Time Series Analysis
时间序列分析

Yongfeng Zhang, Tsinghua University
zhangyf07@gmail.com

Outline

- 什么是时间序列分析(Time Series Analysis)
- 常见模型和基本手段
  - 趋势(Trend Component)
  - 周期性(Seasonal Component)
  - 随机性(Random Component)
- 简单示例
  - Modeling a Time Series
- 常用模型 – ARMA
  - AR (Auto Regressive)
  - MA (Moving Average)
  - ARIMA (Auto Regressive Integrated Moving Average)
- 应用示例
  - Google Trends

Outline

- 什么是时间序列分析(Time Series Analysis)

- 常见模型和基本手段
  - 趋势(Trend Component)
  - 周期性(Seasonal Component)
  - 随机性(Random Component)

- 简单示例

- 常用模型 – ARMA
  - AR (Auto Regressive)
  - MA (Moving Average)
  - ARIMA (Auto Regressive Integrated Moving Average)

- 应用示例
  - Google Trends

Time Series

- What is Time Series

A time-series plot is a two-dimensional plot of time series data

- the vertical axis measures the variable of interest
- the horizontal axis corresponds to the time periods

Time Series (cont.)

- Other Examples
  - Governments forecast **unemployment rates, income taxes** for policy purposes.
  - Marketing executives forecast **demand, sales, and consumer preferences** for strategic planning
  - etc…

Types of Time Series

- Types of time series
  1. **continuous**
  2. **discrete**
- **Discrete** means that observations are recorded in discrete times - it says nothing about the nature of the observed variable
- The time intervals can be annually, quarterly, monthly, weekly, daily, hourly, etc.
- **Continuous** means that observations are recorded continuously -e.g. temperature and/or humidity in some laboratory
- Again, time series can be continuous *regardless* of the nature of the observed variable

Types of Time Series (cont.)

- Discrete time series can result when continuous time series are **sampled**.
- Sometimes quantities that don't have an instantaneous value get **aggregated** also resulting in a discrete time series e.g. daily rainfall
- We will mostly study discrete time series in this course. Note that discrete time series are often the result of discretization of continuous time series (e.g. monthly rainfall)

Example

Year	2000	2001	2002	2003	2004
Sales	75.3	74.2	78.5	79.7	80.2

Time Series Analysis – Objectives

- Objectives of time series analysis:
  - **description** - summary statistics, graphs
  - **analysis and interpretation** - find a model to describe the time dependence in the data, can we interpret the model?
  - **forecasting or prediction** - given a sample from the series, forecast the next value, or the next few values
  - **control** - adjust various control parameters to make the series fit closer to a target

Outline

- 什么是时间序列分析(Time Series Analysis)
- 常见模型和基本手段
  - 数值变换(Transformations)
  - 趋势(Trend Component)
  - 季节性(Seasonal Component)
  - 周期性(Cyclical Component)
  - 随机性(Random Component)
- 简单示例
- 常用模型 – ARMA
  - AR (Auto Regressive)
  - MA (Moving Average)
  - ARIMA (Auto Regressive Integrated Moving Average)
- 应用示例
  - Google Trends

Transformation

1. To stabilize the variance. For example, if the standard deviation is proportional to the mean, log transform can be used
2. To make the seasonal effect additive.
   - multiplicative vs additive noise- if the noise is also multiplicative, the transformation will also help stabilize the variance
3. To make the data normally distributed - useful for a variety of reasons to be discussed later

Example

logarithmic, square root, reciprocal square root, Box-Cox as a general approach

4. In general, transforming the data is usually not a great idea except where doing so makes physical sense. Example: percentage data transformed using a log transform

Major Types of Variation

Time Series

Trend Component	Seasonal Component	Cyclical Component	Irregular Component
Overall, persistent, long-term movement	Regular periodic fluctuations, usually within a 12-month period	Repeating swings or movements over more than one year	Erratic or residual fluctuations

Major Types of Variation (cont.)

Types of Variation

1. Seasonal variation: sales figures and temperature readings exhibit variation that is annual in period.

Example

Unemployment is typically "high" in winter and "lower" in summer.

2. Cyclic variation:
   1. variation at other fixed periods.

Example

Daily variation in temperature "high" at noon, "low" at night.

   2. Some time series exhibit oscillations without a fixed period, they are predictable to some extent.

Example

Economic data are affected by business cycles.

Major Types of Variation (cont.)

