南京大学
NANJING UNIVERSITY

Improving prediction of exchange rates using Differential EMD

Phase II

141292018 桑梓洲

141110018 赵诚宇

Contents

南京大学 NANJING UNIVERSITY

1. 文献回顾及实验思路
2. 核心方法回顾：EMD 与 SVR
3. 实验具体操作
4. 实验结果及反思

文献回顾

南京大学 NANJING UNIVERSITY

- 文章特点：侧重信号处理的**技术应用**而非金融
- 文章概述
  - 作者选取了21个相关金融序列作为自变量试图对**欧元-美元汇率**序列进行预测，其中dEMD去噪声+SVR预测效果(67.81%)优于单纯的SVR预测(53.07%)、及当前较先进的MS回归(50.48%)以及MS-GARCH(47.69%)
- 数据选取
  - 98-10年：汇率、股指、利率、期货四大类21个相关实例
  - 有些并不直观，如澳洲国债、铂铜锌锡镍咖啡可可期货
- 衡量指标：预测序列与真实序列**上下变动的一致性**

实验思路

南京大学 NANJING UNIVERSITY

- 实验目的
  - 了解EMD、SVR并掌握其初级应用
  - 考虑到文章特点，我们更**侧重核心方法**相对深入的研究而非简单的沿作者思路全盘照做一遍
  - 我们希望通过我们的报告能让大家入门文中涉及到的技术并在日后需要时派上用场
- 数据
  - 时间：2009.6.1-2017.11.30，共2076组观测值(Wind)
  - 变量选取：**人民币-美元汇率**为因变量，人民币-欧元汇率，中国10年期国债利率，美国10年期国债利率，标普500指数，沪深300指数，恒生指数，沪金期货为自变量

