Presented at Audio and Music Computing for Multimedia, ACM Multimedia 2006.

Predicting success from music sales data – a statistical and adaptive approach

Song Hui Chon
CCRMA
Stanford University
Stanford, CA 94305
1 650 723 4971
shchon@ccrma.stanford.edu

Malcolm Slaney
Yahoo! Research
Sunnyvale, CA 94089
1 408 349 4008
malcolm@ieee.org

Jonathan Berger
CCRMA
Stanford University
Stanford, CA 94305
1 650 723 4971
brg@ccrma.stanford.edu

ABSTRACT
Everyone has different musical taste and a person's preference changes over time in unpredictable ways. However, the general public as a collection of individuals may reveal an entirely different pattern of musical taste, and furthermore they may be even predictable. The goal of this paper is to find these patterns which may help us understand how the general public, who can be regarded as a group of "average" people in a statistical sense, will respond to new stimuli.

This paper addresses three questions. One is to find statistically meaningful patterns within the data. The next question is if we can predict how long an album will stay in chart, given the first few weeks' sales data, using statistical patterns found from the first question. The last question is to see if a new album's position in chart can be predicted on a certain week in the future (such as the 5th week or 12th week), with the first few weeks' sales data. For this, we used LMS (least mean square) algorithm, a well known adaptive algorithm.

This paper uses published bi-weekly sales data from the Billboard magazine, more specifically, the Top Jazz chart. The results show some interesting correlations, one of which emphasizes the role of marketing. According to our findings, it is probably worth a good investment on marketing before starting sales of an album, since the data shows that the higher the starting position of an album is, the longer it is likely to stay in chart.

Categories and Subject Descriptors
I.5.1 [**Pattern Recognition**]: Models – *statistical models.*

General Terms
Algorithms, Measurement

Keywords
Music preference, sales data, statistical analysis, adaptive algorithm, time-series prediction.

1. INTRODUCTION
Over the years the music industry has seen a growing number of artists and recording companies, as well as many attempts to figure out the secret recipe of a hit song. Numerous analyses have been conducted from temporal, acoustical and lyrical perspectives, some of which concentrate on musical similarity and classification [1][2][3][4], though the authors have not encountered a research paper on this matter from a statistical point of view. The motivation of this paper is to detect any statistical patterns in general public's taste of music, by using Billboard Charts data and furthermore to predict an album's success.

In this paper, we use over 3 years of Billboard Top Jazz charts [5] that reflect the sales popularity of Jazz albums. The objectives here are to see if there is a statistically generic lifecycle model for an album in the genre of jazz and if there is a correlation between different parameters, such as an album's starting position in the chart with its lifespan. Also, we tried to predict an album's future, such as how long an album's lifespan will be and what will be the position of an album on a certain week in future.

Two terms should be defined before proceeding further – lifecycle and lifespan – which will be used throughout this paper. A lifecycle of an album is a trajectory of the album's weekly positions from the very first week to the very last week in Top Jazz chart, within the time period that was considered for this project. If the album happens to be off-chart for a number of weeks before coming back in to the chart, those off-chart weeks are also considered to be a part of the lifecycle. A lifespan is defined to be how long the lifecycle of an album is, in terms of the number of weeks. Again, if an album happens to be off-chart during one or more times within its lifecycle, the lifespan also includes those weeks off-chart.

