【统计理论与方法】

高维面板数据降维与变量选择方法研究

张 波[1],方国斌[1,2]
(1. 中国人民大学 统计学院,北京 100872;2. 安徽财经大学 统计与应用数学学院,安徽 蚌埠 233030)

摘要:从介绍高维面板数据的一般特征入手,在总结高维面板数据在实际应用中所表现出的各种不同类型及其研究理论与方法的同时,主要介绍高维面板数据因子模型和混合效应模型;对混合效应模型随机效应和边际效应中的高维协方差矩阵以及经济数据中出现的多指标大维数据的研究进展进行述评;针对高维面板数据未来的发展方向、理论与应用中尚待解决的一些关键问题进行分析与展望。

关键词:高维;面板数据;降维;变量选择
中图分类号:O212;F222.3 文献标志码:A 文章编号:1007-3116(2012)06-0021-08

一、引 言

在社会现象观测和科学实验过程中经常会产生面板数据。这类数据通过对多个个体在不同时间点上进行重复测度,得到每个个体在不同样本点上的多重观测值,形成时间序列和横截面相结合的数据,也就是所谓的"面板数据"。由于应用背景的不同,面板数据有时也称作纵向数据(longitudinal data)。面板数据广泛产生于经济学、管理学、生物学、心理学、健康科学等诸多领域。

随着信息技术的高速发展,数据采集、存储和处理能力不断提高,所谓的高维数据分析问题不断涌现。对于多元统计分析而言,高维问题一般指如下两种情形:一种是变量个数 p 较大而样本量 n 相对较小,例如药物试验中有成千上万个观测指标而可用于实验观测的病人个数较少;另一种是变量个数 p 不大但是样本个数 n 较多,例如一项全国调查牵涉到大量的调查对象,而观测指标个数相对较少。面板数据高维问题较多元(时序)高维问题更为复杂,因为面板数据至少包括两个维度:时间和横截面。在实际应用中,不同个体在不同时间进行观测时可以获得多个指标值。为了以下论述的方便,用 p 表示指标个数,T 表示观测期长度,N 表示个体(individual)或主题(subject)个数。数理统计中所提到的高维(大维)问题,通常是指个体数 N、时期长度 T 或指标个数 p 这三个变量中的一个或多个可以趋向于无穷。具体应用中,只要 N、T 和 p 中有一个或多个大于某个给定的临界值,都称为高维问题。

本文主要研究两种基本类型的高维面板问题:一类为面板数据分析中解释变量个数 p 非常多,超过个体数 N 和时期数 T,比如零售商业网点成千上万种商品扫描数据,央行和国家统计部门得到的多个指标在不同个体宏观经济观测数据等;另一类是混合效应模型中随机效应和固定效应设定时方差协方差矩阵所需确定的参数个数较多,某些参数的值趋向于零,要对方差协方差矩阵进行变量选择,此时针对固定效应和随机效应可以采用不同的变量选择策略。

二、高维面板数据因子模型

大型数据集构成的社会经济面板的特点是具有

收稿日期:2011-11-14;修复日期:2012-04-22
基金项目:中国人民大学科学研究基金项目(中央高校基本科研业务费专项资金资助)《基于高频和超高维数据的中国金融市场若干重大问题研究》(10XNL007);国家自然科学基金项目《基于高频数据的股市极端风险测度及其防范研究》(71071155)
作者简介:张 波,男,黑龙江拜泉人,教授,博士生导师,研究方向:随机分析,高频数据分析;
方国斌,男,安徽宿松人,博士生,副教授,研究方向:高维数据分析,金融数据统计分析。

成百上千个观测指标，也就是具有所谓的高维特征。由于这种特征的存在，采用经典统计计量分析方法很难进行处理。因子模型(factor model)不仅可以有效降低数据的维度，而且可以充分体现面板数据内部的序列相依性和截面相依性，因此可以针对不同的应用领域建立相应的因子模型对高维面板数据进行分析。例如构建套利定价模型时，将多个证券的投资组合用公因子表示，进行收益率预测；研究经济周期变动，尤其是重大事件对经济发展影响时，将各经济体的产出指标用几个公因子表示，用因子模型分析各经济体同步变动情况以及重大事件对各经济体的冲击大小，等等。

面板数据因子模型是对解释变量或者误差成分项进行因子分解后所建立的模型。实际应用中，当模型中解释变量的个数较多，例如 p 大于 N，就可以对解释变量进行因子分解，用少数几个公因子和与之对应的因子载荷表示大量解释变量，从而起到降维的效果。对误差成分进行因子分解主要是为了体现个体或时间的共同趋势和交互效应，其中因子分解的方法一般采用多元统计分析中的主成分法，为了进一步研究的需要，有时候还要采用极大似然法或者回归法计算因子得分，并将因子得分代入模型进行估计。

