# Appendix to "Shapes and Transitions"

Revised: Februrary, 2017

# A Classification for the U.S. Treasury yield curves

## A.1 Fundamentals of yields and yield curves

The central concept in term structure analysis is the yield to maturity, henceforth YTM. As the most widely referenced interest return on a bond associated with a specific maturity date, the YTM measure implies an average rate of return if the investors' hold the bond to maturity. This implied interest rate is "backed out" from the market price of the corresponding security.

In theory, the YTM satisfies the bond pricing equation:

$$P = \sum_{t=1}^{T} \frac{CF_t}{(1+Y_t)^t} \quad or \quad P = \sum_{t=1}^{T} CF_t * e^{-Y_t t} \tag{A.1}$$

where $P$ is the market price of the security, $Y$ is the YTM corresponding to time $t$ cash flow, the cash flow $CF_t$ equals coupon payment at time $t$ before maturity $T$ and it equals coupon plus principal payment at maturity date $T$. The two different versions differ in discrete or continuous compounding practice.

Relating YTM on a security to its maturity $T$ constitutes the term structure analysis. However, this is easier said than done—since bond yield (inversely related to the price) is also influenced by many other risk factors such as creditworthiness, liquidity, callable features, tax treatment, coupon payment schemes, etc.. In practice, it is never possible to hold all other factors constant and isolate the relationship between yields and their maturities, therefore, pure term structure of interest rates never exists and is unobservable. For instance, bonds within the same credit class trade with various degrees of liquidity. More often, due to the differences in coupon payment, securities with the same maturity can carry different yields (the higher the coupon rate, the higher the price, all else equal), let alone different

maturities. Thus, while economists can model the pure term structure through zero coupon bond in theory, practitioners need to "bootstrap" the discount factors successively from the coupon bearing bonds of different maturities to ensure less pricing error.

To approximate term structure of interest rates, market practitioners and researchers find proxies from two types of most widely traded financial instruments—government securities and swaps—since they are the least contaminated alternatives. Both are commonly used interest-rate benchmark for pricing and setting yields in all other sectors of the debt markets. For a thorough comparison between the two, refer to Fabozzi (2016, pp. 119-120).[1] In this paper, we analyze yields on U.S. Treasury securities. At the short end, Treasury bills are money market assets with maturities of 1 year or less, sold at a discount from par and do not bear periodic interest payments. Treasury notes are median term coupon securities with maturities from 2 to 10 years. Treasury bonds have maturities more than 10 years. Treasury notes and bonds are capital market assets carrying periodic coupon payments. Yields derived from all these securities are therefore regarded as short term rates, median term rates and long term rates, respectively.

The market yields are calculated from composite of quotations obtained by the Federal Reserve Bank of New York, more specifically, the closing market yields on actively traded Treasury securities in over-the-counter market. At any specific time during a trading day, Treasury yields of various maturities constitute the term structure of interest rates. By market convention, plotting the cross sectional yields against their corresponding maturities produces a simple yield curve, which help to visualize the term structure of interest rates at that specific time. The methodology and techniques to construct more reliable market yield curves have been improving significantly over the past decades. The current Treasury daily yield curve is fitted by a quasi-cubic hermite spline function. Readers may find a detailed description on the methodology from the U.S. Treasury Department website.[2] A more technical approach to fitting and estimating the U.S. Treasury yield curve can be found

[1] Frank J. Fabozzi. 2016. Bond markets, analysis, and strategies. 9th Edition. Pearson. Page 119-120.
[2] https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/yieldmethod.aspx

in Gürkaynak, et. al. (2007).

## A.2 Data description for the U.S. Treasury yields

The Federal Reserve Board data download program provides the most comprehensive yield data on U.S. Treasury securities with various frequencies (daily, weekly, monthly, yearly) in H.15 statistics dating back to 1953. The dataset contains four instruments with specific identifiers: (1) TB—secondary market Treasury bills with four maturities; (2) TCMNOM—Treasury constant maturities nominal securities including 11 maturities; (3) TCMII—Treasury constant maturities inflation indexed securities containing five maturities; (4) LTAVG—long term average inflation-indexed securities longer than 10-year.

In the paper, the dataset U.S. Treasury constant maturities nominal securities (TCMNOM) is used for three reasons: (1) It covers all available maturities ranging from 1-month to 30-year (whereas TB only covers four); (2) It consists of nominal yields which incorporate market expected inflation rates (whereas TCMII reflect more of real rates); (3) It represents the most liquid Treasury security classes, whereas TCMII are highly illiquid. The term structure of interest rates hence described by the dataset TCMNOM is subject less to various sources of measurement error among the four.

Table A1 lists data availability information for all frequencies of the TCMNOM yields. Daily yield series is based on the closing market bid yields on the business days. Weekly series are represented by the yields on Fridays. Monthly and annual yields are simply the averages of the daily yields. However, the monthly series covers a much longer period for 20-year and 30-year maturities than daily and weekly series.

For classification purposes, high frequency data are preferred because more detailed patterns of yield curves are retained. For modeling and forecasting purposes, monthly yield data are ideal because macroeconomic conditions are commonly measured on a monthly basis. The annual series misses out a vast amount of information and thereby will not be chosen in subsequent analysis.

Table A1: Data availability summary on U.S. Treasury constant maturity nominal yields

| | Data frequency and availability description | | | |
|---|---|---|---|---|
| Maturity | Business day | Weekly (Friday) | Monthly | Annual |
| 1-month | 2001-07-31 : present | 2001-08-03 : present | 2001-07 : present | 2001 : present |
| 3-month | 1982-01-04 : present | 1982-01-08 : present | 1982-01 : present | 1982 : present |
| 6-month | 1982-01-04 : present | 1982-01-08 : present | 1982-01 : present | 1982 : present |
| 1-year | 1962-01-02 : present | 1962-01-05 : present | 1953-04 : present | 1962 : present |
| 2-year | 1976-06-01 : present | 1976-06-04 : present | 1976-06 : present | 1976 : present |
| 3-year | 1962-01-02 : present | 1962-01-05 : present | 1953-04 : present | 1962 : present |
| 5-year | 1962-01-02 : present | 1962-01-05 : present | 1953-04 : present | 1962 : present |
| 7-year | 1969-07-01 : present | 1969-07-04 : present | 1969-07 : present | 1969 : present |
| 10-year | 1962-01-02 : present | 1962-01-05 : present | 1953-04 : present | 1962 : present |
| 20-year | 1962-01-02 : 1986-12-31 | 1962-01-05 : 1987-01-02 | 1953-04 : 1986-12 | 1962: 1986 |
| | 1993-10-01 : present | 1993-10-01 : present | 1993-10 : present | 1993 : present |
| 30-year | 1977-02-15 : 2002-08-27 | 1977-02-18 : 2002-02-15 | 1977-02 : 2002-02 | 1977 : 2002 |
| | 2008-06-30 : present | 2006-02-10 : present | 2006-02 : present | 2006 : present |

Data source: Federal Reserve Board data download program H.15 interest rate statistics.

Table A2: U.S. Treasury yield to maturity statistics (1953.4–2016.3)

| YTM | Obs. | Mean | S.D. | Tr.M. | Med. | M.A.D. | Min. | Max. | Skew. | Kurt. |
|---|---|---|---|---|---|---|---|---|---|---|
| 1-month | 177 | 1.30 | 1.62 | 1.03 | 0.26 | 0.37 | 0.00 | 5.21 | 1.13 | -0.02 |
| 3-month | 411 | 4.18 | 3.16 | 3.98 | 4.63 | 3.99 | 0.01 | 14.28 | 0.36 | -0.41 |
| 6-month | 411 | 4.37 | 3.26 | 4.15 | 4.80 | 4.05 | 0.04 | 14.81 | 0.40 | -0.31 |
| 1-year | 756 | 4.99 | 3.28 | 4.75 | 4.92 | 2.92 | 0.10 | 16.72 | 0.75 | 0.77 |
| 2-year | 478 | 5.61 | 3.76 | 5.36 | 5.67 | 3.94 | 0.21 | 16.46 | 0.45 | -0.28 |
| 3-year | 756 | 5.41 | 3.14 | 5.19 | 5.14 | 2.91 | 0.33 | 16.22 | 0.70 | 0.52 |
| 5-year | 756 | 5.66 | 3.00 | 5.42 | 5.38 | 2.83 | 0.62 | 15.93 | 0.77 | 0.52 |
| 7-year | 561 | 6.49 | 3.07 | 6.32 | 6.47 | 2.80 | 0.98 | 15.65 | 0.44 | 0.02 |
| 10-year | 756 | 5.97 | 2.82 | 5.68 | 5.55 | 2.64 | 1.53 | 15.32 | 0.89 | 0.51 |
| 20-year | 675 | 5.97 | 2.75 | 5.60 | 5.30 | 2.21 | 2.20 | 15.13 | 1.14 | 0.89 |
| 30-year | 423 | 7.15 | 2.91 | 6.96 | 6.94 | 2.98 | 2.46 | 14.68 | 0.47 | -0.45 |

Data source: Federal Reserve Board data download program H.15 interest rate statistics. Reported are total observations (Obs.), standard deviation (S.D.), trimmed 10% mean (Tr.M.), median (Med.), median absolute deviation (M.A.D.), minimum (Min.), maximum (Max.), skewness (Skew.), and excess kurtosis (Kurt.).

The descriptive statistics for monthly YTMs are shown in Table A2. Four YTMs are available throughout the entire sample—1-, 3-, 5-, and 10-year; 1-month YTM is only available since August, 2001; 3- and 6-month since January, 1982; 2-year since June, 1976; 7-year since July, 1969; 20- and 30-year were discontinued for years and become more available in recent years. Thus, one needs to be cautious when comparing their sampling statistics

due to the non-synchronous nature of the data. Statistics of yields to different maturities reported include the mean, standard deviation (S.D.), trimmed 10% mean (Tr.M.), median (Med.), median absolute deviation (M.A.D.), minimum (Min.), maximum (Max.), skewness (Skew.), and excess kurtosis (Kurt.). As each of these values captures certain aspect of the sampling distribution, some are more robust than the other and can be used for more reliable inference. First, mean, trimmed mean, and median yields all seem to indicate that YTM is an increasing function of term to maturity; but non-monotonicity happens to 2-, 7-, and 20-year yields because of their unavailability in certain periods in the sample. Second, yield volatility, measured by standard deviation and median absolute deviation, is not a monotonic function of time to maturity. The yield volatility first increases then declines, and seems to hump before the 2-year maturity. Third, the sampling distribution of each YTM is highly non-symmetric, all skewed to the right (but the relationship between mean and median seem to indicate that short term yields tend to left skewed). Fourth, the sampling distribution of short yields tend to have thin tails (except 1-year) whereas long yields tend to have heavy tails (except 30-year).

## A.3 Classification method description

Standard textbook classification is based on the slope and curvature of the yield curves, more specifically, the yield spreads among various maturities (e.g., Mishkin 2015; Fabozzi 2010). But the problem is which yield and which spread should be chosen—since there are 11 Treasury yields and hence 55 spreads! Though data availability restricts our choice of yields in the spread calculation, special concerns shall be given to avoid a lengthy data mining process and any underutilization of data information.

A better classification algorithm, therefore, starts with a reduction in the dimension of cross sectional yields without omitting any information in the sample period. Denote the average yield on short term Treasury bills as $Y_s$, the average yield on median term notes $Y_m$ and the average yield on long term bonds as $Y_l$, it is sufficient to condense all yield information

into these three vectors and perform subsequent slope and curvature approximation. Though the short term yield $Y_s$ is a simple average of four bill yields, it could just equal the one-year yield when one-month, three-month and six-month yield data are not available in the averaging period. Similar strategy applies for calculating median term yield $Y_m$ and long term yield $Y_l$—whenever time series data of any specific YTM is not available, the average of the available ones is used in the category.

Composing the yield through simple averaging may cause bias toward the available YTM in the sample. For instance, at the short end, one-month yield is only available after July, 2001 and one-year yield series starts from January 2, 1967; whereas at the long end, the thirty-year bond yield were introduced after February, 1977 but were discontinued from February, 2002 to February, 2006. The averaged short term yield $Y_s$ would be biased upward in the 1962 to 2001 sample and the averaged long term yield $Y_l$ is biased downward in the 1962 to 1977 and 2002 to 2006 sample. Yield data of other frequencies also suffer similar measurement problem. But the bias arising from dimension reduction remains secondary in the analysis so long as the classification result matches the original yield curve shapes in the sample period and generates consistent states for Markov chain modeling.

Table A3: Dimension reduced monthly yields statistics (1953.4-2016.3)

| | Averaged yields and their spreads (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| Key statistics | $Y_s$ | $Y_m$ | $Y_l$ | $Y_m - Y_s$ | $Y_l - Y_m$ | $Y_l - Y_s$ | $Y_m - \frac{Y_s+Y_l}{2}$ |
| Mean | 4.88 | 5.64 | 6.20 | 0.76 | 0.56 | 1.33 | 0.10 |
| Median | 4.89 | 5.32 | 5.75 | 0.73 | 0.42 | 1.10 | 0.11 |
| Minimum | 0.04 | 0.74 | 2.33 | -2.19 | -1.25 | -3.40 | -0.66 |
| Maximum | 16.72 | 15.92 | 14.90 | 8.64 | 9.05 | 9.11 | 9.10 |
| Skewness | 0.73 | 0.78 | 0.90 | -0.19 | 0.63 | 0.12 | -0.05 |
| Ex.kurtosis | 0.82 | 0.53 | 0.42 | -0.13 | -0.24 | -0.48 | 0.41 |
| St.deviation | 3.26 | 3.00 | 2.67 | 0.81 | 0.77 | 1.47 | 0.30 |

Data source: Federal Reserve Board data download program H.15 interest rate statistics.
Notes: $Y_s$, $Y_m$, $Y_l$ are the average of Treasury bill, note, and bond yields, respectively.

To confirm the point made above, a summary statistics of these averaged yields and their spreads is presented in Table A3, which covers a sample of 756 observations. Data frequency

is monthly and the dimension of cross sectional yields has been reduced from 11 to 3. The constructed yield and spread measures deliver information on the level, slope and curvature of the yield curve in the sample. From the table, the means of short yields, median yields and long yields are not significantly different from each other due to their relatively large standard deviation. All their corresponding yield spreads are also not significantly distinct from zero. Formal statistical tests are thus required to make formal inference.

Table A4: Two Sample $t$ tests and Wilcoxon tests for the null hypothesis $\mu_i = \mu_j$

| Tests | $\mu_s = \mu_m$ | $\mu_m = \mu_l$ | $\mu_s = \mu_l$ | Assumptions and explanation |
|---|---|---|---|---|
| Type 1 - $t$ | 4.722 | 3.864 | 8.646 | Normality; independent sampling; homogeneity variances |
| | (0.000) | (0.000) | (0.000) | |
| Type 2 - $t$ | 25.686 | 20.119 | 24.808 | Normality; dependent sampling; homogeneity variances |
| | (0.000) | (0.000) | (0.000) | |
| Type 3 - $t$ | 4.722 | 3.864 | 8.646 | Normality; independent sampling; heterogeneity variances |
| | (0.000) | (0.000) | (0.000) | |
| Type 4 - $t$ | 25.686 | 20.119 | 24.808 | Normality; dependent sampling; heterogeneity variances |
| | (0.000) | (0.000) | (0.000) | |
| Wilcox 1 | 244789 | 322186.5 | 213301 | Nonparametric; independent sampling; rank sum test |
| | (0.000) | (0.000) | (0.000) | |
| Wilcox 2 | 26205.5 | 245193 | 27334.5 | Nonparametric; dependent sampling; signed rank sum |
| | (0.000) | (0.000) | (0.000) | |
| Fisher's $F$ | 1.181 | 1.262 | 1.491 | $H_0$ : homoskedaticity variances; $F_{(0.95,755,755)}$=1.127 |
| | (0.022) | (0.001) | (0.000) | |

Note: $t-$test statistics reported are of various degree of freedom; $P-$value in parenthesis.

Various tests are performed and results are summarized in Table A4. The null hypothesis is that the true means of the averaged yields are drawn from the same population. To illustrate, the averaged short yields and median yields are chosen for testing the hypothesis $H_0 : \mu_s = \mu_m$, the two population means are equal. Depending on the assumption of independent sampling from a normally distributed population, and the assumption of homogeneity of variances, four types of two sample $t$ tests are performed. Fisher's F-test of homoskedasticity is applied in choosing the type of $t$ tests but the evidence is not strong since it fails to reject homoskedastic variances at 0.01 significance level. A more important issue is the the non-normality feature of the bond yields, thus, we perform two types of nonparametric Wilcoxon rank sum test in comparison with the $t$ tests. The results from formal

testing procedure confirm the validity of the dimension reduction strategy. Four types of two sample $t$ test and two nonparametric tests all strongly reject the null of same population mean. Rejection conclusion also applies to the difference between the averaged median yields and long yields as well as the difference between the averaged long and short yields. Therefore, the simple averaged yields are statistically reliable for our classification problem.

Based on actual observations, it is straightforward to classify yield curves into different shapes: upward, downward, flat, hump, and bowl yield curves. However, with monthly yield data, there are three outlier observations that complicate the classification problem. In February 1970, the averaged short yield equals averaged median yield of 7.59 but both are greater than the averaged long yield of 6.67; in August 1973, the averaged median yield is the same as the averaged long yield of 7.61 but both are smaller than the averaged short yield of 8.82; and in November 1970, the averaged median yield equates the averaged long yield of 6.58 but both larger than the averaged short yield of 5.51. None of the three cases is strictly monotonically sloping. Weakly upward or downward-sloping is therefore a good solution, i.e., an upward yield curve could simply satisfy $Y_s \leqq Y_m < Y_l$ or $Y_s < Y_m \leqq Y_l$ whereas a downward yield curve should just need to meet $Y_s \geqq Y_m > Y_l$ or $Y_s > Y_m \geqq Y_l$.

Outliers aside, the classification algorithm must be effective such that all observed yield curve shapes are precisely identified, mutually exclusive, and exhaustive for any given sample period with any data frequencies. If the shape of the yield curve on a particular observation cannot be identified, the algorithm is flawed. It causes problems when modeling the dynamic transition of yield curves by a Markov chain—a non-identifiable yield curve does not map onto any Markov chain states. As a consequence, the stochastic matrix (transition probabilities) of the Markov chain cannot be estimated. Hence, an effective algorithm must also necessitate the estimation of the Markov chain.

Last but not least, a threshold value (the difference between two averaged yields) shall be chosen to separate various types of yield curve. It is relatively easy to start with the case of a flat yield curve, in the sample there is not any single observation that satisfies the condition

all cross-sectional yields being equal. An approximation is thus necessary. Heuristically, a flat yield curve shall satisfy $Y_s \approxeq Y_m \approxeq Y_l$ . By how much shall these averaged yields be different from one another? 1%, 0.5%, 0.25% or 0.1%? A relatively small value necessarily narrows down the flat type whereas a relatively large value will group fewer yield curves into other categories. Moreover, there may exist multiple tradeoffs in this classification problem. The choice of threshold value may also depend on the data frequency, classification result robustness and the research problems at hand. Therefore, while selecting the precise cutoff value awaits theoretical rigor, a relatively small value of 10 basis points (0.1 percent) is chosen to illustrate the methodology in the paper.

## A.4 Classification results for U.S. Treasury yield curves

The table below completes classification results for yield data of three frequencies.