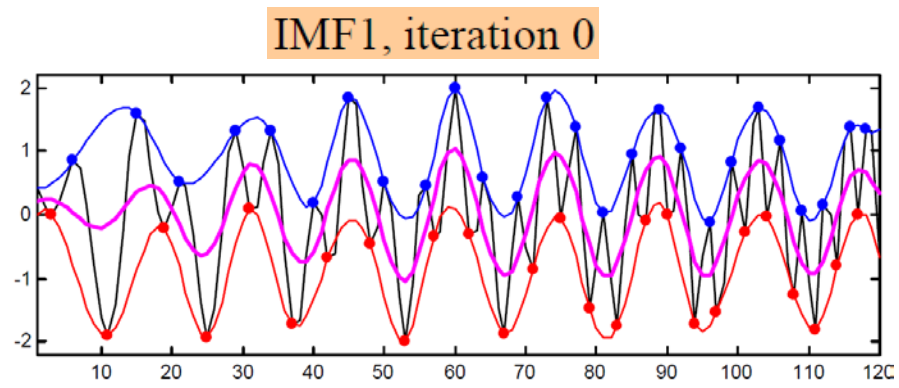
实验前我们的一些疑问

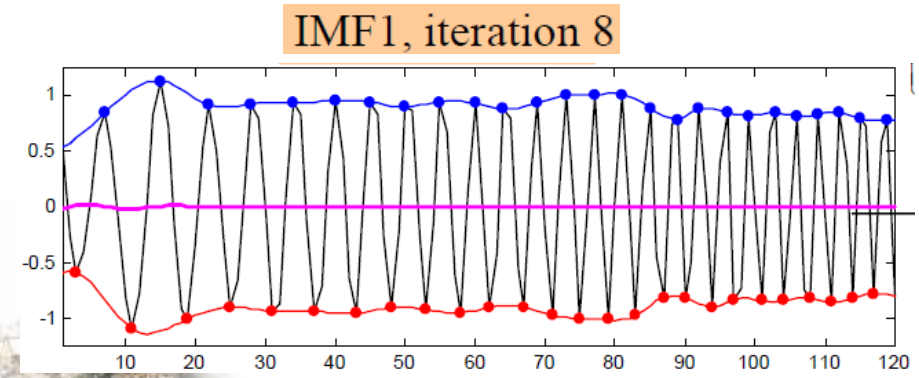
南京大学 NANJING UNIVERSITY

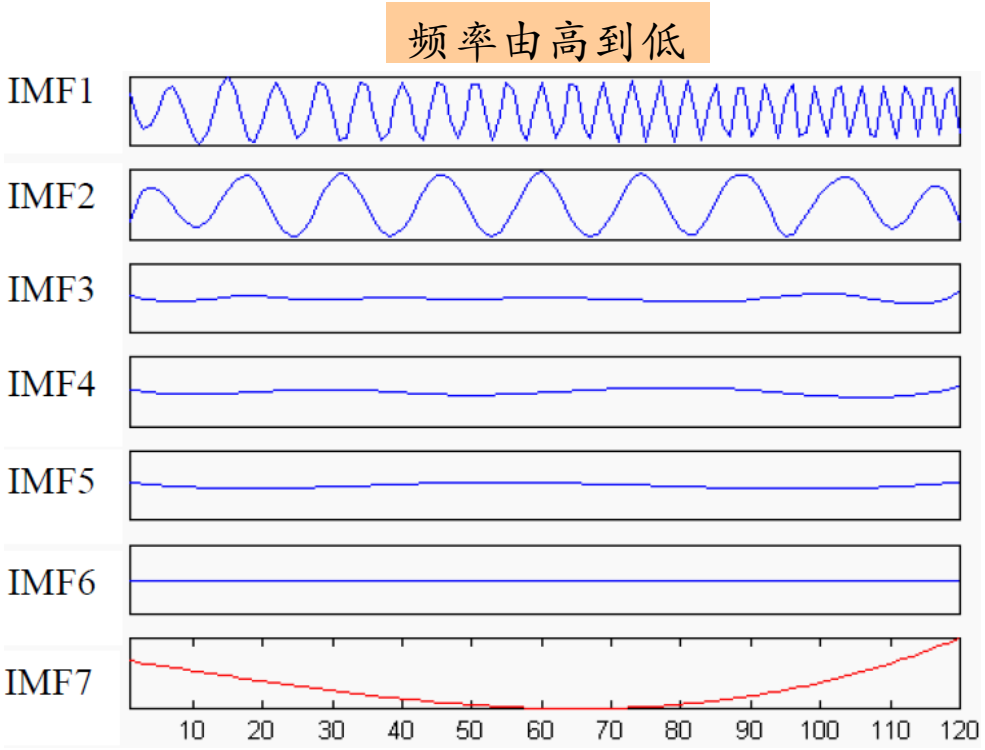
- 数据处理上，全文没有一处提到”归一化”; 作者直接采用**原始汇率**而未采用收益率序列**作为 y**
- 降噪算法上，作者在EMD算法基础上自创了dEMD(微分-EMD)，从**收益率序列**提取出了最近似服从正态分布的一个IMF(可看作趋势项)作为白噪声却用**原始序列与之相减**，其中含义让人费解
- 核心算法上，作者只是**简单运用**了SVR而未对参数调优做足够工作和说明，图表信息量不高
- 衡量标准上，单纯的变动方向预测全蒙向上也有近50%的准确率，预测**真实变动**上可能并不理想

EMD算法回顾

- HHT核心部分，原文被引15988，应用于各领域
- 核心思想：将一个不规则的信号通过不断筛选提取，分解为多个相对规则的信号







SVM摘要

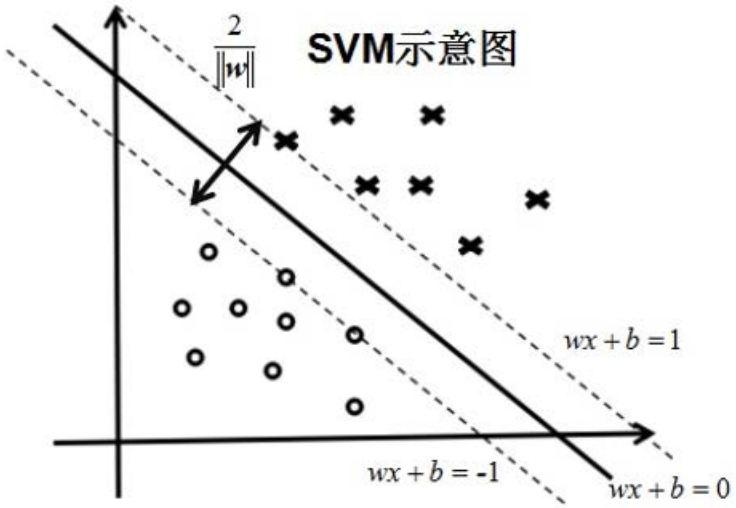
南京大学 NANJING UNIVERSITY

- SVM两大类：分类机(SVC) | 回归机(SVR)
- 核心思想
  - SVC: 找到一组超平面，使两组样本间隔最大(分得最开)
  - SVR: 使所有样本落在平面$\pm\varepsilon$的弹性管道内(非线性回归)
  - 线性不可分$\mapsto$高维线性可分空间，核函数简化高维计算
  - 松弛变量和惩罚函数: 允许适当误差，防止过度拟合
  - 优化求解: 拉格朗日对偶问题与KKT条件(凸优化理论)
  - 优点: 效果最好*; 全局最优; 适合小样本、高维数据
  - 几个层次: 入门级-能用软件; 专家级-根据问题构造核函数

*believed by many to be the best "off-the-shelf" supervised learning algorithm, Eric Xing, CMU