2. EXPERIMENTS

TOP JAZZ

THIS WEEK	LAST WEEK	WEEKS ON CHT	ARTIST TITLE IMPRINT & NUMBER / DISTRIBUTING LABEL	CERT
1	1	20	#1 (10 WKS) MICHAEL BUBLE IT'S TIME 143/REPRISE 48946/WARNER BROS. ⊕	●
2	2	3	PAUL ANKA ROCK SWINGS VERVE 004751/VG	
3	3	41	MADELEINE PEYROUX CARELESS LOVE ROUNDER 613192	
4	4	39	CHRIS BOTTI WHEN I FALL IN LOVE COLUMBIA 92872/SONY MUSIC ⊕	●
5	5	3	JOHN SCOFIELD THAT'S WHAT I SAY: JOHN SCOFIELD PLAYS THE MUSIC OF RAY CHARLES VERVE 004360/VG	
6	6	2	HARRY CONNICK, JR. OCCASION MARSALIS 613313/ROUNDER	
7	11	5	VERA LEE 83 AND STILL PLAYING WITH THE BOYS S.D.E.G. 1954	
8	7	62	DIANA KRALL THE GIRL IN THE OTHER ROOM VERVE 001826/VG ⊕	●
9	8	3	VARIOUS ARTISTS PLAYBOY JAZZ: AFTER DARK II CONCORD JAZZ 2751/CONCORD	
10	17	57	RENEE OLSTEAD RENEE OLSTEAD 143/REPRISE 48704/WARNER BROS.	
11	12	5	JOSHUA REDMAN ELASTIC BAND MOMENTUM NONESUCH 79864/WARNER BROS.	
12	9	11	TORD GUSTAVSEN TRIO THE GROUND ECM 004123/UNIVERSAL CLASSICS GROUP	
13	10	2	WAYNE SHORTER QUARTET BEYOND THE SOUND BARRIER VERVE 004518/VG	
14	14	73	HARRY CONNICK, JR. ONLY YOU COLUMBIA 90551/SONY MUSIC	■
15	13	2	EDDIE PALMIERI LISTEN HERE! CONCORD 2276	
16	20	21	VARIOUS ARTISTS PUTUMAYO PRESENTS: NEW ORLEANS PUTUMAYO 0232	
17	16	24	DAVID SANBORN CLOSER VERVE 003095/VG	
18	22	12	VARIOUS ARTISTS VERVE//UNMIXED3 VERVE 004302/VG	
19	15	3	TERENCE BLANCHARD FLOW BLUE NOTE 78273	
20	18	42	JANE MONHEIT TAKING A CHANCE ON LOVE SONY CLASSICAL 92495/SONY MUSIC	
21	24	8	KEITH JARRETT RADIANCE ECM 004314/UNIVERSAL CLASSICS GROUP	
22	RE-ENTRY		ARTURO SANDOVAL LIVE AT THE BLUE NOTE HALF NOTE 4522 ⊕	
23	RE-ENTRY		ELDAR DJANGIROV ELDAR SONY CLASSICAL 92593/SONY MUSIC	
24	19	40	VARIOUS ARTISTS 20 BEST OF JAZZ MADACY SPECIAL PRODUCTS 5328/MADACY	
25	RE-ENTRY		BOBBIE EAKES SOMETHING BEAUTIFUL BCI 40960	

Figure 1. An example of Billboard Top Jazz chart

For this experiment, we used albums on Billboard Top Jazz charts from August 31, 2002 to June 24, 2006. The chart is a bi-weekly list of No.1 through No.25 in terms of sales rank. An example of Top Jazz chart is shown in Figure 1. The reason for concentrating on one genre was that it was believed to yield cleaner results that could provide a better insight. Also, the Jazz is a genre with unique characteristics, with a specific audience with a rather well-defined taste and that jazz audience is a more knowledgeable group, in comparison with other more popular genres, like Pop or R&B.

291 albums were considered that started and ended their lifecycle during our study period. Of course, an album that falls off the chart before June 24, 2006 can still come back to the chart, hence continuing its lifecycle. But there had to be a limit in the data set, especially since this is a first attempt to find patterns in this type of data. Perhaps this period of consideration can be expanded in future researches.

2.1 Statistical Analysis

The 291 albums showed lifespans that ranged from minimum 1 week to maximum 105 weeks. The average was 16.6 weeks and median 10 weeks. The histogram of lifespans is shown in Figure 2. Out of 291 albums, 66 albums had lifespan of 1−2 weeks and 7 albums had lifespan of over 99 weeks. We grouped the albums into 6 (octave) groups, which are 1−2, 3−8, 9−16, 17−32, 33−64, 65+ weeks.