Table A5: Classification result for daily, weekly and monthly yield curves

| Shapes | Ocurrence | $\overline{Y}_s$ | $\overline{Y}_m$ | $\overline{Y}_l$ | $\overline{Y}_m - \overline{Y}_s$ | $\overline{Y}_l - \overline{Y}_m$ | $\overline{Y}_l - \overline{Y}_s$ | $\overline{Y}_m - \frac{\overline{Y}_s + \overline{Y}_l}{2}$ |
|---|---|---|---|---|---|---|---|---|
| **Daily yield curves and key statistics (1962.1.2–2016.5.13)** | | | | | | | | |
| Upward (U) | 9891 (72.88%) | 4.18 | 5.34 | 6.29 | 1.16 | 0.94 | 2.10 | 0.11 |
| | | (2.84) | (2.81) | (2.44) | (0.63) | (0.69) | (1.17) | (0.31) |
| Humped (H) | 1208 (8.90%) | 7.67 | 8.16 | 7.90 | 0.48 | -0.26 | 0.22 | 0.37 |
| | | (2.84) | (2.96) | (2.94) | (0.45) | (0.18) | (0.52) | (0.22) |
| Flat (F) | 245 (1.80%) | 5.73 | 5.75 | 5.75 | 0.02 | 0.00 | 0.02 | 0.01 |
| | | (1.69) | (1.65) | (1.66) | (0.06) | (0.06) | (0.08) | (0.04) |
| Bowl (B) | 855 (6.30%) | 6.71 | 6.30 | 6.56 | -0.41 | 0.26 | -0.15 | -0.33 |
| | | (1.64) | (1.51) | (1.47) | (0.30) | (0.20) | (0.41) | (0.16) |
| Downward (D) | 1378 (10.14%) | 9.32 | 8.73 | 8.27 | -0.59 | -0.46 | -1.05 | -0.06 |
| | | (3.73) | (3.34) | (3.16) | (0.54) | (0.34) | (0.76) | (0.25) |
| All sample obs | 13577 (100%) | 5.20 | 6.01 | 6.64 | 0.80 | 0.63 | 1.44 | 0.09 |
| | | (3.37) | (3.06) | (2.61) | (0.86) | (0.80) | (1.54) | (0.32) |
| **Weekly yield curves and key statistics (1962.1.5–2016.5.13)** | | | | | | | | |
| Upward (U) | 2069 (72.93%) | 4.18 | 5.33 | 6.28 | 1.16 | 0.94 | 2.10 | 0.11 |
| | | (2.83) | (2.80) | (2.43) | (0.63) | (0.69) | (1.17) | (0.31) |
| Humped (H) | 254 (8.95%) | 7.68 | 8.17 | 7.91 | 0.48 | -0.25 | 0.23 | 0.37 |
| | | (2.81) | (2.95) | (2.93) | (0.46) | (0.18) | (0.53) | (0.23) |
| Flat (F) | 51 (1.80%) | 5.88 | 5.90 | 5.89 | 0.03 | -0.01 | 0.02 | 0.02 |
| | | (1.69) | (1.65) | (1.66) | (0.06) | (0.06) | (0.08) | (0.04) |
| Bowl (B) | 174 (6.13%) | 6.76 | 6.34 | 6.59 | -0.42 | 0.25 | -0.17 | -0.34 |
| | | (1.61) | (1.47) | (1.44) | (0.31) | (0.19) | (0.40) | (0.16) |
| Downward (D) | 289 (10.19%) | 9.36 | 8.77 | 8.30 | -0.59 | -0.46 | -1.05 | -0.06 |
| | | (3.74) | (3.36) | (3.18) | (0.54) | (0.33) | (0.75) | (0.24) |
| All sample obs | 2873 (100%) | 5.21 | 6.01 | 6.64 | 0.80 | 0.63 | 1.43 | 0.09 |
| | | (3.38) | (3.06) | (2.61) | (0.86) | (0.80) | (1.54) | (0.32) |
| **Monthly yield curves and key statistics (1953.4–2016.3)** | | | | | | | | |
| Upward (U) | 546 (72.22%) | 3.97 | 5.07 | 5.93 | 1.10 | 0.86 | 1.96 | 0.12 |
| | | (2.76) | (2.78) | (2.52) | (0.61) | (0.68) | (1.15) | (0.29) |
| Humped (H) | 78 (10.32%) | 6.34 | 6.72 | 6.49 | 0.38 | -0.23 | 0.15 | 0.31 |
| | | (3.02) | (3.16) | (3.11) | (0.39) | (0.15) | (0.43) | (0.21) |
| Flat (F) | 16 (2.12%) | 5.96 | 5.99 | 5.98 | 0.03 | -0.01 | 0.02 | 0.02 |
| | | (2.13) | (2.11) | (2.11) | (0.06) | (0.07) | (0.09) | (0.04) |
| Bowl (B) | 40 (5.29%) | 6.67 | 6.29 | 6.53 | -0.38 | 0.24 | -0.14 | -0.31 |
| | | (1.62) | (1.50) | (1.49) | (0.28) | (0.16) | (0.34) | (0.15) |
| Downward (D) | 76 (10.05%) | 8.75 | 8.22 | 7.77 | -0.53 | -0.45 | -0.99 | -0.04 |
| | | (3.79) | (3.40) | (3.25) | (0.52) | (0.31) | (0.70) | (0.25) |
| Full sample | 756 (100%) | 4.88 | 5.64 | 6.20 | 0.76 | 0.56 | 1.33 | 0.10 |
| | | (3.26) | (3.00) | (2.67) | (0.81) | (0.77) | (1.47) | (0.30) |

Notes: Data are from the Federal Reserve Board H.15 statistics. $\overline{Y}_s$, $\overline{Y}_m$, $\overline{Y}_l$ are the sample means of the averaged Treasury bill, note, and bond yields, respectively. Standard deviations in the parentheses.

Applying the classification algorithm to the yield data generates frequency counts for yield curve shapes. For each type of yield curve, sample mean and standard deviation of average yields and their corresponding spreads are reported, along with key information on the level, slope, curvature, and the volatility of yields (spreads) conditional on each type of yield curve. The following patterns are observed for the monthly yield curves.

First, an upward yield curve is the most common one—it accounts for 72.22% of the total sample observation; the hump shape and downward sloping yield curve are much less often detected, observed about 10% of the time in the sample; the bowl shape and flat yield curves are least frequently observed types—with only 5.29% and 2.12% of the time. These relative frequencies can be viewed as the unconditional distribution of monthly yield curve shape in the sample period.

Second, yield curve levels, measured by three average yields ($\overline{Y}_s$, $\overline{Y}_m$, $\overline{Y}_l$), are most likely to be low when yield curve shapes upward; they are most likely to be high when yield curve shapes downward. (The yield levels of the upward yield curve are the lowest compared with those of other types—a sample mean 5.93% of the averaged long yields is smaller than all other sample means of the averaged yields of other types. The yield levels of the downward yield curve are the highest among those of other types—a sample mean 7.77% of the averaged long yields is higher than all other sample means of the averaged yields of other types.) The flat yield curve, on average, happens at an averaged yield level of around 6%. The hump and bowl yield curves are, on average, observed at levels higher than the flat yield curve, and they seem to "mirror" each other. Results from daily and weekly yield data with a shorter time span (1962.1.2–2016.5.13) suggest that the hump yield curve on average remains at a higher yield level than the bowl yield curve. However, the yield levels are lower in the pre-1962 period than after.

Third, yield curve slopes, gauged by three averaged spreads, display different degrees of steepness for different types. Yield spreads are most likely to be small when the yield curve is flat (by definition), yield spreads are most likely to be large when the yield curve

shapes upward, and they all turn negative when the yield curve shapes downward. While the upward yield curves, on average, are much steeper than the downward yield curves—almost twice in absolute value, the hump and bowl yield curves, interestingly, share almost the same sizes of the spreads in absolute value. The flat yield curve, by definition, has least significant averaged spreads.

Fourth, yield curve curvature, approximated by the difference between averaged median yields and averaged long-short yields, is the least important factor to explain yield curve variations among the three. A curvature measure of 0.10 using all sample observations is much smaller than its level and spread counterparts. While the upward yield curves tend to bow up (0.12), the downward yield curves seem to slightly bow down (-0.04), though both are not statistically significant. Interestingly, the hump and bowl yield curves, on average, share exactly the same curvature, and both are statistically significant.

Lastly, yield levels are least volatile when the yield curve exhibits bowl shape, most volatile when it is downward-sloping; yield spreads are least volatile when the yield curve is flat, most volatile when it is upward-sloping. Similar patterns are observed in daily and weekly data.