因子模型中采用较广泛的是动态因子模型(dynamic factor model)，这主要是因为动态因子模型能够较好体现变量前后时期之间的相关性，便于进行外推预测，体现序列的内在结构。面板数据动态因子模型的一般形式如下：

$$\boldsymbol{X}_{it} = \boldsymbol{\Lambda}'_i \boldsymbol{F}_t + \boldsymbol{e}_{it} \quad (1)$$

$$\boldsymbol{Y}_{it+h} = \boldsymbol{\beta}'_F \boldsymbol{F}_t + \boldsymbol{\beta}'_w \boldsymbol{Z}_{it} + \boldsymbol{\varepsilon}_{it+h} \quad (2)$$

其中 $\boldsymbol{X}_{it}$ 表示第 i 个横截面单元在第 t 时刻的解释变量(协变量)的观测向量($i = 1,2,\ldots,N$；$t = 1,2,\ldots,T$)；$\boldsymbol{Y}_{it+h}$ 表示第 i 个横截面单元在第 $t+h$ 时刻的被解释变量(响应变量)的观测(预测)值，若 $h \neq 0$，则模型(2)为一个预测模型；$\boldsymbol{F}_t$ 是 $r \times 1$ 维潜在因子向量；$\boldsymbol{\Lambda}'$ 是 $p \times r$ 维因子载荷向量；$\boldsymbol{e}_{it}$ 是 $\boldsymbol{X}_{it}$ 的特质性成分；$\boldsymbol{Z}_{it}$ 表示已观测变量(例如 Y_t 的滞后变量)或不可观测的 $\boldsymbol{F}_t$ 的滞后项组成的 $q \times 1$ 维向量；$\boldsymbol{\beta}_F$ 和 $\boldsymbol{\beta}_w$ 分别是 $r \times 1$ 维和 $q \times 1$ 维向量，表示潜在因子和已观测变量的系数；$\boldsymbol{\varepsilon}_{it+h}$ 表示模型(2)的随机(预测)误差。一般称模型(1)为因子模型，模型(2)为动态模型。

动态因子模型在对解释变量(协变量)进行降维的同时，尽可能用较少公因子体现解释变量的大部分信息。对于社会经济现象中大量存在的高维面板数据而言，动态因子模型提供了高维问题降维的一种思路。相比较其他统计建模方法而言，动态因子模型充分考虑到横截面相关和序列相关对面板数据建模的影响，正确揭示了面板数据内部相依特征，能够更加合理地解释某些社会经济现象的变化规律。在动态因子模型估计和检验过程中，通过对统计量的渐近性和协方差矩阵的结构特征进行研究，推动了诸如随机矩阵理论、谱分解理论、高维变量选择等理论的进一步发展。近年来，动态因子模型已逐渐运用于大型宏观数据集的分析中。研究者分别从动态因子模型形式的设定、协方差结构和潜在因子的估计等方面进行了理论探讨，同时相关的应用研究也正在逐步展开。

(一)动态因子模型的设定和估计

高维面板数据集普遍存在序列相关和(弱)截面相关，Stock 和 Watson 提出在因子模型中加入观测变量的滞后项进行前向预测，从而充分考虑时间序列的相关性(动态性)[1]。他们在时齐因子模型的基础上采用时变因子载荷刻画序列和截面相依。在对美国联邦储备委员会工业产品指数的预测中，该模型与自回归模型(AR)和向量自回归模型(VAR)相比预测误差(MSE)相对较小。Stock 和 Watson 进一步将 VAR 和动态因子模型相结合，运用这种近似因子模型研究货币政策冲击对宏观经济的影响，讨论动态因子个数估计和 VAR 基础上的因子约束检验问题[2]；Pesaran 和 Chudik 在无限维向量自回归模型中采用动态因子，以体现具有显著效果的某个变量或截面单元对当期和滞后期其他变量的影响[3]；Song、Hardle、Ritov 考虑到时间序列中往往存在非平稳性和可能的周期性，提出了一种两步估计方法[4]：第一步，采用分组 LASSO(最小绝对收缩和选择算子)类型的技术选择时间基函数，运用平滑函数主成分分析选择空间基函数；第二步，运用动态因子模型获得一个去除趋势(又称退势)的低维随机过程，并将这种广义动态半参数因子模型应用于气温、核磁共振和隐含波动面数据的分析中。

动态因子载荷的估计也得到了进一步的研究。Forni 等人提出了一种两阶段“广义主成分”估计方法，第一步估计公共成分的协方差，第二步确定主成分分析的权重，这种分析放宽了对特性因子的结构约束[5]；Deistler 和 Zinner 讨论了广义线性动态因子模型的结构特征，包括可识别性，模型估计等一系列问题[6]；因子载荷阵用随机游走表示显然缺乏实