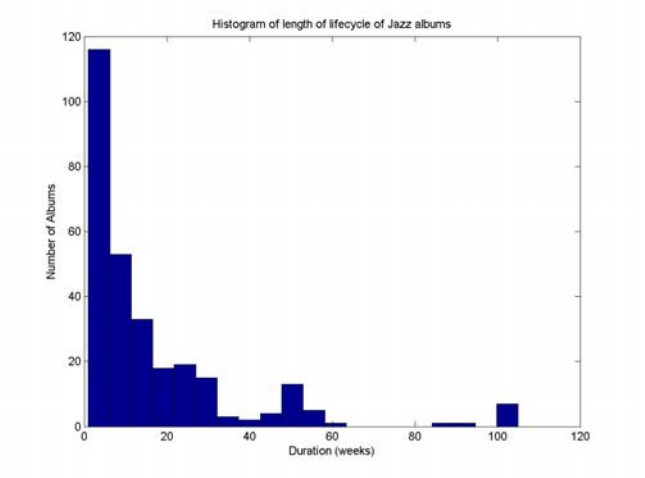

Figure 2. Histogram of lifespans of Jazz albums

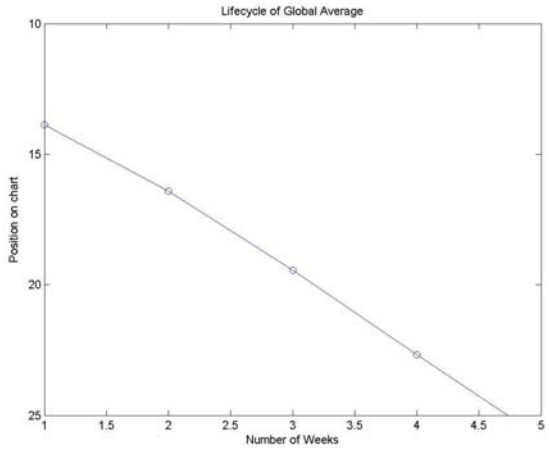

Figure 3. Lifecycle of Global Average

We used Microsoft Excel and MathWorks MATLAB™ programs for statistical analysis. Before the experiments, the authors hypothesized that we would see something close to a Gaussian curve for the generic lifecycle of an album, starting at a low position in the chart, climbing to higher places before dropping down to off chart. However, Figure 3 shows the global average lifecycle, which starts at its peak position and falls linearly down the chart. The global average lifecycle was obtained simply by taking an average of all 291 albums' weekly positions. There is a strong correlation between the starting position of an album in the chart and the duration of its lifecycle. For example, as can be seen in Figure 4, almost a half of albums that had only 1 week lifespan started at or below 20, while 3 out of the 7 albums that had over 99 weeks of lifespan started at position 1.

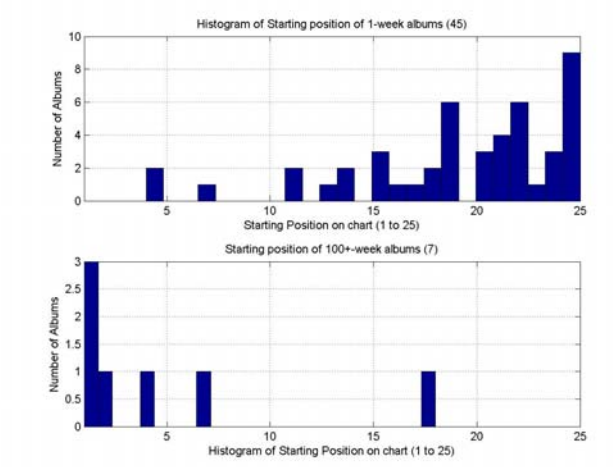

Figure 4. Histogram of starting positions of 1-week lifespan and 100+ week lifespan albums

After the 291 albums were binned into six groups according to their lifespan, we obtained each group's average lifecycle. Figure 5 illustrates average lifecycles of the groups. Note that the starting point of the lifecycle curve decrease as the lifespan increase. It is also noteworthy that nearly all the lifecycle curves exhibit a linear descent and that even when there is a deviation from the trend of linear decay (i.e. a temporary upward move in the chart position), it is never as high as the peak position that was achieved during the first few weeks of lifecycle.