SVM实例

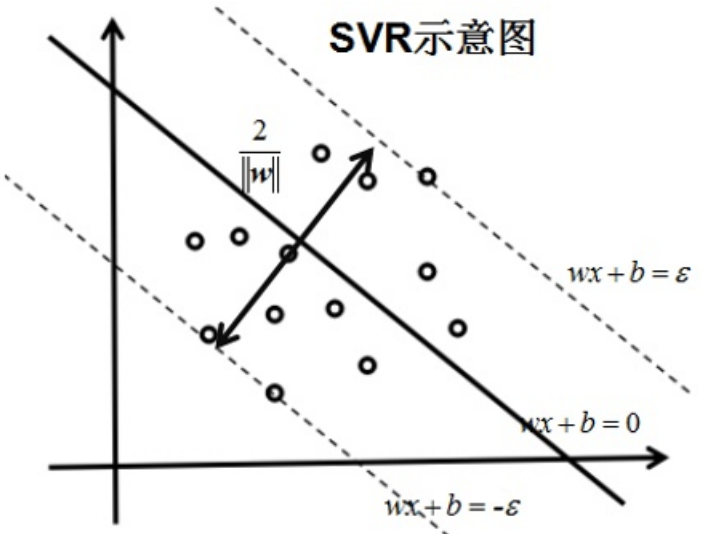
C-SVC



$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t. y_i(w^T x + b) \geq 1 - \xi_i, \xi_i \geq 0$$

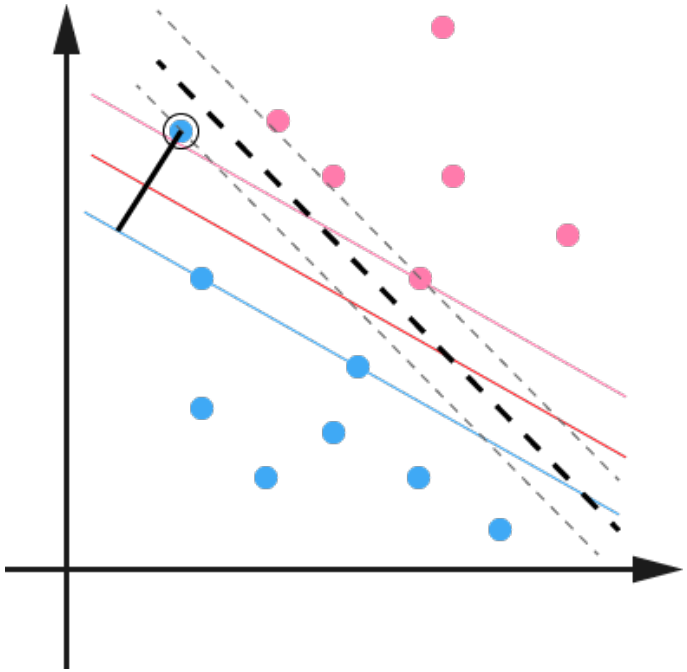
ε-SVR

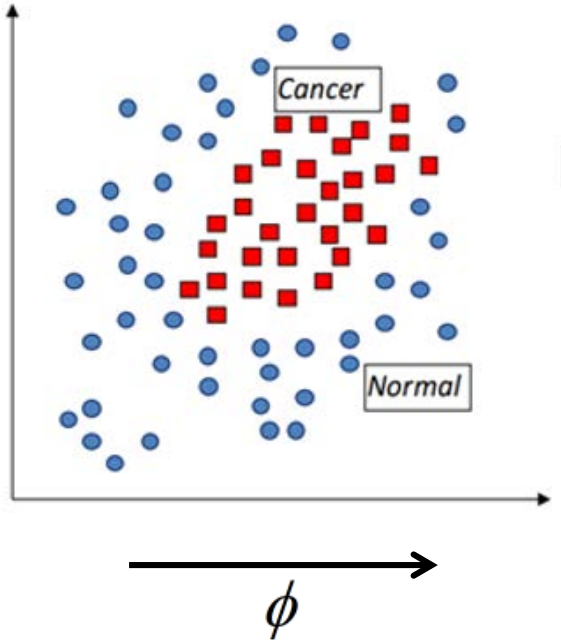


$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$s.t. \begin{cases} y_i - w^T x_i - b \leq \varepsilon + \xi_i \\ w^T x_i + b - y \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

松弛变量和映射到高维

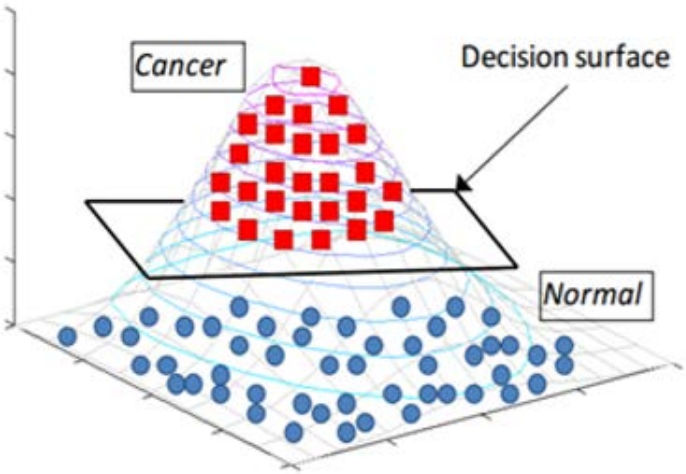




徐普
数学,计算机,加班

29 人赞同了该回答

f(x,y)=2x+3y是线性的,
f(x,y)=2x+3y+4xy是非线性的,
f(x,y,xy)=2x+3y+4xy是线性的.