际证据，Banerjee 和 Marcellino 研究表明运用因子载荷中的时间变动进行预测效果较差，尤其是小样本情形[7]；传统的假设要求特性因子的结构为对角矩阵，然而由于因子载荷中可能存在结构突变，这一条件很难得到满足，Breitung 和 Eickmeier 提出构造 LR、LM 和 Wald 统计量对静态和动态因子模型结构突变进行检验，并将其运用于美国和欧元区国家经济增长模式转变的研究[8]。

因子个数的选择是因子分析必须考虑的问题之一。在高维动态因子模型中，因子个数的选择可以不依赖于复杂的协方差矩阵；Bai 和 Ng 提出了高维面板数据选择因子个数的一种准则，这种准则考虑由因子模型的类型来决定因子个数，而不是采用数据驱动的方法[9]；Hallin 和 Liska 运用谱密度矩阵的特征值识别广义动态因子模型的因子个数[10]；动态因子模型不仅要确定因子个数，还要确定解释变量的滞后阶数，Harding 和 Nair 对传统的碎石图(scree plot)方法予以了推广，并运用随机矩阵理论和 Stieltjes 变换对特征值的分布进行分析，得出了基于矩的因子个数和滞后阶数的一致估计方法[11]。

高维面板数据分析中，因子个数的多少决定了最终维数的大小，同时也决定了因子模型解释能力的大小。在尽量减少原有信息损失的同时，选择合理的公因子个数将是一个长期讨论的问题。

(二)因子载荷阵协方差结构和潜在因子估计

在金融学的套利定价理论中，多因子模型可以用于减少维度和估计协方差矩阵。好的协方差矩阵估计量可以避免过度放大估计误差，协方差矩阵的最小和最大特征值对应于证券投资组合的极小和极大的方差，协方差矩阵的特征向量可用于优化投资组合。应用因子模型的协方差矩阵在进行证券投资组合选择时，所包含的统计含义和实际意义比较明显，而估计高维协方差矩阵则相对比较困难，Fan、Fan、Lv 研究了高维因子模型的维数对协方差矩阵估计的影响，并通过对样本协方差矩阵估计和基于因子模型估计进行比较，得出了协方差矩阵的逆矩阵更有利于揭示因子结构的结论[12]；由于投资组合的优化配置和投资组合方差的减少都与协方差矩阵的逆矩阵有关，因此在优化投资组合配置中研究因子结构具有重要意义，但其风险评价效果欠佳，Hautsch 和 Kyj 基于已实现协方差多重标度谱分解(Multi－scale spetral decomposition)分析高维动态协方差，将该原理运用于标准普尔 500 股票全局最小方差(GMV)投资组合的构建，检验基于协方差矩阵的投资组合样本外预测的效果[13]。

协方差矩阵结构的研究目前主要运用于投资组合的构建，已有研究主要从协方差矩阵的特征根和特征向量以及协方差矩阵的逆矩阵出发，而对于高维情形，协方差矩阵的估计受维度影响。

潜在因子(latent factor)，又称隐性因子或公因子，潜在因子的估计主要是指因子载荷矩阵的估计。一般通过对解释变量(协变量)的 $N\times N$ 阶非负定矩阵的特征分析进行因子载荷矩阵和因子过程的估计。解释变量的个数(N)和时期长度(T)之间长度往往不一致，对于高维数据而言，如果 $N>T$，可以采用 Bai 提出的最小二乘法进行潜在因子的估计[14]；对于合适的变量个数 N 和非平稳因子估计，Pan 和 Yao 通过求解几个非线性规划问题来解决[15]；Lam，Yao、Bathi 研究表明：当所有因子都比较强大并且因子载荷矩阵每一列的范数都是 N 的 $1/2$ 次方阶数时，因子载荷矩阵估计的弱一致 L_2 范数与 N 的收敛比率独立，并运用这种估计方法进行了三支股票的隐含波动面建模分析[16]。

潜在因子的估计主要基于因子载荷矩阵的分析。由于潜在因子既代表解释变量的共同行为，又是因子模型分析基础，高维数据分析中潜在因子的估计方法将决定协方差矩阵结构特征的刻画。

三、高维面板数据内部相依性的刻画

面板数据内部相依包括序列相依和截面相依。高维面板数据分析中，横截面相依对模型的估计和检验影响较大。近年来，截面相依的处理逐渐得到重视，包括相依类型刻画和度量等。由于序列相依和横截面相依经常同时出现，所以在讨论横截面相关时通常也会考虑序列相依。

在空间相依存在的情况下，也就是存在个体的异质性，处理这种相依性的一般方法就是进行空间加权和引入空间滞后算子建立空间滞后模型。假设对如下简单的混合回归模型进行估计：

