



# 动态因子模型在高维数据中的应用

## 有监督的动态因子模型预测

汪利军（导师：张荣茂）

数学科学学院

March 8, 2018





# 内容概要

研究背景

数学定义

方法综述

研究内容





- 在宏观经济学中，用于预测的实际序列的个数  $N$  非常大，经常远大于观测值个数  $T$ ，即  $N \gg T$ 。这种高维问题可以用少量的潜在因子来建模，采用动态因子模型能够起到降维目的(Stock and Watson, 2002)；
- 估计动态因子模型中的参数有多种方法，如传统的主成分方法 (PC)(Stock and Watson, 2002)，以及更适用于异方差情形的广义最小二乘 (GLS)(Breitung and Tenhofen, 2011)。



# 动态因子模型

令  $x_{it}$  为在时间  $t = 1, \dots, T$  第  $i = 1, \dots, N$  个观测值。因子模型由下式给出(Breitung and Tenhofen, 2011)

$$x_{it} = \lambda_i' F_t + e_{it}$$

其中  $F_t = [f_{1t}, \dots, f_{rt}]'$  是  $r$  维公共因子的向量,  $\lambda_i$  是对应的  $r$  维因子载荷向量。用矩阵表示, 模型可写成

$$\mathbf{X} = \mathbf{F}\mathbf{\Lambda}' + \mathbf{e} \quad (1)$$





# 动态因子模型

其中

- $\mathbf{X} = [X_1, \dots, X_T]'$  是  $T \times N$  的观测矩阵, 行向量为  $X'_t = [x_{1t}, \dots, x_{Nt}]$ ;
- $\mathbf{e} = [\mathbf{e}_1, \dots, \mathbf{e}_T]'$  是  $T \times N$  的特质 (idiosyncratic) 误差矩阵, 行向量为  $\mathbf{e}'_t = [e_{1t}, \dots, e_{Nt}]$ ;
- $\mathbf{F} = [F_1, \dots, F_T]'$
- $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N]'$





# 预测问题

用时间序列向量  $X_t$  对单时间序列变量  $y_t$  进行预测。

一般地，预测问题可以分两步求解(Stock and Watson, 2002)

- ① 根据  $X_t$  估计出因子  $F_t$ ;
- ② 估计  $y_t$  和因子  $F_t$  之间的关系。





# 主成分估计

在  $T^{-1}\mathbf{F}\mathbf{F}' = \mathbf{I}_r$  约束下以及一定的假设条件下, 最小化

$$S(\mathbf{F}, \mathbf{\Lambda}) = \text{tr}[(\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')'(\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')] \quad (2)$$

得到估计

$$\hat{\mathbf{F}} = \sqrt{T}\hat{\mathbf{V}}_r$$

其中  $\hat{\mathbf{V}}_r$  是  $\mathbf{X}\mathbf{X}'$  的前  $r$  个最大的特征值对应的特征向量组成的矩阵。(Stock and Watson, 2002)



# GLS 估计

考虑异方差，提出 GLS 估计，对加权残差平方和进行最小化

$$S(\mathbf{F}, \mathbf{\Lambda}, \mathbf{\Omega}) = \text{tr}[\mathbf{\Omega}^{-1}(\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')'(\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')] \quad (3)$$

根据  $\mathbf{\Omega}$  的不同，衍生了不同的方法

- ①  $\mathbf{\Omega} = \text{diag}[E(e_{1t}^2), \dots, E(e_{Nt}^2)]$  (Choi, 2012)
- ② 任意阵 (Forni et al., 2005; Choi, 2012)
- ③  $\mathbf{\Omega} = \sigma^2(\mathbf{I}_N - \varrho \mathbf{W}_N)^{-1}(\mathbf{I}_N - \varrho \mathbf{W}_N')^{-1}$  (Chudik et al., 2011)
- ④ 同时考虑异方差和自相关性 (Breitung and Tenhofen, 2011)





