Forecasting the Sales of Music Albums:
A Functional Data Analysis of Demand and Supply Side P2P Data

Il-Horn Hann[•], JooHee Oh[⊗], and Gareth James[⊗]

Abstract

We predict the sales of music albums by utilizing demand and supply side P2P data using a functional data analysis (FDA) approach. We find that the characteristics of the functional form of downloading behavior explain first-week sales by more than 60% after controlling for album characteristics. By updating our forecasts from 4 weeks to 1 week prior to the album release date, we examine the dynamic changes across different quantiles of the sales-distribution for the demand- and supply-side P2P data. We find that the gap between downloading effect on sales among high-quantile vs. low-quantile albums reach the highest level one week before the release date.

Introduction

A recent article in The Economist (Simmonds 2008) reports how firms can benefit from P2P piracy by gaining a real-time glimpse on the consumers' music tastes and preferences. Our own interactions with industry experts have shown an ambivalent thinking regarding P2P networks. While many executives almost reflexively see P2P networks as 'the enemy,' some are seeking ways to use them to their advantage. For example, movie studios have started to seed P2P networks with movie trailers as part of their viral marketing strategy. In or work, we take advantage of the appearance of music files well before the official launch date to predict the success of music albums.

Pre-release forecasts of album sales have historically been poor. Good predictors for music sales are hard to come by, historically, the best predictors were 'known quantities' such

[•] R.H. Smith School of Business, University of Maryland

[⊗] Marshall School of Business, University of Southern California

as the reputation of an artist and the number of gold and platinum albums (see Lee, Boatwright, and Kamakura 2003). This makes the forecast of new releases especially difficult. In our research, we utilize longitudinal P2P traffic data to forecast album sales.[1] For this purpose, we employ functional data analysis (FDA), an empirical method developed by Ramsay and Silverman (2005), to examine the functional characteristics of demand- and supply-side P2P data. By using functional data analysis, we capture the similarities and differences of functional shapes such as trends or curvatures across downloading histories. Our approach is similar to Foutz and Jank (2007) who used online virtual stock market prices to predict movie box-office.

Data

We have three different sources of data. Our data consists of downloading data from the Ares peer-to-peer network, sales data for newly released albums in the Billboard's Top 200 albums, and album specific characteristics. We used downloading data from P2P network and album specific characteristics to generate first-week sales of newly released albums.

Our P2P data comes from a leading P2P anti-piracy and marketing solutions provider. The company actively monitors all major P2P networks and collects data on downloading and sharing activities and provides services to all major record labels and movie studios. We obtained downloading and sharing data from the Ares peer-to-peer network for the time period of April 2007 to September 2007. Ares was chosen primarily for three reasons: popularity of the P2P network, breadth of coverage of the P2P network, and ability to monitor downloading and sharing activities. The raw data for the network is about 60GB per month; this includes data for downloading activities as well as for sharing. The data includes lists of the name of file, title of album, artist name, genre of music, unique hash of the file, and user IP address. We have two types of file-sharing data, referred as hashes and sources. Each song file is identified by several unique hash codes. Hash-based downloads data is based on daily hash-requests number for each song files. While hash-based downloading measure represent demand-side song downloads, source-based downloading measure represent supply-side of

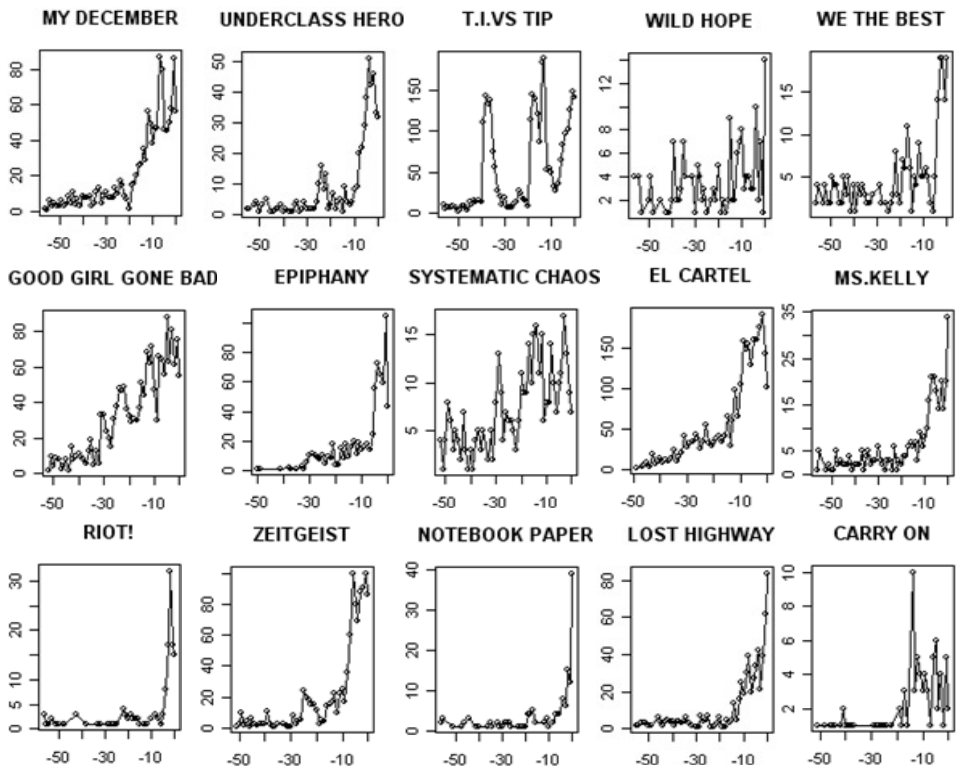
[1] Bhattacharjee et al. (2007) use longitudinal data for a survival analysis.

downloads. Source-based downloading measure is based on total units of global nodes where song-file is available and provided for file requests. Hash-based and Source-based file downloads data is processed as daily time-series for each album from as early as two-months prior to the release date. Figure 1-1 and Figure 1-2 each illustrates average number of daily hash-based and source-based downloads of songs in the albums from two-months prior to the release.

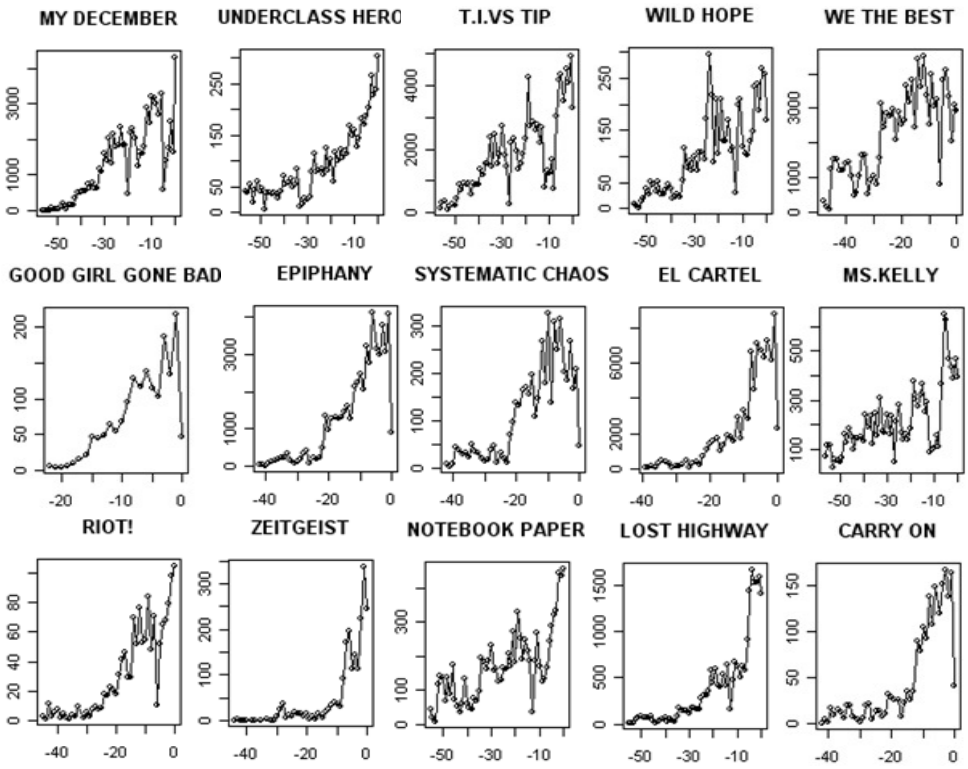
For the weekly sales of newly-released albums, we selected new albums from the Billboard 200 albums chart in the time period of May 2007 to September 2007 provided by SoundScan Inc. [2] Among around 1,000 albums that appears at least once in the chart from May to July 15, 2007, we selected only newly-released albums on that week, this represents about 20 albums per week. We deleted movie soundtracks or re-entered albums due to their atypical sales patterns. We included albums that has minimum three days of downloads prior to the release date from the Ares Peer-to-Peer network. Here, we study total 172 albums of sales prediction using P2P downloading pattern and album characteristics from two months before the release. We ended up from 75 to 152 albums from a month to one week prior release for the pre-release forecasting. We included all the albums that were downloaded more than three days from two-months prior release and updates the album set every week that appears new in every following week. At one month prior to the release, we have around 75 albums, 102 at three-weeks prior stage, 129 at two-weeks prior and 152 albums a week before release date. Our data includes albums which result first-week sales range from as low as 3,772 units to as high as 622,827 units.