## A.5 Graphs for classified shapes of monthly yield curves

Figure A1: Upward U.S. Treasury monthly yield curve 1982.7

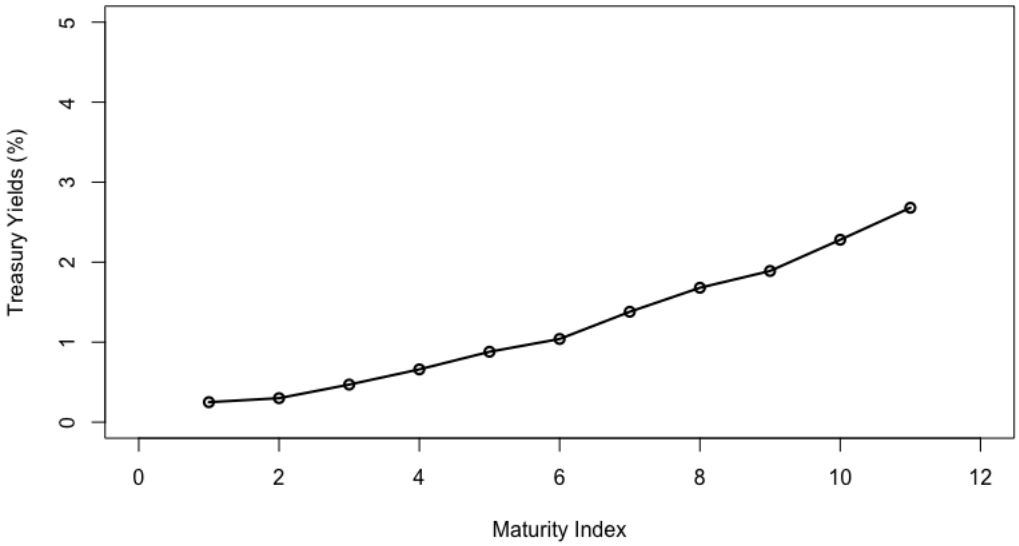



Figure A2: All upward U.S. Treasury monthly yield curves

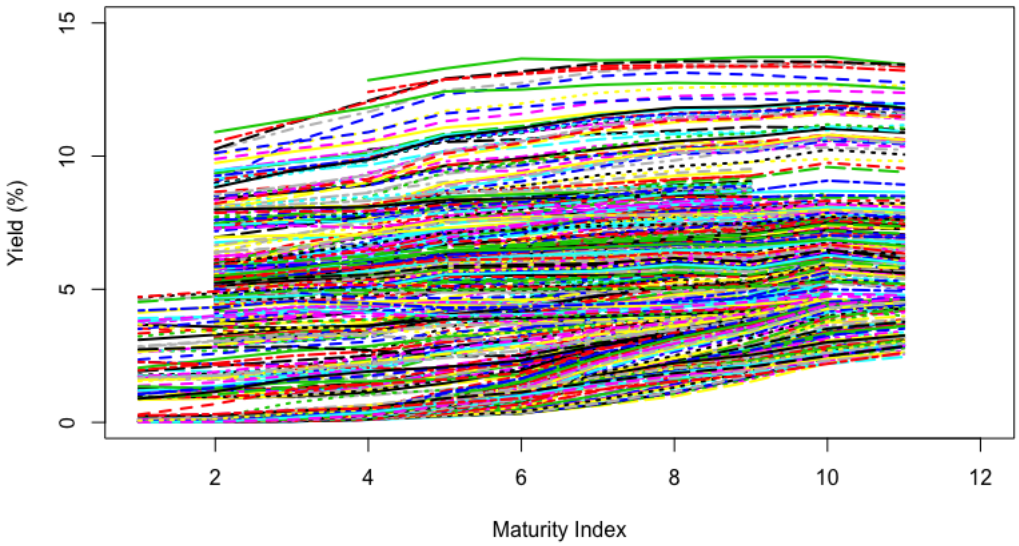


Figure A3: Hump U.S. Treasury monthly yield curve 1982.7

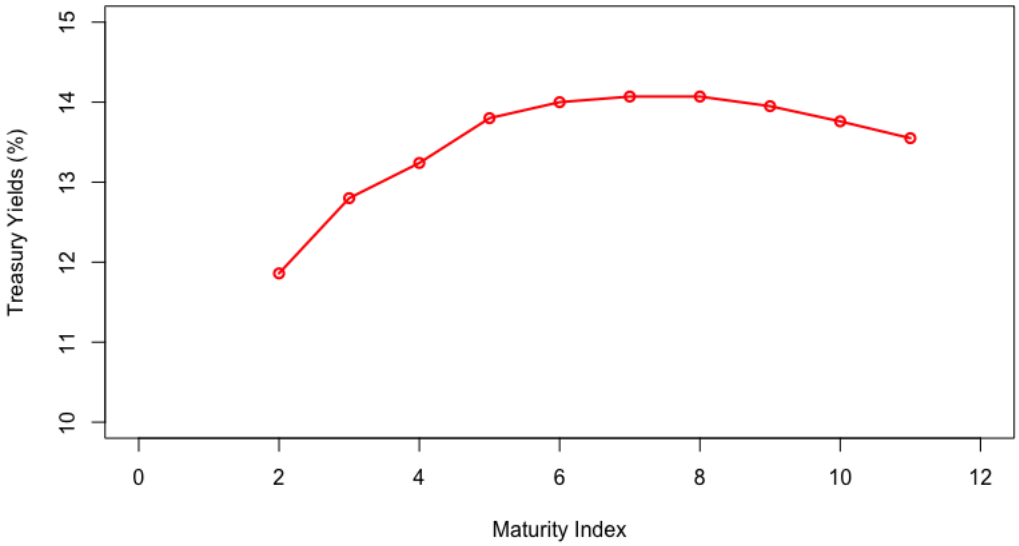



Figure A4: All hump U.S. Treasury monthly yield curves

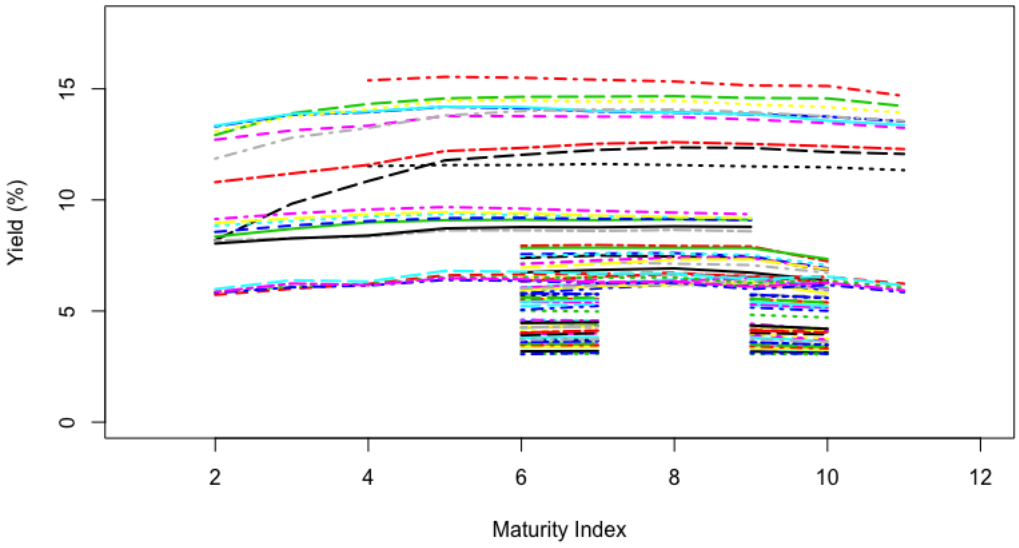


Figure A5: Flat U.S. Treasury monthly yield curve 2006.2

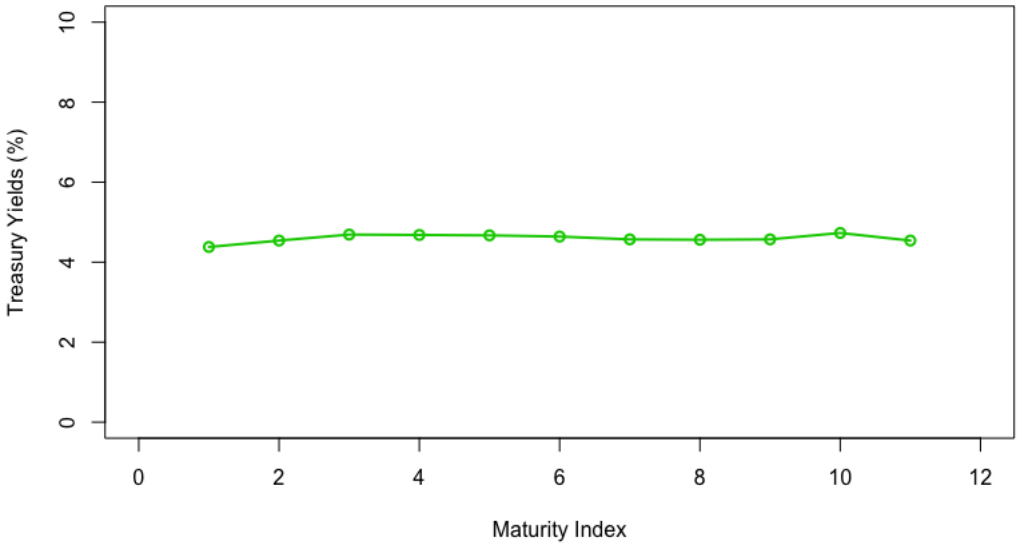



Figure A6: All flat U.S. Treasury monthly yield curves

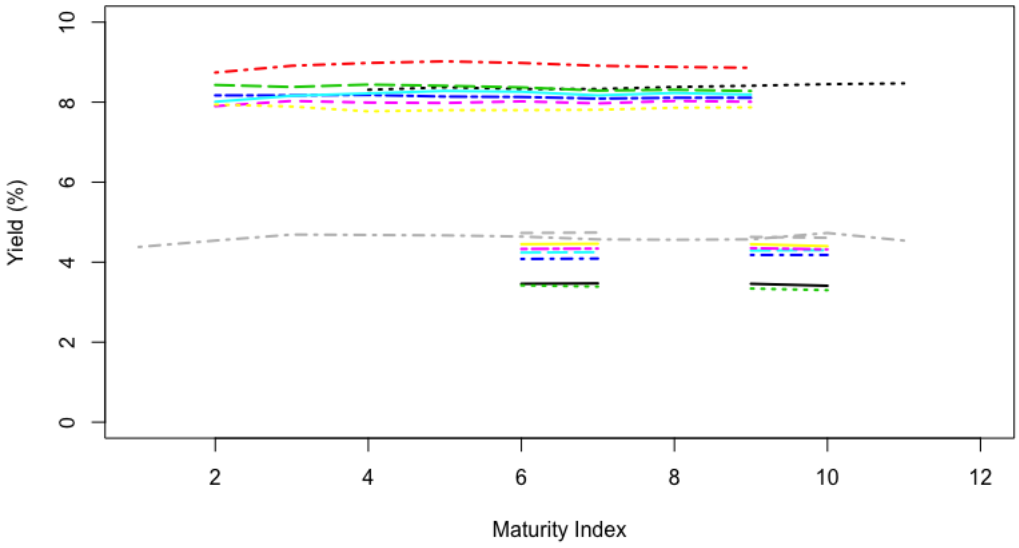


Figure A7: Bowl U.S. Treasury monthly yield curve 2006.8

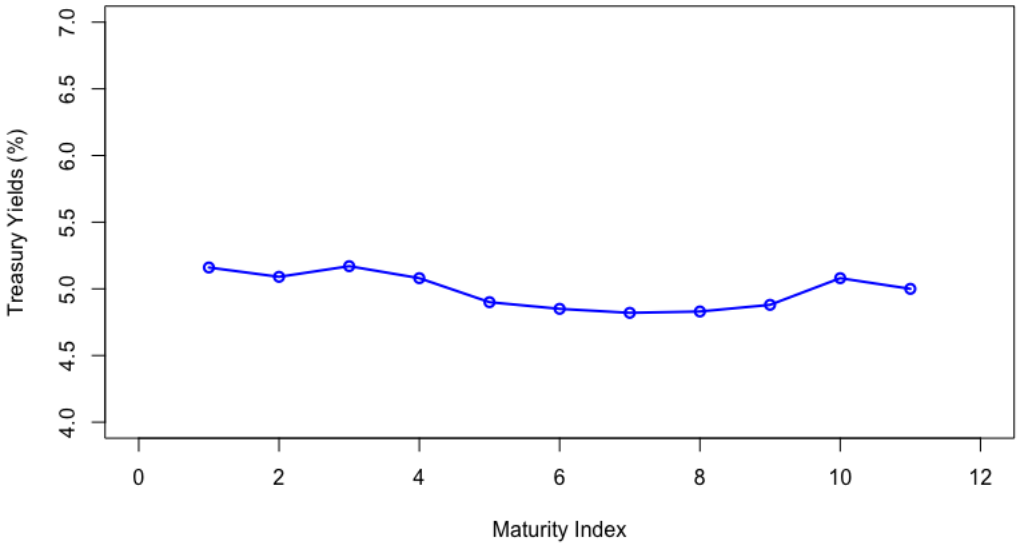



Figure A8: All bowl U.S. Treasury monthly yield curves

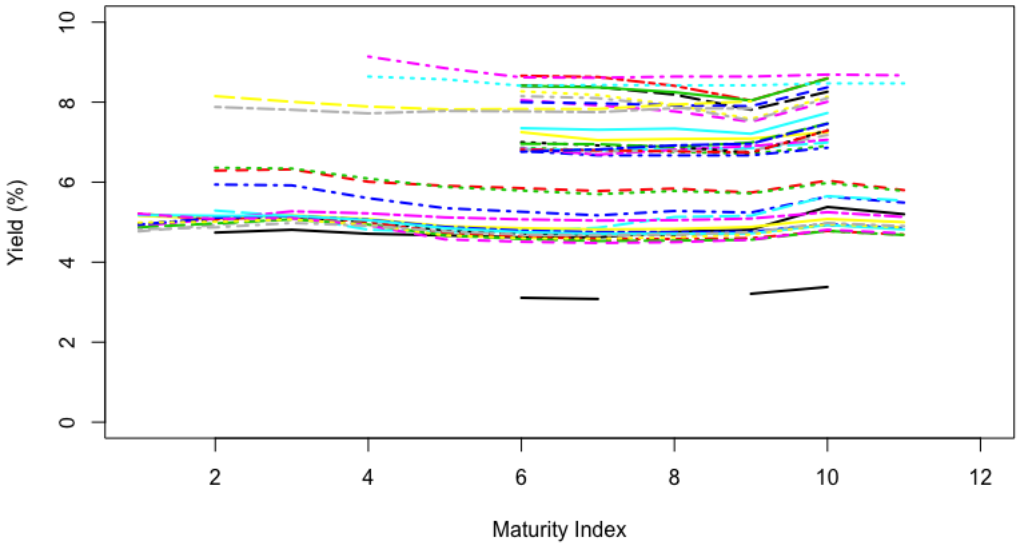


Figure A9: Downward U.S. Treasury monthly yield curve 1982.2

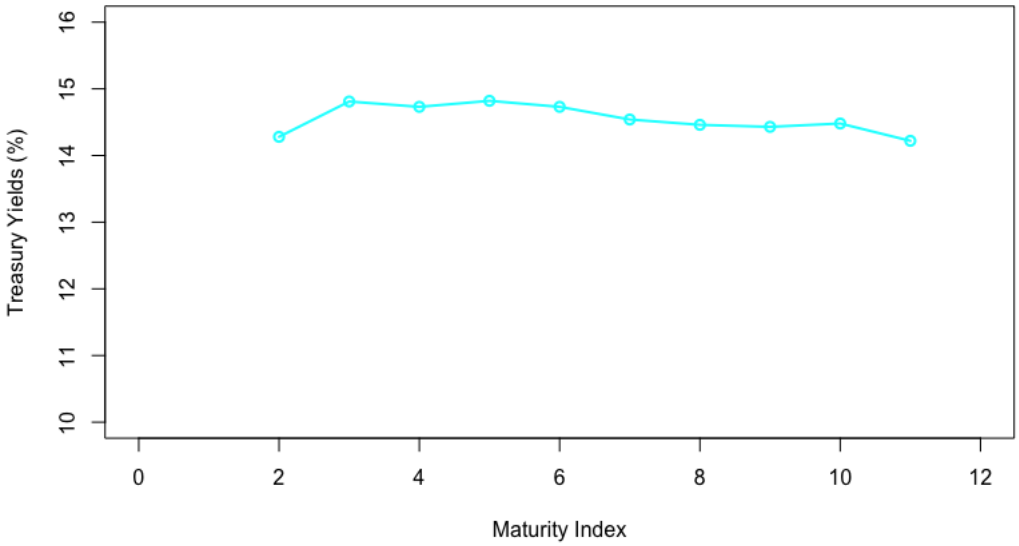
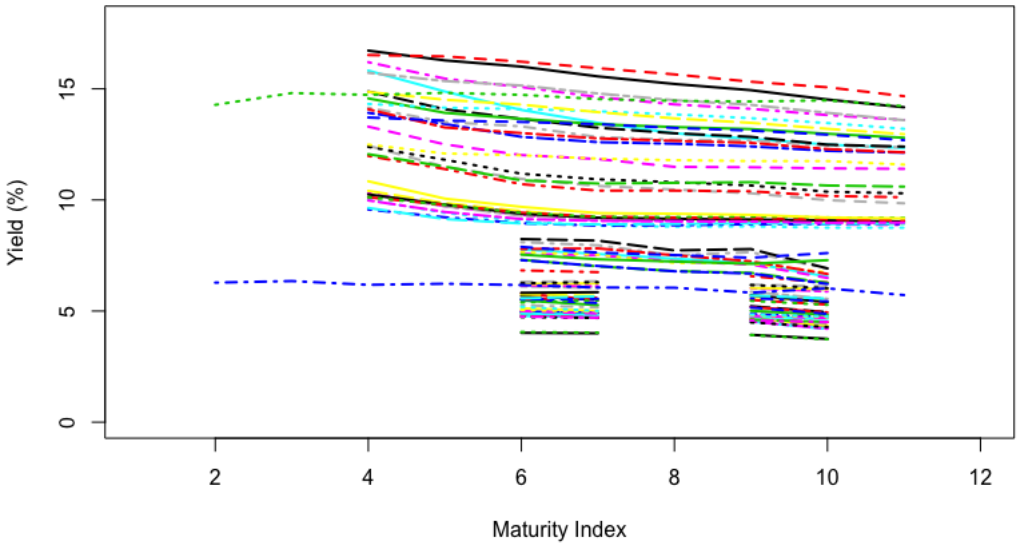



Figure A10: All downward U.S. Treasury monthly yield curve



# B Yield curve shapes and macroeconomic states

With more robust statistics, Table B1 complements Table 3 in Section 4.1 and Table B2 complements Table 4 in Section 4.2.

Table B1: Average yield statistics over different business cycle stages

| Statistics | Mean | Median | Trim. | S.D. | M.A.D | Min. | Max. | Skew. | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| | All pre-recession 18-month periods (155 months) | | | | | | | | |
| $Y_s$ | 6.71 | 5.98 | 6.37 | 2.95 | 2.83 | 2.49 | 16.20 | 1.05 | 0.73 |
| $Y_m$ | 6.63 | 6.18 | 6.35 | 2.61 | 2.45 | 2.75 | 14.71 | 0.96 | 0.60 |
| $Y_l$ | 6.52 | 6.03 | 6.29 | 2.48 | 2.73 | 2.91 | 13.71 | 0.82 | 0.25 |
| $Y_m - Y_s$ | -0.08 | 0.00 | -0.04 | 0.56 | 0.47 | -1.67 | 1.26 | -0.52 | 0.21 |
| $Y_l - Y_m$ | -0.11 | -0.08 | -0.09 | 0.32 | 0.28 | -1.00 | 0.66 | -0.42 | 0.30 |
| $Y_l - Y_s$ | -0.19 | -0.08 | -0.13 | 0.76 | 0.56 | -2.52 | 1.43 | -0.80 | 0.72 |
| $Y_m - \frac{Y_s+Y_l}{2}$ | 0.02 | 0.07 | 0.02 | 0.26 | 0.24 | -0.66 | 0.60 | -0.23 | -0.41 |
| | All recession periods (121 months) | | | | | | | | |
| $Y_s$ | 5.94 | 5.98 | 5.52 | 4.32 | 4.49 | 0.20 | 16.72 | 0.69 | -0.40 |
| $Y_m$ | 6.51 | 6.78 | 6.08 | 3.95 | 4.65 | 1.54 | 15.92 | 0.75 | -0.46 |
| $Y_l$ | 6.91 | 6.67 | 6.53 | 3.49 | 3.62 | 2.60 | 14.90 | 0.77 | -0.47 |
| $Y_m - Y_s$ | 0.57 | 0.68 | 0.61 | 0.82 | 0.69 | -2.19 | 2.53 | -0.46 | 0.55 |
| $Y_l - Y_m$ | 0.40 | 0.39 | 0.38 | 0.76 | 0.63 | -1.25 | 2.12 | 0.18 | -0.37 |
| $Y_l - Y_s$ | 0.97 | 1.01 | 0.99 | 1.42 | 1.44 | -3.40 | 4.24 | -0.16 | -0.02 |
| $Y_m - \frac{Y_s+Y_l}{2}$ | 0.09 | 0.10 | 0.08 | 0.35 | 0.27 | -0.55 | 1.28 | 0.40 | 1.04 |
| | All post-recession 12-month periods (101 months) | | | | | | | | |
| $Y_s$ | 3.92 | 3.18 | 3.72 | 2.85 | 3.06 | 0.14 | 10.14 | 0.48 | -0.88 |
| $Y_m$ | 5.31 | 4.32 | 5.04 | 2.80 | 3.02 | 1.87 | 11.52 | 0.57 | -0.81 |
| $Y_l$ | 6.12 | 5.81 | 5.91 | 2.61 | 3.02 | 2.57 | 11.89 | 0.56 | -0.68 |
| $Y_m - Y_s$ | 1.38 | 1.29 | 1.36 | 0.54 | 0.63 | 0.59 | 2.51 | 0.32 | -1.86 |
| $Y_l - Y_m$ | 0.81 | 0.55 | 0.75 | 0.70 | 0.49 | -0.43 | 2.26 | 0.75 | -0.71 |
| $Y_l - Y_s$ | 2.19 | 1.82 | 2.13 | 1.17 | 1.20 | 0.34 | 4.41 | 0.48 | -1.13 |
| $Y_m - \frac{Y_s+Y_l}{2}$ | 0.29 | 0.30 | 0.30 | 0.22 | 0.19 | -0.39 | 0.66 | -0.75 | 0.47 |
| | All remaining periods (307 months)) | | | | | | | | |
| $Y_s$ | 4.94 | 4.97 | 4.86 | 2.07 | 1.82 | 0.96 | 11.36 | 0.51 | 0.67 |
| $Y_m$ | 5.90 | 5.60 | 5.67 | 2.22 | 2.13 | 2.24 | 13.34 | 1.00 | 0.97 |
| $Y_l$ | 6.50 | 6.16 | 6.26 | 2.22 | 2.19 | 2.86 | 13.49 | 0.90 | 0.55 |
| $Y_m - Y_s$ | 0.95 | 0.87 | 0.92 | 0.71 | 0.88 | -0.33 | 2.51 | 0.32 | -1.09 |
| $Y_l - Y_m$ | 0.61 | 0.51 | 0.54 | 0.61 | 0.45 | -0.55 | 2.18 | 0.97 | 0.47 |
| $Y_l - Y_s$ | 1.56 | 1.31 | 1.48 | 1.23 | 1.28 | -0.88 | 4.41 | 0.50 | -0.57 |
| $Y_m - \frac{Y_s+Y_l}{2}$ | 0.17 | 0.12 | 0.16 | 0.24 | 0.16 | -0.42 | 1.00 | 0.63 | 0.52 |

Note: Over the entire 1953 to 2010 U.S. business cycle, four stages are divided following the NBER business cycle chronology and adjusted for two overlapping periods. Data cover 1953.7 to 2010.6, a total of 10 U.S. business cycles. Trim. mean drops the top and bottom 10% fraction in the sample. SD is the standard deviation. MAD is the median absolute deviation. There are two adjustments for overlaps in between two short recession periods. Oct. 1958 to Mar. 1960 is counted as a pre-recession 18-month, which overlaps the 1958 post-recession period. Aug. 1980 to Jun. 1981 (11-month) is counted pre-recession period due to its overlap with the 1980 recession.

Table B2: Statistics for macroeconomic state variables conditional on yield curve shapes

| | U1: Usual upward shaped (301) | | | | | | | | | U2: Steep upward shaped (245) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Median | Trim. | S.D. | M.A.D | Min. | Max. | Skew. | Kurto. | Mean | Median | Trim. | S.D. | M.A.D | Min. | Max. | Skew. | Kurto. |
| CPIA | 3.01 | 2.51 | 2.65 | 2.55 | 1.78 | -0.86 | 13.48 | 1.58 | 2.97 | 2.58 | 2.71 | 2.61 | 1.49 | 1.45 | -1.98 | 6.72 | -0.16 | 0.34 |
| CPIC | 3.42 | 2.57 | 3.02 | 2.41 | 1.50 | 0.65 | 12.76 | 1.67 | 2.68 | 2.81 | 2.30 | 2.68 | 1.38 | 1.10 | 0.60 | 6.82 | 0.79 | -0.21 |
| PPIA | 2.46 | 1.80 | 1.99 | 3.35 | 2.55 | -2.32 | 17.81 | 1.69 | 3.67 | 1.79 | 1.94 | 1.86 | 2.70 | 2.09 | -6.71 | 9.25 | -0.22 | 0.58 |
| PPIC | 2.76 | 1.75 | 2.04 | 3.70 | 2.06 | -0.92 | 20.48 | 2.73 | 8.97 | 1.77 | 1.70 | 1.64 | 1.37 | 0.93 | -1.00 | 7.21 | 1.22 | 2.70 |
| RGDP | 3.56 | 4.00 | 3.74 | 2.51 | 2.08 | -2.90 | 9.20 | -0.59 | 0.04 | 2.51 | 2.70 | 2.63 | 2.17 | 1.63 | -4.10 | 8.60 | -0.43 | 1.89 |
| UNEM | 5.74 | 5.60 | 5.63 | 1.39 | 1.19 | 2.50 | 10.80 | 0.82 | 1.20 | 7.00 | 6.90 | 6.89 | 1.42 | 1.48 | 4.30 | 10.80 | 0.56 | -0.47 |
| INPR | 3.26 | 4.42 | 3.76 | 5.90 | 4.05 | -13.25 | 16.83 | -0.79 | 0.51 | 1.48 | 2.26 | 1.99 | 4.93 | 2.69 | -16.72 | 11.50 | -1.33 | 2.86 |
| CUR | 0.81 | 0.82 | 0.81 | 0.03 | 0.03 | 0.71 | 0.89 | -0.71 | 0.42 | 0.78 | 0.78 | 0.78 | 0.04 | 0.03 | 0.67 | 0.84 | -0.63 | 0.54 |
| | H: Hump shaped (78) | | | | | | | | | F: Flat yield curve (16) | | | | | | | | |
| CPIA | 3.82 | 3.54 | 3.70 | 2.05 | 1.53 | 0.14 | 12.02 | 1.02 | 2.39 | 3.64 | 3.50 | 3.54 | 1.65 | 1.53 | 1.19 | 7.56 | 0.44 | -0.23 |
| CPIC | 4.42 | 4.29 | 4.16 | 2.25 | 2.40 | 1.52 | 11.34 | 0.94 | 0.49 | 3.27 | 3.68 | 3.11 | 1.76 | 1.76 | 1.22 | 7.21 | 0.45 | -0.68 |
| PPIA | 3.09 | 3.37 | 3.05 | 1.88 | 0.96 | -0.60 | 12.52 | 1.27 | 6.58 | 3.91 | 4.29 | 3.88 | 1.88 | 1.70 | 0.60 | 7.67 | -0.04 | -0.67 |
| PPIC | 4.03 | 3.66 | 3.95 | 2.36 | 2.08 | -0.30 | 10.06 | 0.40 | -0.44 | 3.65 | 3.79 | 3.55 | 2.38 | 3.31 | 0.90 | 7.80 | 0.61 | -0.93 |
| RGDP | 2.91 | 3.00 | 2.96 | 2.73 | 2.37 | -2.90 | 9.20 | -0.12 | -0.41 | 4.39 | 3.00 | 4.27 | 2.35 | 0.96 | 2.00 | 8.50 | 0.86 | -1.01 |
| UNEM | 5.18 | 5.00 | 4.89 | 1.62 | 1.33 | 3.40 | 10.10 | 1.57 | 1.83 | 4.79 | 4.90 | 4.78 | 0.63 | 0.67 | 3.80 | 5.90 | -0.09 | -1.40 |
| INPR | 2.71 | 2.53 | 2.33 | 5.51 | 3.97 | -7.06 | 19.61 | 0.76 | 0.93 | 3.77 | 2.13 | 3.49 | 3.95 | 2.97 | -0.30 | 11.82 | 0.70 | -1.06 |
| CUR | 0.82 | 0.82 | 0.82 | 0.04 | 0.04 | 0.73 | 0.89 | -0.47 | -0.63 | 0.83 | 0.83 | 0.83 | 0.02 | 0.01 | 0.80 | 0.86 | -0.27 | -0.88 |
| | B: Bowl shaped (40) | | | | | | | | | D: Downward shaped (76) | | | | | | | | |
| CPIA | 5.74 | 4.83 | 5.59 | 3.27 | 3.94 | 1.40 | 11.29 | 0.32 | -1.48 | 6.63 | 5.53 | 6.52 | 3.73 | 4.49 | 1.17 | 13.62 | 0.28 | -1.33 |
| CPIC | 4.39 | 3.14 | 4.10 | 2.37 | 1.22 | 2.14 | 10.06 | 1.05 | -0.23 | 6.36 | 5.72 | 6.29 | 3.37 | 4.77 | 1.22 | 12.30 | 0.14 | -1.42 |
| PPIA | 7.00 | 6.06 | 6.92 | 4.96 | 5.98 | -1.13 | 16.69 | 0.18 | -1.27 | 6.06 | 3.88 | 6.00 | 4.03 | 3.83 | -0.60 | 13.10 | 0.25 | -1.38 |
| PPIC | 4.92 | 3.11 | 4.01 | 4.80 | 2.73 | -0.51 | 19.27 | 1.49 | 1.41 | 5.47 | 3.88 | 5.35 | 3.29 | 3.64 | 0.92 | 11.44 | 0.32 | -1.46 |
| RGDP | 1.97 | 2.25 | 1.92 | 2.44 | 2.37 | -2.90 | 6.70 | 0.16 | -0.59 | 3.26 | 3.00 | 3.21 | 2.23 | 2.37 | -1.20 | 8.50 | 0.19 | -0.53 |
| UNEM | 4.88 | 4.85 | 4.86 | 0.54 | 0.52 | 3.90 | 6.00 | 0.28 | -0.45 | 5.06 | 4.80 | 4.95 | 1.50 | 1.63 | 3.40 | 8.90 | 0.51 | -1.04 |
| INPR | 2.85 | 2.35 | 2.79 | 3.13 | 1.88 | -7.07 | 9.12 | -0.17 | 1.13 | 4.24 | 4.86 | 4.38 | 3.49 | 3.68 | -3.14 | 9.56 | -0.35 | -1.03 |
| CUR | 0.84 | 0.85 | 0.84 | 0.03 | 0.04 | 0.79 | 0.89 | 0.06 | -1.65 | 0.84 | 0.85 | 0.85 | 0.03 | 0.04 | 0.76 | 0.88 | -0.54 | -0.83 |

[1] Source: FRED - St. Louis. CPIA–CPI all items inflation rate, CPIC–core CPI inflation rate, PPIA–PPI all item inflation rate, PPIC–core PPI inflation rate, RGDP–real GDP growth rate, UNEM–unemployment rate, INPR–industrial production growth rate, CUR–capacity utilization rate.