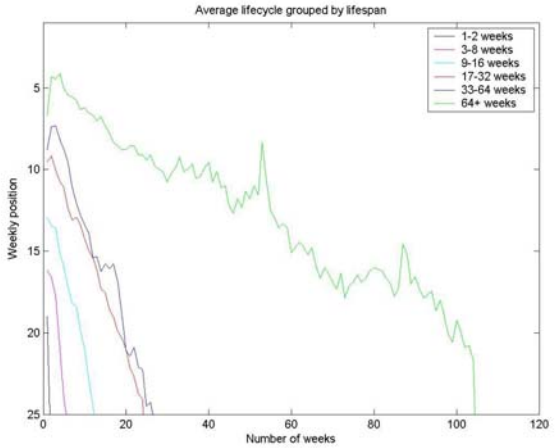

Figure 5. Average lifecycle of six groups

2.2 Nearest Neighbor Algorithm and Lifespan Prediction

One of the questions this paper answers is whether we can predict the lifespan of a new album, given its first few weeks of sales data. For this, a nearest neighbor algorithm was used. This algorithm calculates Euclidean distances between the first few weeks' sales history of a new album considered and the same number of weeks from the average lifecycle patterns from the six groups and determines the expected number of lifespan with the minimum distance.

For example, consider an album whose lifecycle (its position on the chart) was [9 15 13 11] and we are interested in how long it will be on the chart. The first group (of albums with lifespan of 1−2 weeks) is excluded from the comparison, since the length of the given data vector is 4. For each of the other groups, the test vector is compared to the first four positions of the average vector and the minimum distance obtained using Euclidean distance, along with the index of the corresponding average vector. The index points to songs with similar history, hence gives a good estimation of how long this new album will be in chart. Going back to our example, the index of the minimum distance turns out to be 2, which points to the group of 3−8 week lifespan albums

If the input vector is longer, say [9 15 13 11 10 11 8 11 12 15], it can be compared to average vectors in groups 3 to 6. The nearest neighbor algorithm predicts that it will be on chart for 9−16 weeks. The album's actual lifespan was 12 weeks (*Season* by Jane Monheit, the album on the third row of Table 1).

Table 1. Eight albums considered for lifespan estimation and their first 10-week lifecycles

Album	1	2	3	4	5	6	7	8	9	10
Got You on My Mind By Madeleine Peyroux	16	23	-	-	-	-	-	-	-	-
The Contennial Collection By Glenn Miller	14	21	18	20	25	25	-	-	-	-
Season By Jane Monheit	9	15	13	11	10	11	8	11	12	15
What Goes Around By Dave Holland Big Band	5	7	8	7	8	8	11	10	11	18
Glamoured By Cassandra Wilson	2	2	2	3	3	4	4	4	6	8
Trumpet Evolution By Arturo Sandoval	6	7	7	8	10	10	11	11	13	16
Paganini: After a Dreami By Regina Carter	9	6	7	1	4	5	4	5	6	9
Peter Cincotti By Peter Cincotti	2	2	2	2	3	1	1	2	3	2
To Love Again By Chris Botti	1	1	2	3	3	4	4	4	3	5

Table 1 lists the first ten positions of nine albums we tested for the lifespan estimation. Table 2 shows the estimated lifespans for the albums. With the first four weeks' positions only, the algorithm predicted correctly for 5 out of 8 albums (excluding the last album whose lifecycle is still ongoing), thus showing over 60% of hit rate. For the last album listed in Table 2, *To Love Again* by Chris Botti, the algorithm predicts that it will perform very well (65+ weeks) and currently it is on 34th week on chart (as of the last week of June 2006).