[Figure 1-1: P2P Downloads of Albums based on Hash-Request]

[2] The data is based on sales from 14,000 retail outlets, including 40 different chains, 11 mass merchandisers, and over 600 independent retail locations. The average weekly transaction amount to 9-10 million albums and the data are sent via model from point-of-sale registers in the stores.


MY DECEMBER
UNDERCLASS HERO
T.I.VS TIP
WILD HOPE
WE THE BEST
GOOD GIRL GONE BAD
EPIPHANY
SYSTEMATIC CHAOS
EL CARTEL
MS.KELLY
RIOT!
ZEITGEIST
NOTEBOOK PAPER
LOST HIGHWAY
CARRY ON

[Figure 1-2: P2P Downloads of Albums based on Global-Source]



We also collected following set of album characteristics variables that have been known as influential to forecasting of music sales. They are genre category of an album, number of daily comments from the YouTube website before release date, gender of the artist, total number of albums released by an artist, the existence of single albums before release. While previous literature used weekly number of Radio Airplay as a proxy for marketing variables, we utilized daily number of comments from the YouTube website. We also created a dummy for the albums where the single album exists prior to the release. Total number of previous albums of the artist, genre category of music, and gender of artists variables have been known to be influential on music sales. We separated genre variable as Rock, Rap, R&B, Pop, Country and Gospel. Gender variable of the artist is based on male, female and group. While we also collected average ratings from All Music Guide (AMG) website, a mixture of professional and commons ratings for the album, we excluded the variable because the ratings were not available prior to the release. Table 1 summarizes characteristics of the data.

[Table 1: Summary of Album on P2P down2loads, First-week Sales and Characteristics]

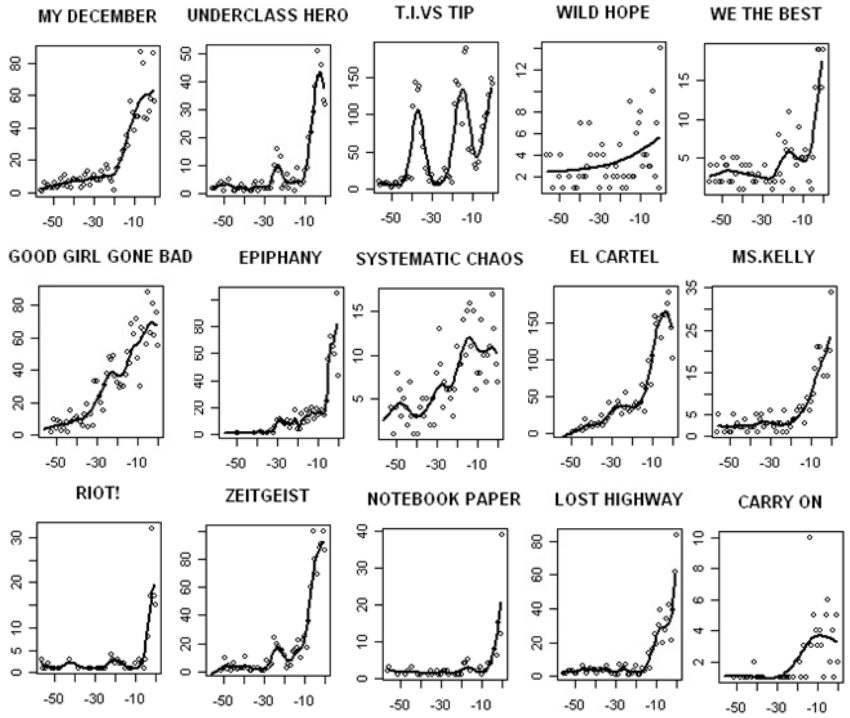
Variable	*Obs.*	*Mean*	*Std dev.*	*Min*	*Max*	*Median*
Rank	172	51.37	49.10	1	198	40
Sales	172	51738.77	86138.94	3772	622827	17054
Ave. hash	172	6.95	18.32	1	221.94	2.46
Ave. source	172	228.99	839.57	1	9540.78	21.93
YouTube	172	15.36	45.872	0	489.74	2.48
Total albums	172	7.62	8.73	1	63	5
Single	172	0.36	0.48	0	1	0
AMG rating	140	3.53	0.64	2	4.5	3.5
Rock	172	0.65	0.48	0	1	1
Rap/R&B/Pop	172	0.31	0.46	0	1	0
Country/Gospel	172	0.04	0.20	0	1	0
Male	172	0.311	0.47	0	1	0
Group	172	0.58	0.50	0	1	1
Female	172	0.11	0.32	0	1	0

Empirical Pattern of Piracy Curves

This study attempts to examine functional characteristics of piracy curves based on P2P downloading data for pre-release forecasting purpose. We analyze shapes of piracy curves for the newly launching albums to understand different types of album information. In particular, we relate individual album level information grounded on characteristics of piracy curves to its implication on first-week sales. Measuring dynamics of piracy dispersion in level, speed, and acceleration along the time helps understanding shape characteristics of individual curves. We compare shapes of individual curves to the shape of aggregated market-level piracy curve. Information on shape characteristics of piracy-curves of high-sales albums compare to the aggregated market-level is useful for forecasting. Also information on downloading patterns of low-sales albums with respect to the market-level realizes not identical to that of high-sales albums. Based on Functional Data Analysis (FDA), which involves Functional Principal Component Analysis (FPCA), our analysis heavily relies on the characteristics of shape and covariation degree of individual curve with respect to market level piracy curve. Our approach involves three major steps based on the Ramsay and Silverman (2005).

First, we derive smooth downloads paths from the observed downloading history data before release. We analyze functional process of each hash and source based P2P traffic as a measure of early information dispersion for newly launching albums. Despite their continuous nature, limitations in measurement capability allow us to record only discrete, such as daily, observations of these curves. Thus, the first step is to recover, from the observed data, the underlying continuous functional objects by using smoothing methods. We have $x(t_i) = \mu(t) + f(t_i) + \varepsilon(t_i),$ where ε_i is the unobserved error component and the sequence of x_i are the observed outcomes of the P2P downloads. A major underlying premise is that the underlying process, $f(t_i)$, is differentiable function to some order. We applied a flexible and computationally efficient technique called the penalized smoothing spline (Ruppert et al. , 2003; Ramsay and Silverman, 2005). The penalized smoothing spline f minimizes the penalized square error (PNSSE), $PNSSE_\lambda(f \mid x) = \sum_i [x_i - f(t_i)]^2 + \lambda_i PEN_2(f)$. The degree of departure from a straight line is measured by defining a roughness penalty $PEN_k(f) = \int \{D^k f(t)\}^2 dt,$ where $D^k f$, $k = 1,2,\cdots,$ denotes the k-th derivative of the function f. The smoothing parameter λ controls the trade-off between the data-fit, as measured by the summation on the left-hand side of above equation, and the local variability of the function f, measured by the roughness penalty PEN_k.

[Figure 2: Penalized smoothed function of P2P Downloads]

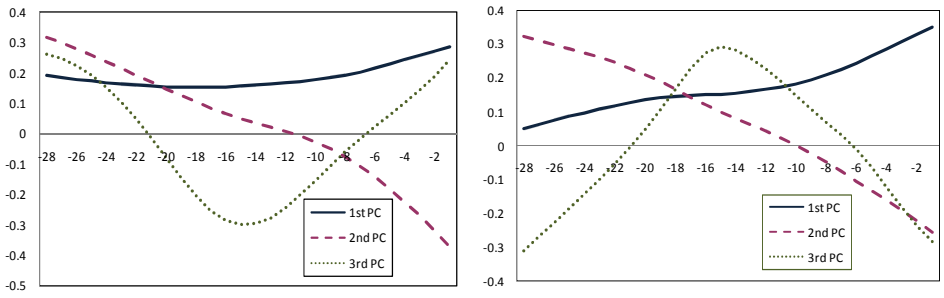


Second, we extract the key shape characteristics that capture the similarities and differences across these downloading histories from the functional Principal Components Analysis (FPCA). We summarize characteristics of level, first-derivative, and second-derivatives of hash-based download function and source-based download function. Functional principal component analysis is similar to the principal component analysis except for that we now operate on a set of continuous curves rather than discrete vectors. For downloading curves of albums, $x_1(s), \ldots, x_n(s)$, we can decompose in the following form,

$x_i(t) = \mu(t) + \sum_{j=1}^{\infty} s_{ij} PC_j(t),\ i = 1, \ldots, n$ using principal components curves. We find a corresponding set of PC curves $PC_j(s)$ that maximize the variance along each component and are orthogonal to one another. We find the PC function $PC_j(s)$ whose PCS $S_{i1} = \int PC_1(s) x_i(s)\, ds$ maximize $\sum_i S_{i1}^2$ subject to $\int PC_1^2 ds = \|PC_1\|^2 = 1$. The next step involves finding $PC_2(s)$ for which the PCS $S_{i2} = \int PC_2(s) x_i(s)\, ds$ maximize $\sum_i S_{i2}^2$ subject to $\|PC_2\|^2 = 1$ and the constraint $\int PC_2(s)\, PC_1(s) ds = 0$.