3. Trend: long-term change in the mean level "long term" relative to the number of observations.

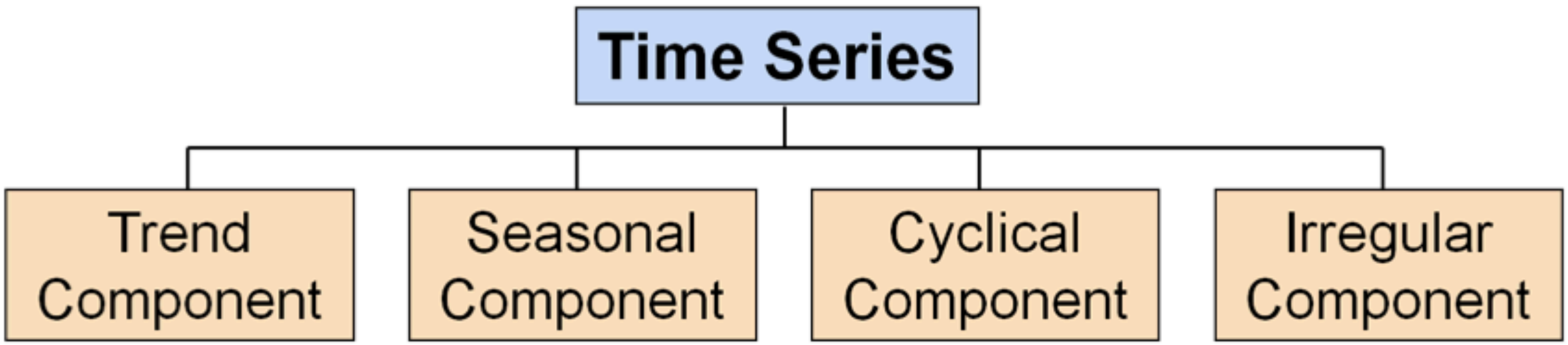
Example

Climate variables exhibit cyclic variation over long periods.

4. Irregular fluctuations: after trend and cyclic variations have been removed, a series of residuals may or may not be "random".
   1. any cyclic variation is still left.
   2. Probability models such as moving average (MA) or autoregressive (AR).

Stationary Time Series: If there is no systematic change in mean (no trend), variance and if periodic variations have been removed.

Major Types of Variation (cont.)



Overall, persistent, long-term movement

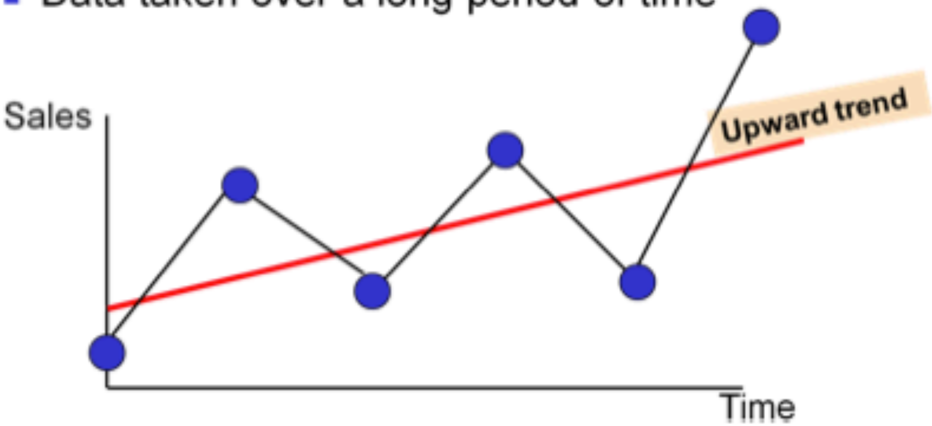
Regular periodic fluctuations, usually within a 12-month period

Repeating swings or movements over more than one year

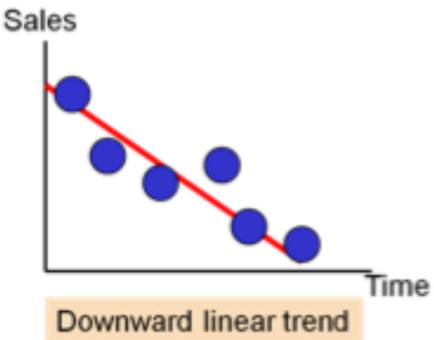
Erratic or residual fluctuations

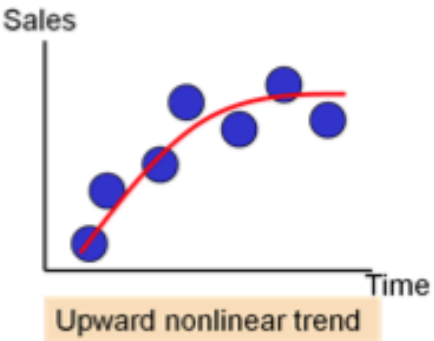
Trend Component

- Long-run increase or decrease over time
  (overall upward or downward movement)
- Data taken over a long period of time



