Inc Pub.

Womack, K. (2007). Authorship and the Beatles. *College Literature*, *34*(3), 161–182.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *68*(1), 49–67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *67*(2), 301–320.

Lennon-authored Song	McCartney Probability
All I've Got To Do	0.008
Doctor Robert	0.012
I'm Happy Just to Dance With You	0.038
No Reply	0.041
Girl	0.047
I'll Be Back	0.048
I'm Only Sleeping	0.049
There's a Place	0.064
I'll Cry Instead	0.065
When I Get Home	0.066
And Your Bird Can Sing	0.067
Help!	0.071
We Can Work It Out (Bridge)	0.071
You're Going to Lose that Girl	0.076
I'm a Loser	0.100
Run For Your Life	0.109
It's Only Love	0.111
This Boy	0.128
I Call Your Name	0.148
It Won't Be Long	0.178
Please Please Me	0.185
You Can't Do That	0.231
Ticket to Ride	0.244
A Hard Day's Night (Verse/Chorus)	0.279
Day Tripper	0.294
I Don't Want to Spoil the Party	0.332
Tomorrow Never Knows	0.378
Not a Second Time	0.390
Tell Me Why	0.438
Nowhere Man	0.445
You've Got to Hide Your Love Away	0.524
If I Fell	0.574
Any Time At All	0.588
I Feel Fine	0.598
I Should Have Known Better	0.615
Norwegian Wood (Verse/Chorus)	0.666
Yes It Is	0.802
She Said She Said	0.836
What Goes On (Verse/Chorus)	0.944

Table 2. Songs or song fragments known to be written by John Lennon, rank ordered according to the out-of-sample probability (second column) that is attributed to Paul McCartney.

McCartney-authored Song	McCartney Probability
You Won't See Me	0.069
And I Love Her (Verse/Chorus)	0.105
For No One	0.184
Here There and Everywhere	0.202
PS I Love You	0.282
I'll Follow the Sun	0.284
Can't Buy Me Love	0.440
Got to Get You Into My Life	0.448
Eight Days a Week	0.528
Eleanor Rigby	0.570
I'm Down	0.606
Hold Me Tight	0.606
She's a Woman	0.660
I've Just Seen a Face	0.668
Tell Me What You See	0.668
What You're Doing	0.679
Drive My Car	0.688
Yesterday	0.689
The Night Before	0.715
All My Loving	0.719
Yellow Submarine	0.734
Every Little Thing	0.806
We Can Work It Out (Verse/Chorus)	0.866
Michelle (Verse/Chorus)	0.912
Things We Said Today	0.938
Good Day Sunshine	0.953
I'm Looking Through You	0.957
Another Girl	0.964
I Saw Her Standing There	0.979
I Wanna Be Your Man	0.986
Love Me Do	0.989

Table 3. Songs or song fragments known to be written by Paul McCartney, rank ordered according to the out-of-sample probability (second column) that is attributed to Paul McCartney.

Feature	Coefficient	c-statistic
Intercept	–0.796	—
Chord: V	1.096	0.806
Chord: iii	–0.350	0.842
Note: Flat 2	–0.874	0.817
Note: Flat 3	0.603	0.828
Note: 4th	1.347	0.788
Note: 6th	0.046	0.825
Chord transition: between I and vi	–0.315	0.823
Chord transition: between ii and iii	–0.255	0.846
Chord transition: between ii and IV	1.428	0.795
Chord transition: between ii and V	–0.291	0.830
Chord transition: non-diatonic to diatonic	–0.096	0.833
Melodic transition: down from 4th to flat 3rd	0.481	0.849
Melodic transition: down from flat 3rd to tonic	1.206	0.778
Melodic transition: down 1 note on diatonic scale, not incl. 1 or $4 \to 5/5 \to 4$	–0.348	0.824
Melodic transition: down 1 half step from non-diatonic to diatonic	1.030	0.797
Melodic transition: phrase end on 5th	–0.633	0.808
Melodic transition: pair of notes on the 6th	–0.218	0.825
Melodic transition: up 1 note on diatonic scale, not incl. 1 or $4 \to 5/5 \to 4$	–0.576	0.821
Melodic transition: up 1 half step from non-diatonic to diatonic	–1.232	0.798
Melodic transition: up from tonic to flat 3rd	0.376	0.833
Melodic transition: from 3rd to tonic	0.284	0.829
Melodic transition: from 4th to 5th	–0.653	0.816
Melodic transition: up from or to a non-diatonic note	1.135	0.806
Contour: (Up, Up, Down)	–0.098	0.841
Contour: (Down, Down, Same)	0.535	0.824
Contour: (Up, Same, Same)	–0.098	0.835
Contour: (Down, Up, Same)	–0.938	0.825
Contour: (Same, Down, Up)	–0.501	0.812
Contour: (Up, Down, Up)	–0.555	0.826