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \tag{3}$$

其中 $\boldsymbol{y}$ 是 $NT\times 1$ 向量，$\boldsymbol{X}$ 是 $NT\times K$ 矩阵，$\boldsymbol{\beta}$ 是 $K\times 1$ 向量，$\boldsymbol{\varepsilon}$ 是 $NT\times 1$ 向量。在考虑横截面相依的条件下，各个个体的相依关系通过空间加权矩阵来表示。按照相依结构的不同，空间相依又可以分成两类：第一类是解释变量的个体相依，称之为空间滞后模型；第二类是误差项的空间相依，称之为空间误差模型。

(一) 空间权重的设定

空间权重的设定是空间经济学中的一个重要问

题，一般空间权重都是预先设定的。计量经济分析中，空间权重可采用经济距离表示，也可采用分块权重(block weights)，例如将中国一个省内的多个地区各看作一个分块。Anselin 提出一种空间滞后模型，或称混合空间自回归模型[17]，其特点是在模型的右端项设置一个空间滞后解释变量，虽然这种方法针对的是截面情形，但是通过堆栈(stacked) 的方法很容易运用于面板建模，即用如下模型：

$$\boldsymbol{y} = \rho(\boldsymbol{I}_t \otimes W_N)\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

其中 ρ 是空间自回归参数，$\boldsymbol{I}_t$ 是 T 阶单位阵，$\boldsymbol{W}_N$ 是 N 阶权重矩阵，$\otimes$ 表示 Kronecker 积，其他字母和符号的含义如前。

空间滞后模型在一些社会或空间交互效应的文献中得到应用。Brueckner 和 Jan 分别将其运用于空间反应函数(spatial reaction function)和社会乘子(social multiplier)的参数估计当中[18]；Anselin 进一步提出所谓的空间乘子(spatial multiplier)，并将其用于空间体系中设定被解释变量为解释变量和随机误差项的函数[19]。

(二)空间误差模型

与空间滞后模型相比，空间误差模型并不要求建立一个空间交互作用的理论模型，而是考虑非球形误差项协方差矩阵。空间误差模型除了直接表示协方差结构以外，还可以采用空间误差过程、空间误差成分和公因子(common factors，或称共同因子)模型，其中公因子模型是当前正在发展的一种主流方法，尤其适用于高维面板数据的分析。空间误差模型使用加权矩阵来表示相对位置和近邻程度，模型中相邻关系的设定不同于协方差矩阵的空间相依范围的设定。通过对模型误差项结构的分析，Anselin、Bera 和 Anselin 提出了两种常用的空间误差模型：空间自回归(SAR)模型和空间移动平均(SMA)模型。这两种模型分别运用于讨论误差项存在横截面误差自相关和共同变动情形[19-20]。空间误差成分模型(SEC)由 Kelejian 和 Robinson 提出，与 SAR 和 SMA 不同，SEC 的误差项被分解成局部效应(local effect)和溢出效应(spillover effect)两部分[21]。在异质性面板的误差成分模型中，时间成分被表示成不可观测的共同效应或因子(factor)，它包含了所有的横截面单元。与标准的误差成分不同的是，每一个横截面单元在这个因子上有不同的因子载荷。最简单的形式是所谓的单因子结构，这时误差项可以表示为：

$$\varepsilon_{it} = \delta_i f_t + u_{it} \tag{5}$$

其中 δ_i 表示因子 f_t 在横截面上的载荷，u_{it} 是均值为 0 的独立同分布的误差项。共同因子模型已经推广到多重因子情形。

四、高维面板数据混合效应模型的变量选择

(一) 面板数据混合效应模型

混合效应模型是面板数据研究中最重要的模型之一，该类模型的研究已比较充分[22]。此类模型包括线性和非线性参数混合效应模型、半(非) 参数混合效应模型、广义线性混合效应模型。线性和非线性参数混合效应模型是两种参数混合效应模型，从贝叶斯的角度看，这两种模型分别是分层线性和非线性模型。线性混合效应模型是指响应(被解释) 变量和协变量(解释变量) 为线性关系，线性混合效应模型(LME) 一般可表示为：

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{6}$$

其中 $\boldsymbol{b}_i \sim N(0,\boldsymbol{D})$，$\boldsymbol{\varepsilon}_i \sim N(0,\boldsymbol{R}_i)$，$i = 1,\cdots,n$，$\boldsymbol{y}_i$ 和 $\boldsymbol{\varepsilon}_i$ 分别是第 i 个个体的解释变量向量和测度误差，$\boldsymbol{\beta}$ 和 $\boldsymbol{b}_i$ 分别是固定效应(总体参数) 和随机效应(个体参数)，$\boldsymbol{X}_i$ 和 $\boldsymbol{Z}_i$ 是相关的固定效应和随机效应的设计阵。固定效应部分对应总体参数估计，随机效应部分对应个体参数估计。