Principal component (PC) curves illustrate the direction of greatest variability in the curves about their mean. From the shape of PC curves, we can contrasts the pattern of variation level in album downloads across the period. For example, in Figure 3, first PC curve from hash-based downloading curve depicts increasing convex curve around 2 weeks before release following after declining trend. First PC curve from source-based downloading curve captures increasing convex curve around 2 weeks before release following after increasing concave trend. First PC curve of demand-side downloading curve shows that variation of daily song file demand across albums, starting from positive level, slowly decrease until two weeks prior to the release where it turns to increase towards the launching in average. While the variation of daily demand for the song files in the early-stage capture the interest-gap among early-adopters who represent unique preference on yet unknown songs, as more information available approaching two weeks prior stage, these preference-oriented demands variation decreases. Variation among song file demands start to rapidly increase from two-weeks prior to the release date as the early information on songs diffuse with externality on its popularity. In turn, variation in supply-side of song-files increases over time with different speed. While we have increasing variety of song files on supply-side as time passes, variation across song-file in supply-side increase with high-speed at last two-week period compare to the early period. As demand-side externality for popular song-files speed up during last two week, steeper curve slope of supplying-side of song files in last two week shows the cumulative degree of expansion for the gap of file dispersion by their popularity.

[Figure 3: First three PC of Demand- and Supply-side of Piracy Level curves]



The decomposed vectors of downloading variation in Figure 3 allow us to explain general shape of downloading curves and the direction of curve variability across time. While the first

principal component curve explains almost 70% of variation, the second and third components carry information regarding remained direction of different variability in the downloading curves. The second PC curves contrast the variation of downloads for song files in early and late period. It depicts decreasing pattern of downloading curves at early vs. late stage or period in both demand- and supply-side. The third PC curve illustrates variability of downloading patterns for the mid-term around three weeks to two weeks prior to the release. While demand side for song files downloads curve depict convex shape hitting the bottom around mid-term, supply side for song downloads curve shows concave shape hitting the top around two weeks prior to the release. Demand-side file downloading curves, grounded on hash-requests queries, show that demand-side information flow on album remains high at the early and last stage where the variation of interests across albums is also high. Supply-side file downloading curves, grounded on units of source nodes on the network, illustrate the shape of cumulative amount of information available for downloadable song files at each period where the variation of interests across albums remain high or low. Each PC curves in Figure 3 represent the respective direction of variability of downloading curves in early, mid- and late-term across time periods.

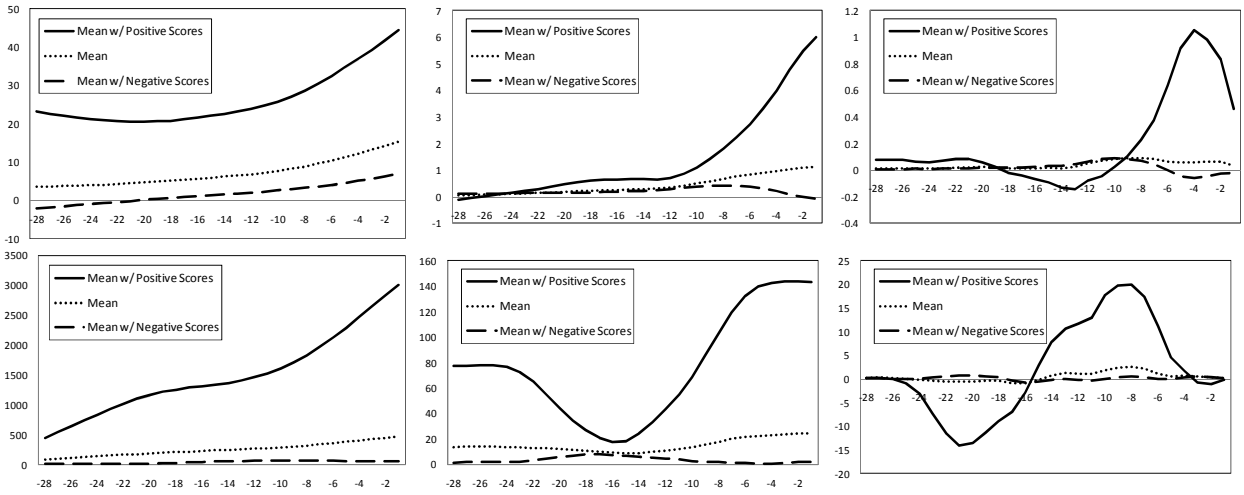
In Figure 4, we show an alternative way to visualize PC curves accompanied with mean curve. We illustrate mean curve and first principal component (PC) of downloading curves with mean separately on albums that have positive value vs. negative value of scores, $\mu(t) \pm \bar{s}_j PC_j(t)$ where $\bar{s}_j$ is the mean scores of positive value vs. negative value separately. In each chart, we illustrate mean curve which depicts market-level downloading curve, albums with positive value of PCSs and albums with negative value of scores. Real line in Figure 4 designates a typical shape of downloading curves for the albums with positive value of score. Dashed line captures the typical shape of downloading curves for the albums with negative value of scores.

The value of mean positive scores (PCS) is represented as the gap between real-line and dotted-line (mean-curve) in each chart. Also the average value of negative scores is the gap between dashed line and dotted-line (mean-curve) in each chart. In this aspect, real line and dashed line at Figure 4 reveal the average level of PCSs for the albums with positive vs. negative

value separately in the direction of first PC curve. The value of score (PCS) designates the degree of covariation of downloading curves with respect to the direction of PC curve. This means that albums with higher scores accompany downloading curves that follows highly covariating patterns with respect to the first principal component. Hence, downloading curves of albums with positive scores follow the pattern of first principal component with positive scale of covariation. In contrast, downloading curves of albums with negative scores covariate with respect to the opposite direction of the first PC curve. In the following part, we relate the value of scores (PCS) to the first week sales. If the relationship between PCS of level function and the sales is positive, we can analyze the result as that albums with positive value of score will result high sale after release and vice versa. Hence the value of PCS designates first week sales.

The first raw panel of Figure 4 illustrates the raw (Level), first-derivative (Velocity), and second-derivative (Accelerator) of demand-side downloading curves. The second raw panel curves shows the same based on supply-side downloading data. Real line and dashed line in each chart can be analyzed in three aspects: respective level of downloading amount, the shape of curve change across the period, and the size of incremental gap with respect to the mean curve of the market.

[Figure 4: First PC curve of Level, Velocity and Accelerator of Demand- and Supply-side Piracy curves]



The first column of charts shows the general shape of downloading curves in level. Albums with positive value of score, in general, have higher demand and supply in downloading all over the period while albums with negative value of score, have lower demand and supply in downloading. Albums with positive value of scores have increasing pattern following after decreasing towards the release while albums with negative value of scores remain relatively stagnant. Compare to the mean curve of the market, deviation in file-demands and supply across albums remains high for the albums with positive value of score. The degree of variation in file-demand across albums follow increasing pattern before release following after decreasing pattern as captured by the incremental gap with mean curve. Supply-side of downloads variation follow increasing pattern over the period as captured by the incremental gap in the second raw. This result implies that if the sales coefficient of level PCS is positive, then albums with positive scores will result high sales and the deviation between high-sales albums will be larger than the sales deviation of low-sales albums. The predicted sales gap among albums with high PCS vs. low PCS shows positive and slightly U-patterned difference in demand-side of downloads. In contrast, supply-side downloads curves predict sales gap based on value of PCS will show increasing pattern towards release date. For this difference in information that PC curves capture from demand- and supply-side data, we combine the PCS from both demand- and supply-side PC curves in the model.