Table 4. Coefficient estimates in the final logistic regression in the second column, and ROC analysis c-statistics in the third column. The c-statistics are computed from the 70 leave-one-out probabilities with the variable removed from the prediction algorithm; thus smaller c-statistics indicate greater variable importance.

Song	McCartney Probability (95% confidence interval)
Ask Me Why	0.057 (0.018, 0.080)
Do You Want to Know a Secret	0.080 (0.033, 0.097)
A Hard Day's Night (Bridge)	0.069 (0.016, 0.135)
Michelle (Bridge)	0.199 (0.109, 0.300)
Wait	0.391 (0.275, 0.540)
What Goes On (Bridge)	0.235 (0.088, 0.255)
In My Life (Verse)	0.189 (0.079, 0.307)
In My Life (Bridge)	0.435 (0.270, 0.692)

Table 5. Probability estimates for eight songs or song fragments of disputed or unknown authorship with 95% confidence intervals based on a leave-one-out analysis being attributable to McCartney.

Song	McCartney Probability (95% confidence interval)
Misery	0.310 (0.245, 0.451)
And I Love Her (Bridge)	0.263 (0.110, 0.315)
Norwegian Wood (Bridge)	0.330 (0.135, 0.408)
Little Child	0.337 (0.175, 0.417)
Baby's in Black	0.920 (0.822, 0.977)
The Word	0.976 (0.899, 0.994)
From Me To You	0.606 (0.510, 0.721)
Thank You Girl	0.106 (0.036, 0.202)
She Loves You	0.616 (0.515, 0.733)
I'll Get You	0.062 (0.016, 0.107)
I Want to Hold Your Hand	0.115 (0.065, 0.182)

Table 6. Probability estimates for 11 collaborative songs or song fragments with 95% confidence intervals based on a leave-one-out analysis being attributable to McCartney.