[2] Note: Trim. mean drops the top and bottom 10% fraction in the sample. SD is the standard deviation. MAD is the median absolute deviation. Steep upward yield curve has a long-short spread no less than 200 basis points. Except for unemployment rate and capacity utilization rate, all other economic indicator variables are expressed in annual percentage rate calculated from seasonally adjusted data.

# C Markov chain models and estimation

## C.1 Markov chain elements and properties

The definitions and notations adopted here are similar to chapter 2 of Kulkarni (2011) and chapter 11 of Evans et. al. (2010).

**Definition 1.** *(Markov Chain) A stochastic process $\{X_t, t \geq 0\}$ on state space $S$ is said to be a discrete-time Markov chain (DTMC) if, for all $i$ and $j$ in $S$, the conditional probability satisfies*

$$P(X_{t+1} = j | X_t = i, X_{t-1}, ..., X_0) = P(X_{t+1} = j | X_t = i) = P_{ij}. \tag{C.1}$$

A DTMC $\{X_t, t \geq 0\}$ is said to be time homogenous if, for all $t = 0, 1, 2, ...,$

$$P(X_{t+1} = j | X_t = i) = P(X_1 = j | X_0 = i). \tag{C.2}$$

Consider the yield data and types of yield curve classified, a DTMC with finite state space $S = \{U, H, F, B, D\}$. The assumption of time homogeneity impose a strong constrain on any successive conditional probabilities—they are all equal and independent of the time index. It is this assumption that greatly reduce the complexity of the chain and facilitates the estimation problem. Note that for 5 states of nature there are 25 such one-step conditional probabilities. Arrange them in a five by five matrix forms a convenient expression of the transition probability matrix or stochastic transition matrix.

**Definition 2.** *(Transition Probability Matrix) A square matrix that inputs all possible transition probabilities of corresponding states in a discrete-time Markov chain.*

Note that the rows of the matrix correspond to the starting states and columns ending states of a one-step transition. The diagonal elements signal the probability of staying in the same position over time. And the elements in each rows must be non-negative $P_{ij} \geq 0$

and must add up to one: $\sum_{j \in S} P_{ij} = 1$ for all $i$ and $j$ in $S$.

$$\boldsymbol{P} = \begin{array}{c} \\ U \\ H \\ F \\ B \\ D \end{array} \begin{array}{c} \begin{array}{ccccc} U & H & F & B & D \end{array} \\ \begin{pmatrix} P_{uu} & P_{uh} & P_{uf} & P_{ub} & P_{ud} \\ P_{hu} & P_{hh} & P_{hf} & P_{hb} & P_{hd} \\ P_{fu} & P_{fh} & P_{ff} & P_{fb} & P_{fd} \\ P_{bu} & P_{bh} & P_{bf} & P_{bb} & P_{bd} \\ P_{du} & P_{dh} & P_{df} & P_{db} & P_{dd} \end{pmatrix} \end{array} \tag{C.3}$$

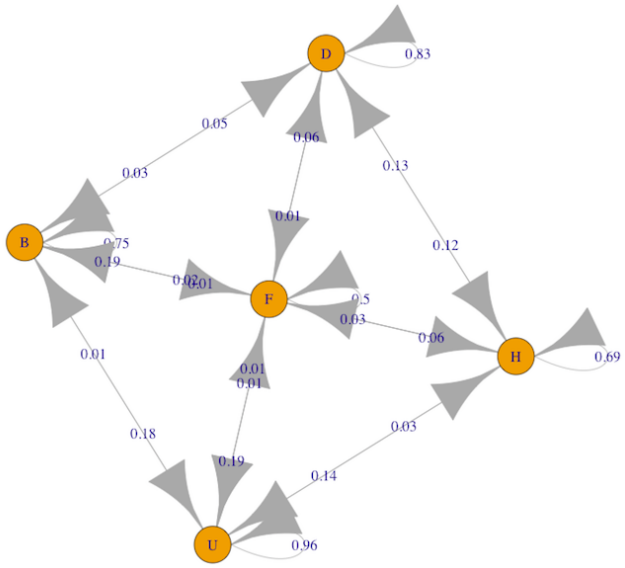
The information about the stochastic matrix can also be graphed in a transition diagram of the DTMC. In Figure A11, for the yield curve dynamics we are modeling, five circles represent five states of the chain and arrows connect the ending states from the starting states in a transition. The fractional numbers are estimated probabilities using monthly yield data.

Some of the transition probabilities can be zero if two states are not communicative in a single step. Estimation techniques of these probabilities will be discussed in detail in the next subsection. Before we move to analyze the long run properties of DTMC, it is essential to specify an initial distribution to complete the introduction to the basic components of a Markov chain.

**Definition 3.** *(Initial Distribution) The starting probabilities of each state in a Markov chain. Notationally, an initial distribution is the set* $\{\pi_i^{(0)} : i \in S\}$*, where* $\pi_i^{(0)} = P(X_0 = i)$.

Similar to transition probabilities, initial state probability must be non-negative $\pi_i^{(0)} \geq 0$ and add up to one $\sum_{i \in S} \pi_i^{(0)} = 1$. In the context of yield curve dynamics, initial distribution can be written in a vector containing the starting probabilities of each state $\boldsymbol{\pi^{(0)}} = (\pi_u^{(0)}, \pi_h^{(0)}, \pi_f^{(0)}, \pi_b^{(0)}, \pi_d^{(0)})$. Hence, a Markov chain is a sequence $\{X_t\}$ of random variables, with transition probabilities $\{P_{ij}\}$ such that $P(X_{t+1} = j | X_t = i) = P_{ij}$, and with an initial distribution $\{\pi_i^{(0)}\}$ such that $P(X_0 = i) = \pi_i^{(0)}$. With the stochastic matrix and a

Figure A11: Transition diagram of the Markov chain with estimated transition probabilities




given initial distribution of the Markov chain, we can calculate the probability of the chain entering into a particular state in the future. These future state probabilities are called transient distribution, which are of major interest in prediction.

**Definition 4.** *(Transient Distribution) The probability distribution of the states in a Markov chain at time* $t \geq 0$. *Notationally, transient distribution is the set* $\{\pi_j^{(t)} : j \in S\}$, *where* $\pi_j^{(t)} = P(X_t = j)$.

The following theorem sheds light on how to compute transient probabilities of our interest.

**Theorem 1.** $\pi_j^{(t)} = P(X_t = j) = \sum_i^S P(X_0 = i)P(X_t = j|X_0 = i) = \sum_i^S \pi_i^{(0)} P_{ij}^{(t)} = \sum_i^S \pi_i^{(0)} P_{ij}^t$.

This theorem is a direct application of the law of total probabilities but contains more information on how to compute $t$-step transition probabilities of the DTMC. The $t$-step transition probability matrix is equivalent to raising the one-step transition matrix to the power of $t$. In matrix notation, the transient distribution vector after $t$ steps $\boldsymbol{\pi^{(t)}} = (\pi_u^{(t)}, \pi_h^{(t)}, \pi_f^{(t)}, \pi_b^{(t)}, \pi_d^{(t)})$, so that

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(0)} * \boldsymbol{P}^{(t)} = \boldsymbol{\pi}^{(0)} * \boldsymbol{P}^{t} \tag{C.4}$$

A proof of the theorem employs the basic Markov property and time homogeneity assumption.

To investigate an equilibrium outcome and the long run behavior of a discrete time Markov chain (DTMC), we include two useful definitions. It turns out that if the states of a DTMC satisfy some nice properties, there is a unique solution to the equilibrium analysis.

**Definition 5.** *(Stationary Distribution) The distribution* $\{\pi_i : i \in S\}$ *is stationary for a Markov chain with transition probabilities* $\{P_{ij}\}$ *on a state space* $S$ *if* $\sum_{i \in S} \pi_i P_{ij} = \pi_j$ *for all* $j \in S$.

This definition indicates that if a Markov chain has a stationary distribution and begins with this distribution, then it will always run on this probabilities, i.e., if $P(X_t = i) = \pi_i$ for all $i$ for some $t$, then $P(X_T = i) = \pi_i$ for all $i$ for all $T > t$. It is also straightforward to solve a set of linear equations to find out the stationary distribution and this distribution may not be unique.

Writing the definition of stationary distribution in matrix form, we have

$$\boldsymbol{\pi} * \boldsymbol{P} = \boldsymbol{\pi} \tag{C.5}$$

together with the restriction

$$||\boldsymbol{\pi}||_{l1} = \sum_{i \in S} \pi_i = 1 \tag{C.6}$$

Equations (C.5) and (C.6) are the well known balance equations ( or the steady-state equations) and normalizing equation in which a stationary distribution must satisfy.

**Definition 6.** *(Limiting Distribution) As step* $t$ *move forward, the limiting distribution of a Markov chain is defined as* $\lim_{t \to \infty} P(X_t = i) = \pi_i^*$ *for each* $i \in S$..

In general, limiting distribution may depend on the initial distribution $\{\pi_i^{(0)}\}$, as explained in theorem 1. It is also easy to check that a limiting distribution, when it exists, is also a stationary distribution because it satisfies the same balance and normalizing equations. However, limiting and stationary distributions are not identical, as the way to calculate them differ; and more importantly, they bear different meanings as to the equilibrium analysis of the DTMC. The stationary distribution is a "timeless" concept whereas limiting distribution takes the infinite time horizon.

Since two distributions coincide if balancing equation share the same solution with the limiting distribution, what sort of Markov chain could have such solution? Moreover, is the solution unique if they exist? Two more definitions are introduced to answer these questions.

**Definition 7.** *(Irreducible Chain) A Markov chain is irreducible if it is possible for the chain to move from any state to any other state in one or more steps. Equivalently, irreducibility requires that there exists a positive integer $K$ such that $P_{ij}^{(K)} = P(X_K = j|X_0 = i) > 0$ for any $i, j \in S$.*

**Definition 8.** *(Aperiodic Chain) Given Markov chain transition probabilities $\{P_{ij}\}$ on a state space $S$, and a state $i \in S$, the period of $i$ is the greatest common divisor of the times at which it is possible to travel from $i$ to $i$. A Markov chain is aperiodic if the period of each state is equal to one.*

Finally, the following theorem establishes the relationship between stationary and limiting distribution, guaranteeing the unique solution to the Markov chain equilibrium distribution.

**Theorem 2.** *Suppose a Markov chain is irreducible and aperiodic and has a stationary distribution $\{\pi\}$. Then regardless of its initial distribution $\boldsymbol{\pi}^{(0)}$, we have $\lim_{t\to\infty} P(X_t = i) = \pi_i$ for all $i \in S$. And the solution is unique.*

It says that stationary and limiting distributions will be identical if the DTMC is irreducible and aperiodic. A proof of this theorem can be found in Rosenthal (2006)[3]. Theorem

[3]J. S. Rosenthal, pages 92 to 93 of A First Look at Rigorous Probability, 2nd Edition. World Scientific Publishing, Singapore.

2 also shows that for irreducible and aperiodic chains the state probabilities will converge to the stationary distribution eventually. Hence, the stationary distribution characterizes the long run behavior of a DTMC.

Therefore, if one can obtain reliable estimates of transition probabilities matrix from the yield data, by examining whether the matrix meets the irreducibility and aperiodicity properties, one may calculate, in theory, the unique stationary distribution of the chain, which is also the long run equilibrium of the state distribution.

## C.2 Model extension

The first-order Markov chain assumes that the future state only depends on present state, independent of all past memory. However, given any Markov chain, one is not sure about its underlying dependence structure. To explore the potential existence of further dependence, the first-order Markov chain can be extended to its higher-order counterparts. Though the complexity increases, it might capture more features of the data and potentially be a better model. The selection of Markov chain order can be assigned to statistical methods testing their forecast performance.

Introduced in Ching, et al. (2013), a generalized version of Raftery (1985) higher-order Markov chain model follows:

$$\boldsymbol{\pi}^{(t+N+1)} = \sum_{n=1}^{N} \lambda_n \boldsymbol{\pi}^{(t+N+1-n)} \boldsymbol{P_n}. \tag{C.7}$$

where time $t+N+1$ state distribution $\boldsymbol{\pi}^{(t+N+1)}$ is a weighted average of the past $N$ state distribution vectors $\boldsymbol{\pi}^{(t+N)}$, $\boldsymbol{\pi}^{(t+N-1)}$, ..., $\boldsymbol{\pi}^{(t+1)}$. Here, $N$ is the highest order of the Markov chain. The weight parameter is $\lambda_n$ and the $n$-step transition matrix is $\boldsymbol{P_n}$. The higher-order dependence of $\boldsymbol{\pi}^{(t+N+1)}$ on $\boldsymbol{\pi}^{(t+N+1-n)}$ is relayed by $\lambda_n$ and $P_n$ in the model. The model also assumes that the weight $\lambda_n$ is non-negative and $\sum_{n=1}^{N} \lambda_n = 1$. An advantage of the model is that it nests all lower-order Markov chains up to $N$.

Similar to the first-order Markov chain, the following theorem guarantees the existence of the unique stationary distribution.

**Theorem 3.** *If* $\boldsymbol{P_n}$ *is irreducible and aperiodic,* $\lambda_1$, $\lambda_N$*>0 and* $\sum_{n=1}^{N} \lambda_n = 1$*, then the higher-order Markov chain has a stationary distribution satisfies* $\boldsymbol{\pi^*}(I - \sum_n^N \lambda_n \boldsymbol{P_n}) = \mathbf{0}$ *with* $\boldsymbol{\pi^* 1^T}$*=1. This distribution is unique and equal to the limiting distribution* $\lim_{t\to\infty} \boldsymbol{\pi^{(t)}} = \boldsymbol{\pi^*}$.

Depending on the order $N$, the total number of parameters is $(N + NS^2)$. There are $N$ weight parameters associated with each stochastic matrix and $S^2$ transition probabilities within each stochastic matrix. While we can construct the $n$-step transition matrix by computing the relative transition frequency within each state, the estimation of weight parameters $\lambda_n$ must resort to constrained optimization implied by the theorem above. Consider the following $l_1$-norm minimization problem:

$$\underset{\lambda}{\text{Min}}\{\sum_{s=1}^{S} |(\sum_{n=1}^{N} \lambda_n \hat{\boldsymbol{\pi}} \hat{\boldsymbol{P}}_{\boldsymbol{n}} - \hat{\boldsymbol{\pi}})|_1\}$$

subject to $\sum_{n=1}^{N} \lambda_n = 1$ and $\lambda_n \geq 0$, $\forall n$. $\hat{\boldsymbol{\pi}}$ is estimated from the proportion of the occurrence of each state in the Markov chain sequence. $\hat{\boldsymbol{P}}_{\boldsymbol{n}}$ is estimated from the contingency counts. The problem can also be formulated as a linear programming problem. Solution for parameters $\lambda_n$ is then obtained. Readers may refer to Ching, et al. (2013) for details.

## C.3 Estimation strategies

The most common technique for estimating the MC transition matrix is the maximum likelihood estimation (MLE). In practice, depending on data availability and research purposes, additional techniques are used to enhance the performance of regular MLE. For instance, the bootstrap method can help determine the sampling distribution of the the MLE by resampling from the estimated MC but its performance is subject to limited data problem. The Laplace smoothing adds a strictly positive parameter to the MLE so that all state transition probabilities become positive; this method overcomes limited sample estimation bias when

the true transition is most likely to happen given a large sample. The Bayesian method combines prior knowledge of the transition density and sample information to generate an estimator that is usually more precise in terms of smaller estimation error.

### C.3.1 Regular MLE

Following the previous notation, a DTMC $\{X_t, t = 0, 1, 2, ..., T\}$ evolving in a finite state space $\{i, i \in S\}$ with transition probabilities $P_{ij} = P(X_t = j | X_{t-1} = i) = P_{ij}(t) = P_{ij}(t+1)$ for all $t$, and $0 \leq P_{ij} \leq 1$ and $\sum_j P_{ij} = 1$ for all $i \in S$. The joint probability of an ordered sequence for such a Markov chain may be written as

$$P(X_0, X_1, X_2, ..., X_T) = P(X_0) \prod_t P(X_t | X_{t-1}) = P(X_0) \prod_t P_{ij}(t) = P(X_0) \prod_{i,j} P_{ij}^{n_{ij}} \quad \text{(C.8)}$$

where $n_{ij}$ is the total number of events over $t$ for which $X_{t-1} = i$ and $X_t = j$. Equation (C.8) hence derived is the likelihood function of an observed trajectory $(X_0, X_1, X_2, ..., X_T)$. By maximizing (C.8) or just the product term $\prod_{i,j} P_{ij}^{n_{ij}}$ with respect to the $P_{ij}$ subject to probability constraints $0 \leq P_{ij} \leq 1$ and $\sum_j P_{ij} = 1$, the Lagrange multipliers method can be applied to solve the optimization problem. Take logs to equation (C.8) and form the objective function

$$\mathcal{L}(P_{ij}) = logP(X_0) + \sum_{i,j} n_{ij} * logP_{ij} - \sum_i \lambda_i (\sum_j P_{ij} - 1) \quad \text{(C.9)}$$

Differentiate with respect to $\lambda_i$ and $P_{ij}$ respectively and setting them to zero

$$\frac{n_{ij}}{P_{ij}} - \lambda_i = 0$$

Together with the probability constraint $\sum_j P_{ij} = 1$

$$\sum_j \frac{n_{ij}}{\lambda_i} = 1$$

Thus the maximum likelihood estimator of the transition probabilities $P_{ij}$ is

$$\hat{\boldsymbol{P}} = (\hat{P}_{ij}) = (\frac{n_{ij}}{\lambda_i}) = (\frac{n_{ij}}{\sum_j n_{ij}}) \geq 0 \tag{C.10}$$

Interestingly, the MLE solution turns out to be very simple and intuitive. The transition probability from state $i$ to state $j$ can be computed from the proportion of its transition count to the total transition count from $i$ to all other states. The standard error of the MLE can be derived from the information matrix

$$\hat{\sigma}(P_{ij}) = \frac{\hat{P}_{ij}^{MLE}}{\sqrt{n_{ij}}} \tag{C.11}$$

The ML estimator is consistent and converges almost surely to the true transition probability. $\hat{\boldsymbol{P}}$ can also be shown to be asymptotically normal with a rate of convergence $\sqrt{T}$. Refer to Athreya and Fuh (1992) for asymptotic properties of the MLE.

### C.3.2 MLE bootstrap

To determine the exact (sampling) distribution of the MLE for computing confidence intervals and testing hypotheses is infeasible in practice since the sampling distribution depends on the population distribution, which is unknown in most cases. A popular technique known as the bootstrap method, introduced by Bradley Efron (1979), can be used when the statistical distribution is unknown or the normality assumptions are not met. The bootstrap procedure offers a simple way to obtain a good approximate sample distribution of the MLE estimator, conditional on the observed data.