核心软件包简介

南京大学 NANJING UNIVERSITY

- EMD：台湾中央大学包
  - 正宗原班人马打造
  - HHT变换提出者黄锷院士、EEMD论文第一作者吴召华教授等编写
  - EEMD是EMD升级版，EMD一键分解：

```
>>rslt=eemd(y,1,0);
```

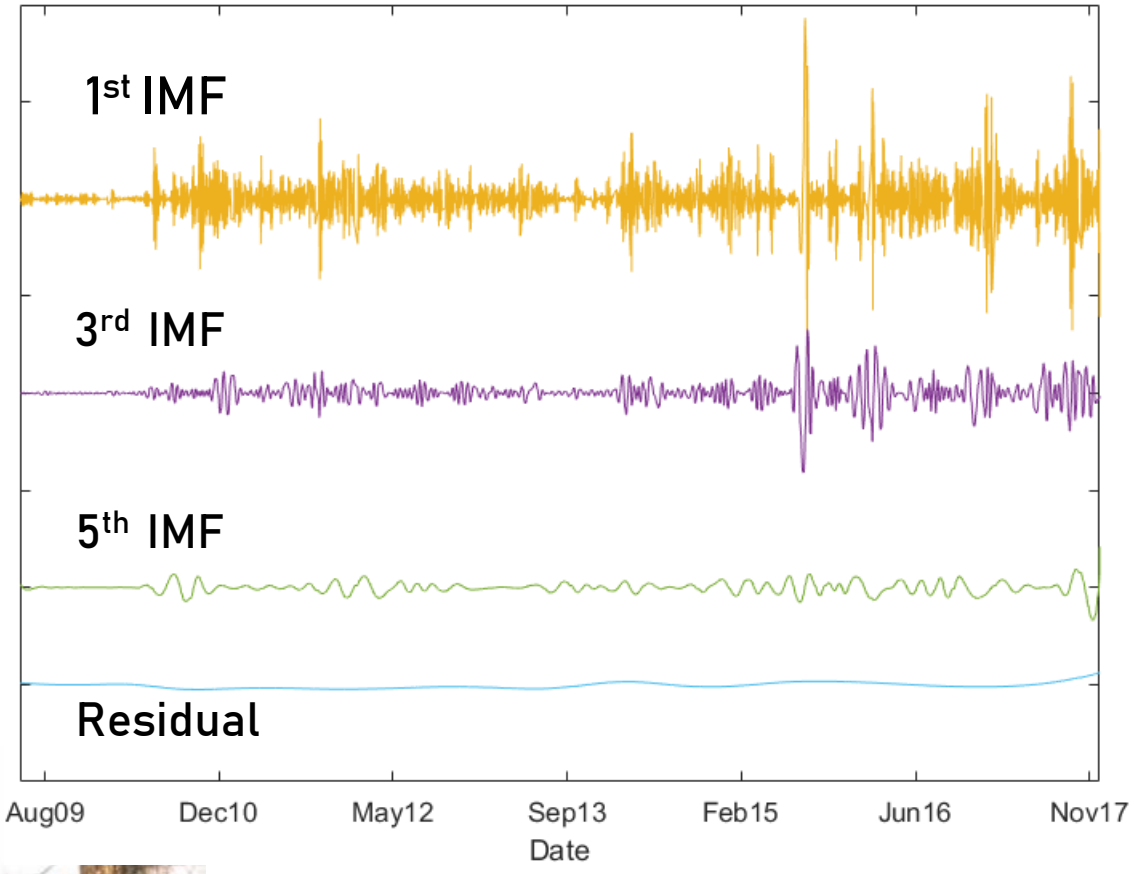
  - 返回矩阵中第一列是原数据，其余为IMF

- SVM：LibSVM
  - 台湾大学林智仁教授编写
  - 涵盖C, Java, Python等
  - Matlab使用Set Path添加libsvm/windows文件夹即可
  - 相比Matlab自带的SVM，LibSVM支持SVR及多种分类模型，且参数可调
  - 三步走：读入-训练-预测
  - 采用SMO算法计算最优

EMD

南京大学 NANJING UNIVERSITY

分解人民币-美元汇率收益率




```
hold on;
for i=2:2:6;
    %the 1st is original so 2nd is 1st IMF
    %plot 2,4,6 column -> 1st,3rd,5th IMF
    plot(date,rslt(:,i)-0.005*(i-1));
end;
plot(date,sum(rslt(:,7:12),2)-0.03);
% sum of all other IMF, totalled 12
set(gca,'yTickLabel',[]);
% let y label be blank
datetick('x',12,'keepticks');
xlim([733928,737029]);
```

与作者得到的结论不同，我们12个IMF经Matlab检验均非正态分布，Lilliefors test, Komogorov-Smirmov test, Jarque-Bera test p值均显著小于0.0001. 因此dEMD失效

LIBSVM入门

南京大学 NANJING UNIVERSITY

- 三个核心函数

```
>> [y, x]=libsvmread(filePath);
>> model=svmtrain(y,x,'Parameters');
>> py=svmpredict(y,x,model);
```

- 输入文件格式(为什么?)