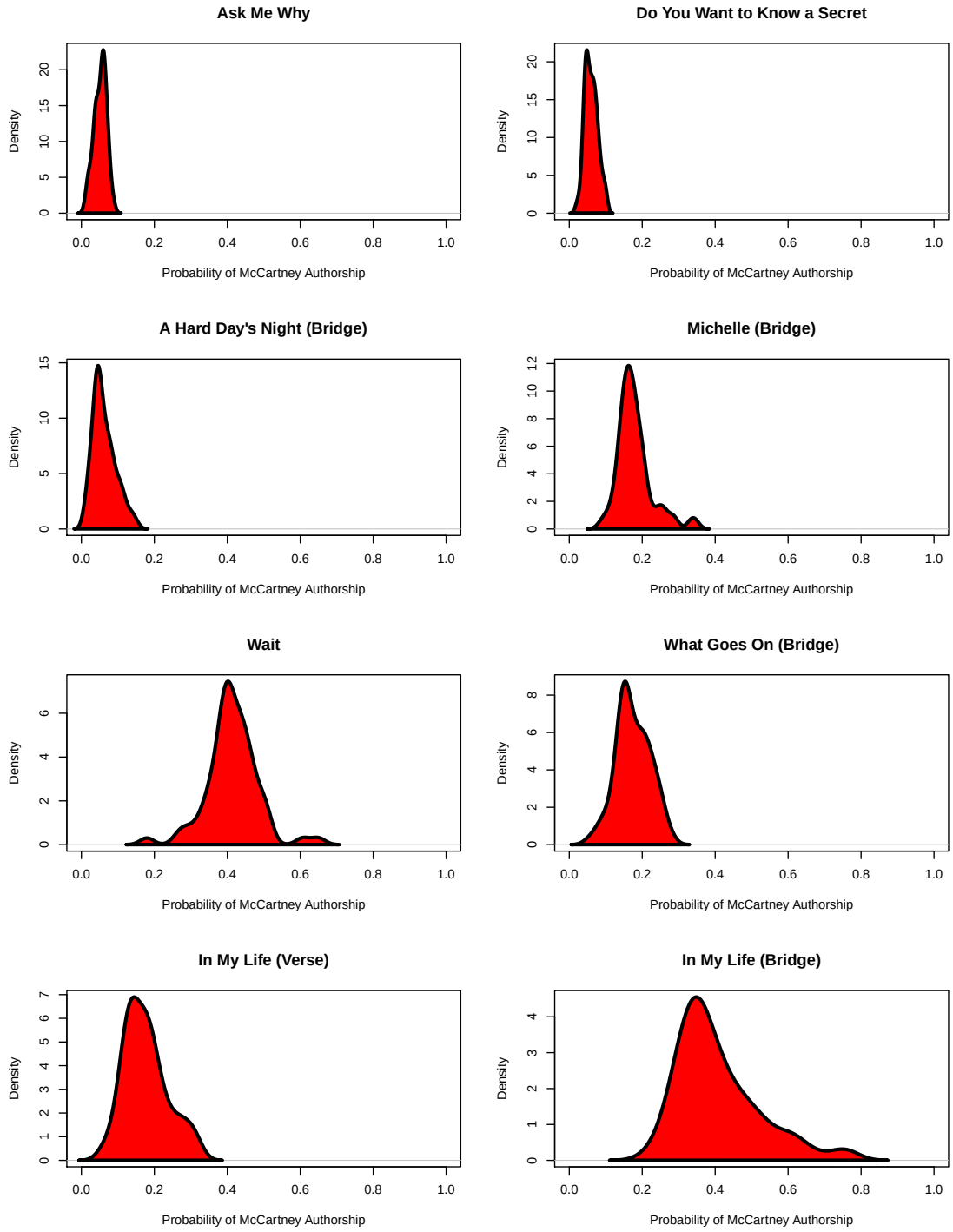

Figure 4. Density plots of the leave-one-out probability predictions for the eight songs of disputed authorship.

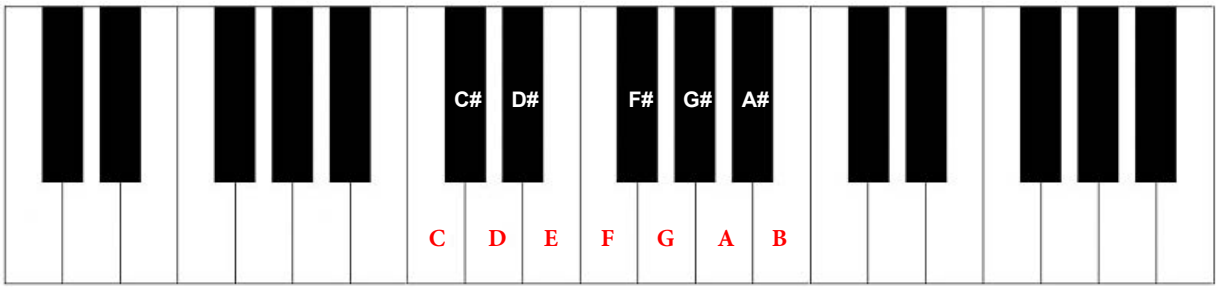

Figure 5. Chromatic scale notes appearing on a piano diagram.