When applied to estimating Markov chains, the bootstrap method computes $\hat{\boldsymbol{P}}$ from the original chain and then resamples $B$ new samples (chains) based on $\hat{\boldsymbol{P}}$ with uniform probability distribution adopted as their initial state distribution. For each new generated resample, a maximum likelihood estimator $\hat{\boldsymbol{P}}_{\boldsymbol{b}}^{*}, b = 1, ..., B$ is obtained for constructing bootstrap estimator of the true transition probabilities matrix $\boldsymbol{P}$ and its sampling (empirical)

distribution. For instance, after obtaining $B$ resampling MLE $\hat{\boldsymbol{P}}_{\boldsymbol{1}}^{*}$, ..., $\hat{\boldsymbol{P}}_{\boldsymbol{B}}^{*}$, estimators for $E(\boldsymbol{P}_{vec})$ and $Cov(\boldsymbol{P}_{vec})$ can be computed as follows:

$$\widehat{E(\boldsymbol{P}_{\boldsymbol{vec}})} = \sum_{b=1}^{B} \hat{\boldsymbol{P}}_{\boldsymbol{b}}^{*}$$

$$\widehat{Cov(\boldsymbol{P}_{\boldsymbol{vec}})} = \frac{1}{B-1} \sum_{b=1}^{B} [\hat{\boldsymbol{P}}_{\boldsymbol{b}}^{*} - \widehat{E(\boldsymbol{P}_{\boldsymbol{vec}})}][\hat{\boldsymbol{P}}_{\boldsymbol{b}}^{*} - \widehat{E(\boldsymbol{P}_{\boldsymbol{vec}})}]'$$

In the two equation above, $\boldsymbol{P}_{\boldsymbol{vec}}$ and $\hat{\boldsymbol{P}}_{\boldsymbol{b}}^{*}$ are the column stacked vectors such that $E(\boldsymbol{P}_{vec})$ is $S^2 \times 1$ vector and $Cov(\boldsymbol{P}_{vec})$ is a $S^2 \times S^2$ matrix, so are their estimators.

The bootstrap maximum likelihood estimator has similar asymptotic behavior as the distribution of the maximum likelihood estimator (Kulperger et. al, 1990). However, the bootstrap method may not perform well if the MLE $\hat{\boldsymbol{P}}$ does not have a structure close to the true $\boldsymbol{P}$. This would most likely occur when only a limited data sequence is available.

### C.3.3 Laplace smoothing

When a sparse estimator $\hat{\boldsymbol{P}}$ of stochastic matrix results from limited sample sequence, i.e., the transition probabilities $\hat{P}_{ij} = 0$ for some $i$ and $j$, then a transition from state $i$ or $j$ never happens in the original sample. The same problem occurs also in bootstrap procedures, causing the bootstrap to perform poorly, as noticed in Guerra et. al. (1997). However, the probability of this transition may in reality be greater than zero when sample size becomes large enough. Whereas we can perform MLE to larger samples available, e.g., a higher frequency or longer span dataset, another solution to the sparse estimation problem is the Laplace smoothing method, particularly useful when large sample dataset is not feasible.

To guarantee the positive entries of the stochastic matrix, a smoothed version $\widetilde{\boldsymbol{P}}$ is given by

$$\tilde{P}_{ij} = \frac{1}{\omega}[\hat{P}_{ij} + T^{-\alpha}] \tag{C.12}$$

where

$$\omega = \sum_{j=1}^{S} [\hat{P}_{ij} + T^{-\alpha}] = \sum_{j=1}^{S} \hat{P}_{ij} + \sum_{j=1}^{S} T^{-\alpha} = 1 + S * T^{-\alpha}$$

and $\alpha > 0$ is a positive smoothing parameter, $S$ is the total number of states in the DTMC. Note that the smoothed transition probabilities satifies

$$\sum_{j=1}^{S} \tilde{P}_{ij} = \frac{1 + S * T^{-\alpha}}{\omega} = 1$$

therefore, the smoothed stochastic matrix $\widetilde{\boldsymbol{P}}$ is a valid specification.

The difficulty then rests upon the choice of the smoothing parameter $\alpha$. Consider a simpler version of the equation (C.12) and compare it with the regular maximum likelihood estimator,

$$\tilde{P}_{ij}^{L} = \frac{n_{ij} + \alpha}{\sum_{k=1}^{S} n_{ik} + \alpha} \tag{C.13}$$

where $n_{ij}$ is the total number of events over $t$ for which $X_{t-1} = i$ and $X_t = j$. Equation (C.13) is the Laplace version smoothing, which extends the regular MLE by inserting a smoothing parameter $\alpha$.

The parameter $\alpha$ acts as a "weight" in estimating the stochastic matrix. It weights smaller $n_{ij}$ more heavily than larger $n_{ij}$. Suppose with a small sample for a particular transition $i \to j$ in the observed period, no observation is made, so in effect total number of events $n_{ij}$ is zero. A positive $\alpha$ first of all ensures positive entry of the stochastic matrix. Second, when $n_{ij}$ is small, e.g., close to zero, the bigger the $\alpha$, the higher the estimate for $P_{ij}$. The opposite effect occurs to the big $n_{ik}$ for $k \neq j$. Finally, when $\alpha$ tends to infinity, $P_{ij}$ tends to $1/S$ for all $j$, a uniform smoothing case.

Technically, the choice of smoothing parameter depends on a performance criterion for $\widetilde{\boldsymbol{P}}$ in terms of some measure of performance of the resulting bootstrapping method. Another criterion applied in simulation studies is to select it to produce a consistent estimator of the true stochastic matrix at the same convergence rate as $\hat{\boldsymbol{P}}$. It can be showed that a choice of

$\alpha \geq 0.5$ will ensure that $\widetilde{\boldsymbol{P}}_T$ preserves the asymptotic consistency property of $\hat{\boldsymbol{P}}_T$.

The smoothed estimator may be more accurate than the regular MLE since it carries some information on the low transition probabilities. In a simulation study by Teodorescu (2009), the bootstrap method based on the smoothed estimation with the smallest smoothing parameter ($u = 0.5$) also shows better coverage performance of the confidence intervals. Refer to the article for further discussion and asymptotic properties of smoothed estimators.

### C.3.4 Bayesian estimation

Bayesian methods can also be applied to estimate the transition probabilities of a Markov chain. Lee, Judge, and Zellner (1968) compare the sampling properties of several Markov chain estimators and, through Monte Carlo experiments, find Bayesian methods superior to maximum likelihood and weighted least squares in terms of smaller root mean square error.

Based on the Bayes' theorem, both sample information and prior knowledge about transition densities can be combined to develop an estimator for the transition densities.[4] Considering the characteristics of the $P_{ij}$, the multivariate beta distribution (Dirichlet) is commonly chosen for the prior in Markov chain estimation. It is also known as conjugate prior because the posterior distribution hence derived would have exactly the same form as the prior up to a normalizing constant and distribution parameters. For a given state $i$ for $i \in S$, i.e., the $i^{th}$ row of the transition matrix, the prior density can be specified as

$$Prior(i) = f(P_{i1}, P_{i2}, ..., P_{iS}) = \frac{\Gamma(\sum_j a_{ij})}{\Gamma(a_{i1})\Gamma(a_{i2})...\Gamma(a_{iS})} \prod_j^S P_{ij}^{a_{ij}-1} \tag{C.14}$$

where $\Gamma(.)$ means the gamma function and $a_{i1}, a_{i2}, ..., a_{iS}$ are hyperparameters determining the shape of the prior distribution. Then the prior for the entire matrix is $Prior(ij) =$

[4]Bayes' theorem states that $P(\theta|D) = \frac{P(\theta D)}{P(D)} = \frac{P(\theta)P(\theta|D)}{P(D)}$, where $\theta$ indicates the parameter of interest and $D$ represents data. In Bayesian analysis, $P(\theta)$ reflects prior knowledge about the probability distribution of the parameter, the probability of observing a given data sample $P(D|\theta)$ is the likelihood functon, $P(D) = \int P(\theta)P(D|\theta)\, d\theta$, which can be considered as a normalizing constant. The posterior probability $P(\theta|D)$ is then computed.

$\prod_{i \in S} P(i)$.

In estimating the Markov chain, the prior can be non-informative (platykurtic prior specification), equivalent to a uniform distribution over [0, 1]. On the other hand, the prior can be quite accurate and sharp if their true values are known, which is infeasible in practice. A eclectic approach is to infer from the data sample and choose the prior based on the 'relation $a_{ij} = n_{ij} + 1$, serving as a fake count' to reflect knowledge of the data.[5] A comparison of sampling results for the Bayesian estimators with different priors is discussed in a simulation study by Lee, Judge, and Zellner (1968).

Given the prior specification and data, assuming independence, the likelihood function becomes

$$P(D|ij) = \prod_{j}^{S} \prod_{i}^{S} P_{ij}^{n_{ij}} \tag{C.15}$$

Knowing the prior and likelihood, the data evidence can be integrated out from their joint density as $P(D) = \int P(\theta) P(D|\theta)\, d\theta$, simplifying the expression further

$$P(D) = \prod_{i}^{S} \{ \frac{\Gamma(\sum_j a_{ij})}{\Gamma(a_{i1})\Gamma(a_{i2})...\Gamma(a_{iS})} \frac{\prod_j \Gamma(n_{ij} + a_{ij})}{\Gamma(n_i + \sum_j a_{ij})} \} \tag{C.16}$$

Applying Bayes' theorem, the posterior distribution can be shown to mimic the form of the prior

$$P(ij|D) = \prod_{i}^{S} \{ \frac{\Gamma(n_i + \sum_j a_{ij})}{\prod_j \Gamma(n_{ij} + a_{ij})} \prod_{j}^{S} P_{ij}^{n_{ij} + a_{ij} - 1} \} \tag{C.17}$$

Maximizing the posterior yields the MAP (maximum a posteriori) estimator,[6]

$$\hat{P}_{ij} = \{ \frac{n_{ij} + a_{ij} - 1}{n_i + \sum_j a_{ij} - S}, i \in S, j \in S \} \tag{C.18}$$

A more detailed mathematical derivation of this MAP estimation can be found in Lee, Judge

[5] With respect to the prior distribution, the mode of a parameter is proportional to one less than the corresponding hyper-parameter.

[6] The method of maximum a posteriori estimation estimates $P_{ij}$ as the mode of the posterior distribution; refer to Wikipedia page "Maximum a posteriori estimation" for detail.

and Zellner (1968). However, the MAP estimation only provides point estimates. More often, posterior mean and variance are typical choices for Bayesian estimators[7], which are given by

$$E(P_{ij}|D) = \{\frac{n_{ij} + a_{ij}}{n_i + \sum_j a_{ij}}, i \in S, j \in S\} \tag{C.19}$$

$$Var(P_{ij}|D) = \{\frac{n_{ij} + a_{ij}}{(n_i + \sum_j a_{ij})^2} \frac{n_i + \sum_j a_{ij} - n_{ij} - a_{ij}}{n_i + \sum_j a_{ij} + 1}, i \in S, j \in S\} \tag{C.20}$$

Taking the square root of (C.20) produces the standard error and the confidence intervals are constructed by computing the inverse of the beta integral.

# D Estimation results and transition dynamics

## D.1 Estimation procedure

For monthly yield data, the classification algorithm produces a time series of categorized yield curves. The Markov chain is therefore represented by a sequence of categorical variable changing from one state to another over time. In total, there are 25 possible transition outcomes characterized by a $5 \times 5$ matrix. The assumption of homogenous transition in each step simplifies this Markov process and its estimation. Using the four methods, we implement the estimation with the R package "Markovchain" written by Spedicato (2015).

The MLE estimator of the transition probabilities of state $i$ to $j$ is simply the proportion of this transition count to the total transition count from $i$ to all other states given the sample sequence. Hence, we can build a contingency table documenting the counts of all transition outcomes and compute all corresponding transition probabilities. In the following matrix, a total of 756 observations is in the sample.

[7]When squared error loss function is used, the Bayes estimator is the posterior mean. This is the result from minimizing $E[L(\theta, \hat{\theta})|D] = \int L(\theta, \hat{\theta})) * P(\theta|D)d\theta$; the MAP estimation is a limit of Bayes estimators under 0-1 loss function.

$$
\begin{array}{c} \\ U \\ H \\ F \\ B \\ D \end{array}
\begin{array}{c}
\begin{array}{ccccc} U & H & F & B & D \end{array} \\
\left(\begin{array}{ccccc}
523 & 14 & 4 & 4 & 0 \\
11 & 54 & 2 & 1 & 10 \\
3 & 1 & 8 & 3 & 1 \\
7 & 0 & 1 & 30 & 2 \\
1 & 9 & 1 & 2 & 63
\end{array}\right)
\end{array}
$$

First, each row sums up to the total occurrence of each observed state. Second, the diagonal counts are the greatest in each row, an indication of "transition inertia". Once getting into a particular state, it is more likely to stay there in the next step. Moreover, there are two zero counts: An upward yield curve never ensues a downward yield curve, neither would a bowl yield curve precede a hump yield curve. Finally, there are many one counts—transition that occurs only once in the sample period.

Now the MLE estimates are just the corresponding proportions in the contingency table. Thus, the transition probabilities matrix estimated by MLE is sparse due to two zero entries: $P_{ud} = 0$ and $P_{bh} = 0$. Note that, from the ML estimators, the standard error of them would also be zero. The bootstrap method based on such sparse MLE estimator would yield exactly the same estimates for the transition density and its standard error, no matter the number of bootstrap resamples we choose. Other than that, the bootstrap estimates for the standard errors of the transition densities shall expected to be smaller than the regular MLE. In our estimation, we consider resample length $B = 10$ and $B = 1000$.

The null transition probability estimates may be biased simply because of the existence of rare transitions in the observed sample sequence. Smoothed MLE estimator is proposed to remedy the problem by appropriately choosing a smoothing parameter. As discussed before, the asymptotic consistency property requires this parameter greater than one half. However, as the smoothing parameter $\alpha$ increases, all transition probabilities would smooth

out evenly, leading to uniform distribution over all states. While the exact choice of this parameter presents some difficulty in theory, a smaller value is preferred since it matches the empirical evidence.

In applying Bayesian method, the hyper parameters of the prior distribution (Dirichlet) are chosen such that $a_{ij} = n_{ij} + 1$ is inferred from the data sample. It turns out that, under such prior distribution, no significant difference is observed compared with the regular MLE.

## D.2 A comparison of different estimation methods

The results are organized in Table D1 for comparison. Overall, the point estimates of the transition probabilities are quite stable across four MLE methods. The regular MLE method serves as a benchmark; except for the null events and the rare events (happen only once in the sample), it provides non-zero estimates for the transition densities with statistical significance. A noticeable pattern in estimated stochastic matrix shows up in the diagonal elements: The transition probabilities $\hat{P}_{ii}$ $\forall i \in S$ dominate non-diagonal transition probabilities, which implies high momentum in the shape transitions.

The bootstrapped MLE with larger resamples performs better than small resamples. The results are closer to the MLE point estimates and much smaller variance, on average. However, as mentioned above, it may be biased toward zero, as in the case of regular MLE, when the true transition density is not zero. In that case, a Laplace smoothing is proposed. As the smoothing parameter will increase the point estimates for the small/null probability transitions, relatively small values of smoothing parameter, $\alpha = 0.5$ and $\alpha = 1$, are chosen in estimation.

The null estimates in MLE and bootstrapped MLE become positive using smoothing method. Moreover, the smoothing method tends to increase the small transition probabilities and decrease large ones. A bigger smoothing parameter has stronger weighting effect on small events. However, the smoothing estimates may also be biased if the true transition density is zero. Given a long MC sequence spanning more than 60 years, there are in fact two null

Table D1: Estimates of transition probabilities matrix for monthly sequence

| States | U | H | F | B | D |
|---|---|---|---|---|---|
| | Maximum Likelihood Estimation (MLE) | | | | |
| U | 0.9596 (0.0420) | 0.0257 (0.0068) | 0.0073 (0.0037) | 0.0073 (0.0037) | 0.0000 (0.0000) |
| H | 0.1410 (0.0425) | 0.6923 (0.0942) | 0.0256 (0.0181) | 0.0128 (0.0128) | 0.1282 (0.0405) |
| F | 0.1875 (0.1083) | 0.0625 (0.0625) | 0.5000 (0.1768) | 0.1875 (0.1083) | 0.0625 (0.0625) |
| B | 0.1750 (0.0661) | 0.0000 (0.0000) | 0.0250 (0.0250) | 0.7500 (0.1369) | 0.0500 (0.0354) |
| D | 0.0132 (0.0132) | 0.1184 (0.0395) | 0.0132 (0.0132) | 0.0263 (0.0186) | 0.8289 (0.1044) |
| | MLE bootstrap ($\#B = 10$) | | | | |
| U | 0.9542 (0.0021) | 0.0293 (0.0016) | 0.0088 (0.0014) | 0.0015 (0.0015) | 0.0000 (0.0000) |
| H | 0.1456 (0.0091) | 0.6729 (0.0135) | 0.0379 (0.0069) | 0.0214 (0.0042) | 0.1223 (0.0113) |
| F | 0.1985 (0.0296) | 0.0463 (0.0152) | 0.5046 (0.0436) | 0.1796 (0.0275) | 0.0710 (0.0149) |
| B | 0.1987 (0.0186) | 0.0000 (0.0000) | 0.0325 (0.0074) | 0.7259 (0.0212) | 0.0429 (0.0099) |
| D | 0.0172 (0.0052) | 0.1310 (0.0110) | 0.0079 (0.0043) | 0.0196 (0.0059) | 0.8242 (0.0122) |
| | MLE bootstrap ($\#B = 1000$) | | | | |
| U | 0.9591 (0.0003) | 0.0261 (0.0002) | 0.0075 (0.0001) | 0.0073 (0.0001) | 0.0000 (0.0000) |
| H | 0.1470 (0.0014) | 0.6835 (0.0018) | 0.0261 (0.0006) | 0.0128 (0.0004) | 0.1305 (0.0013) |
| F | 0.2043 (0.0039) | 0.0640 (0.0023) | 0.4646 (0.0046) | 0.2018 (0.0038) | 0.0654 (0.0023) |
| B | 0.1898 (0.0023) | 0.0000 (0.0000) | 0.0247 (0.0008) | 0.7330 (0.0026) | 0.0525 (0.0013) |
| D | 0.0146 (0.0005) | 0.1254 (0.0014) | 0.0152 (0.0005) | 0.0299 (0.0007) | 0.8148 (0.0017) |
| | MLE Laplace smoothing $\alpha = 0.5 \mid \alpha = 1$ | | | | |
| U | 0.9562 \| 0.9527 | 0.0265 \| 0.0273 | 0.0082 \| 0.0091 | 0.0082 \| 0.0091 | 0.0009 \| 0.0018 |
| H | 0.1429 \| 0.1446 | 0.6770 \| 0.6627 | 0.0312 \| 0.0361 | 0.0186 \| 0.0241 | 0.1304 \| 0.1325 |
| F | 0.1892 \| 0.1905 | 0.0811 \| 0.0952 | 0.4595 \| 0.4286 | 0.1892 \| 0.1905 | 0.0811 \| 0.0952 |
| B | 0.1765 \| 0.1778 | 0.0118 \| 0.0222 | 0.0353 \| 0.0444 | 0.7176 \| 0.6889 | 0.0588 \| 0.0667 |
| D | 0.0191 \| 0.0247 | 0.1210 \| 0.1235 | 0.0191 \| 0.0247 | 0.0318 \| 0.0370 | 0.8089 \| 0.7901 |
| | Bayesian estimates with prior inferred from data | | | | |
| U | 0.9596 (0.0062) | 0.0257 (0.0049) | 0.0073 (0.0027) | 0.0073 (0.0027) | 0.0000 (0.0009) |
| H | 0.1410 (0.0275) | 0.6923 (0.0367) | 0.0256 (0.0136) | 0.0128 (0.0106) | 0.1282 (0.0265) |
| F | 0.1875 (0.0635) | 0.0625 (0.0443) | 0.5000 (0.0808) | 0.1875 (0.0635) | 0.0625 (0.0443) |
| B | 0.1750 (0.0411) | 0.0000 (0.0116) | 0.0250 (0.0199) | 0.7500 (0.0485) | 0.0500 (0.0254) |
| D | 0.0132 (0.0109) | 0.1184 (0.0259) | 0.0132 (0.0109) | 0.0263 (0.0140) | 0.8289 (0.0313) |

Note: the monthly yield curve Markov chain in estimation is a categorical five-state sequence classified by the effective algorithm introduced in section 3. The original monthly yields data are downloaded from the Federal Reserve Board H.15 Treasury nominal yield statistics. Four MLE based methods are introduced in section 4. Standard errors are in parentheses.

transition events observed. Hence, the regular MLE estimates for the zero entries are still more reliable.