Table 2. Comparison of actual and estimated lifespan with the same input of different lengths considered

Album	Actual Lifespan (weeks)	Estimated Lifespan		
		With first 4 weeks	With first 10 weeks	With first 20 weeks

Got You on My Mind By Madeleine Peyroux	2	1−2	1−2	1−2
The Contennial Collection By Glenn Miller	6	17−32	3−8	3−8
Season By Jane Monheit	12	17−32	17−32	9−16
What Goes Around By Dave Holland Big Band	20	17−32	17−32	17−32
Glamoured By Cassandra Wilson	30	65+	65+	65+
Trumpet Evolution By Arturo Sandoval	44	33−64	33−64	33−64
Paganini: After a Dreami By Regina Carter	58	33−64	33−64	33−64
Peter Cincotti By Peter Cincotti	100	65+	65+	65+
To Love Again By Chris Botti	34 (as of July 2006)	65+	65+	65+

Table 2 shows that the accuracy of estimation increases as the length of the input vector increases. The album *Glamoured* by Cassandra Wilson was consistently mispredicted. Though it stayed on chart for 30 weeks in reality, the algorithm predicts that it will continue to be on chart for 65+ weeks, even when the first 20 positions were considered. The reason for this seems to be that the first trajectory of this particular album is much closer to the average lifecycle of 65+ week lifespan albums than that of 17−32 week lifespan albums. If you compare its trajectory with the trajectory on the 8th row, *Peter Cincotti* by Peter Cincotti, which lasted 100 weeks on chart, obviously the two albums show very similar positions for the first ten weeks.

2.3 Adaptive Algorithm and Lifecycle Prediction

We also considered whether we can predict on a particular album's performance can be predicted for the next week, given the first few weeks of sales data. We used the Least Mean Square (LMS) algorithm [6], a well-known adaptive algorithm. An adaptive FIR filter with 105 adjustable weights (which corresponds to the longest lifespan amongst the 291 albums) was trained with the data of 291 albums, with a parameter of the desired week (for example, week 5 or week 30), hence creating 105 different models for each of 105 weeks (given as the parameter). The results vary slightly with different desired week values considered, but the model predicted that by week 30 most albums will be off chart. The last seven albums in Table 1 were considered for this experiment. The results are shown in Table 3.

Table 3. LMS prediction of album positions on the 5th, 10th and 30th weeks

Album	5th week		10th week		30th week	
	Est.	Actual	Est.	Actual	Est.	Actual
Season By Jane Monheit	19	8	23	18	Off	Off
What Goes Around By Dave Holland Big Band	19	10	23	15	Off	Off
Glamoured By Cassandra Wilson	18	3	22	8	Off	19
Trumpet Evolution By Arturo Sandoval	19	10	23	16	Off	Skipped
Paganini: After a Dreami By Regina Carter	18	4	22	9	Off	15
Peter Cincotti By Peter Cincotti	18	3	22	2	Off	9
To Love Again By Chris Botti	18	3	22	5	Off	3

For most albums, the prediction is pessimistic. Still, this is understandable considering the fact that the median of the lifespans of all albums is 10 weeks, so the model expects most albums to be near the cutoff position (that is 25) by the 10th week and off chart by the 30th. For the last album (*To Love Again* by Chris Botti), which is still very strong in chart, the model expected it to be at 18th by week 5, at 22th by week 10 and off chart by week 30, which is quite different from real data.

The discrepancy between what's expected from the system and what's observed in real sales data probably comes from the fact that the model which was built for the experiment was not complex enough to handle the overall complexity of the real data. We presume that a better estimation may be possible with a different adaptive technique such as machine learning. We also plan to extend the scope of this project to other charts and see if similar patterns can be found.

3. ANALYSIS OF RESULTS

We performed a statistical and adaptive analysis to find patterns in Billboard chart data as a measure of general public's response over time. We considered 46 months' data from Billboard Top Jazz chart. This paper focused on one genre, jazz, since we expected cleaner and more coherent patterns would give better insights.

There were a couple of assumptions made about our data. Since the exact off-chart positions cannot be obtained, the chart data were linearly extrapolated for those positions. Also the 291 albums are considered to have ended their lifecycle by June 24, 2006, though some of them may still come back to the chart after skipping a number of weeks.