非线性混合效应模型(NLME) 中响应变量和协变量是非线性形式，模型中非线性函数已知，只有非参数是未知的。分层非线性模型或 NLME 模型的一般形式可表示为[22]60-61：

$$\boldsymbol{y}_i = f(\boldsymbol{X}_i,\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i \tag{7}$$

其中

$$\boldsymbol{\beta}_i = d(\boldsymbol{A}_i,\boldsymbol{B}_i,\boldsymbol{\beta},\boldsymbol{b}_i)$$

$$\boldsymbol{b}_i \sim N(0,\boldsymbol{D}) \quad \boldsymbol{\varepsilon}_i \sim N(0,\boldsymbol{R}_i) \quad i = 1,\cdots,n$$

其中 $f(\cdot)$ 是已知函数，$f(\boldsymbol{X}_i,\boldsymbol{\beta}_i) = [f(\boldsymbol{X}_{i1},\boldsymbol{\beta}_i),\cdots, f(\boldsymbol{X}_{in},\boldsymbol{\beta}_i)]^T$，$\boldsymbol{X}_i = [x_{i1},\cdots,x_{in}]$ 是设计阵，$\boldsymbol{\beta}_i$ 是第 i 个个体的特有参数。在非线性混合效应模型中，$d(\cdot)$ 是设计阵 $\boldsymbol{A}_i$ 和 $\boldsymbol{B}_i$ 的已知函数，$\boldsymbol{\beta}$ 和 $\boldsymbol{b}_i$ 分别是固定效应和随机效应向量。

面板研究中，通常认为来自不同个体的数据相互独立，而来自同一个体的数据是相关的，这种相关可能是由于个体间的异质性，也可能是由于测度误差的序列相关所致，而忽略这些相关性可能导致估计结果并非有效。面板分析的核心问题就是选择合适的模型和正确估计方差协方差成分的方法，这也是面板数据分析与其他类型的数据分析都面临的主要问题。选择线性模型还是选择非线性模型，主要

根据响应变量和协变量之间的关系，并需要根据不同的应用背景以及图形的直观解释，如果假定响应变量和协变量之间没有任何非线性关系，就可以采用非参数方法进行研究。

在估计混合效应模型随机效应和固定效应方差协方差成分的时候，由于待估参数较多，所以有时需要进行变量选择，相对而言固定效应变量选择比较直观，随机效应变量选择难度稍大，因为其方差结构较为复杂。Chen 和 Dunson 提出了采用分层贝叶斯模型识别 0 方差的随机效应，通过再参数化混合模型使得随机效应分布的协方差参数函数与回归系数结合成标准正态潜变量，以选择随机效应方差的混合先验进行多个随机效应的变量选择[23]；Vaida 和 Blanchard 提出了采用条件赤池信息准则（cAIC）对混合效应模型进行变量选择的方法[24]；显著的随机效应选择依赖于协方差选择策略，Dziak 等人对纵向数据的变量选择方法进行了综述[25]。

（二）高维面板混合效应模型的变量选择

面板数据分析中经常存在很多变量，这些潜在的预测子（potential predictors）个数可能很大，尤其是为了减少可能的建模偏差而引入非线性项和协变量的交互效应时。事实上通常在模型中包含着一个重要变量的子集，也就是所谓的最优子集（best subset），它能够增强模型的可预测性，并且能够使得模型更加精简，变量选择的终极目标也就是找到这个最优子集。线性回归模型中存在很多子集选择准则，一些传统的变量选择方法（如 Mallows 信息准则（Cp）、赤池信息准则（AIC）、舒瓦茨信息准则（BIC））也已推广到面板数据中，而更多的是考虑采用惩罚似然的方法，例如在线性混合效应模型（6）的变量选择中，令 $\ell_i(\beta,\theta)$ 为给定 $\boldsymbol{x}_i$ 和 $\boldsymbol{z}_i$ 时 $\boldsymbol{y}_i$ 的条件似然函数的对数，定义惩罚条件对数似然函数为：

$$\frac{1}{n}\sum_{i=1}^{n}\ell_i(\beta,\theta)-\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|) \tag{8}$$

其中 $p_{\lambda_j}(\cdot)$ 是带有正则化参数 λ 的惩罚函数，最大化上式得出惩罚似然估计量，λ 控制模型的惩罚性，可以设成固定值或者通过数据驱动的选择方法，例如采用广义交叉验证（GCV）；惩罚函数 $p_{\lambda_j}(\cdot)$ 的选择在罚似然变量选择中非常重要，不恰当的罚函数达不到应有的效果。若令惩罚函数为熵或者 L_0 惩罚，即：

$$p_{\lambda_j}(|\beta_j|)=\frac{1}{2}\lambda^2 I(|\beta_j|\neq 0) \tag{9}$$

其中 $I(\cdot)$ 是示性函数，所有的 $\lambda_j=\lambda$，带有熵惩罚的惩罚似然函数可以写作：

$$\frac{1}{n}\sum_{i=1}^{n}\ell_i(\beta,\theta)-\frac{1}{2}\lambda^2\,|M| \tag{10}$$