With monthly yield curve sequence, the five-state Markov chain model estimated through MLE and Bayesian method yield the same transition probability estimates: the nonzero transition probabilities are all statistically significant and the two zero estimates are derived from the empirical null transition events. In large samples, both estimators are consistent and asymptotically normal. In small samples, the Bayesian estimator performs better than MLE in terms of smaller root mean square error. Since predicting state distribution just involve a multiplication of initial state distribution and transition probabilities matrix, reliable estimates of the transition densities are necessary for the purpose of forecasting.

In principle, while sampling properties of different estimators can be compared through Monte Carlo studies as in Lee, Judge, and Zellner (1968), there is not any well-established procedure to evaluate the forecast performance in Markov chain modeling because the transient probability distribution depends not only on the estimated stochastic matrix but also on the initial distribution chosen at the beginning of the forecast.

While Bayesian methods with Dirichlet prior inferred from the data share the same point estimates with the regular MLE, it provides smaller standard errors and hence better inference. However, for the null transition events $U \rightarrow D$ and $B \rightarrow H$, the transition probability estimates in both cases (MLE and Bayesian) are zero, but Bayesian method reports non-zero standard error estimates. For the purpose of forecasting, one is more concerned about obtaining reliable non-zero transition probability estimates. Hence, regular MLE estimates will suffice.

## D.3 The equilibrium distribution of the yield curve Markov chain

What are the stationary distribution and limiting distribution of the chain? If it exists, is it unique? By definition, solving the system of equations can partially answer one of the questions. An examination of the estimated matrix can offer more interesting insights. From

the theorem, if the chain is irreducible and aperiodic, then it will have a unique stationary and limiting distribution, irrespective of its initial distribution. From the stochastic matrix, it is easy to check irreducibility since all states are communicative either in one or at most two transition steps (e.g. $U$ can reach $D$ through first entering $H$ and then $D$). Because the chain will never settle down into any single state or sub-state classes due to $0 < P_{ij} < 1$ for any $i \neq j$, it is never reducible. Checking the aperiodicity is also straightforward. Since $P_{ii} > 0$, clearly the period of state $i$ is 1 for all states. A Markov chain is aperiodic if the period of each state is one.

Thus, the unique stationary and limiting distribution can be solved from the balance equations or the limiting distribution: $P_u = 0.7219, P_h = 0.1033, P_f = 0.0212, P_b = 0.0530, P_d = 0.1007$. It is not surprising that the result is very close to the in occurrence frequency distribution ($P_u = 0.7222, P_h = 0.1032, P_f = 0.0212, P_b = 0.0529, P_d = 0.1005$). This distribution can be regarded as the long run equilibrium distribution of the chain. As demonstrated in forecast exercise, regardless of the initial distribution, the chain will eventually converge. And it could take a long time to reach equilibrium if starting from some other distributions. On the other hand, if the chain begins with the stationary distribution, it will remain there forever, unless a systematic shock hits.

## D.4 Transition dynamics: Two hypotheses

In monthly estimation, two zero transition probabilities are $P_{ud}$ and $P_{bh}$. They translate into two null events in the transition sample space: An upward yield curve never leads downward yield curve; neither does a hump yield curve lag a bowl yield curve. Nevertheless, the two opposite transitions could happen, though rarely observable ($P_{du} = 0.013$ and $P_{hb} = 0.013$). In April to May of 1980, we witness one of the sharpest decline in yield levels in the transition of D to U (around 400 basis points at the short maturity). The transition of H to B in November to December 1957 also shifts down, but to a less extent, in yield levels of all maturities (ranging from 20 to 60 basis points).

Two null transitions aside ($P_{ud} = 0$, $P_{bh} = 0$), there are a significant number of "trivially" positive transition probabilities—$P_{uh}$, $P_{uf}$, $P_{ub}$, $P_{hf}$, $P_{hb}$, $P_{bf}$, $P_{bd}$, $P_{du}$, $P_{df}$, $P_{db}$. To connect the dots between these estimates and yield curve shapes, we must examine carefully how corresponding transitions occur in the sample. The classification result for the monthly yield curve shows that different shapes of yield curves on average position themselves at different yield levels. In a descending order of yield levels, we can observe the downward, the hump, the bowl, the flat, and the upward yield curves. Thus, there exist certain intermediate transition phrases between a downward and an upward yield curve. Although possible, yield curves are less likely to jump through to non-adjacent states than to adjacent states.[8] Given the two null transitions and their opposites, we propose this feature of yield curve transition dynamics as "no-jump" hypothesis. Further, if there exists any rare jump, it seems more likely to be a downward direction than a upward direction, as observed for the case from D to U and H to B. We propose this characteristic of transition as an "easy fall" hypothesis.

Consider the transitions beginning from upward yield curves. At first glance, the transitions of U to H seems to violate both "no-jump" and "easy fall" hypotheses, given the spatial distribution of five types of yield curves. At most, this transition probability should also be smaller than $U \to F$ or $U \to B$. How can the upward yield curve jump over two intermediate phrases to the hump shape yield curve with a higher probability than the other way around? To provide answers, each of these "jumps" is documented in Table D2. Out of these 14 "jumps", a majority are "pseudo-jumps" and only one levels up around 100 basis points. Moreover, there are 5 "jumps" that do not rise in levels. Instead, they shift down from an upward yield curve to a hump shape or cross each other. To understand such "exotic" dynamics, one needs to bear in mind that an upward yield curve displays "hump" shape (concave) most often due to the price-yield convexity effect (long-term bond prices are more sensitive to yield changes). But an upward yield curve could also display convex shape, which is observed in recent years. Since a bowl yield curve can never precedes a hump yield

[8]This implication is derived from Prof. Liuren Wu's comments on the term structure transition dynamics.

curve in the sample, it must be the case that the 14 transitions from U to H are within the "hump" family and there are just relative increases in median yield to short and long yields, not jumping through the bowl shape. Hence, there is no jump in the transition from U to H. The observed "jumps" are due to their proximity in position.

Compared with others, the transition dynamics from a hump yield curve to others seem "well behaved": It shift down in yield levels to bowl shape once (November 1957), and to slight hump shape twice (December 1956 and April 1989). The transitions from downward yield curves to others also support the two transition hypotheses. A seemingly puzzling "jump" happens to the transitions from a bowl to downward yield curves—the hump shape is not an intermediate phrase. They occur three times when yield levels are relatively high (from 7% to 9%), and each time with relatively larger increases in the short and median yields than in the long yield. Daily yield data also show that a bowl yield curve could pass through a flat type when median yields rise fast enough relative to short yields, and then it becomes downward sloping (i.e., in June 5 to 9, 1989 and August 8 to 22, 1989).

These facts imply that a yield curve can never experience a sudden level jump from U to D, or from B to H, without passing through intermediate phrases. But it can "jump" from a B to a D—the H shape is not even a transition phrase between the two. Yield data of daily frequency also confirm this point. On the contrary, a yield curve seems to experience a sudden level drop but this is an extreme event recorded in monthly data. Indeed, such a sudden decline is a measurement error due to over-smoothing in monthly data. No transition like $D \to U$ or $H \to B$ is recorded in daily data. Hence, there is no jump at all in yield curve transitions, regardless of the direction.

Table D2: Transitions analyzed in estimation results for monthly sequence

| Transition | $Y_s$ | $Y_m$ | $Y_l$ | Shift in levels |
|---|---|---|---|---|
| $U \to H$ ($P_{uh} = 0.026$) | | | | |
| 1956.03 : 04 | 2.61 : 2.92 | 2.93 : 3.19 | 2.99 : 3.10 | Yes, up. |
| 1956.06 : 07 | 2.74 : 2.76 | 2.97 : 3.10 | 3.00 : 3.08 | Yes, up. |
| 1958.12 : 1959.01 | 3.29 : 3.36 | 3.80 : 3.98 | 3.86 : 3.95 | Yes, up. |
| 1959.03 : 04 | 3.61 : 3.72 | 3.98 : 4.09 | 3.99 : 4.06 | Yes, up. |
| 1967.06 : 07 | 4.48 : 5.01 | 4.97 : 5.15 | 4.99 : 5.01 | Yes, up. |
| 1971.05 : 06 | 5.04 : 5.64 | 6.27 : 6.51 | 6.32 : 6.38 | Yes, up. |
| 1972.07 : 08 | 4.96 : 4.98 | 5.99 : 6.05 | 6.01 : 5.94 | No, cross. |
| 1980.08 : 09 | 10.2 : 11.5 | 10.8 : 11.6 | 11.0 : 11.4 | Yes, up. |
| 1981.12 : 1982.01 | 12.8 : 13.7 | 13.6 : 14.6 | 13.6 : 14.4 | Yes, up 100 bsp. |
| 1982.08 : 09 | 10.3 : 9.62 | 12.8 : 12.2 | 12.8 : 12.1 | Yes, down 70 bsp. |
| 1984.08 : 09 | 11.4 : 11.2 | 12.6 : 12.4 | 12.6 : 12.4 | Yes, down. |
| 1988.11 : 12 | 8.27 : 8.68 | 8.81 : 9.12 | 9.02 : 9.01 | No, cross. |
| 1990.02 : 03 | 8.08 : 8.27 | 8.43 : 8.62 | 8.50 : 8.56 | Yes, up. |
| 2000.01 : 02 | 5.79 : 5.98 | 6.57 : 6.64 | 6.75 : 6.39 | No, cross. |
| $U \to B$ ($P_{ub} = 0.007$) | | | | |
| 1973.02 : 03 | 6.19 : 6.85 | 6.61 : 6.78 | 6.88 : 6.91 | Yes, up. |
| 1998.08 : 09 | 5.13 : 4.75 | 5.30 : 4.70 | 5.60 : 5.29 | Yes, down. |
| 2006:06 : 07 | 4.99 : 5.12 | 5.09 : 5.07 | 5.22 : 5.19 | No, cross. |
| 2007.06 : 07 | 4.79 : 4.95 | 5.03 : 4.89 | 5.25 : 5.15 | No, cross. |
| $U \to F$ ($P_{uf} = 0.007$) | | | | |
| 1964.11 : 12 | 3.91 : 4.02 | 4.07 : 4.12 | 4.17 : 4.18 | Yes, up. |
| 1965.08 : 09 | 4.07 : 4.20 | 4.20 : 4.26 | 4.25 : 4.30 | Yes, up. |
| 1978.07 : 08 | 8.39 : 8.31 | 8.55 : 8.36 | 8.67 : 8.46 | Yes, down. |
| 2006.01 : 02 | 4.35 : 4.57 | 4.38 : 4.60 | 4.65 : 4.64 | No, cross. |
| $H \to B$ ($P_{hb} = 0.013$ ) | | | | |
| 1957.11 : 12 | 3.57 : 3.18 | 3.71 : 3.13 | 3.61 : 3.38 | Yes, down. |
| $H \to F$ ($P_{hf} = 0.026$) | | | | |
| 1956.12 : 1957.01 | 3.68 : 3.37 | 3.68 : 3.46 | 3.45 : 3.41 | Yes, down. |
| 1989.04 : 05 | 9.16 : 8.87 | 9.31 : 8.93 | 9.03 : 8.83 | Yes, down. |
| $B \to F$ ($P_{bf} = 0.025$) | | | | |
| 1989.07 : 08 | 8.02 : 8.17 | 7.89 : 8.12 | 8.08 : 8.12 | Yes, up. |
| $B \to D$ ($P_{bd} = 0.050$) | | | | |
| 1973.06 : 07 | 7.31 : 8.39 | 6.81 : 7.31 | 7.06 : 7.29 | Yes, a big up. |
| 1978.10 : 11 | 9.14 : 10.0 | 8.67 : 8.98 | 8.68 : 8.75 | Yes, up. |
| $D \to U$ ($P_{du} = 0.013$) | | | | |
| 1980.04 : 05 | 13.3 : 9.39 | 11.9 : 9.82 | 11.4 :10.4 | Yes, big up. |
| $D \to F$ ($P_{df} = 0.013$) | | | | |
| 1967.01 : 02 | 4.75 : 4.71 | 4.68 : 4.70 | 4.51 : 4.61 | No, cross. |
| $D \to B$ ($P_{db} = 0.026$) | | | | |
| 1973.08 : 09 | 8.82 : 8.31 | 7.61 : 7.12 | 7.61 : 7.25 | Yes, down. |
| 2000.09 : 10 | 6.19 : 6.21 | 5.96 : 5.82 | 5.96 : 5.92 | No, cross. |

## D.5 Other data frequencies

Scrutinizing the estimation result for higher frequency data would serve as a robustness check and provide further insight into the shape transitions. Although the monthly sequence has a large sample size of 756, it smoothes out the variation in daily data; and the estimates may be biased, especially for the zero transition probabilities. In other words, the monthly sequence may contain more null transitions because it is likely to omit higher frequency transition details. Surprisingly, this argument turns out to be wrong: The monthly sequence overestimates the transition density for a null transition event $D \to U$, whereas this event is not observable with daily and weekly data!

The weekly and daily Treasury yields data are described in Section A2. The available monthly yield data span longer horizon from April 1953 to March 2016 than the daily and weekly series, which cover January 1962 to May 2016. The weekly series takes the yields on Fridays as a proxy instead of averaging over the business days in a week. The weekly and daily series are chosen to cover the entire sample period available but the daily series has 607 missing observations across all maturities spreading out in the sample. In estimation, these observations are omitted.

In Table D3, the occurrence frequencies based on the classification algorithm are calculated for daily, weekly and monthly sequences respectively. The daily and weekly results are very close within a maximum of 0.17% difference in the bowl class. The difference between the monthly and the rest is not that significant, with a maximum of 1.42% gap in hump class. The monthly data are observed to have a little more frequent hump and flat yield curves. This may arise from different sample periods since the monthly data has dated 10 years back to 1953, which includes a denser period of hump and flat types. Based on this table, we expect the estimated transition matrices will also be close to each other.

In Table D4, state transition counts are listed for three data frequencies. The transition process displays significant inertia because the diagonal numbers dominate others in each row. Surprisingly, compared with monthly sequence, higher frequency series tell a different

Table D3: Occurrence frequency for daily, weekly, and monthly yield curve sequences

| Frequency | Upward (U) | Hump (H) | Flat (F) | Bowl (B) | Downward (D) | Total Obs |
|---|---|---|---|---|---|---|
| Daily | 9891 (72.85%) | 1208 (8.90%) | 245 (1.80%) | 855 (6.30%) | 1378 (10.14%) | 13577 (100%) |
| Weekly | 2069 (72.93%) | 254 (8.95%) | 51 (1.80%) | 174 (6.13%) | 289 (10.19%) | 2837 (100%) |
| Monthly | 546 (72.22%) | 78 (10.32%) | 16 (2.12%) | 40 (5.29%) | 76 (10.05%) | 756 (100%) |

Note: Author's yield curve classification and calculation. Data are downloaded from Federal Reserve Board H.15 interest rate statistics. Daily sequence from 1962.1.2 to 2016.5.13; weekly sequence from 1962.1.5 to 2016.5.13; monthly sequence from 1953.4 to 2016.3. Yield curve shape notation: U—upward, H—hump, F—flat, B—bowl, D—downward.

transition story for the two events $H \to B$ and $D \to U$, where they disappear in daily observation. A closer examination on the monthly sequence before 1962 reveals that the event $H \to B$ happened once in the transition from November to December 1957 but the event $D \to U$ is not observable in the period. Therefore, the estimated transition probability of event $D \to U$ is biased upward using monthly data. It should be zero, not 1.32%. Whether or not the estimated transition probability (1.28%) of $H \to B$ is biased for the monthly data remains to be answered when daily frequency data become available.

Table D4: Transition contingency counts for the daily, weekly, and monthly yield curves

| | Daily Sequence | | | | | Weekly Sequence | | | | | Monthly Sequence | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | 9813 | 29 | 14 | 34 | 0 | 2037 | 13 | 5 | 13 | 0 | 523 | 14 | 4 | 4 | 0 |
| H | 29 | 1133 | 13 | 0 | 33 | 11 | 222 | 7 | 0 | 14 | 11 | 54 | 2 | 1 | 10 |
| F | 13 | 17 | 197 | 10 | 8 | 5 | 6 | 31 | 3 | 6 | 3 | 1 | 8 | 3 | 1 |
| B | 35 | 0 | 11 | 798 | 11 | 15 | 0 | 2 | 153 | 4 | 7 | 0 | 1 | 30 | 2 |
| D | 0 | 29 | 10 | 13 | 1326 | 0 | 13 | 6 | 5 | 265 | 1 | 9 | 1 | 2 | 63 |

Table D5 reports Bayesian estimates of transition densities for data of three frequencies. Stronger momentum are present in higher-frequency shape transitions. For diagonal elements ($P_{ii}$), daily sequence produces higher estimates than the monthly: More than 10% higher in $P_{hh}$ and $P_{dd}$, almost 20% higher in $P_{bb}$, and 30% higher in $P_{ff}$; the weekly sequence also produces higher estimates than the monthly counterparts, though not as dramatic as the daily estimates. Correspondingly, the off-diagonal estimates with higher frequency data

are smaller than their counterpart estimates with the monthly sequence. For the cross-diagonal elements, daily and weekly sequences produce more zero estimates for $P_{hb}$ and $P_{du}$ since these events are absent in daily and weekly sequence. While such large discrepancies do not necessarily indicate biases in estimation, they might, on the contrary, reflect the estimation robustness and the equilibrium nature of a stationary Markov chain: Higher frequency chain takes more steps (not more calendar time) to converge. Table D6 and D7 include the estimation results using four different estimation methods for the daily and weekly sequences.

Similar forecast procedures can be applied to the daily and weekly sequences. Depending on our forecast horizon and purpose, the choice of data frequency, initial distribution, and sampling transition probabilities could vary. See Appendix F for details.

Table D5: Bayesian estimates of stochastic matrix for daily, weekly and monthly transitions

| | Daily Sequence (1962.1.2 to 2016.5.13) | | | | |
|---|---|---|---|---|---|
| States | U | H | F | B | D |
| U | 0.9922 (0.0006) | 0.0029 (0.0004) | 0.0014 (0.0003) | 0.0034 (0.0004) | 0.0000 (0.0000) |
| H | 0.0240 (0.0031) | 0.9379 (0.0050) | 0.0108 (0.0021) | 0.0000 (0.0004) | 0.0273 (0.0033) |
| F | 0.0531 (0.0102) | 0.0694 (0.0115) | 0.8041 (0.0180) | 0.0408 (0.0091) | 0.0327 (0.0082) |
| B | 0.0409 (0.0048) | 0.0000 (0.0006) | 0.0129 (0.0028) | 0.9333 (0.0061) | 0.0129 (0.0028) |
| D | 0.0000 (0.0004) | 0.0210 (0.0028) | 0.0073 (0.0017) | 0.0094 (0.0019) | 0.9622 (0.0037) |
| | Weekly Sequence (1962.1.5 to 2016.5.13) | | | | |
| U | 0.9850 (0.0019) | 0.0063 (0.0013) | 0.0024 (0.0008) | 0.0063 (0.0013) | 0.0000 (0.0002) |
| H | 0.0433 (0.0091) | 0.8740 (0.0150) | 0.0276 (0.0074) | 0.0000 (0.0019) | 0.0551 (0.0102) |
| F | 0.0980 (0.0292) | 0.1176 (0.0314) | 0.6078 (0.0473) | 0.0588 (0.0238) | 0.1176 (0.0314) |
| B | 0.0862 (0.0150) | 0.0000 (0.0028) | 0.0115 (0.0063) | 0.8793 (0.0179) | 0.0230 (0.0084) |
| D | 0.0000 (0.0017) | 0.0450 (0.0087) | 0.0208 (0.0061) | 0.0173 (0.0056) | 0.9170 (0.0118) |
| | Monthly Sequence (1953.4 to 2016.3) | | | | |
| U | 0.9596 (0.0062) | 0.0257 (0.0049) | 0.0073 (0.0027) | 0.0073 (0.0027) | 0.0000 (0.0009) |
| H | 0.1410 (0.0275) | 0.6923 (0.0367) | 0.0256 (0.0136) | 0.0128 (0.0106) | 0.1282 (0.0265) |
| F | 0.1875 (0.0635) | 0.0625 (0.0443) | 0.5000 (0.0808) | 0.1875 (0.0635) | 0.0625 (0.0443) |
| B | 0.1750 (0.0411) | 0.0000 (0.0116) | 0.0250 (0.0199) | 0.7500 (0.0485) | 0.0500 (0.0254) |
| D | 0.0132 (0.0109) | 0.1184 (0.0259) | 0.0132 (0.0109) | 0.0263 (0.0140) | 0.8289 (0.0313) |

Note: The Markov chain sequence is constructed from the H.15 Treasury nominal yield statistics using the classification algorithm; Bayesian standard errors are in parentheses. Yield curve shape notation: U—upward, H—hump, F—flat, B—bowl, D—downward.