The second column of charts shows the first-derivative (velocity) of downloading curves. Demand-side downloading curves show that albums with positive score grow dramatically faster than average market level downloads. It also shows that demand for song files for albums with negative score remain lower and significantly slower down from the market level from ten days prior to the release than that of albums with positive value of scores. Supply-side downloads growth remains higher than market-level until it slacks down and rapidly back up during last two-weeks. These velocity curves distinctively captures turning point of shapes around two weeks prior to the release date. Incremental gap of velocity curves of album with positive value of score compare to that of market mean curve shows exponentially increasing difference from around 2 weeks prior to the release. These shape characteristics implies that if the sales coefficient of velocity PCS is negative, then albums with positive value of velocity

scores will result lower sales than market average and sales variation among albums predicted to have low-sales will be higher than deviation among albums predicted to result high-sales. While the demand-side velocity curve indicates this information gap between the group of albums with positive vs. negative value of scores from around two-weeks prior to the release, the supply-side velocity curve captures this difference information between groups from very early-stage and late-stage.

The third column of charts illustrate the second-derivative (Accelerator) of demand- and supply side downloading curves. While accelerating speed of market-level downloads remains stagnant for overall period, albums with positive accelerator scores show big jump in both direction of curves. Compare to the level- and velocity- curves, accelerator curves convey information from its big turning shape change around the mean curve to the sales. This implies that if the sales coefficient of accelerator PCS is positive, albums with positive accelerator PCS will result high sales and their 2nd derivative of downloading curve will show dramatically turning shape around the mean-level downloads. Hence one can predict the level of sales by measuring deviation of accelerator curve in absolute term around 0. As we examined from Figure 4, each level-, velocity- and accelerator of downloading curve conveys different type of information based on their characteristics.

Description of Piracy curves on Sales: Functional Regression

We use functional regression to predict first week sales using PCSs obtained from previous steps. Let $x_i(t)$ be the smooth spline representation of the i^{th} downloading curve for the album observed from two months prior release. Let y_i represent the first week sales of an album after release. Functional regression establishes a relationship between predictor, $x_i(t)$, and the item to be predicted, y_i, as $y_i = f(x_i(t)) + \varepsilon_i,\ i = 1,\ldots,n$. While this equation is difficult to work directly because $x_i(t)$ is infinite dimensional, however, for any function f there exists a corresponding function g such that $f(x(t)) = g(s_1, s_2, \cdots)$ where $s_1, s_2 \cdots$ are the principal component scores of $x(t)$. We use this equivalence to perform functional regression with the functional principal component scores as the independent variables. This approach is related

to principal components regression which is often used for non-functional, but high dimensional, data. The simplest choice for g would be a linear functional in which case above equation becomes

$$y_i = \beta_0 + \sum_{j=1}^{D} s_{ij}\beta_j + \varepsilon_i \quad for\ some\ D \geq 1.$$

A somewhat more powerful model is produced by assuming that g is an additive but non-linear, function (Hastie and Tibshirani, 1990). In this case, above equation becomes as follows where g_j's are non-linear functions that are estimated as part of the fitting procedure.

$$y_i = \beta_0 + \sum_{j=1}^{D} g_j(s_{ij}) + \varepsilon_i$$

One advantage of using above equations to implement a functional regression is that once the scores (PCS) variables, s_{ij}'s have been computed via the functional PCA, we can then use standard linear or additive regression to relate y_i to the principal component scores. We can also extend the model by augmenting covariates that contain information about the scores beyond the principal components, such as album characteristics or marketing variables. In the forecasting part, we develop two-types of functional regression by augmenting market information on the principal component scores.

We apply median and nonlinear functional regression on selected variables to identify shape characteristics that have significant impact on first week of music sales. In this last step, we selected first PCS (principal components scores) which explain more than 70% of variation calculated from the raw and derivative functions as independent variables in the model. We combined PCS variables with album characteristic variables such as the total number of previous albums and average daily number of YouTube comments to explain first-week sales of newly launching album. We generate pre-release forecasts of sales weekly from as early as a month to a week before release date. To compare our forecasting models, we calculated mean absolute percentage error (MAPE) with one-album cross-validation (CV) for each model. Instead of separating training set as ten-fold methods, we cross-validated each album in the album set by treating remaining set as training set. This individual album-based cross-

validation enabled us to generate unique and invariable measure of MAPE with respect to the sample.

Estimation Result

We present estimation result of median regression based on albums characteristics and scores (PCS) obtained from functional principal component analysis (FPCA). We selected average daily number of the YouTube comments, total number of previous albums that an artist has, and dummy variable for the existence of single albums as album characteristics which turned out to be significant on sales prediction. From four weeks prior release to a week prior stage, the adjusted R-square and Psuedo R-square of our model increase significantly. While adjusted R-square of the model about a month prior release explain first week sales 56%, it increase to 73% at a week prior release.

In Table 2, characteristics variables for the albums show highly positive relationship to the first week sales at each stage. All variables turn out to be highly significant at most stage in the period except YouTube variable at three weeks prior model and Single variable at four weeks prior to the release. We have six principal component score (PCS) variables in the model each representing demand- and supply side information of level, velocity and accelerator downloading curves. All PCS variables that capture shape characteristics grounded on FPCA turned out to be significant and influential on first-week sales at each stage after controlling the effect of album characteristics information.

Score (PCS) variable based on level curve of both demand- and supply-side downloading data shows positive relationship on the sales. Degree of covariation to the same direction implied by corresponding PC represents our unit increase of independent variable based on PCS. A scalar increase of covariation level of demand-side downloading curve with respect to the same direction of level PC curve increase about 380 units of sales a month prior. In turn, a unit increase of covariation of supply-side downloading curve on the direction of PC-curve increases 36 units of sales a month prior.

The effects of unit-change on PCS variables based on demand-side downloading curves and velocity of supply-side downloading curves on sales show increasing pattern from 4 weeks prior to 2 weeks prior stage and turn down at a week prior release. In contrast, a unit-change of supply-side downloads level curve on sales decrease from 4 weeks prior to 2 weeks prior stage and shows upturn to a week prior release. While the sales coefficient of median regression in Table 2 measures effects of unit-change of each PCS variables on sales, it cannot capture overall influence of each PCS variables on sales. Value of each PCS variables also affect to the sales multiplied by sales coefficient. To compare the effects of PCS variables on sales, Table 3 summarizes the effect of each PCS variable on sales around average value of score (PCS).

[Table 2: Estimation Result of the model at each stage prior to the Release]

	Variables	*Week-4*	*Week-3*	*Week-2*	*Week-1*
R-square	Psuedo R2	0.287	0.273	0.36	0.432
	R2 (OLS)	0.602	0.536	0.596	0.749
	Adj. R2	0.56	0.501	0.573	0.733
Album Charac.	Cons	32805.889 (3.58**)	24361.58 (4.41**)	21993.93 (10.07**)	29256.34 (11.19**)
	YouTube	701.508 (3.04**)	95.50087 (0.84)	396.976 (6.84**)	213.5689 (4.03**)
	Total#	2035.346 (2.97**)	1575.133 (3.95**)	1708.043 (9.95**)	1003.294 (5.06**)
	Single	4972.167 (0.45)	27304.61 (4.31**)	16013.85 (5.57**)	9628.87 (2.61**)
Hash	Level	382.07 (1.57*)	579.6362 (3.53**)	936.449 (22.01**)	398.908 (9.83**)
	Velocity	-11267.95 (-1.98*)	-9153.208 (-3.34**)	-8970.01 (-27.58**)	-13425.32 (-12.42**)
	Accelerator	NA	NA	NA	5885.876 (14.54**)
Source	Level	36.440 (9.55**)	10.44 (4.72**)	2.074 (2.53*)	12.75649 (18.48**)
	Velocity	-719.526 (-8.36**)	-173.92 (-5.34**)	-14.98 (-1.93*)	-57.37263 (-14.74**)
	Accelerator	NA	NA	NA	3.466532 (1.3)

Table 3 summarizes the effect of each PCS variable on sales by multiplying mean value of score (independent variables) on the unit-effect of score value increase (beta-coefficient). In the early stage from 4 weeks to 3 weeks prior-release, PCS obtained from supply-side downloading curves are more influential on sales compare to the PCS based on demand-side downloading information. However, as we approach to the release date from around 2 weeks prior to the release, demand-side downloading curves affect first-weeks sales more than supply-side information around average level of score value. In a week prior to the release, instead of downloading level curves, velocity curve of downloading data reveals the most influential on sales. The sign of effects of PCS variables on sales is negligible in Table 3. Average mean value of PCS variables were very small number close to the 0 which range from a big negative value to a big positive value. For this reason, we compared absolute term of its effect on sales to compare relative influence of PCS variables on sales.