SVR

```
1 1:1 2:1 3:1 4:1 5:1 6:1 7:1
1.0006 1:1.0107 2:0.99967 3:0.98383 4:1.002 5:1.0024 6:0.97355 7:0.99437
1.0006 1:0.99977 2:1.0346 3:0.95957 4:0.98822 5:1.0284 6:0.98348 7:1.0054
1.001 1:1.0025 2:1.0517 3:1.0027 4:0.99957 5:1.0334 6:0.97957 7:1.0054
1.001 1:0.98762 2:1.0192 3:1.035 4:0.99705 5:1.0283 6:0.98893 7:0.99806
1.0016 1:0.98279 2:1.0307 3:1.0539 4:0.99604 5:1.0315 6:0.96637 7:0.97406
1.0013 1:0.99443 2:1.0252 3:1.0404 4:0.99953 5:1.0358 6:0.95605 7:0.96944
1.001 1:0.98712 2:1.0417 3:1.0728 4:0.99605 5:1.0459 6:0.99455 7:0.96944
1.0014 1:0.99739 2:1.017 3:1.0458 4:1.0021 5:1.0361 6:0.99483 7:0.96944
1.0013 1:0.99142 2:1.0106 3:1.027 4:1.0015 5:1.0168 6:1.0001 7:0.97175
1.0014 1:0.97532 2:1.0265 3:1.0135 4:0.99979 5:1.0377 6:0.97937 7:0.97175
1.0011 1:0.978 2:1.0071 3:1.0022 4:0.96235 5:1.036 6:0.96172 7:0.95093
1.0015 1:0.9862 2:1.0206 3:0.99191 4:0.96589 5:1.0533 6:0.95744 7:0.95093
1.0012 1:0.98299 2:1.0301 3:1.0404 4:0.97402 5:1.0697 6:0.94113 7:0.95093
1.0014 1:0.98583 2:1.0299 3:1.0216 4:0.97705 5:1.0775 6:0.94877 7:1.0064
1.0011 1:0.97999 2:1.278 3:1.0027 4:0.94715 5:1.0784 6:0.95611 7:0.95679
1.0012 1:0.99508 2:1.0938 3:0.98383 4:0.94934 5:1.0789 6:0.92852 7:0.93519
1.0009 1:0.98389 2:1.045 3:1.0027 4:0.95553 5:1.0918 6:0.94725 7:0.94336
1.0012 1:0.98899 2:1.053 3:0.95687 4:0.97602 5:1.0908 6:0.96752 7:0.95148
1.001 1:0.99323 2:1.1732 3:0.94879 4:0.97458 5:1.0945 6:0.98474 7:0.95896
1.001 1:0.99479 2:1.0532 3:0.94609 4:0.98341 5:1.1125 6:0.98094 7:0.95236
```

SVC

```
+1 1:0.708333 2:1 3:1 4:-0.320755 5:-0.105023
-1 1:0.583333 2:-1 3:0.333333 4:-0.603774 5:1
+1 1:0.166667 2:1 3:-0.333333 4:-0.433962 5:-
-1 1:0.458333 2:1 3:1 4:-0.358491 5:-0.374429
-1 1:0.875 2:-1 3:-0.333333 4:-0.509434 5:-0.3
-1 1:0.5 2:1 3:1 4:-0.509434 5:-0.767123 6:-1
+1 1:0.125 2:1 3:0.333333 4:-0.320755 5:-0.40
+1 1:0.25 2:1 3:1 4:-0.698113 5:-0.484018 6:-
+1 1:0.291667 2:1 3:1 4:-0.132075 5:-0.237443
+1 1:0.416667 2:1 3:1 4:0.0566038 5:0.283105
-1 1:0.25 2:1 3:1 4:-0.06415 5:-0.506849 6:-
-1 2:1 3:1 4:-0.0943396 5:-0.543379 6:-1 7:1
-1 1:-0.375 2:1 3:0.333333 4:-0.132075 5:-0.5
+1 1:0.333333 2:1 3:-1 4:-0.245283 5:-0.50684
-1 1:0.166667 2:-1 3:1 4:-0.358491 5:-0.19178
-1 1:0.75 2:-1 3:1 4:-0.660377 5:-0.894977 6:-
+1 1:-0.291667 2:1 3:1 4:-0.132075 5:-0.15525
+1 2:1 3:1 4:-0.132075 5:-0.648402 6:1 7:1 8:
-1 1:0.458333 2:1 3:-1 4:-0.698113 5:-0.61187
-1 1:-0.541667 2:1 3:-1 4:-0.132075 5:-0.66666
+1 1:0.583333 2:1 3:1 4:-0.509434 5:-0.52968
```