Table D6: Estimates of transition probabilities matrix for daily sequence

| States | U | H | F | B | D |
|---|---|---|---|---|---|
| | Maximum Likelihood Estimation (MLE) | | | | |
| U | 0.9922 (0.0100) | 0.0029 (0.0005) | 0.0014 (0.0004) | 0.0034 (0.0006) | 0.0000 (0.0000) |
| H | 0.0240 (0.0045) | 0.9379 (0.0279) | 0.0108 (0.0030) | 0.0000 (0.0000) | 0.0273 (0.0048) |
| F | 0.0531 (0.0147) | 0.0694 (0.0168) | 0.8041 (0.0573) | 0.0408 (0.0129) | 0.0327 (0.0115) |
| B | 0.0409 (0.0069) | 0.0000 (0.0000) | 0.0129 (0.0039) | 0.9333 (0.0330) | 0.0129 (0.0039) |
| D | 0.0000 (0.0000) | 0.0210 (0.0039) | 0.0073 (0.0023) | 0.0094 (0.0026) | 0.9622 (0.0264) |
| | MLE bootstrap ($\#B = 10$) | | | | |
| U | 0.9923 (0.0003) | 0.0029 (0.0002) | 0.0013 (0.0001) | 0.0035 (0.0002) | 0.0000 (0.0000) |
| H | 0.0263 (0.0020) | 0.9375 (0.0030) | 0.0101 (0.0076) | 0.0000 (0.0000) | 0.0261 (0.0015) |
| F | 0.0548 (0.0056) | 0.0616 (0.0028) | 0.8048 (0.0076) | 0.0432 (0.0041) | 0.0356 (0.0044) |
| B | 0.0422 (0.0024) | 0.0000 (0.0000) | 0.0106 (0.0015) | 0.9360 (0.0033) | 0.0112 (0.0011) |
| D | 0.0000 (0.0000) | 0.0217 (0.0015) | 0.0068 (0.0007) | 0.0097 (0.0007) | 0.9618 (0.0021) |
| | MLE bootstrap ($\#B = 1000$) | | | | |
| U | 0.9922 (0.0000) | 0.0029 (0.0000) | 0.0014 (0.0000) | 0.0035 (0.0000) | 0.0000 (0.0000) |
| H | 0.0246 (0.0001) | 0.9370 (0.0002) | 0.0109 (0.0001) | 0.0000 (0.0000) | 0.0275 (0.0000) |
| F | 0.0531 (0.0004) | 0.0706 (0.0005) | 0.8024 (0.0008) | 0.0413 (0.0004) | 0.0326 (0.0004) |
| B | 0.0413 (0.0002) | 0.0000 (0.0000) | 0.0130 (0.0001) | 0.9326 (0.0003) | 0.0131 (0.0001) |
| D | 0.0000 (0.0000) | 0.0212 (0.0001) | 0.0074 (0.0001) | 0.0096 (0.0001) | 0.9617 (0.0002) |
| | MLE Laplace smoothing $\alpha = 0.5 \mid \alpha = 1$ | | | | |
| U | 0.9920 \| 0.9918 | 0.0030 \| 0.0030 | 0.0015 \| 0.0015 | 0.0035 \| 0.0035 | 0.0001 \| 0.0001 |
| H | 0.0244 \| 0.0247 | 0.9364 \| 0.9349 | 0.0112 \| 0.0115 | 0.0004 \| 0.0008 | 0.0277 \| 0.0280 |
| F | 0.0545 \| 0.0560 | 0.0707 \| 0.0720 | 0.7980 \| 0.7920 | 0.0424 \| 0.0440 | 0.0343 \| 0.0360 |
| B | 0.4140 \| 0.0419 | 0.0006 \| 0.0012 | 0.0134 \| 0.0134 | 0.9312 \| 0.9291 | 0.0134 \| 0.0140 |
| D | 0.0004 \| 0.0007 | 0.0214 \| 0.0217 | 0.0076 \| 0.0079 | 0.0098 \| 0.0101 | 0.9609 \| 0.9595 |
| | Bayesian estimates with prior inferred from data | | | | |
| U | 0.9922 (0.0006) | 0.0029 (0.0004) | 0.0014 (0.0003) | 0.0034 (0.0004) | 0.0000 (0.0000) |
| H | 0.0240 (0.0031) | 0.9379 (0.0050) | 0.0108 (0.0021) | 0.0000 (0.0004) | 0.0273 (0.0033) |
| F | 0.0531 (0.0102) | 0.0694 (0.0115) | 0.8041 (0.0180) | 0.0408 (0.0091) | 0.0327 (0.0082) |
| B | 0.0409 (0.0048) | 0.0000 (0.0006) | 0.0129 (0.0028) | 0.9333 (0.0061) | 0.0129 (0.0028) |
| D | 0.0000 (0.0004) | 0.0210 (0.0028) | 0.0073 (0.0017) | 0.0094 (0.0019) | 0.9622 (0.0037) |

Note: The daily yield curve Markov chain (1962.1.2 to 2016.5.13) in estimation is a categorical five-state sequence classified by an effective algorithm introduced in section 3. The original daily yields data are downloaded from the Federal Reserve Board H.15 Treasury nominal yield statistics. There are 607 missing observations in the daily data. It is assumed the sequence is continuous on a business day basis. Four MLE based methods are introduced in section 4. Standard errors are in parentheses. Yield curve shape notation: U—upward, H—hump, F—flat, B—bowl, D—downward.

Table D7: Estimates of transition probabilities matrix for weekly sequence

| States | U | H | F | B | D |
|---|---|---|---|---|---|
| | Maximum Likelihood Estimation (MLE) | | | | |
| U | 0.9850 (0.0218) | 0.0063 (0.0017) | 0.0024 (0.0011) | 0.0063 (0.0017) | 0.0000 (0.0000) |
| H | 0.0433 (0.0131) | 0.8740 (0.0587) | 0.0276 (0.0104) | 0.0000 (0.0000) | 0.0551 (0.0563) |
| F | 0.0980 (0.0438) | 0.1176 (0.0480) | 0.6078 (0.1092) | 0.0588 (0.0240) | 0.1176 (0.0480) |
| B | 0.0862 (0.0223) | 0.0000 (0.0000) | 0.0115 (0.0081) | 0.8793 (0.0711) | 0.0230 (0.0115) |
| D | 0.0000 (0.0000) | 0.0450 (0.0125) | 0.0208 (0.0085) | 0.0173 (0.0077) | 0.9170 (0.0563) |
| | MLE bootstrap ($\#B = 10$) | | | | |
| U | 0.9866 (0.0008) | 0.0062 (0.0005) | 0.0016 (0.0004) | 0.0056 (0.0004) | 0.0000 (0.0000) |
| H | 0.0412 (0.0031) | 0.8729 (0.0067) | 0.0361 (0.0034) | 0.0000 (0.0000) | 0.0498 (0.0037) |
| F | 0.1291 (0.0173) | 0.1323 (0.0213) | 0.5710 (0.0413) | 0.0692 (0.0112) | 0.0984 (0.0126) |
| B | 0.0866 (0.0090) | 0.0000 (0.0000) | 0.0113 (0.0032) | 0.8796 (0.0091) | 0.0226 (0.0022) |
| D | 0.0000 (0.0000) | 0.0543 (0.0061) | 0.0196 (0.0036) | 0.0213 (0.0038) | 0.9048 (0.0050) |
| | MLE bootstrap ($\#B = 1000$) | | | | |
| U | 0.9849 (0.0001) | 0.0063 (0.0001) | 0.0024 (0.0000) | 0.0064 (0.0001) | 0.0000 (0.0000) |
| H | 0.0453 (0.0004) | 0.8691 (0.0007) | 0.0286 (0.0003) | 0.0000 (0.0000) | 0.0569 (0.0005) |
| F | 0.1022 (0.0014) | 0.1186 (0.0016) | 0.5979 (0.0022) | 0.0604 (0.0011) | 0.1209 (0.0014) |
| B | 0.0897 (0.0008) | 0.0000 (0.0000) | 0.0116 (0.0003) | 0.8750 (0.0009) | 0.0237 (0.0004) |
| D | 0.0000 (0.0000) | 0.0470 (0.0005) | 0.0215 (0.0003) | 0.0181 (0.0003) | 0.9134 (0.0006) |
| | MLE Laplace smoothing $\alpha = 0.5 \mid \alpha = 1$ | | | | |
| U | 0.9841 \| 0.9831 | 0.0065 \| 0.0068 | 0.0027 \| 0.0029 | 0.0065 \| 0.0068 | 0.0002 \| 0.0005 |
| H | 0.0448 \| 0.0463 | 0.8674 \| 0.8610 | 0.0292 \| 0.0309 | 0.0019 \| 0.0039 | 0.0565 \| 0.0579 |
| F | 0.1028 \| 0.1071 | 0.1215 \| 0.1250 | 0.5888 \| 0.5714 | 0.0654 \| 0.0714 | 0.1215 \| 0.1250 |
| B | 0.0878 \| 0.0894 | 0.0028 \| 0.0056 | 0.0142 \| 0.0168 | 0.8697 \| 0.8603 | 0.0255 \| 0.0279 |
| D | 0.0017 \| 0.0034 | 0.0463 \| 0.0476 | 0.0223 \| 0.0238 | 0.0189 \| 0.0204 | 0.9108 \| 0.9048 |
| | Bayesian estimates with prior inferred from data | | | | |
| U | 0.9850 (0.0019) | 0.0063 (0.0013) | 0.0024 (0.0008) | 0.0063 (0.0013) | 0.0000 (0.0002) |
| H | 0.0433 (0.0091) | 0.8740 (0.0150) | 0.0276 (0.0074) | 0.0000 (0.0019) | 0.0551 (0.0102) |
| F | 0.0980 (0.0292) | 0.1176 (0.0314) | 0.6078 (0.0473) | 0.0588 (0.0238) | 0.1176 (0.0314) |
| B | 0.0862 (0.0150) | 0.0000 (0.0028) | 0.0115 (0.0063) | 0.8793 (0.0179) | 0.0230 (0.0084) |
| D | 0.0000 (0.0017) | 0.0450 (0.0087) | 0.0208 (0.0061) | 0.0173 (0.0056) | 0.9170 (0.0118) |

Note: The weekly yield curve Markov chain (1962.1.5 to 2016.5.13) in estimation is a categorical five-state sequence classified by an effective algorithm introduced in section 3. The original weekly yields data are downloaded from the Federal Reserve Board H.15 Treasury nominal yield statistics. Four MLE based methods are introduced in section 4. Standard errors are in parentheses. Yield curve shape notation: U—upward, H—hump, F—flat, B—bowl, D—downward.

# E Higher-order Markov chain k-fold cross validation

Applying k-fold cross-validation to Markov chain models of different order, one can compare average prediction error rates associated with different orders and choose the optimal model. A summary of k-fold C.V. algorithm is laid out for evaluating Markov chain models: (1) Split the Markov chain sequence into k sequential folds with approximately equal sample size; (2) Estimate the transition probabilities matrix using k-1 sequential folds and compute the average test error rate using the hold-out fold, assuming deterministic state vectors of 0-1 entries; (3) Repeat step 2. for each hold-out dataset: estimate transition probabilities matrix $k$ times and use them to compute the corresponding $k$ average test error rates; (4) Calculate the mean of all k-fold average test error rates for the first order Markov chain model; (5) Repeat step 1 to 4 for higher order Markov chain models, and select the model with the lowest mean prediction error rate.

Two cases of C.V. include the in-sample and out-of-sample forecast, as commonly practiced in econometric model evaluation. When $k = 1$, k-fold C.V. is equivalent to in-sample prediction using all the observations. When $k = 2$, it nests out-of-sample forecast and sub-sample estimation. Further, when $k = n$, the total number of sample observations, it is the most expensive case and amounts to training and testing the model $n$ times with each time leaving out one data point. This is therefore called leave-one-out cross validation (LOOCV). In practice, the choice of $k$ should balance the bias-variance in test errors. Typically, one chooses $k = 5$ or 10, as these values performs well in empirical estimation (James et al., 2013, chapter 5.1). K-fold C.V. estimation results can also be used for sub-sample analysis when estimation is performed for the corresponding training data.

When k=1, the cross-validation result amounts to the full-sample test. The average prediction error rate is computed for each model by $Ave(I(S_t \neq \hat{S}_t)) = \frac{1}{T}\sum_{t=1}^{T} I(S_t \neq \hat{S}_t)$, where $\hat{S}_t$ is the predicted state for the $t$th observation in the chain, $T$ is the number of data points in the test set. And $I(S_t \neq \hat{S}_t)$ is an indicator variable that equals 1 if $S_t \neq \hat{S}_t$ and 0 if $S_t = \hat{S}_t$. Hence this formula computes the fraction of incorrect predictions in the test set.

When k=2, the cross-validation result nests the sub-sample analysis and out-of-sample test. The data are divided into two folds each with equal number of observations. The first sub-sample 1953.04 to 1984.09 serves as training set for estimation and the second sub-sample 1984.10 to 2016.03 as test set for validation. The average prediction error rate is calculated from the validation set for each model. Then, using the second sub-sample as the training set in estimation, one can calculate the average prediction error rate for the first sub-sample. To evaluate the model, one computes the mean of these two average prediction error rates. The sub-sample analysis shows that the estimated transition probability matrices using two samples are significantly different. Moreover, for Markov chain models up to three orders, first sub-sample estimation predicts second sub-sample states much more precise than the other way around. Results from 5-fold C.V. also indicate that the first sub-sample states are much harder to predict.

Table E1: K=1 full-sample mean prediction error rate calculation

1st-order MC prediction equation: $\boldsymbol{\pi}^{(t+1)}=\lambda_1\boldsymbol{\pi}^{(t)}\boldsymbol{P_1}$. Average prediction error rate: 10.20%.

2nd-order MC prediction equation: $\boldsymbol{\pi}^{(t+1)}=\lambda_1\boldsymbol{\pi}^{(t)}\boldsymbol{P_1}+\lambda_2\boldsymbol{\pi}^{(t-1)}\boldsymbol{P_2}$. Average prediction error rate: 11.54%.

3rd-order MC prediction equation: $\boldsymbol{\pi}^{(t+1)}=\lambda_1\boldsymbol{\pi}^{(t)}\boldsymbol{P_1}+\lambda_2\boldsymbol{\pi}^{(t-1)}\boldsymbol{P_2}+\lambda_3\boldsymbol{\pi}^{(t-2)}\boldsymbol{P_3}$. Average P.E.R.: 12.61%.

| | $P_1$ transition matrix | | | | | $P_2$ transition matrix | | | | | $P_3$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .960 | .026 | .007 | .007 | .000 | .936 | .039 | .011 | .011 | .004 | .921 | .046 | .013 | .015 | .006 |
| H | .141 | .692 | .026 | .013 | .128 | .218 | .564 | .051 | .013 | .154 | .244 | .474 | .051 | .038 | .192 |
| F | .188 | .063 | .500 | .188 | .063 | .313 | .125 | .250 | .188 | .125 | .375 | .186 | .186 | .186 | .063 |
| B | .175 | .000 | .025 | .750 | .050 | .225 | .000 | .025 | .650 | .100 | .275 | .025 | .025 | .576 | .100 |
| D | .013 | .118 | .013 | .026 | .829 | .053 | .145 | .013 | .053 | .737 | .092 | .158 | .013 | .066 | .671 |

First order MC $\lambda_1 = 1$; Second-order MC $\lambda_1 = \lambda_2 = 0.5$; Third-order MC $\lambda_1 = \lambda_2 = \lambda_3 = 0.3333$.

Note: The monthly yield curve Markov chain (1953.4 to 2016.3) in estimation is a categorical five-state sequence classified by the effective algorithm. Prediction error is an indicator function equals one when the model predicted state is not the same as the observed state. Average prediction error rate is defined as the percentage of total prediction errors in the sample. Data are from Federal Reserve Board H.15 interest rate statistics.

Table E2: K=2 fold C.V. mean prediction error rate calculation

1st-order MC mean prediction error rate: 12.57%=(5.56+19.58)/2.

2nd-order MC mean prediction error rate: 19.84%=(7.98+31.76)/2.

3rd-order MC mean prediction error rate: 21.69%=(8.73+34.66)/2.

Training set 1 (estimation): 1953.04–1984.09 (378 obs); Test set 1 (validation): 1984.10–2016.03 (378 obs)

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}P_1^1}$. Average prediction error rate: 5.56%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}P_1^1}+\lambda_2\boldsymbol{\pi^{(t-1)}P_2^1}$. Average prediction error rate: 7.98%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}P_1^1}+\lambda_2\boldsymbol{\pi^{(t-1)}P_2^1}+\lambda_3\boldsymbol{\pi^{(t-2)}P_3^1}$. Average P.E.R.: 8.73%.

| | $P_1^1$ transition matrix | | | | | $P_2^1$ transition matrix | | | | | $P_3^1$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .929 | .052 | .014 | .005 | .000 | .885 | .072 | .019 | .014 | .010 | .856 | .087 | .024 | .019 | .014 |
| H | .141 | .688 | .016 | .016 | .141 | .219 | .578 | .031 | .016 | .156 | .219 | .578 | .031 | .016 | .156 |
| F | .250 | .125 | .375 | .125 | .125 | .250 | .250 | .125 | .125 | .250 | .250 | .375 | .000 | .375 | .000 |
| B | .095 | .000 | .000 | .810 | .095 | .143 | .000 | .000 | .667 | .190 | .190 | .000 | .000 | .619 | .190 |
| D | .014 | .122 | .014 | .014 | .838 | .054 | .149 | .014 | .027 | .757 | .095 | .162 | .014 | .041 | .689 |

Training set 2 (estimation): 1984.10–2016.03 (378 obs); Test set 2(validation): 1953.04–1984.09 (378 obs)

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}P_1^2}$. Average prediction error rate: 19.58%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}P_1^2}+\lambda_2\boldsymbol{\pi^{(t-1)}P_2^2}$. Average prediction error rate: 31.76%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}P_1^2}+\lambda_2\boldsymbol{\pi^{(t-1)}P_2^2}+\lambda_3\boldsymbol{\pi^{(t-2)}P_3^2}$. Average P.E.R.: 34.66%.

| | $P_1^2$ transition matrix | | | | | $P_2^2$ transition matrix | | | | | $P_3^2$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .979 | .009 | .003 | .009 | .000 | .967 | .018 | .006 | .009 | .000 | .961 | .021 | .006 | .012 | .000 |
| H | .077 | .769 | .077 | .000 | .077 | .154 | .538 | .154 | .000 | .154 | .154 | .385 | .154 | .154 | .154 |
| F | .125 | .000. | .625 | .250 | .000 | .375 | .000 | .375 | .250 | .000 | .500 | .000 | .375 | .250 | .000 |
| B | .263 | .000 | .053 | .684 | .000 | .316 | .000 | .053 | .632 | .000 | .368 | .053 | .053 | .526 | .000 |
| D | .000 | .000 | .000 | .500 | .500 | .000 | .000 | .000 | .000 | 1.00 | .000 | .000 | .000 | .000 | 1.00 |

First order MC $\lambda_1 = 1$; Second-order MC $\lambda_1 = \lambda_2 = 0.5$; Third-order MC $\lambda_1 = \lambda_2 = \lambda_3 = 0.3333$.