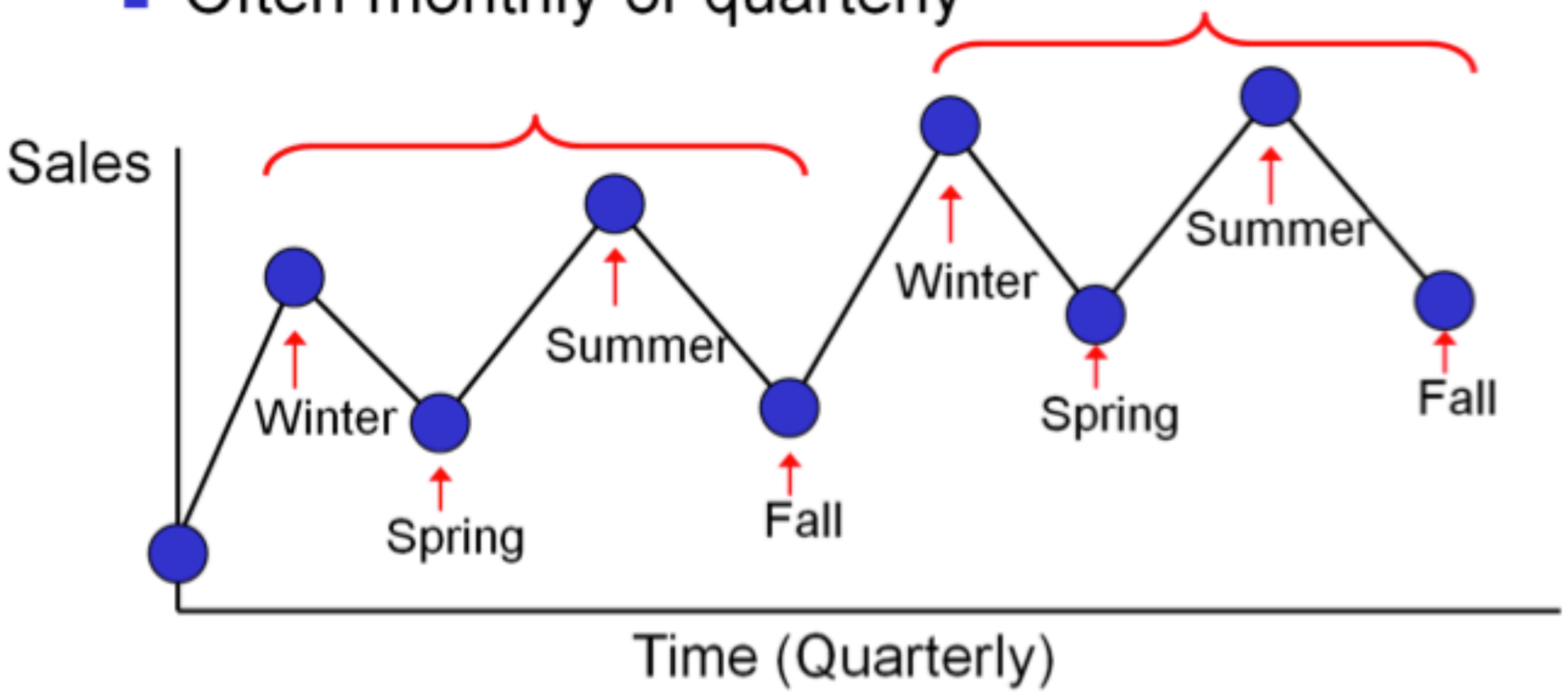
- Trend can be upward or downward
- Trend can be linear or non-linear





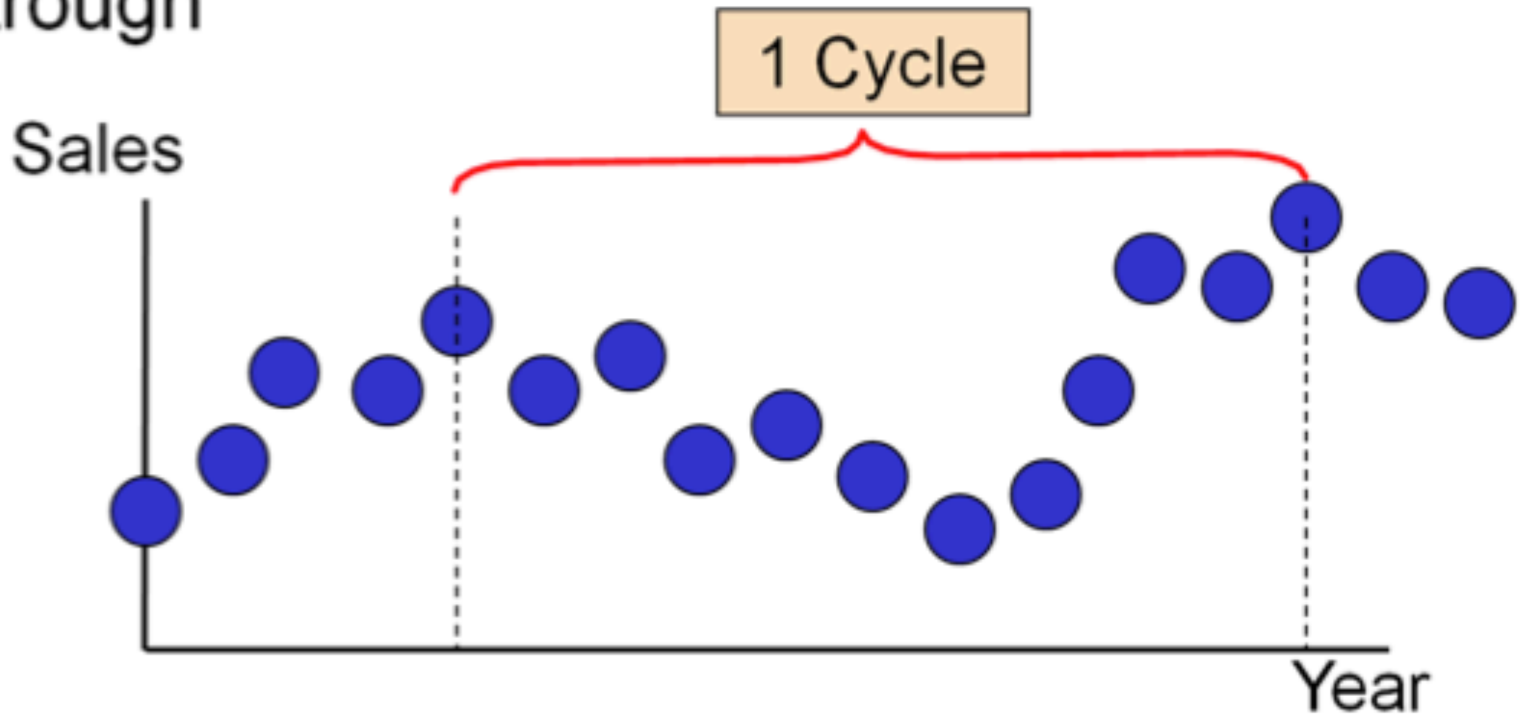
Seasonal Component

- Short-term regular wave-like patterns
- Observed within 1 year
- Often monthly or quarterly