SVM应用

南京大学 NANJING UNIVERSITY

- 二分类问题
  - 根据身高体重分男女生
  - 根据化学成分分酒品种
  - …
- 文本情感分类(初级)
  - 新闻标题：积极 | 消极
  - 分词包将句子分为词语
  - 词语转数值(Tf-idf)
  - 每个词语看作一维向量
  - 贴上标签放到SVM里跑

- 文字转LibSVM (Java)

建筑与工程：转型改革两条主线
异动股扫描
多利好促山东路桥涨停
建筑股表现不俗 山东路桥等两股涨停
山东3P模式再推进 概念股受益
山东路桥以拼搏精神闯天下
山东路桥以人为本 打造一流的国家化企业
山东路桥以先进科技的经营理念打造国际型企业
1月28日晚间数据揭秘
山东路桥严把质量关，用品质赢得市场

事项, 1, 0. 427001844
涨停, 2, 0. 421816409
业务, 3, 0. 340702939
机构, 4, 0. 33
跌停, 5, 0. 31
新, 6, 0. 278848355
上涨, 7, 0. 256540486
净利, 8, 0. 256540486
遭, 9, 0. 242529412
发展, 10, 0. 239673531
定制, 11, 0. 239673531
新股, 12, 0. 239673531
项目, 13, 0. 231586
交易, 14, 0. 211934297
宜, 15, 0. 211934297

```
+1 67:0.137059087
+1 2:0.421816409 139:0.1019382
+1 2:0.421816409
+1 87:0.125922501 107:0.114241625 110:0.114241
+1 35:0.167752801 74:0.130561857
+1 35:0.167752801 74:0.130561857
+1 17:0.208164799 37:0.167752801
+1 113:0.114241625
+1 2:0.421816409 52:0.141662814 53:0.137059087 63:0.137059087
+1 2:0.421816409 17:0.208164799 102:0.114241625
+1 125:0.114241625 136:0.1019382
+1 35:0.167752801 78:0.125922501
+1 53:0.137059087 102:0.114241625 136:0.1019382
+1 114:0.114241625
+1 8:0.256540486 48:0.147712125 113:0.114241625 142:0.1019382
+1 10:0.239673531
+1 2:0.421816409 6:0.278848355
+1 2:0.421816409 6:0.278848355 17:0.208164799
```

LIBSVM参数设置

南京大学 NANJING UNIVERSITY

- 示例格式 `model=svmtrain(y,x,'-s 3 -t 0 -c 0.3 -g 0.1 -p 0.02');`
- 主要参数说明
  - -s: SVM 类型, 0为C-SVC, 3为ε-SVR (1号ν-SVC与0类似)
  - -t: 核函数类型, 0为线性, 1为多项式, 2为RBF
  - -c: cost, 惩罚系数, 越大则对误差要求越严格
  - -g: gamma, 核函数系数 (核函数形式自带文档中给出)
  - -p: ε-SVR中的ε值, 越大则允许的回归管道宽度越大
  - -v: 交叉验证次数, 随机分成n份交互检验Acc/MSE
  - **填错或不填均按默认值处理，因此不必担心崩溃**
- 预测效果衡量 `[py,acc/mse,a]=svmpredict(y,x,model);`
  - 回归返回MSE，分类返回Accuracy

参数调优

南京大学 NANJING UNIVERSITY

- 很明显，论文作者在参数调优没下太多功夫

Table 5
Comparison of support vector regression as estimator between the Differential EMD and 'Dependent variables'.

Parameters/input signals	*Dependent variables*	Differential EMD
Kernel function	Polynomial	Laplace
Support vectors	1176	374
Epsilon	1	1
Cost	1	1
Sigma	1	0.0895736571
Degree	1	n/a
Scale	n/a	n/a
Offset	n/a	n/a
Order	n/a	n/a
Objective Function	(144.1071)	(34.4059)
Training error	0.035348	0.004379
Prediction (%)	53.07	67.81

实验操作流程

南京大学 NANJING UNIVERSITY

- Matlab处理
  - price2ret转为收益率
  - ret2price实现归一化
  - dlmwrite保存为txt
  - libsvmread读入数据
  - 手动调参大体确定
  - Faruto包精确调优
  - 建立模型并预测
  - 结果反馈
- 收益率效果较差略去展示