Note: The monthly yield curve Markov chain (1953.4 to 2016.3) in estimation is a categorical five-state sequence classified by the effective algorithm. Prediction error is an indicator function equals one when the model predicted state is not the same as the observed state. Average prediction error rate is defined as the percentage of total prediction errors in the sample. Data are from Federal Reserve Board H.15 interest rate statistics.

Table E3: K=5 fold C.V. mean prediction error rate calculation

1st-order MC mean prediction error rate: 10.20%=(11.92+15.23+15.89+3.97+3.97)/5.

2nd-order MC mean prediction error rate: 12.72%=(19.87+15.89+18.54+4.64+4.64)/5.

3rd-order MC mean prediction error rate: 14.44%=(21.19+19.87+21.19+5.30+4.64)/5.

Test set 1 (validation): 1953.04–1965.10 (151 obs); Training set 1 (estimation): remaining 605 observation.

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^1}$. Average prediction error rate: 11.92%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^1}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^1}$. Average prediction error rate: 19.87%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^1}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^1}+\lambda_3\boldsymbol{\pi^{(t-2)}}\boldsymbol{P_3^1}$. Average P.E.R.: 21.19%.

First order MC $\lambda_1 = 1$; Second-order MC $\lambda_1 = 0, \lambda_2 = 1$; Third-order MC $\lambda_1 = \lambda_2 = 0, \lambda_3 = 1$

| | $P_1^1$ transition matrix | | | | | $P_2^1$ transition matrix | | | | | $P_3^1$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .963 | .023 | .004 | .009 | .000 | .940 | .035 | .007 | .014 | .005 | .923 | .044 | .007 | .019 | .007 |
| H | .151 | .679 | .019 | .000 | .151 | .226 | .585 | .038 | .000 | .151 | .264 | .453 | .038 | .057 | .189 |
| F | .182 | .000 | .455 | .273 | .091 | .364 | .000 | .273 | .091 | .273 | .455 | .091 | .273 | .091 | .091 |
| B | .154 | .000 | .026 | .769 | .051 | .205 | .000 | .026 | .667 | .103 | .256 | .026 | .026 | .590 | .103 |
| D | .014 | .103 | .015 | .029 | .838 | .059 | .103 | .015 | .059 | .765 | .103 | .118 | .015 | .059 | .706 |

Test set 2 (validation): 1965.11–1978.05 (151 obs); Training set 2 (estimation): remaining 605 observation.

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^2}$. Average prediction error rate: 15.23%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^2}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^2}$. Average prediction error rate: 15.89%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^2}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^2}+\lambda_3\boldsymbol{\pi^{(t-2)}}\boldsymbol{P_3^2}$. Average P.E.R.: 19.87%.

First order MC $\lambda_1 = 1$; Second-order MC $\lambda_1 = \lambda_2 = 0.5$; Third-order MC $\lambda_1 = \lambda_2 = \lambda_3 = 0.3333$

| | $P_1^2$ transition matrix | | | | | $P_2^2$ transition matrix | | | | | $P_3^2$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .962 | .023 | .008 | .006 | .000 | .944 | .031 | .013 | .008 | .004 | .937 | .034 | .010 | .013 | .006 |
| H | .167 | .667 | .042 | .021 | .104 | .229 | .521 | .083 | .021 | .146 | .229 | .479 | .083 | .042 | .167 |
| F | .214 | .071 | .500 | .214 | .000 | .429 | .143 | .214 | .214 | .000 | .429 | .143 | .286 | .071 | .071 |
| B | .273 | .000 | .045 | .636 | .045 | .318 | .000 | .045 | .545 | .091 | .364 | .045 | .045 | .455 | .091 |
| D | .024 | .098 | .000 | .024 | .854 | .073 | .146 | .000 | .049 | .732 | .122 | .146 | .000 | .073 | .659 |

Note: The monthly yield curve Markov chain (1953.4 to 2016.3) in estimation is a categorical five-state sequence classified by the effective algorithm. Prediction error is an indicator function equals one when the model predicted state is not the same as the observed state. Average prediction error rate is defined as the percentage of total prediction errors in the sample. Data are from Federal Reserve Board H.15 interest rate statistics.

Table 3 continued...

Test set 3 (validation): 1978.06–1990.12 (151 obs); training set 3 (estimation): remaining 605 observation.

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^3}$. Average prediction error rate: 15.89%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^3}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^3}$. Average prediction error rate: 18.54%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^3}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^3}+\lambda_3\boldsymbol{\pi^{(t-2)}}\boldsymbol{P_3^3}$. Average P.E.R.: 21.19%.

| | $P_1^3$ transition matrix | | | | | $P_2^3$ transition matrix | | | | | $P_3^3$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .967 | .018 | .007 | .009 | .000 | .947 | .031 | .011 | .011 | .000 | .934 | .040 | .013 | .013 | .000 |
| H | .098 | .738 | .016 | .016 | .131 | .180 | .590 | .033 | .016 | .180 | .213 | .492 | .033 | .033 | .230 |
| F | .333 | .111 | .444 | .000 | .111 | .444 | .222 | .111 | .000 | .222 | .444 | .333 | .000 | .000 | .222 |
| B | .167 | .000 | .000 | .806 | .028 | .222 | .000 | .000 | .722 | .056 | .013 | .000 | .000 | .639 | .056 |
| D | .000 | .156 | .022 | .044 | .778 | .022 | .200 | .022 | .089 | .667 | .044 | .222 | .022 | .111 | .600 |

Test set 4: 1991.01–2003.07 (151 obs); training set 4: remaining 605 observation.

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^4}$. Average prediction error rate: 3.97%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^4}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^4}$. Average prediction error rate: 4.64%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^4}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^4}+\lambda_3\boldsymbol{\pi^{(t-2)}}\boldsymbol{P_3^4}$. Average P.E.R.: 5.30%.

| | $P_1^4$ transition matrix | | | | | $P_2^4$ transition matrix | | | | | $P_3^4$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .951 | .032 | .010 | .007 | .000 | .921 | .047 | .015 | .012 | .005 | .904 | .054 | .017 | .017 | .007 |
| H | .153 | .681 | .028 | .014 | .125 | .236 | .556 | .056 | .014 | .139 | .264 | .472 | .056 | .028 | .181 |
| F | .188 | .063 | .500 | .186 | .063 | .313 | .125 | .250 | .188 | .125 | .375 | .188 | .188 | .063 | .188 |
| B | .143 | .000 | .029 | .771 | .057 | .171 | .000 | .029 | .686 | .114 | .200 | .029 | .029 | .629 | .114 |
| D | .014 | .122 | .014 | .014 | .838 | .054 | .149 | .014 | .027 | .757 | .095 | .162 | .014 | .041 | .689 |

Test set 5: 2003.08–2016.03 (151 obs); training set 5: remaining 605 observation.

1st-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^5}$. Average prediction error rate: 3.97%.

2nd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^5}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^5}$. Average prediction error rate: 4.64%.

3rd-order MC prediction equation: $\boldsymbol{\pi^{(t+1)}}=\lambda_1\boldsymbol{\pi^{(t)}}\boldsymbol{P_1^5}+\lambda_2\boldsymbol{\pi^{(t-1)}}\boldsymbol{P_2^5}+\lambda_3\boldsymbol{\pi^{(t-2)}}\boldsymbol{P_3^5}$. Average P.E.R.: 4.64%.

| | $P_1^5$ transition matrix | | | | | $P_2^5$ transition matrix | | | | | $P_3^5$ transition matrix | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | U | H | F | B | D | U | H | F | B | D | U | H | F | B | D |
| U | .953 | .034 | .007 | .005 | .000 | .924 | .052 | .010 | .010 | .005 | .906 | .062 | .012 | .012 | .007 |
| H | .141 | .692 | .025 | .013 | .128 | .218 | .564 | .051 | .013 | .154 | .244 | .474 | .051 | .038 | .192 |
| F | .143 | .071 | .500 | .214 | .071 | .214 | .143 | .286 | .214 | .143 | .286 | .214 | .214 | .071 | .214 |
| B | .179 | .000 | .036 | .714 | .071 | .250 | .000 | .036 | .571 | .143 | .286 | .035 | .035 | .500 | .143 |
| D | .013 | .118 | .013 | .026 | .829 | .053 | .145 | .013 | .053 | .737 | .092 | .158 | .013 | .066 | .671 |

First order MC $\lambda_1 = 1$; Second-order MC $\lambda_1 = \lambda_2 = 0.5$; Third-order MC $\lambda_1 = \lambda_2 = \lambda_3 = 0.3333$.

# F Full sample forecast: Daily and weekly yield sequence

Table F1: Full-sample daily forecast with deterministic initial distributions

| Forecast | U | H | F | B | D | U | H | F | B | D |
|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | $\boldsymbol{\pi^{(0)}} = (1,0,0,0,0)$ | | | | | $\boldsymbol{\pi^{(0)}} = (0,1,0,0,0)$ | | | | |
| 1-week | 0.9643 | 0.0136 | 0.0053 | 0.0153 | 0.0015 | 0.1093 | 0.7368 | 0.0333 | 0.0061 | 0.1144 |
| 2-week | 0.9352 | 0.0245 | 0.0080 | 0.0265 | 0.0058 | 0.1951 | 0.5617 | 0.0397 | 0.0193 | 0.1841 |
| 1-month | 0.8906 | 0.0402 | 0.0109 | 0.0405 | 0.0178 | 0.3185 | 0.3576 | 0.0364 | 0.0438 | 0.2438 |
| 2-month | 0.8324 | 0.0586 | 0.0138 | 0.0525 | 0.0428 | 0.4662 | 0.1993 | 0.0289 | 0.0675 | 0.2382 |
| 6-month | 0.7485 | 0.0832 | 0.0172 | 0.0614 | 0.0898 | 0.6772 | 0.1040 | 0.0201 | 0.0670 | 0.1317 |
| 1-year | 0.7302 | 0.0885 | 0.0180 | 0.0628 | 0.1005 | 0.7241 | 0.0903 | 0.0182 | 0.0633 | 0.1041 |
| 2-year | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| | $\boldsymbol{\pi^{(0)}} = (0,0,1,0,0)$ | | | | | $\boldsymbol{\pi^{(0)}} = (0,0,0,1,0)$ | | | | |
| 1-week | 0.2028 | 0.2096 | 0.3462 | 0.1219 | 0.1194 | 0.1820 | 0.0099 | 0.0384 | 0.7143 | 0.0554 |
| 2-week | 0.3123 | 0.2417 | 0.1356 | 0.1385 | 0.1719 | 0.3150 | 0.0298 | 0.0434 | 0.5200 | 0.0918 |
| 1-month | 0.4325 | 0.2056 | 0.0424 | 0.1152 | 0.2043 | 0.4817 | 0.0640 | 0.0353 | 0.2914 | 0.1276 |
| 2-month | 0.5512 | 0.1475 | 0.0258 | 0.0834 | 0.1921 | 0.6216 | 0.0934 | 0.0242 | 0.1228 | 0.1379 |
| 6-month | 0.6954 | 0.0987 | 0.0194 | 0.0657 | 0.1210 | 0.7137 | 0.0932 | 0.0186 | 0.0644 | 0.1101 |
| 1-year | 0.7256 | 0.0898 | 0.0182 | 0.0632 | 0.1032 | 0.7272 | 0.0893 | 0.0181 | 0.0631 | 0.1022 |
| 2-year | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| | $\boldsymbol{\pi^{(0)}} = (0,0,0,0,1)$ | | | | | $\boldsymbol{\pi^{(0)}} = \boldsymbol{\pi^{*}} = \boldsymbol{\pi}(\infty)$ | | | | |
| 1-week | 0.0114 | 0.0899 | 0.0251 | 0.0405 | 0.8331 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| 2-week | 0.0427 | 0.1469 | 0.0342 | 0.0664 | 0.7096 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| 1-month | 0.1306 | 0.1981 | 0.0380 | 0.0904 | 0.5429 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| 2-month | 0.3103 | 0.1973 | 0.0341 | 0.0938 | 0.3646 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| 6-month | 0.6453 | 0.1132 | 0.0214 | 0.0697 | 0.1505 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| 1-year | 0.7213 | 0.0911 | 0.0183 | 0.0636 | 0.1057 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |
| 2-year | 0.7284 | 0.0890 | 0.0180 | 0.0630 | 0.1015 | 0.7285 | 0.0890 | 0.0180 | 0.0630 | 0.1015 |

Note: Forecast equation: $\boldsymbol{\pi^{(t)}} = \boldsymbol{\pi^{(0)}} * \boldsymbol{P^{t}}$. First-order Markov chain model with regular MLE transition probabilities matrix. $\boldsymbol{\pi^{0}}$, $\boldsymbol{\pi^{*}}$, $\boldsymbol{\pi^{(\infty)}}$ stand for the initial, stationary, and limiting distributions, respectively. Time conversion: 1 week=5 business days, 1 month=20 business days, 2 months=40 business days, 6 months=120 business days, 1 year=240 business days, 2 years=480 business days.

Table F2: Full-sample weekly forecast with deterministic initial distributions

| Forecast | U | H | F | B | D | U | H | F | B | D |
|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | $\boldsymbol{\pi^{(0)}} = (1,0,0,0,0)$ | | | | | $\boldsymbol{\pi^{(0)}} = (0,1,0,0,0)$ | | | | |
| 1-month | 0.9470 | 0.0217 | 0.0063 | 0.0211 | 0.0039 | 0.1529 | 0.6098 | 0.0520 | 0.0125 | 0.1728 |
| 2-month | 0.9079 | 0.0360 | 0.0087 | 0.0340 | 0.0134 | 0.2596 | 0.4112 | 0.0488 | 0.0338 | 0.2465 |
| 3-month | 0.8711 | 0.0479 | 0.0106 | 0.0434 | 0.0269 | 0.3542 | 0.2834 | 0.0411 | 0.0535 | 0.2678 |
| 6-month | 0.8103 | 0.0658 | 0.0137 | 0.0536 | 0.0565 | 0.5086 | 0.1663 | 0.0301 | 0.0712 | 0.2240 |
| 1-year | 0.7569 | 0.0814 | 0.0165 | 0.0590 | 0.0861 | 0.6520 | 0.1127 | 0.0221 | 0.0674 | 0.1458 |
| 2-year | 0.7325 | 0.0886 | 0.0178 | 0.0611 | 0.1000 | 0.7200 | 0.0923 | 0.0185 | 0.0621 | 0.1071 |
| 5-year | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 | 0.7292 | 0.0896 | 0.0180 | 0.0613 | 0.1019 |
| | $\boldsymbol{\pi^{(0)}} = (0,0,1,0,0)$ | | | | | $\boldsymbol{\pi^{(0)}} = (0,0,0,1,0)$ | | | | |
| 1-month | 0.2547 | 0.2244 | 0.1575 | 0.1123 | 0.2512 | 0.2860 | 0.0132 | 0.0227 | 0.6052 | 0.0729 |
| 2-month | 0.3547 | 0.2144 | 0.0518 | 0.1078 | 0.2714 | 0.4538 | 0.0376 | 0.0230 | 0.3791 | 0.1065 |
| 3-month | 0.4339 | 0.1841 | 0.0346 | 0.0939 | 0.2534 | 0.5705 | 0.0620 | 0.0215 | 0.2249 | 0.1212 |
| 6-month | 0.5603 | 0.1402 | 0.0269 | 0.0765 | 0.1963 | 0.6797 | 0.0891 | 0.0197 | 0.0919 | 0.1197 |
| 1-year | 0.6713 | 0.1067 | 0.0210 | 0.0662 | 0.1348 | 0.7177 | 0.0924 | 0.0186 | 0.0631 | 0.1082 |
| 2-year | 0.7223 | 0.0916 | 0.0183 | 0.0619 | 0.1058 | 0.7279 | 0.0900 | 0.0181 | 0.0615 | 0.1026 |
| 5-year | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| | $\boldsymbol{\pi^{(0)}} = (0,0,0,0,1)$ | | | | | $\boldsymbol{\pi^{(0)}} = \boldsymbol{\pi^{*}} = \boldsymbol{\pi}(\infty)$ | | | | |
| 1-month | 0.0276 | 0.1403 | 0.0445 | 0.7320 | 0.0556 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| 2-month | 0.0950 | 0.1995 | 0.0483 | 0.0817 | 0.5754 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| 3-month | 0.1913 | 0.2169 | 0.0453 | 0.0924 | 0.4542 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| 6-month | 0.4026 | 0.1832 | 0.0351 | 0.0869 | 0.2922 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| 1-year | 0.6157 | 0.1232 | 0.0239 | 0.0707 | 0.1664 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| 2-year | 0.7157 | 0.0936 | 0.0187 | 0.0625 | 0.1096 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |
| 5-year | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 | 0.7292 | 0.0896 | 0.0180 | 0.0614 | 0.1019 |

Note: Forecast equation: $\boldsymbol{\pi^{(t)}} = \boldsymbol{\pi^{(0)}} * \boldsymbol{P^{t}}$. First-order Markov chain model with regular MLE transition probabilities matrix. $\boldsymbol{\pi^{0}}$, $\boldsymbol{\pi^{*}}$, $\boldsymbol{\pi^{(\infty)}}$ stand for the initial, stationary, and limiting distributions, respectively. Time conversion: 1 month=4 weeks, 2 months=8 weeks, 3 months=13 weeks, 6 months=26 weeks, 1 year=52 weeks, 2 years=104 weeks, 5 years=260 weeks.

# References


Athreya, K. B., and Fuh, C. D. (1992), *Bootstrap Markov chains. Exploring the limits of bootstrap,* (ed. R. Lepage and L. Billard), 49-64, New York: Wiley.

Ching, W. K., Huang, X., Ng, M. K., and Siu, T. K. (2013), *"Markov chains–models, algorithms and applications. International series in operations research & management science 189"*, Springer Science.

Efron, B. (1979), "Bootstrap methods: Another look at the Jackknife," *The Annals of Statistics*, 7(1), 1-26.

Evans, M. J., and Rosenthal, J. S. (2010), *Probability and statistics: The science of uncertainty.* 2nd Edition. W. H. Freeman and Company.

Fabozzi, F. J. (2016). *Bond markets, analysis, and strategies.* 9th Edition. Pearson.

Guerra, R., Polansky, A. M., and Schucany, W. R. (1997), "Smoothed bootstrap confidence intervals with discrete data,". *Computational Statistics and Data Analysis*, 26, 163-176.

Gürkaynak, R. S., Sack, B., and Wright, J. H. (2007), "The U.S. Treasury yield curve: 1961 to the present," *Journal of Monetary Economics*, 54, 2291-2304.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning with applications in R.* Springer.

Kulperger, R. J., Prakasa Rao, and B. L. S. (1990), "Bootstrapping a finite state Markov chain," *Sankhya Series A*, 51, 178-191.

Mishkin, F. 2015. *The economics of money, banking, and financial markets.* 11th edition. Pearson.

Lee, T. C., Judge, G. G., and Zellner, A. (1968), "Maximum likelihood and Bayesian estimation of transition probabilities". *Journal of the American Statistical Association*, 63(324), 1162-1179.

Raftery, A. E. (1985), "A model for higher-order Markov chains," *Journal of the Royal Statistical Society Series B*, 47(3), 528-539.

Ross, S. M. (2014), *Introduction to probability models.* 11th Edition. Elsevier.

Rosenthal, J. S. (2006), *A first look at rigorous probability.* 2nd Edition. Singapore: World Scientific Publishing.

Spedicato, G. A. (2015), "Markovchain: An R package to easily handle discrete Markov chains."

Teodorescu, J. (2009), "Maximum likelihood estimation for Markov chains," *Research gate paper.*

Vulkarni, V. G. (2011), *Introduction to modeling and analysis of stochastic systems.* 2nd Edition. New York: Springer.