[Table 3: Mean effects of PCS variables on Sales]

Mean Effect (x 10^{-4})	*Variables*	*Week-4*	*Week-3*	*Week-2*	*Week-1*
Hash	Level	1.36	-5.65	0.82	-0.03
	Velocity	-3.57	1.60	-0.64	-4.34
	Accelerator	NA	NA	NA	1.24
Source	Level	-7.62	11.93	0.67	-0.42
	Velocity	-0.41	-3.13	-0.43	-0.24
	Accelerator	NA	NA	NA	0.03

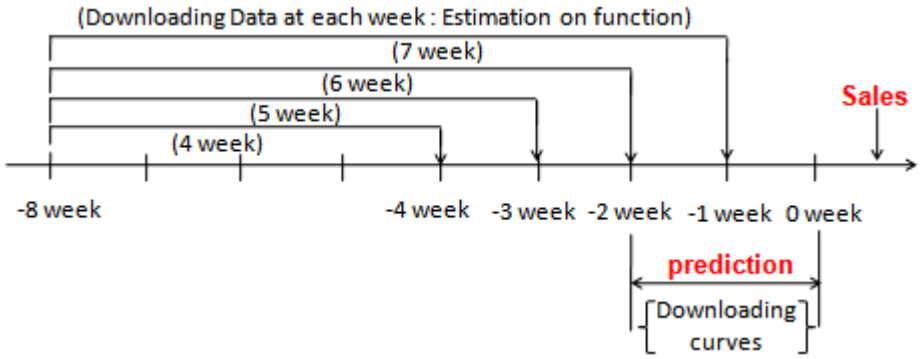
Pre-launch Forecasting with Cross-Validation

We present the performance of our proposed model based on augmented functional regression using cross-validated measure of MAPE. Table 4 presents out-of-sample forecasting results based on MAPE (Mean Absolute Percentage Error) using median regression by updating the information from one month to a week prior to the release date with cross-validation. We also applied non-linear regression based on generalized additive model (GAM) and ordinary least squares (OLS), however, median-regression outperformed the others in forecasting based on MAPE. Overshooting was severe in two other methods. We present and compare all our forecasting result based on cross-validated measure of MAPE for hold-out sample. Cross-validation method based on an album rotates sampling each and every album in the data set as

hold-out and treat remaining set as training sample. This unique album-based cross-validation enables us to obtain hold-out set invariant, unique measure of errors to calculate MAPE.

Figure 5 illustrates our forecasting model. At each stage from four weeks to a week prior to the release date, we use available set of downloading data to estimate coefficients of download functions. Instead of using estimated curves of past data, we used downloading curves for last two-weeks period predicted from the downloading functions from available data.

[Figure 5: Description on Forecasting Model]



We found the last two-week period of downloading data more informative to generate sales forecasts. At a month prior to the release, however, downloading data for last two-week period are not available. Hence, we predicted downloading data for the last two week period using the estimated functions based on available data from 8weeks prior to a month prior stage. Table 4-1 summarizes our pre-release forecasting results at each week starting from a month prior to the release. It shows decreasing pattern of MAPE based on average measure as we update information from downloading data every week. While average measure of MAPE is as high as from 100% to 200% of sales, median-measure of MAPE ranges as low as from 48% to 73%. High value of average-measure of MAPE reveals the variety of albums with sales that ranges from very low to very high in our data sample. In our data sample, we have albums that realize MAPE from as low as 0.2% to as high as 1900% of sales. This range of MAPE implies that any MAPE forecasting result that is not based on our cross-validation method for each album generate only an instance of MAPE in this range depending on the holdout set. Our unique and

holdout-invariable measure of MAPE result shows as low as 48. 4% based on median from two weeks prior to the release.

[Table 4-1: Proposed Model - Augmented Functional Regression]

Cross Validated MAPE				
	Hold-Out sample			
	Min	Max	Median	Average
Week -4	0.017	13.182	0.687	2.026
Week -3	0.038	19.777	0.737	1.822
Week -2	0.004	12.948	0.484	1.312
Week -1	0.002	7.988	0.509	1.057

Table 4-2 supports our forecasting model illustrated in Figure 5. Benchmark model in Table 4-2 results better prediction for the sales which used real downloading data for last two-week period although the data is not really available at 4 weeks prior stage. Our forecasting model is built on this benchmark model using predicted data for the period where the downloading data seems the most informative in forecasting sales. A week prior release, our model using predicted data, performs even slightly better than using real data at the last stage. This result might imply as we approach to the release date, at the last week, noise in downloading data dramatically increase which create difference in forecasting result with our model based on prediction using the real data until two weeks prior stage.

[Table 4-2: Benchmark for Proposed model based on Real data]

Cross Validated MAPE				
	Benchmark: Hold-Out sample			
	Min	Max	Median	Average
Week -4	0.008	14.824	0.535	1.714
Week -3	0.014	17.234	0.592	1.446
Week -2	0.054	17.468	0.611	1.254
Week -1	0.007	16.11	0.578	1.169

Comparing Alternative Models

To fully understand the advantages of FDA, we compare two models of FDA with 3 non-Functional models. Two functional models are named as Functional Regression and Augmented Functional Regression and the three non functional models are Album Characteristic, Meta Bass

and Augmented Bass. Table 4 classifies all the models based on their use of information across curves and the types of variables.

The functional regression approach has three main strengths. First it is able to incorporate information from other types of albums to improve prediction accuracy. Second, it implements a non-parametric fitting procedure so it is not restricted by parametric assumptions. Third, it utilizes the functional nature of the downloading curves. We chose the three comparison models to gain an understanding of the gains from each of these strengths. For example, Album Characteristics are parametric and non-functional and doesn't incorporate information from characteristics of other albums. In contrast, Meta bass and Augmented Meta Bass models extend Class Bass can incorporate information from characteristics of other albums but are parametric and non-functional so they illustrate the improvement from borrowing strength across curves. Functional Regression and Augmented Functional Regression incorporate information from other albums, and are flexible to be non-parametric and functional.

[Table 4: Comparison of Alternative Models]

Album Characteristics	YouTube comments, Total number of albums, Single
Meta Bass Model	Bass Parameters: market potential (m), innovator coeff. (p), imitators coeff. (q)
Augmented Bass Model	Bass Parameters (m,p,q) and album characteristics
Functional Regression	Principal Component Scores (PCS) from level, velocity and accelerator curves
Augmented Functional Regression Model	Principal Component Scores (PCS) from level, velocity and accelerator curves and album characteristics variables
Median Reg.	Non-parametric linear regression around the median (50% quantile) value of sales by minimizing sum of absolute value of errors
GAM	Non-linear regression based on Generalized Additive Model (GAM)

Album Characteristics Model is just based on the three albums characteristics variable, described in the Table 4. This is base-line model that doesn't utilize any shape information compare to the Bass-related model or functional-type model.

Meta-Bass model is an extended model of Class Bass Model. The Classic Bass Model (Bass, 1969) fits each curve in the sample separately by estimating the following models:

$$s(t) = m\left[F(t) - F(t-1)\right] + \varepsilon_t, \qquad F(t) = \frac{1 - e^{-(p+q)t}}{1 + (q/p)e^{-(p+q)t}}$$

where t = time period, $s(t)$ = sales at time t, p = coefficient of innovation, q = coefficient of imitation, m = cumulative market potential. We estimate the model via the genetic algorithm because Venkatesan et al. (2004) provide convincing evidence that the genetic algorithm provides the bests method for fitting the Bass model relative to all prior estimation methods. For each curve, we use the downloading curves using smooth spline to estimate the three Bass parameters, m, p, and q. We then extend the Classic Bass model to use information across curves. To do so, we first estimate m, p, and q for each curve using the genetic algorithm, as outlined above. Then, for each item to be predicted, we fit the non-linear additive model,

$$y_i = \beta_0 + g_1(m_i) + g_2(p_i) + g_3(q_i) + \varepsilon_i ,$$

to the estimation sample where, g_1, g_2, and g_3 are smoothing splines as defined previously. We used the estimated parameters from this additive model and the estimates of m, p, and q for each curve in the holdout sample to compute the corresponding item to be predicted for each of the holdout curves. Not that the estimation of m, p, and q can also be done using a Bayesian formulation with a prior on $\{m, p, q\}$.

The Augmented Meta-Bass is the same non-linear additive model used for the Meta-Bass except that we add on album characteristics for each the albums, thus:

$$y_i = \beta_0 + g_1(m_i) + g_2(p_i) + g_3(q_i) + \sum_{k=1}^{k} g_k(z_{ik}) + \varepsilon_i ,$$

where z_{ik} denotes k^{th} variable of album characteristics for each album. Note that the Meta-Bass and Augmented Meta-Bass are extensions of the Classic Bass that make use of all of the information across curves, rather than just utilizing each curve individually. Since, using information across curves is an essential feature of functional regression, doing so puts the Meta Bass and the Augmented Bass on the same platform as the FDA models.

For the Functional Regression model, we compute three principal component scores, the first each on the downloading curves, $x_i(t)$, on the velocity curves, $x_i'(t)$, and on the accelerator curves, $x_i''(t)$. The principal component scores on the velocity and accelerator curves are computed in an identical fashion to that for the piracy curves except that we utilize the derivative of $x_i(t)$. We then use these three scores as the independent variables in an additive regression model, as shown in the equation at previous parts. We then used the estimated parameters of this equation and the data from curves in the holdout sample, to compute the items to be predicted in the holdout sample.