- Java转格式

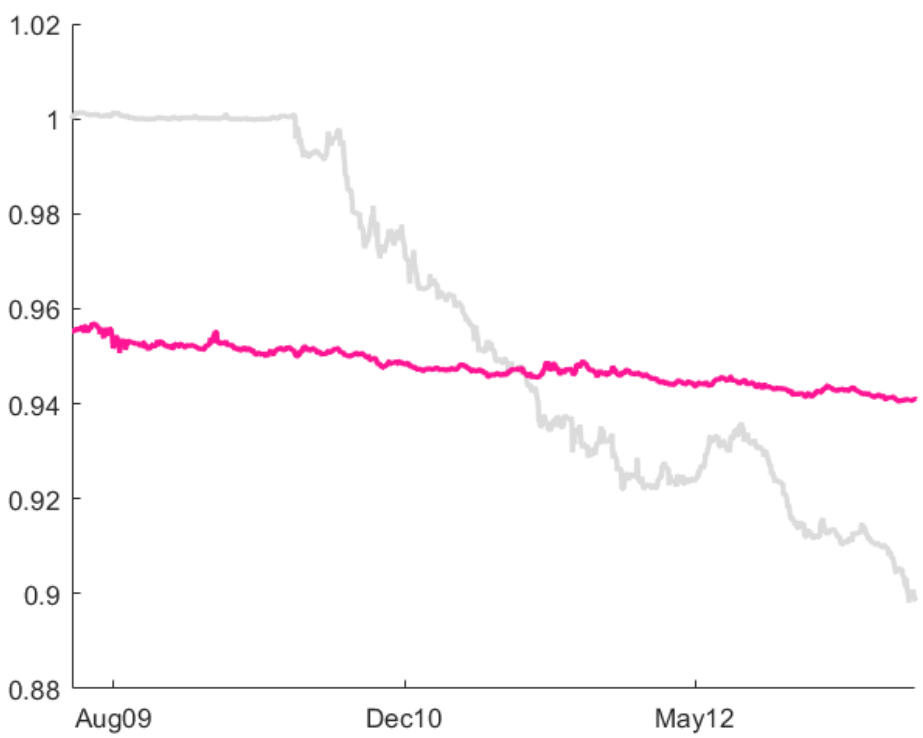
```
FileReader fr=new FileReader("f:/all.txt");
BufferedReader br = new BufferedReader(fr);
FileWriter fw=new FileWriter("f:/asvm.txt");
BufferedWriter bw =new BufferedWriter(fw);
String s;
String[]str;
int i=0;
while ((s=br.readLine())!=null){
    str=s.split(" ");
    bw.flush();
    StringBuilder newStr=new
StringBuilder(str[0]);
    for(int j=1;j<=7;++j){
        str[j]=" "+j+":"+str[j];
        newStr.append(str[j]);
    }
    bw.write(newStr.toString());
    bw.newLine();

}
fr.close();
fw.close();
```

Trainset参数粗调: ε

南京大学 NANJING UNIVERSITY

'-s 3 -t 0 **-p 0.05**' '-s 3 -t 0 **-p 0.005**'



Mean squared error = 0.00110066 (regression)
Squared correlation coefficient = 0.803248 (regression)

Mean squared error = 0.000138524 (regression)
Squared correlation coefficient = 0.907776 (regression)

Trainset参数粗调: 核函数类型

南京大学 NANJING UNIVERSITY

'-s 3 **-t 0** -p 0.01'

Mean squared error = 0.000136746 (regression)
Squared correlation coefficient = 0.908446 (regression)

'-s 3 **-t 2** -p 0.01'

Mean squared error = 5.36887e-05 (regression)
Squared correlation coefficient = 0.964608 (regression)

Trainset参数粗调: 成本函数

南京大学 NANJING UNIVERSITY

'-c 0.1 -t 2 -p 0.01'

Mean squared error = 8.94353e-05 (regression)
Squared correlation coefficient = 0.940772 (regression)

'-c 100 -t 2 -p 0.01'

Mean squared error = 2.82864e-05 (regression)
Squared correlation coefficient = 0.981346 (regression)