其中 $|M|=\sum_j I(|\beta_j|\neq 0)$ 代表候选模型的参数个数。

在误差项独立同分布假设下，进行线性回归模型惩罚最小二乘估计时，一些其他类型的惩罚被引入。惩罚函数 $p_{\lambda_j}(\cdot)$ 的形式决定了估计量的优劣。定义 L_p 惩罚为 $p_{\lambda_j}(|\beta_j|)=\lambda_j p^{-1}|\beta_j|^p, p>0$，这样最小二乘 L_2 惩罚得到脊（ridge）回归估计量；$0<p<2$ 的 L_p 惩罚就是桥（bridge）回归，介于最优子集选择和脊回归之间。L_1 惩罚下，惩罚似然估计量是最小绝对收缩和选择运算子（LASSO）。Fan 和 Li 建议使用平滑切割绝对偏差（SCAD）惩罚，这种方法有两个调整参数，而 SCAD 估计量和 LASSO 估计量很相似，它能得出一个稀疏和连续的解，并且认为 SCAD 比 LASSO 有更低的偏差[26]；Zou 在 LASSO 的基础上提出了适应最小绝对收缩和选择运算子（ALASSO），这种方法具有所谓的神谕（oracle）性质[27]。

Liang 和 Zeger 提出了一种广义估计方程（GEE）的方法对聚类（clustered）或面板数据拟合回归模型，响应变量可以是连续的或离散的[28]，可将这种方法视为拟似然（quasi－likelihood）的一种推广，是一种伪似然（Pseudo－likelihood）方法。GEE 不用假定变量的分布，克服了似然函数不能表示的问题，并且不需要方差独立假设，这些与传统的变量选择方法（比如 Cp，AIC 和 BIC 等）有很大区别，可运用交叉验证（CV）方法选择较小的广义残差平方和（GRSS）或者期望预报偏差（EPB）。SCAD 和 LASSO 与 GEE 相结合，得出惩罚广义估计方程（PGEE），Fu 研究了 L_q 惩罚的 PGEE 的渐近性质以及具体实现，并建议采用广义交叉验证（GCV）选择正则化参数 λ_j[29]。

混合效应模型中方差选择问题的研究文献相对较少，大多数变量过程采用参数或半参数方法研究（不）具有随机效应或不可观测的数据。但是，这些过程主要用来选择显著的固定效应，与之不同的是 Bondell，Krishna、Ghosh 的工作，他们考虑了线性混合效应的选择[30]；Ibrahim 等人使用了一种新颖的再参数化方法，将混合效应的选择看做模型中具有很多缺失数据的分组变量选择，其中的缺失数据代表随机效应[31]；Ni 等人提出了面板数据半参数混合模型中同时进行变量选择和模型估计的双惩罚

似然方法，这种方法将两种惩罚相结合，考虑在普通对数似然上加入两类惩罚：非参数基线函数的粗糙性惩罚和获取模型稀疏性线性系数的非凹收缩惩罚，Ni 等人认为这种方法可以对缺失数据进行正确推断，如果模型设置正确，这种推断更为有效，而且易于计算[32]。

五、研究展望

高维数据变量选择讨论的主要问题是解释变量的个数较多，超过（甚至远大于）个体数情形。对于面板数据而言，这些协变量有可能是实际观测到的解释变量，也可能是模型设定过程中产生的成分（component）变量，例如随机效应成分和固定效应成分。针对这两种不同情形，主要采用高维因子模型和混合效应模型的变量选择方法

在此主要讨论高维面板数据分析和混合效应模型的变量选择问题。高维数据变量选择方法还在不断发展，半参数、贝叶斯统计等方法论已经广泛运用于这类问题中。从生物学和医学角度开展的研究较多，因为大量变量和参数中存在所谓的稀疏性（sparsity），所以变量选择方法很适合于对这类问题的处理。无论是现有的哪种变量选择方法，都很难做到既不损失原有信息，又能正确地决策判断。社会经济应用中，针对大规模数据集的处理，仅仅从降维角度去考虑显然不够，更多地还是要提高模型对数据的拟合效果。所以，高维变量选择技术在经济管理中的应用仍亟待开展。

从未来的发展看，高维面板数据分析主要应该关注以下五个方面的问题：

（一）变量选择技术的发展

对于高维问题而言，首先要解决的问题就是降维。无论是变量选择还是变量替换，其目的都是为了降低数据的维度，然而在实际应用中，甄别各变量对总体的影响，仅从相关性学习的角度分析显然不够。例如大型宏观经济数据集中所研究的各个指标之间可能满足同步关系，也可能是超前或者滞后关系，在对这些非同步关系进行相关分析时可能体现出较小的相关性，这也是了解宏观经济走向不可或缺的重要指示器。