Cyclical Component

- Long-term wave-like patterns
- Regularly occur but may vary in length
- Often measured peak to peak or trough to trough



Irregular Component

- Unpredictable, random, “residual” fluctuations.
- “Noise” in the time series.
- Stochastic factors.

Does The Time-Series have a Trend Component?

- A time-series plot should help answer this question.
- Often it helps if you “smooth” the time series data to help answer this question.
- Two popular smoothing methods are **moving averages** and **exponential smoothing**.

Outline

- 什么是时间序列分析(Time Series Analysis)
- 常见模型和基本手段
  - 数值变换(Transformations)
  - 趋势(Trend Component)
  - 季节性(Seasonal Component)
  - 周期性(Cyclical Component)
  - 随机性(Random Component)
- 简单示例
  - Modeling a Time Series
- 常用模型 – ARMA
  - AR (Auto Regressive)
  - MA (Moving Average)
  - ARIMA (Auto Regressive Integrated Moving Average)
- 应用示例
  - Google Trends

Modeling a Time Series

The simplest model is given by

$$X_t = \alpha + \beta t + \epsilon_t,$$

where $\epsilon_t \sim N\left(0, \sigma_{\epsilon_t}^2\right)$.

$$\text{model} = \text{linear trend} + \text{noise}.$$

The mean level at time t is given by $\mu_t = E(X_t) = \alpha + \beta t$.

Types of Trend

- **Global** trend
  1. **Polynomial**: linear trend – quadratic trend
  2. **Exponential**
  3. **Logistic**

How to describe Trend (Moving Average)

- 1. Curve Fitting
  - Assume a curve function and conduct regression over observed time series.

Approaches to Describe Trend

1 Curve fitting $\rightarrow$ Regression.

Example

Polynomial curve $X_t = \alpha + \beta t \quad \rightarrow \quad X_t = 0.4 + 2t.$

Example

Gompertz curve $\log X_t = a + b \cdot r^t \quad \rightarrow \quad \log X_t = 3 + 2 \cdot 0.5^t.$

Example

Logistic curve $X_t = \frac{a}{1+be^{-ct}} \quad \rightarrow \quad X_t = \frac{0.7}{1+0.3e^{-2t}}.$

How to describe Trend (Moving Average)

- 2. Filtering (Moving Average)
  - measure trend and remove seasonal variation

Linear Filter

$$Y_t = \sum_{r=-q}^{s} a_r \cdot X_{t+r}.$$

Y_t is the linear operator, a_r is the set of weights.

If $\sum a_r = 1 \rightarrow$ smooth out local fluctuations $\rightarrow$ moving average.
MA is often symmetric $s = q$ and $a_j = a_{-j}$.

Example

$a_r = \frac{1}{2q+1}$ for $r = -q, \cdots, +q$. The smoothed value of X_t is given by

$$Y_t = Sm(X_t) = \frac{1}{2q+1} \sum_{r=-q}^{q} X_{t+r}.$$

How to describe Trend (Moving Average)

Used for smoothing a series of arithmetic means over time.

Result dependent upon choice of $L = 2q + 1$ (length of period for computing means).

Example

$$Y_t = Sm(X_t) = \frac{1}{5}(X_{t-2} + X_{t-1} + X_t + X_{t+1} + X_{t+2}).$$

First average:

$$Y_3 = MA(5) = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

Second average:

$$Y_4 = MA(5) = \frac{X_2 + X_3 + X_4 + X_5 + X_6}{5}$$

Outline

- 什么是时间序列分析(Time Series Analysis)

- 常见模型和基本手段
  - 数值变换(Transformations)
  - 趋势(Trend Component)
  - 季节性(Seasonal Component)
  - 周期性(Cyclical Component)
  - 随机性(Random Component)

- 简单示例
  - Modeling a Time Series

- 常用模型 – AR, MA, ARMA, ARIMA
  - AR (Auto Regressive)
  - MA (Moving Average)
  - ARMA (Auto Regressive Moving Average)
  - ARIMA (Auto Regressive Integrated Moving Average)

- 应用示例
  - Google Trends

Basic Definitions

- Stationary

Strictly Stationary

The overall behavior of random process X_t is described by a point distribution function of the process $\{X_{t_1}, X_{t_2}, \cdots, X_{t_k}\}$ at finite number of points $t_1, t_2, \cdots, t_k$ for any positive integer k

This function is

$$F_{t_1,t_2,\cdots,t_k}(X_1, X_2, \cdots, X_k) = P(X_{t_1} < X_1, \cdots, X_{t_k} < X_k).$$

Definition

A time series X_t is **strictly stationary** if $\{X_{t_1}, X_{t_2}, \cdots, X_{t_k}\}$ and $\{X_{t_1+\tau}, X_{t_2+\tau}, \cdots, X_{t_k+\tau}\}$ have the same point distribution for any positive integer $n \geq 1$ and any integer τ $(t_1, t_2, \cdots, t_n, \tau)$, i.e. the joint distribution function is invariant under time shifts.