# 小结

- ① 很多论文研究对象是时间序列向量  $X_t$  (没有响应变量  $y_t$ ), 单纯考虑根据  $X_t$  来估计  $F_t$ , 如Breitung and Tenhofen, 2011
- ② 即使论文研究对象是对  $y_t$  进行预测, 但也是先根据  $X_t$  估计  $F_t$ , 再由  $F_t$  预测  $y_t$ , 如Stock and Watson, 2002





# 小结

- ① 很多论文研究对象是时间序列向量  $X_t$  (没有响应变量  $y_t$ )，单纯考虑根据  $X_t$  来估计  $F_t$ ，如Breitung and Tenhofen, 2011
- ② 即使论文研究对象是对  $y_t$  进行预测，但也是先根据  $X_t$  估计  $F_t$ ，再由  $F_t$  预测  $y_t$ ，如Stock and Watson, 2002

但如果我们提前用到  $y_t$  的信息呢？





# 基本想法

处理高维数据时，首先根据  $y_t$  与  $X_t$  中每个时间序列的关系，提取其中与  $y_t$  相关性较大的时间序列构造新的时间序列向量  $X_t^*$ ，对这个新的序列应用 PC 估计或者 GLS 估计。即

- ① 根据  $y_t$  和  $X_t$  得到初步降维后的  $X_t^*$
- ② 根据  $X_t^*$  估计出因子  $F_t$ ；
- ③ 估计  $y_t$  和因子  $F_t$  之间的关系。

这个想法是受监督主成分方法 (Supervised Principal Components)([Bair et al., 2006](#))的启发。





# 监督主成分

监督主成分的一般算法如下(Bair et al., 2006)

- ① 分别计算每个特征与输出变量之间的单变量回归系数；
- ② 从  $0 \leq \theta_1 \leq \dots \leq \theta_K$  中依次取阈值  $\theta$ 
  - a 提取出原特征矩阵中单变量回归系数的绝对值大于  $\theta$  的特征，  
从新特征中取前  $m$  个主成分
  - b 采用这些主成分对输出变量进行预测
- ③ 通过交叉验证选取  $\theta$  (和  $m$ )





# 监督主成分

监督主成分的一般算法如下(Bair et al., 2006)

- ① 分别计算每个特征与输出变量之间的单变量回归系数；
- ② 从  $0 \leq \theta_1 \leq \dots \leq \theta_K$  中依次取阈值  $\theta$ 
  - a 提取出原特征矩阵中单变量回归系数的绝对值大于  $\theta$  的特征，从新特征中取前  $m$  个主成分
  - b 采用这些主成分对输出变量进行预测
- ③ 通过交叉验证选取  $\theta$  (和  $m$ )

后续研究中，在第 2(b) 步中，尝试除 PC 之外的方法，如 GLS。







# 研究计划

- ① 推导有监督的方法（如有监督的主成分和 GLS）应用于动态因子模型的公式，并给出具体的假设条件，同时尝试讨论其渐近性质；
- ② 通过模拟实验将之与 PC 估计、GSL 估计进行比较；
- ③ 将方法应用于实际数据，如股票数据。








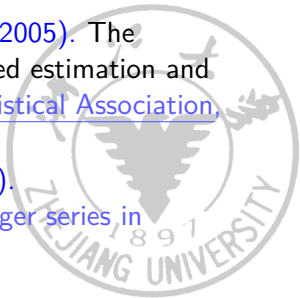
## 参考文献 I

-  Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. [Journal of the American Statistical Association](#), 101(473), 119–137.
-  Breitung, J. & Tenhofen, J. (2011). Gls estimation of dynamic factor models. [Journal of the American Statistical Association](#), 106(495), 1150–1166.
-  Choi, I. (2012). Efficient estimation of factor models. [Econometric Theory](#), 28(2), 274–308.
-  Chudik, A., Pesaran, M. H., & Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. [The Econometrics Journal](#), 14(1).



## 参考文献 II

-  Fan, J., Liao, Y., & Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. Annals of statistics, 39(6), 3320.
-  Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. Journal of the American Statistical Association, 100(471), 830–840.
-  Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York.







## 参考文献 III



Stock, J. H. & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors.  
[Journal of the American statistical association](#), 97(460), 1167–1179.





*Thank you!*