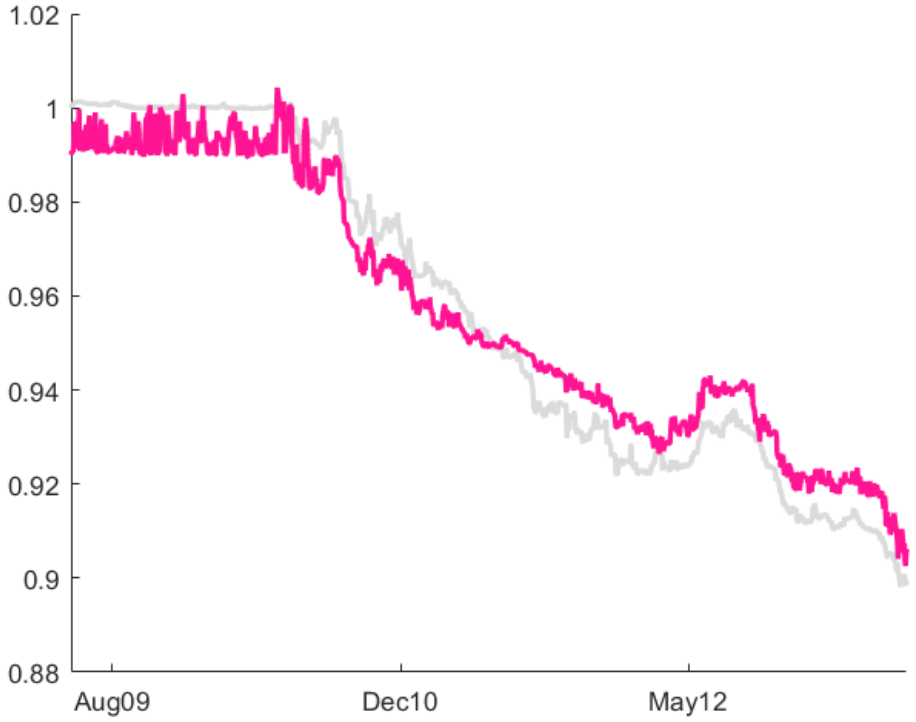
Trainset参数粗调: Gamma

南京大学 NANJING UNIVERSITY

'-c 1 -t 2 **-g 1**'


```
Mean squared error = 2.94287e-05 (regression)
Squared correlation coefficient = 0.981448 (regression)
```

'-c 1 -t 2 **-g 100**'




```
Mean squared error = 5.58583e-05 (regression)
Squared correlation coefficient = 0.985265 (regression)
```

基于遗传算法的参数精调

南京大学 NANJING UNIVERSITY

- LibsvmFarutoUltimate包
  - SVR可寻找最优c,g,p
  - Grid, GA, PSO三种实现
  - GA实现依赖GATBX包

```
ga_option=struct('maxgen',200,'sizepop',20,...
    'ggap',0.9,'cbound',[0,10],...
    'gbound',[0,10],'pbound',[0.005,0.02],'v',5);

[bmse,bc,bg,bp]=gaSVMcgpForRegress(y,x,ga_option);
```

bc	4.7265
bg	2.7770
bmse	9.3346e-06
bp	0.0050

  - 因计算耗时过长，第二轮优化被迫中止

```
'-c 4.73 -t 2 -g 2.78 -p 0.001'
```


```
Mean squared error = 3.0425e-06 (regression)
Squared correlation coefficient = 0.997965 (regression)
```

最终测试集运行结果

南京大学 NANJING UNIVERSITY

- 十分尴尬


```
Mean squared error = 0.0061033 (regression)
Squared correlation coefficient = 0.260817 (regression)
```

- 不如不调 “-s 3 -p 0.01”


```
Mean squared error = 0.00232252 (regression)
Squared correlation coefficient = 0.405525 (regression)
```

对文章的反思

南京大学 NANJING UNIVERSITY

- dEMD算法缺乏理论根基且不具备普适性
  - 并不是每个汇率收益率序列都能找到近似正态的IMF
  - 欧元-美元汇率直接减收益率IMF可以，津巴布韦币呢?
- SVM核心方法等浅尝辄止
  - 我们认为参数调优是SVM的核心工作(包括核函数)
  - 无论未考虑或因最后结果绝口不提都有失妥当
- 变动方向一致的衡量标准值得商榷
  - 绝对意义上偏离过大则方向预测正确率失去意义
  - 作者并未给出预测走势而给了一堆乱七八糟的图表
- 尽管有不解之处，我们非常感谢作者带给我们的启蒙！

对实验的总体反思

南京大学 NANJING UNIVERSITY

- 本次实验充分反映了金融变量的难以预测性
- 我们的主要不足
  - 广度上,作者采用了多种方法证明SVR的相对有效性,我们因水平时间有限无法用其他当今主流方法对比
  - 深度上,SVM理论较复杂,我们只停留在表面应用层次
- 改进空间
  - 去噪声或许可以用某一阶IMF去平滑而不必局限正态
  - SVR参数寻优可能需要大量、多样本实验得到
  - 和其他方法对比寻找相对意义上的最优

参考文献及链接

南京大学 NANJING UNIVERSITY

- 大部分参考来源于网络博客不一一列出

1. Haykin, S., 2008. *Neural Networks and Learning Machines*, 3rd ed., Pearson Education.
2. Chang, C. C., & Lin, C. J. 2011. *LIBSVM: A library for support vector machines*. ACM.
3. 梁循.支持向量机算法及其金融应用. 知识产权出社. 2012
4. 王小川等. MATLAB神经网络43个案例分析. 北京航空航天大学出版社. 2013

- EMD、LibSVM包代码均公开，LibSVM有专业人士写的代码分析，有兴趣的同学可以参考~

谢谢大家!