Basic Definitions (cont.)

- Autocovariance (自协方差)
- Autocorrelation(自相关系数)

- The **autocovariance function (acv.f.)** γ_{t_1,t_2} or $\gamma(t_1, t_2)$ of X_{t_1} with X_{t_2} is defined by

$$\gamma_{t_1,t_2} = E\left[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})\right]$$

$$= \int\int (X_1 - \mu_{t_1})(X_2 - \mu_{t_2}) \cdot f_{t_1,t_2}(X_1, X_2)\, dX_1 dX_2.$$

- When $t = t_1 = t_2$ we get $Var(X_t) = \sigma_t^2$.
- The **autocorrelation function (ac.f.)** ρ_τ is defined by

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}.$$

Basic Definitions (cont.)

For the stationary stochastic process $X(t)$ or X_t we have

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}$$

1. $\rho_0 = 1.$
2. Covariance is symmetric, $\rho_\tau = \rho_{-\tau}$.

$$\gamma_\tau = cov\left(X_t, X_{t+\tau}\right) = \gamma_{-\tau}.$$

   Since X_t is stationary.
3. $|\rho_\tau| \leq 1.$
4. A stochastic process $\Rightarrow$ unique ac.f. The converse is not necessarily true ($\not\Leftarrow$).

Moving Average (MA) Process

- MA(q)

3 Moving average processes MA(q): $\{Z_t\} \sim IID(0, \sigma^2)$.

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \cdots + \beta_q Z_{t-q}.$$

We may rescale Z_t so that $\beta_0 = 1$.

Mean and Variance

$$E(X_t) = 0, \qquad Var(X_t) = \sigma_Z^2 \sum_{i=0}^{q} \beta_i^2.$$

Moving Average (MA) Process (cont.)

- Feature of MA

For $X_t = Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \cdots + \beta_q Z_{t-q}$, show that:

$$\gamma_k = \begin{cases} 0, & k > q. \\ \sigma^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k}, & k = 0, 1, \cdots, q. \\ \gamma_{-k}, & k < 0. \end{cases}$$

$$\rho_k = \begin{cases} 0, & k > q. \\ 1, & k = 0. \\ \frac{\sum_{i=0}^{q-k} \beta_i \beta_{i+k}}{\sum_{i=0}^{q} \beta_i^2}, & k = 1, 2, \cdots, q. \\ \rho_{-k} & k < 0. \end{cases}$$

- The ac.f. cuts off at lag q, a feature/benchmark of MA(q) process.

Auto Regressive (AR) Process

- AR(p)

4 Autoregressive Process AR(p).

Let $\{Z_t\} \overset{iid}{\sim} (0, \sigma_Z^2)$ be the white noise. The **autoregressive process** with parameter p is given by

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t.$$

Auto Regressive (AR) Process (cont.)

- Stationary of AR(p) process
  - The backward shift operator B is defined by $BX_t = X_{t-1}$.
  - $$B^2 X_t = B(BX_t) = BX_{t-1} = X_{t-2}.$$
  - In general,
$$B^j X_t = X_{t-j}, \quad \forall\, j.$$

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t.$$

$$X_t = (\alpha_1 B + \cdots + \alpha_p B^p) X_t + Z_t.$$

Let $\phi(B) = 1 - \alpha_1 B - \cdots - \alpha_p B^p$. Then,

$$\phi(B) \cdot X_t = Z_t.$$

AR(p) process is stationary if the roots of
$\phi(B) = 1 - \alpha_1 B - \cdots - \alpha_p B^p = 0$ are all outside the unit circle.

Auto Regressive Moving Average (ARMA) Process

- The ARMA(p,q) Process

A mixed autoregressive moving average process containing p AR terms and q MA terms is said to be an ARMA process of order (p, q), i.e. ARMA(p, q), and is given by

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}.$$

The ARMA(p, q) can be expressed in terms of the back-shift operator:

$$\phi(B) X_t = \theta(B) Z_t,$$

where $\phi(B) = 1 - \alpha_1 B - \cdots - \alpha_p B^p$, and $\theta(B) = 1 + \beta_1 B + \cdots + \beta_q B^q$.

Auto Regressive Integrated Moving Average (ARIMA) Process

- The ARIMA(p,d,q) Process

X_t is called an autoregressive integrated moving average (ARIMA) process of order (p, d, q) denoted $\{X_t\} \sim ARIMA(p, d, q)$, where $d \geq 1$ is an integer if its d-th difference $W_t = \nabla^d X_t = (1 - B)^d X_t$ is an ARMA(p, q) process, i.e.

$$W_t = \alpha_1 W_{t-1} + \cdots + \alpha_p W_{t-p} + Z_t + \cdots + \beta_q Z_{t-q},$$

or

$$\phi(B) \cdot W_t = \theta(B) \cdot Z_t.$$

$$\phi_p(B) \cdot (1 - B)^d \cdot X_t = \theta_q(B) \cdot Z_t.$$

Auto Regressive Integrated Moving Average (ARIMA) Process

- Why such a definition?