Our second functional approach enhances the power of Functional Regression by adding album characteristics variable for each the albums, as with the Augmented Meta-Bass model. Hence, the Augmented Functional Regression model involves estimating a non-linear additive model on the estimation sample as follows $y_i = \beta_0 + \sum_{j=1}^{3} g_j(s_{ij}) + \sum_{k=1}^{3} g_k(z_{ik}) + \varepsilon_i$ where the g_j's are smoothing splines. We then compute the albums to be predicted for each curve in the holdout sample from the estimated values of the above parameters and the data in each curve in the holdout sample. This model is directly comparable to Augmented Meta Bass as both models use information across curves and from products.

Model Comparison

We present our forecasting results with comparison in five different models on each stage for which Functional Regression and Augmented Functional Regression outperforms each of the other methods. We compare the Functional Regression model to the 3 other non-functional models to assess the ability of functional-based model in prediction. In Table 5, Functional regression and Augmented Functional Regression are superior to the Album characteristics model, Meta Bass model, and Augmented Bass model based on average-measure of MAPE at all period. From a month to a week prior to the release, Functional Regression outperforms for at least 40% based on MAPE at any given week comparisons with three other non-functional models.

[Table 5 : Comparison of Cross-validated MAPE with Benchmarking models]

MAPE Comparison : Holdout sample						
		Only Album characteristics	Meta Bass model	Augmented Bass model	Functional Regression	Augmented Functional Regression
Average	Week -4	1.796	4.673	3.267	1.563	2.026
	Week -3	1.614	1.717	1.541	1.202	1.822
	Week -2	1.383	1.333	1.717	0.924	1.312
	Week -1	1.186	105.82	5.706	1.090	1.057
Median	Week -4	0.627	0.801	0.691	0.724	0.687
	Week -3	0.599	0.878	0.607	0.705	0.737
	Week -2	0.618	0.867	0.709	0.707	0.484
	Week -1	0.667	0.789	0.690	0.605	0.509

The performance of Functional Regression is mixed when compared with a model based on album characteristics in terms of median-based MAPE. From four weeks to three weeks prior release models, Album characteristics model and Augmented Bass model outperforms Functional Models. However, from 2 weeks prior to the release, Augmented Functional Model, our proposed model, outperforms for at least 30% based on median-based MAPE for any given benchmarking models.

It appears that downloading data in early period around a month prior release includes big outliers resulting information difference in downloading curves. Two functional-based models capture the pattern of downloading curves of these outliers that have significantly high level of downloads better than non-functional based model, such as Album Characteristics model, in sales forecasting. That is shown in the forecasting performance gap of functional-related models between average-based MAPE vs. median-based MAPE. Functional-based models clearly outperform all other models measured by average-MAPE. Downloading data convey more information as we approach to the release date around two-week prior stage and Augmented Functional model which combines both albums characteristics information and downloads information across the curves performs the best. It also appears that the Augmented Functional Model is slightly superior to the Functional Model from a week before the release based on average-MAPE. We can interpret that the information acquired across shapes of functional curves embed more noise around a week prior to the release period. Augmented Functional model that incorporate albums characteristics information reduces the

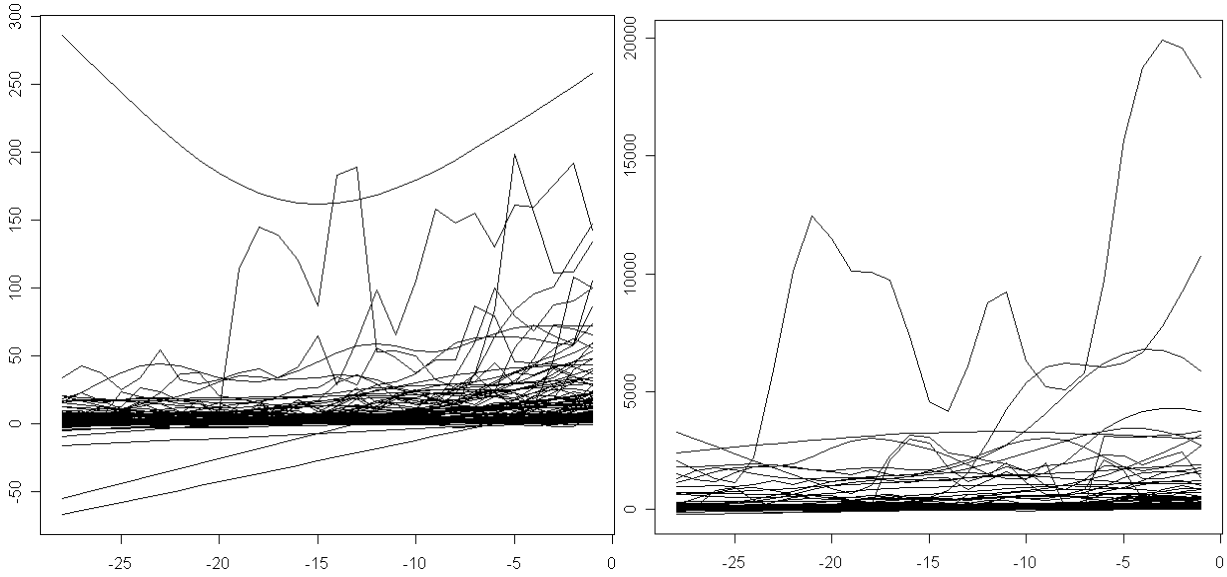
effects of noise obtained form downloading data. From an early stage of prediction around a month prior to the release, these shape information reduces the lower boundary of forecasting error, at the same time, produces overshooting in upper boundary due to the uncertainty. While models relying album characteristics appears to perform well based on median-MAPE in the very early stage of forecasting, from two weeks prior to the release, as more information penetrate in the market, functional-based model outperforms in both mean-based and median-based measure of MAPE.

Appendix

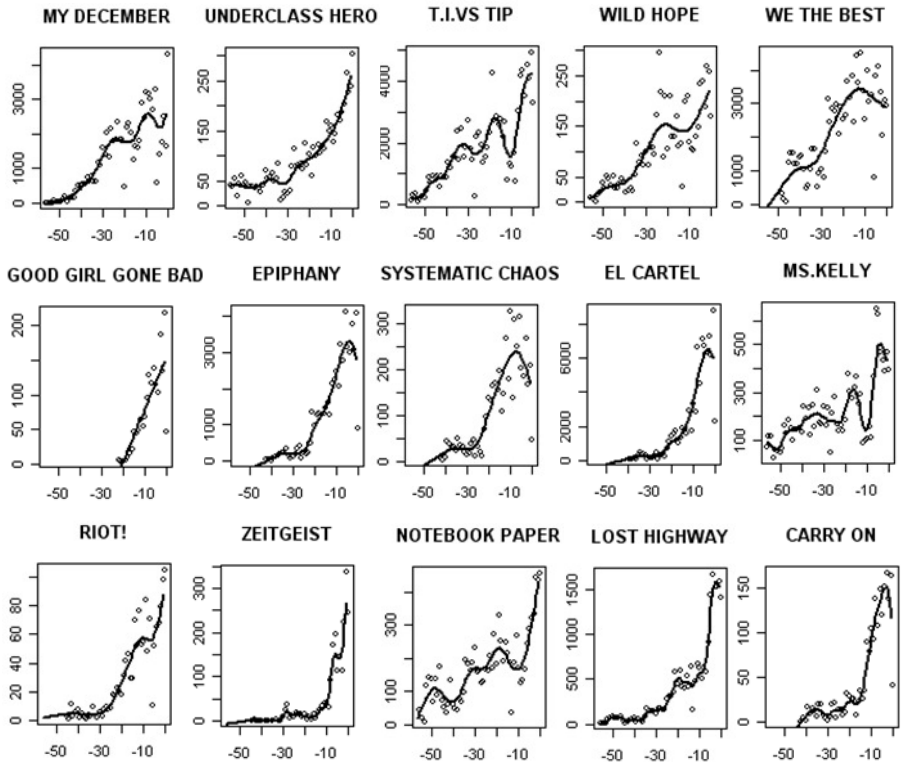
[Table 2: Pair-wise Correlation]

Variable	*Sales*	*Ave. hash*	*Ave. source*	*YouTube*	*Total albums*	*Single*
Sales	1					
Ave. hash	0.6806** (0.0000)	1				
Ave. source	0.6921** (0.0000)	0.9151** (0.0000)	1			
YouTube	0.5468** (0.0000)	0.7846** (0.0000)	0.7500** (0.0000)	1		
Total albums	0.1796* (0.0184)				1	
Single	0.3655** (0.0000)	0.2142** (0.0048)	0.2241** (0.0031)	0.2420** (0.0014)	0.1439* (0.0596)	1

[Figure : Functional shape of Hash- and Source- based P2P downloads]



[Figure 2-2: Penalized smoothed function based on Global-Source]



Album level analysis

We analyze characteristics of downloading curves with examples of individual albums in high-, medium-, and low- range of sales. Table 2 summarizes each three albums from three categories of sales volume where the highest range starts from more than 40K and the lowest volume range limits less than 10K of sales. For the albums summarized in Table 2, Figure 6 contrasts time-changing pattern PCSs of each albums in the sales category from 4 weeks to a week prior to the release.