（二）选择合适的模型

通过降维和变量选择，使高维问题的维度得到了下降，此时还应考虑：采用传统建模方法进行建模是否恰当？能否再建立一套新的建模方法？从现有的发展来看，采用与经典方法不同的建模策略是比较好的选择。无论是惩罚似然估计还是高维因子模型的主成分估计，建模过程根据降维的需要都进行了改进。根据实际应用背景选择合适的模型，不仅是高维问题，也是所有的统计建模过程中需要面对的问题。

（三）改进模型的估计方法

传统模型的估计方法已经有了比较完整的理论体系。对于高维问题而言，现有估计方法是对一些既有方法的改进。例如惩罚似然、LASSO 等方法。在将来的研究中，有可能采用更加复杂的迭代方法，因选择好的算法对于高维问题显得尤为重要。在混合效应模型的变量选择中，一些相对较为复杂的方法需要解决的主要问题还是算法的实现与优化。当然，模拟结果还需要在实证研究中予以验证。

（四）估计和检验统计量的构建与实施

对于一些相对比较复杂的高维问题，如缺失数据，分类数据，分段数据等特殊类型的高维数据，估计和检验统计量的构造还应进一步探索。在追求无偏性、有效性、一致性和充分性的同时，研究稳健统计量是解决特殊类型数据问题的必要条件。合适的统计量应该是能够得出正确结论的统计量，而不仅仅是追求形式上和分布上的一致。神谕（oracle）性质是估计量所要具备的较好特征。

（五）大样本情况下的渐近性质

由于高维问题所研究的数据量往往比较大，而样本容量相对不多，故其渐近性质的讨论与传统的大样本性质分析有一定的区别。随着对高维问题研究的深入，一些不可观测的大样本问题逐渐出现，如重复构造的数据结构、采用再抽样（resampling）方法提取数据等等。这类问题引发的思考是：原始问题并非大样本，因模型转换和参数估计过程中产生的大样本问题，其渐近性质应如何考虑？

参考文献：

[1] Stock J H, Watson M W. Forecasting Using Principal Components from a Large Number of Predictors[J]. Journal of the American Statistical Association, 2002,97(460).

[2] Stock J H, Watson M W. Implications of Dynamic Factor Models for VAR Analysis[R]. NBER Working Paper, 2005.

[3] Pesaran M H, Chudik A. Econometric Analysis of High Dimensional VARs Featuring a Dominant Unit[R]. ECB Working