Most data in reality is non-stationary. If the time series is non-stationary in the mean, we can difference "differentiate" the series

$$\nabla X_t = (1 - B)X_t.$$

$$\nabla^2 X_t = (1 - B)^2 X_t.$$

$$\vdots$$

$$\nabla^d X_t = (1 - B)^d X_t.$$

Example of ARIMA(p,d,q)

Let $W_t = (1-B)^d X_t$. If $W_t \sim ARMA(p,q)$ then $X_t \sim ARIMA(p,d,q)$.

$$\phi(B) \cdot (1-B)^d \cdot X_t = \theta(B) \cdot Z_t.$$

Note that $B = 1$ is one of the roots with multiplicity d.

Example

Random Walk: $X_t = X_{t-1} + Z_t \sim IID(\mu, \sigma^2)$. This is an ARIMA(0,1,0).

$$E(X_t) = t\mu \implies \text{not stationary.}$$

$$X_t = BX_t + Z_t.$$

$$(1-B)^{d=1} \cdot X_t = Z_t.$$

$\phi(B) = 1 = B^0 \to p = 0,\qquad \theta(B) = 1 = B^0 \to q = 0.$

Outline

- 什么是时间序列分析(Time Series Analysis)

- 常见模型和基本手段
  - 数值变换(Transformations)
  - 趋势(Trend Component)
  - 季节性(Seasonal Component)
  - 周期性(Cyclical Component)
  - 随机性(Random Component)

- 简单示例
  - Modeling a Time Series

- 常用模型 – AR, MA, ARMA, ARIMA
  - AR (Auto Regressive)
  - MA (Moving Average)
  - ARMA (Auto Regressive Moving Average)
  - ARIMA (Auto Regressive Integrated Moving Average)

- 应用示例
  - Google Trends

Problem Statement

- Government agencies and other organizations produce monthly reports on economic activity
  - House Sales, Automotive Sales, Unemployment
- Problems with reports
  - Compilation delay of several weeks
- Google Trends releases daily and weekly index of search queries by industry vertical
- Can Google Trends data help predict *current* economic activity?

HongKong Visitor Arrival Statistics

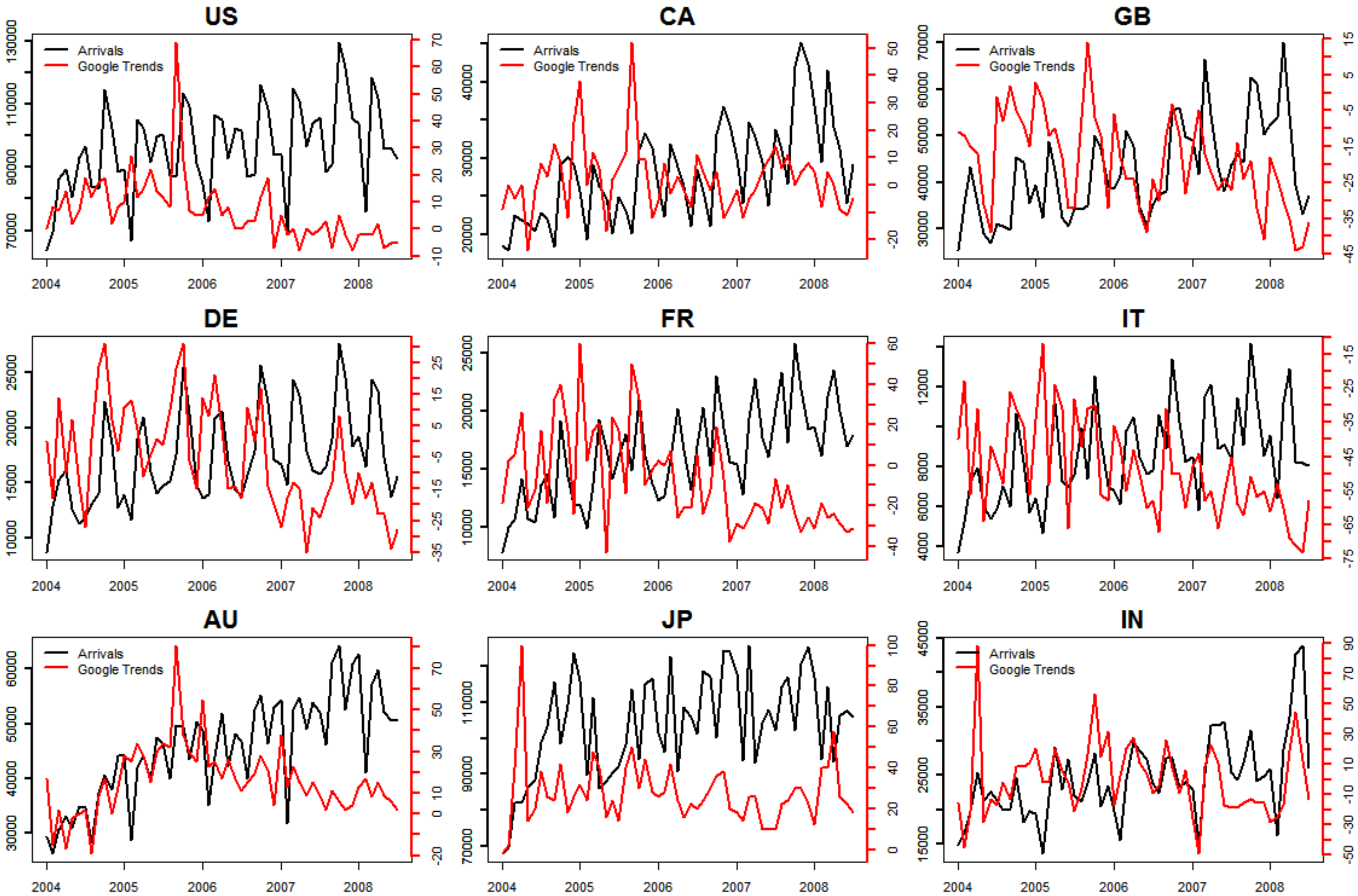
Visitors Arrival Statistics from Hong Kong Tourism Board