Interesting patterns were observed as a result of statistical analysis. The 291 albums considered were categorized into six different groups, according to their lifespan. Then an average lifecycle was calculated for each group. As noted earlier, there is a strong correlation between the starting position of an album and how long it stays on the chart. Another interesting finding was that all groups showed a consistently linear decay in lifecycle, after debuting at their near-peak positions.

Using the statistical analysis result, a new album's lifespan and lifecycle were estimated. To calculate how long an album is likely to stay on the chart given a few weeks of sales data, a nearest neighbor algorithm was used. LMS (least mean square) algorithm, a famous adaptive algorithm, was used to predict the

next week's position of an album. Both estimates turned out to be a bit optimistic, though there were some cases where the estimates were closer to real data than others.

4. CONCLUSION AND FUTURE WORKS

We considered 291 albums from 46 months of Billboard Top Jazz chart. After categorizing the 291 albums into six categories, an average lifecycle was calculated for each category. This became the basis for lifespan and lifecycle predictions

Some interesting patterns were found from the experiments. The most interesting find from this project was that the "average lifecycle" for each group showed a very similar linear decay from the onset, in contrast to our initial hypothesis that the average lifecycle would be something close to a Gaussian bell curve. Also, a strong correlation was found between the starting position of a lifecycle and its lifespan. For example, albums with very short lifespan tend to debut below position 20, while long-lasting albums tend to debut above position 5. These results may imply that marketing before the start of sales of an album is quite important, since these patterns indicate that the higher the starting position is, the longer it will stay in chart.

Using both the nearest neighbor algorithm and the LMS algorithm, a prediction was attempted on a new album's lifecycle and lifespan. With a limited set of data we considered, the predictions were of mixed results, some very different from the real data and some others closer. There are various other adaptive techniques available for analysis. We plan to try other techniques, such as machine learning algorithm techniques, for a better prediction. HMM (hidden markov model) and PCA (principle component analysis) may provide a better insight.

This experiment produced some very interesting result even though it was done with a very limited set of data. We may want to categorize the data into more groups, therefore being able to analyze more detailed patterns in each group. Perhaps it will yield a result that can be fitted with a line.

We believe that this analysis showed a unique characteristic from the fact that the considered genre was Jazz. Jazz is a genre of both vocal and non-vocal music and the popularity in this genre is believed to change more slowly than other genres such as Pop or HipHop. Certainly other genres will have other unique characteristics, which should be left to future work. We do not intend to repeat this analysis on every chart in Billboard magazine, but would like to consider a bigger set of data, from possibly mixed genres. We would also like to further study to see whether there are different patterns in vocal music from instrumental music within the genre of Jazz.

Also, some new factors, which were not considered for this project, may need a consideration such as skip rate (how often an album goes off chart) and start date (the first day when the album enters the chart). These may help improve our models better handle season-specific albums, for Christmas or Valentine's Day, for example. With these improvements, we hope to reveal many more interesting patterns.

5. ACKNOWLEDGMENTS

We thank Professor Bernard Widrow of Electrical Engineering Department at Stanford University for his guidance in adaptive algorithms.

6. REFERENCES

[1] Cano, P., Koppenberger, M., and Wack, N., Content-based audio recommendation, In *Proceedings of the 13th annual ACM international conference on Multimedia,* 2005, 211−212.

[2] Berenzweig, A., Logan, B., Ellis, D.P.W., and Whitman, B., A large-scale evaluation of acoustic and subjective music similarity measures, In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, 2003b, 103−109.

[3] Foote, J.T., Content-based retrieval of music and audio, In *Proceedings of SPIE*, 1997, 138−147.

[4] Dhanaraj, R., and Logan, B., Automatic Prediction of Hit Songs, In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, 2004, 488− 491.

[5] Billboard Magazine, Top Jazz chart, from 08/31/2002 to 01/07/2006.

[6] Widrow, B., and Stearns, S., *Adaptive Signal Processing*, Prentice Hall, 1985.