[Table 1: Hash- and Source-based PCSs by Sales]

	Hash			*Source*			*Sales*
Range	Level	Velocity	Accelerator	Level	Velocity	Accelerator	Mean
Sales > 40K (55)	32.48	0.90	2.47	1498.04	23.67	208.64	128657.40
10K < Sales < 40K (53)	-21.15	-0.64	-1.39	-886.92	-21.75	-124.99	18253.58
Sales < 10K (44)	-15.13	-0.36	-1.41	-804.22	-3.39	-110.24	6516.41

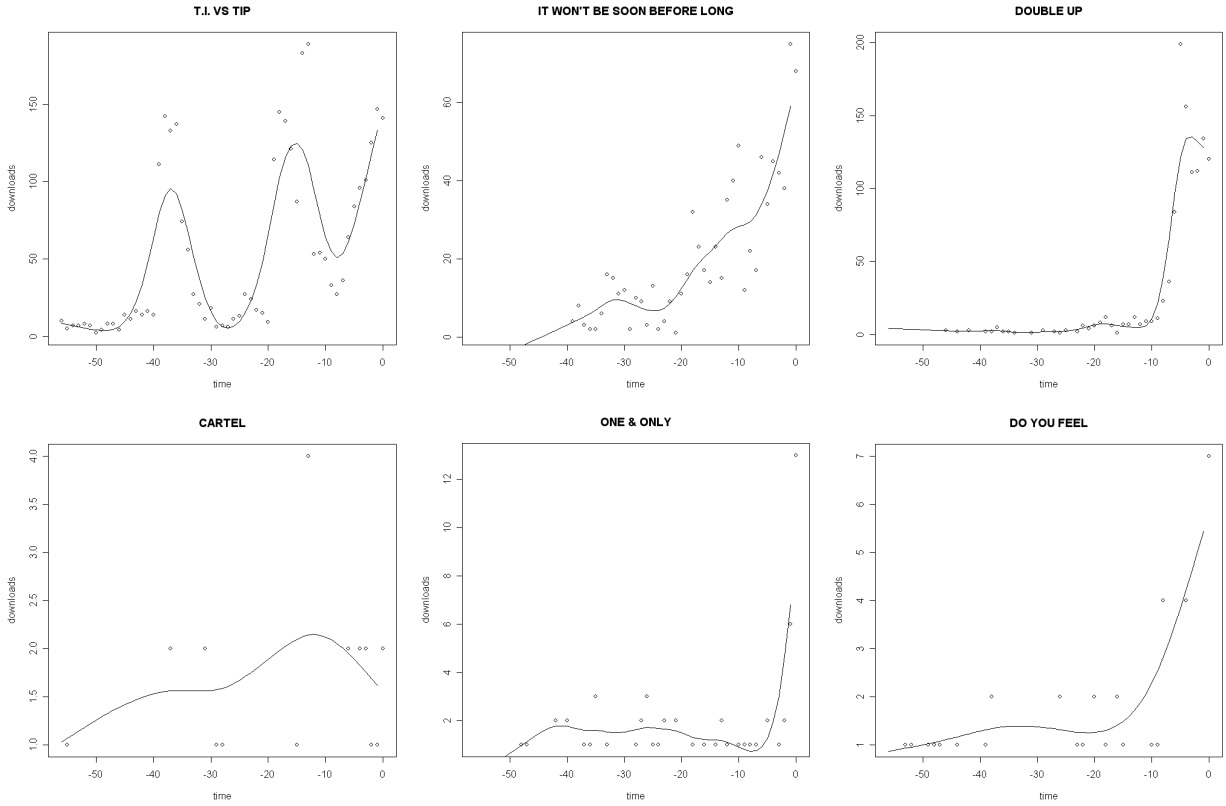
Table 1 summarizes average principal component scores of albums with respect to the first principal components of level, first-derivative, and second derivative of downloading curves by sales. Average principal component score (PCS) of albums with sales more than 40,000 units is 32. 48 for hash-based level curves where that of source-based downloads is 1498. 04. Average PCS value is strictly higher among high-sales albums compare to that of medium or low sales albums where units sold range around from 10,000 to 40,000 or below 10,000. This implies that demand-side downloading curves of high-sales albums covary about 32. 48 times with a unit variation of first PC curve. For the supply-side downloading curves, high-sales albums covary with the shape of first PC curve around 1498. 04 times of unit change. Negative value of average PCS implies opposite direction of covariation with respect to the direction of PC curve. For example, average value of PCS of albums range between 10K and 40K of sales is -0. 64 and that of albums with sales lower than 10K units is -0. 36. Albums with middle range of sales shows higher degree of covariation, 0. 64 times, with first PC of velocity curve in the opposite direction compare to the albums with that of (0. 32 times) low range of sales under 10K. From Table 1, we can confirm that downloading curves of albums predict the level of sales from the degree of covariation with respect to the direction of PC. Albums with higher PCS that covary more with the direction of PC result in higher sales.

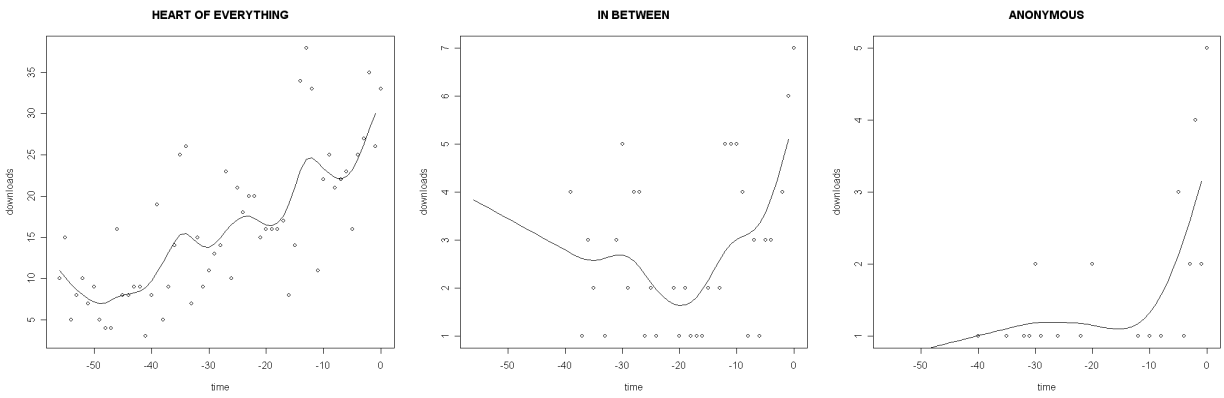
Albums with sales more than 40K shows the highest covariation level in the absolute term based on average value of PCS. It also shows that the level curves indicates the most differences in terms of absolute difference of covaration level (PCS) among downloading curves of albums by sales. In turn, velocity curves indicate the most difference in terms of percentage change of covariation level (PCS) according to the unit change of PC. For example, albums with more than 40K unit sales covary with the direction of PC curve 2384. 96 (1498. 04+886. 92) times with respect to the unit variation based on source-based level curves. In turn, level and accelerator curves show around 160% difference of covariation in terms of percentage change between high-sales and medium-level sales albums. Velocity curve reveal around 190% difference of covariation level between high-sales vs. medium sales group of albums. Hence, each type of curves-level, velocity, and accelerator- reveal unique characteristics as indicator function to predict sales based on the downloading curves.