- Monthly summaries release with 1 month lag
- Reports Country/Territory of Residence of visitors
- Data available 2004-2008

Google Trends Travel by Category

- Hotels & Accommodations
- Air Travel
- Car Rental & Taxi Services
- Cruises & Charters
- Attractions & Activities
- Vacation Destinations
  - Australia
  - Caribbean Islands
  - Hawaii
  - **Hong Kong**
  - Las Vegas
  - Mexico
  - New York City
  - Orlando
- Adventure Travel
- Bus & Rail

Visitors Arrival Statistics vs. Google Trends



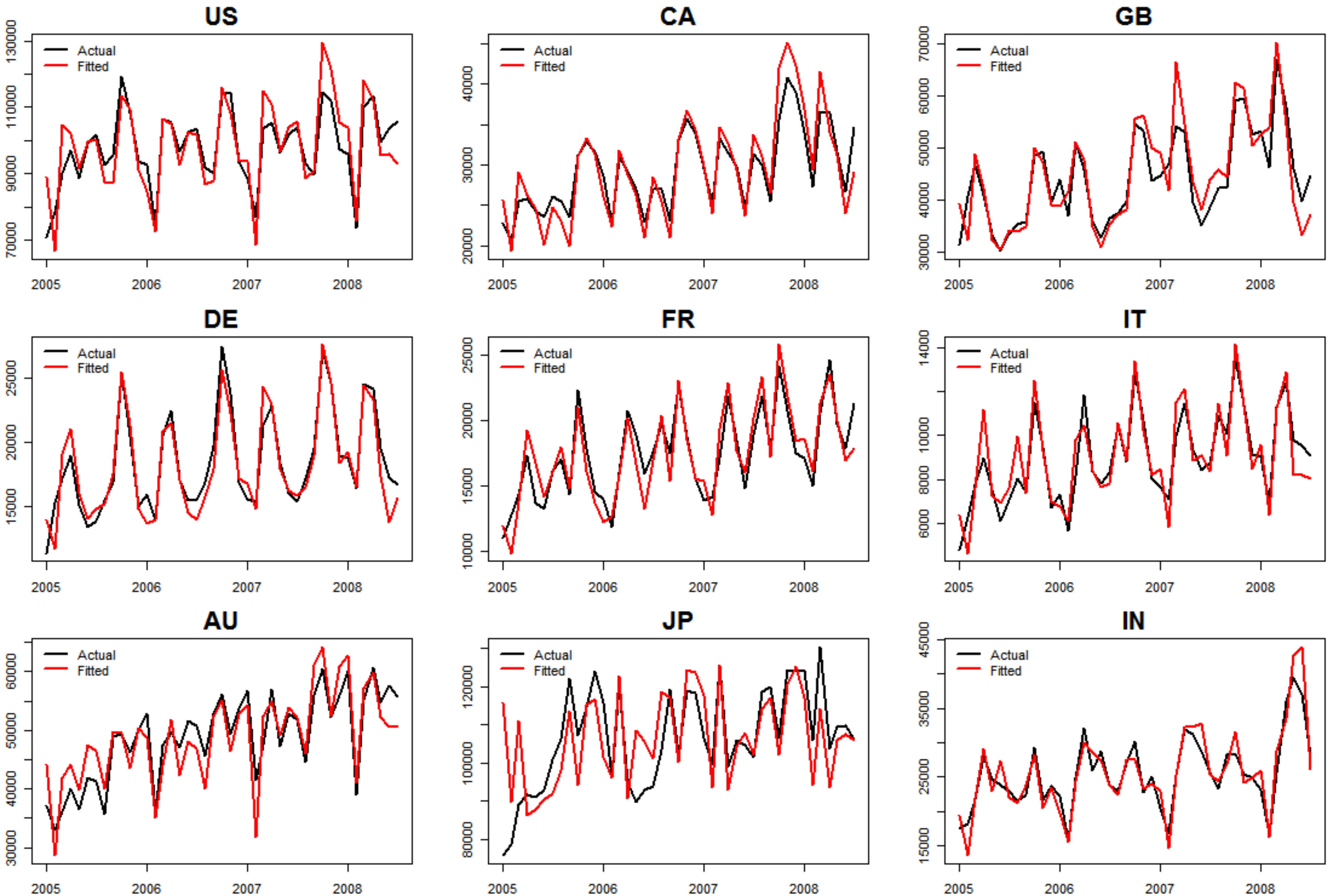
Analysis and Forecasting

- **Model: ARMA(12,0,1)**
- $\log(Y_{i,t}) = 0.664 + 0.113 * \log(Y_{i,t-1}) + 0.828 * \log(Y_{i,t-12}) + 0.001 * X_{i,t,2} + 0.001 * X_{i,t,3} + 0.005 * FXrate_{i,t} + \eta_i, + e_{i,t}$

  $e_{i,t} \sim N(0, 0.0938^2), \eta_i \sim N(0, 0.0228^2)$

  - $Y_{i,t}$ = Arrival to Hong Kong at month t and from i-th country
  - $X_{i,t,1}$ = Google Trend Search at 1st week of month t and from i-th country
  - $X_{i,t,2}$ = Google Trend Search at 2nd week of month t and from i-th country
  - $X_{i,t,3}$ = Google Trend Search at 3rd week of month t and from i-th country
  - $FXrate_{i,t}$ = Hong Kong Dollar per one unit of i-th country's local currency at month t. Average of first week's FX rate is used as a proxy to FX rate per each month.

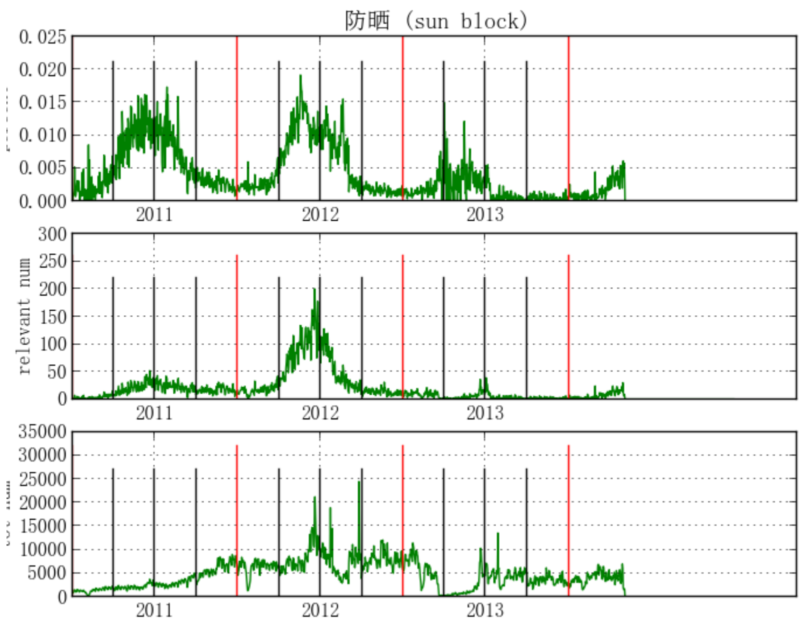
Visitor Arrival Statistics - Actual & Fitted



Share some of Our Finding

京东商城化妆品领域用户评论属性词频率的时间统计

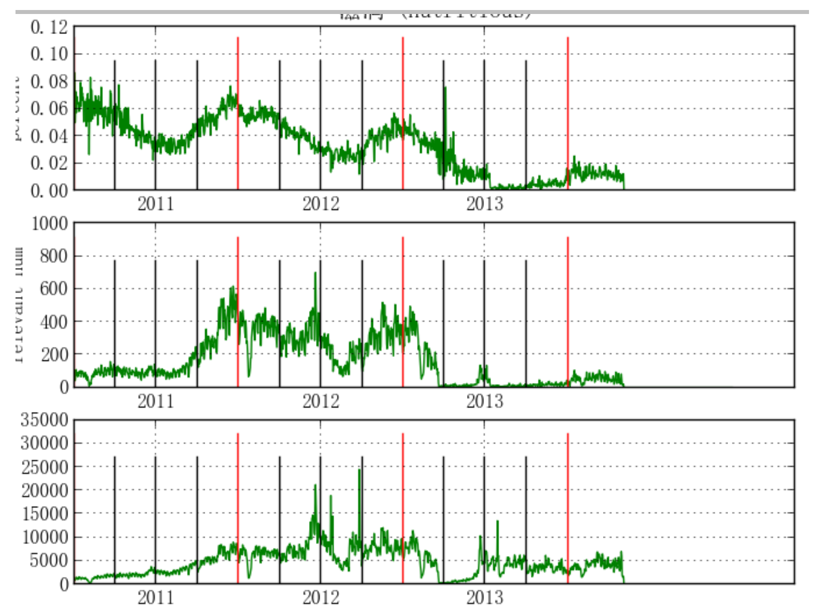
“防晒”



2011年2季度末和3季度(6~9月)

2012年2季度末和3季度(5~9月)

“营养”



2011年3季度末、4季度(9~12月)

Time-Dependent Recommendation

- 用户对物品属性的关心可能具有时间性
  - 周期性(Cyclic)
  - 季节性(Seasonal)
- 目前的推荐策略
  - 根据用户的全部历史评论和评分构建用户模型
  - 过于依赖用户的全部或近期行为
  - 较少考虑特定产品领域内在的规律性(时间和周期性)
- 考虑时间信息的个性化推荐
  - 将领域内的规律性与个性化相结合
  - 因人而异、因时而异

Thanks!