[Table 2 : Illustration of Albums by Sales level]

	Artist	***Album***	***Sales***
High Sales (Sales > 40K)	T. I.	T. I. VS TIP	467,737
	MAROON 5	IT WON'T BE SOON BEFORE LONG	429. 484
	R. Kelly	DOUBLE UP	385,930
Medium Sales (10K < Sales < 40K)	CARTEL	CARTEL	28,079
	LIL WYTE	ONE & ONLY	15,556
	ROCKET SUMMER	DO YOU FEEL	15,088
Low Sales (Sales < 10K)	WITHIN TEMPTATION	HEART OF EVERYTHING	7,001
	PAUL VAN DYK	IN BETWEEN	6,084
	TOMAHAWK	ANONYMOUS	4,906

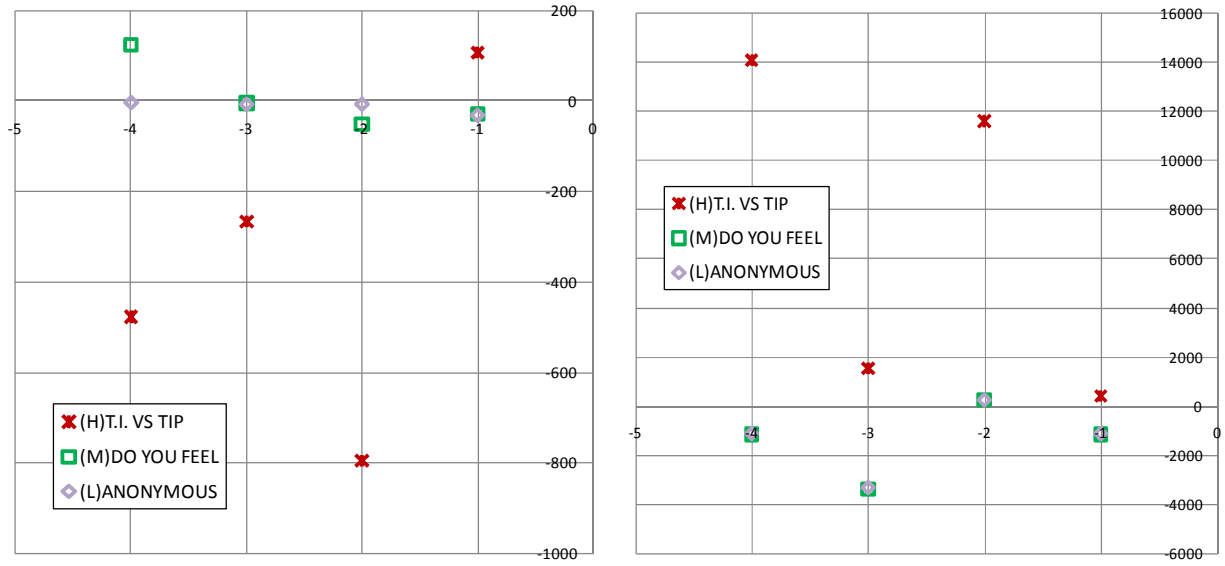
[Figure 5: Illustration of Downloading curvs of Albums]



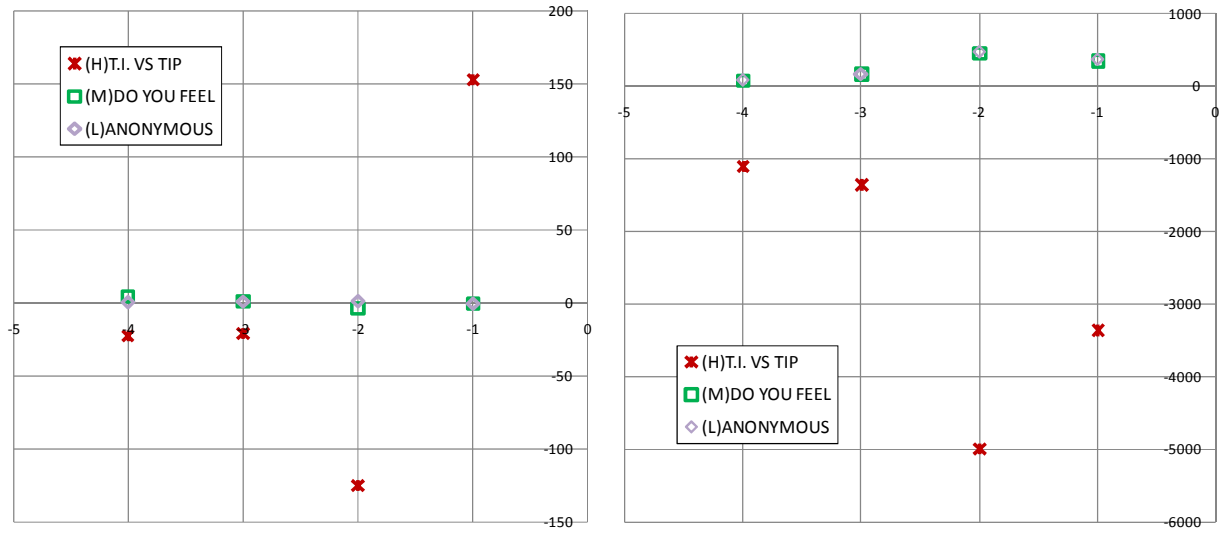


Principal component scores of album titled T. I. VS TIP, with more than 40K units of sales, shows bigger in absolute value, compare to that of albums with less than 40K sales from 4 weeks to a week prior to the release. In figure 6, while PCSs of album titled DO YOU FEEL and ANONYMOUS slightly change around -200 to 200 range for hash-based downloading curves, PCS of T. I. VS TIP realizes in about 4 times bigger range. This shows that our PCS value based on shape of downloading curves for an individual album indicates the sales from four weeks prior to the release by level of covariation, denoted by PCS, with respect to the principal component.

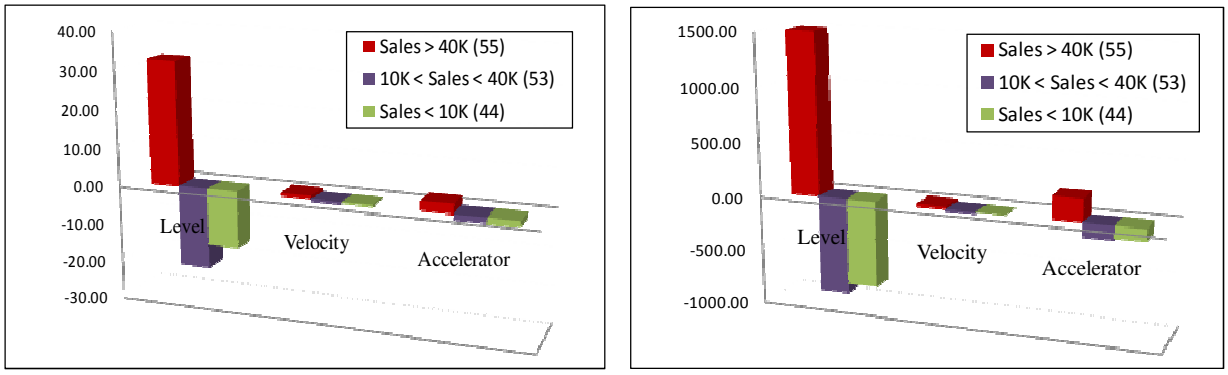
[Figure 6: Hash- and Source-based PCS of Level function]



Hash- and Source-based PCS of Velocity function



[Figure 4 : Hash- and Source-based PC Scores by Sales]



References

Bhattacharjee,S., Gopal, R.D., Lertwachara, K., Marsden, J.R. and Telang,R (2007), The effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts, *Management Science.*

Dellarocas, C. (2003), The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms, *Management Science*, 49(10) 1407-1424

Ramsay, J.O. (2000), Functional components of variation in handwriting, *Journal of the American Statistical Association*, 95(449): 9-15

Ramsay, J.O. and Silverman, B.W. (1996, 2005), *Functional Data Analysis* (First, Second ed.), Springer-Verlag, New York

Koenker, R., and Bassett, G. J. (1978), Regression Quantiles, *Econometrica* 46(2), 33-50

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge University Press, Cambridge

Bass, F. (1995), Empirical generalizations and Marketing Science: A Personal View, *Marketing Science*, 14(2), G6-G19

Sudip Bhattacharjee, Ram D. Gopal, Kaveepan Lertwachara, James R. Marsden, and Rahul Telang (2007), The effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts, *Management Science*

Judith A. Chevalier and Dina Mayzlin (2007), The Effect of Word of Mouth Online: Online Book Reviews, *Journal of Marketing Research*, forthcoming/

Song Hui Chon, Malcolm Slaney, and Jonathan Berger (2006), Predicting Success from Music Sales Data-A statistical and adaptive approach, AMCMM

Chrysanthos Dellarocas (2003), The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms, *Management Science*, 49(10) 1407-1424

Chrysanthos Dellarocas (2005), Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard, Information System Research 16(2), 209-230

Jonathan Lee, Peter Boatwright, Wagner A. Kamakura (2003), A Bayesian Model for Prelaunch Sales forecasting of Recorded Music, *Management Science*, Vol. 49, No.2, pp.179-196

Wendy W. Moe and Peter S. Fader (2002), Using Advance Purchase Orders to Forecast New Product Sales, *Marketing Science*, Vol.21, No. 3, pp.347-364

Alan L. Montgomery and Wendy W. Moe (2002), Should Music Labels Pay for Radio Airplay? Investigating the Relationship Between Album Sales and Radio Airplay,

Oberholzer-Gee, F. and K. Strumpf (2007), The Effect of File Sharing on Record Sales: An Empirical Analysis, *J. of Political Economy*, 115(1):1-42.

Peter E. Rossi, Greg M. Allenby, and Robert McCulloch (2005), Bayesian Statistics and Marketing, Jon Wiley & Sons, Ltd

Rossi, P.E., R.E. McCulloch, G.M. Allenby (1996), The value of purchasing history data in target marketing, *Marketing Science*, 15(4) 321-340