Paper,2010.
[4] Song S, Härdle, W, Ritov Y. Dynamic Factor Models for High Dimensional Nonstationary Time Series[R]. Forthcoming,2010.
[5] Forni M, Hallin M, Lippi M, Reichlin L. The Generalized Dynamic Factor Model: One－Sided Estimation and Forecasting[J]. Journal of the American Statistical Association, 2005,100(471).
[6] Deistler M, Zinner C. Modelling High－Dimensional Time Series by Generalized Linear Dynamic Factor Models: An Introductory Survey[J]. Communications in Information and Systems, 2007,7(2).
[7] Banerjee A, Marcellino M. Factor－Augmented Error Correction Models[C]// Castle J, Shepard N. The Methodology and Practice of Econometrics. Oxford: Oxford University Press,2008.
[8] Breitung J, Eickmeier S. Testing for Structural Breaks in Dynamic Factor Models[J]. Journal of Econometrics, 2011,163(1).
[9] Bai J, Ng S. Determining the Number of Factors in Approximate Factor Models[J]. Econometrica, 2002,70(1).
[10] Hallin M, Liska R. Determining the Number of Factors in the General Dynamic Factor Model[J]. Journal of the American Statistical Association, 2007,102(478).
[11] Harding M,Nair K K. Estimating the Number of Factors and Lags in High Dimensional Dynamic Factor Models[R]. Mimeo,2009.
[12] Fan J, Fan Y, Lv J. High Dimensional Covariance Matrix Estimation Using a Factor Model[J]. Journal of Econometrics,2008,147(1).
[13] Hautsch N, Kyj L M. Forecasting Vast Dimensional Covariances Using a Dynamic Multi－scale Realized Spectral Components Model[R]. Humboldt－Universit at zu Berlin, 2010.
[14] Bai J. Inferential Theory for Factor Models of Large Dimensions[J]. Econometrica, 2003,71(1).
[15] Pan J, Yao Q. Modelling Multiple Time Series Via Common Factors[J]. Biometrika, 2008,95(2).
[16] Lam C, Yao Q, Bathia N. Estimation of Latent Factors for High－Dimensional Time Series[J]. Biometrika 2011,98(4).
[17] Anselin L. A Test for Spatial Autocorrelation in Seemingly Unrelated Regressions[J]. Economics Letters, 1988,28(4).
[18] Brueckner, Jan K. Strategic Interaction Among Governments: An Overview of Empirical Studies[J]. International Regional Science Review, 2003,26(2).
[19] Anselin L. Bera A. Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics[C]// Ullah Amman, Giles David E A. Handbook of Applied Economic Statistics, New York: Marcel Dekker, 1998.
[20] Anselin L. Spatial Externalities, Spatial Multipliers and Spatial Econometrics[J]. International Regional Science Review, 2003,26(2).
[21] Kelejian Harry H, Robinson Dennis P. Spatial Correlation: A Suggested Alternative to the Autoregressive Model[C]// Anselin Luc, Florax Raymond J G M. New Directions in Spatial Econometrics, Berlin: Springer－Verlag, 1995.
[22] Davidian M, Giltinan D M. Nonlinear Models for Repeated Measurement Data[M]. London: Chapman and Hall, 1995.
[23] Chen Z, Dunson D. Random Effects Selection in Linear Mixed Models[J]. Biometrics, 2003,59(4).
[24] Vaida F, Blanchard S. Conditional Akaike Information for Mixed－Effects Models[J]. Biometrika, 2005,92(2).
[25] Dziak, John J, Li R. An Overview on Variable Selection for Longitudinal Data[C]// Hong D. Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques. World Scientific,2010.
[26] Fan J, Li R. Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties[J]. Journal of the American Statistical Association, 2001,96(456).
[27] Zou H. The Adaptive Lasso and Its Oracle Properties[J]. Journal of the American Statistical Association, 2006,101(476).
[28] Liang K Y, Zeger S L. Longitudinal Data Analysis Using Generalized Linear Models[J]. Biometrika, 1986,73(1).
[29] Fu W. Penalized Estimating Equations[J]. Biometrics, 2003,59(1).
[30] Bondell H D, Krishna A, Ghosh S K. Joint Variable Selection for Fixed and Random Effects in Linear Mixed－Effects Models[J]. Biometrics,2010,66(4).
[31] Ibrahim J G, Zhu H, Garcia R I, Guo R. Fixed and Random Effects Selection in Mixed Effects Models[J]. Biometrics, 2010,67(2).
[32] Ni X, Zhang D, Zhang H H. Variable Selection for Semiparametric Mixed Models in Longitudinal Studies[J]. Biometrics,2010,66(1).

【统计理论与方法】

非参数固定效应 Panel Data 模型的分位数回归推断

吕秀梅
(重庆工商大学 财政金融学院，重庆 400067)

摘要：利用分位数回归方法，讨论了非参数固定效应 Panel Data 模型的估计和检验问题，得到了参数估计的渐近正态性及收敛速度。同时，建立一个秩得分(rank score)统计量来检验模型的固定效应，并证明了这个统计量渐近服从标准正态分布。

关键词：分位数回归；渐近正态；固定效应；Panel Data 模型
中图分类号：O212.7 **文献标志码**：A **文章编号**：1007－3116(2012)06－0028－05

一、引 言

Panel Data 是指相同截面上的个体在不同时点重复观测的数据，基于 Panel Data 的回归模型称为 Panel Data 模型，Hsiao 等都对 Panel Data 模型作了详细的阐述[1－3]。近几年，Panel Data 模型的研究主要集中在非参数和半参数模型的估计和检验上。Lin 等使用平滑样条估计和核估计方法研究非参数 Panel Data 模型，推导出了样条估计量与核估计量的渐近偏差和协方差[4]；Li 和 Stengos 借助工具变量对半参数线性 Panel Data 模型进行估计，并且证明当 T 很小，N 很大时，估计量以 $\sqrt{N}$ 一致收敛[5]；Li 和 Hsiao 给出半参数 Panel Data 模型三个检验序列相关的统计量，并且证明这些统计量分别渐近服从正态分布或卡方分布[6]。与此同时，非参数与半参数 Panel Data 模型的广泛应用使它备受理论界和实务界的重视，得到了统计学家和经济计量学家在理论和应用上的深入研究，并且在经济学领域的应用逐渐被经济计量学家所推广。

现有的大多数文献都是使用最小二乘法或 Profile 似然法对 Panel Data 模型进行估计和检验，但是上述方法严重依赖于随机误差项方差的结构，而分位数回归对这一要求较弱，它只要求随机误差项的 $\tau\in[0,1]$ 分位数存在，因此本文采用分位数

收稿日期：2011－12－04
基金项目：教育部科学技术研究重点项目《非线性粘性 Boussinesq 系统的适应性与数值解研究》(109140)
作者简介：吕秀梅，女，四川德阳人，经济学博士，讲师，研究方向：金融计量，经济模型识别。

A Review of Dimensional Deduction and Variable Selection for High Dimensional Panel Data

ZHANG Bo[1], FANG Guo-bin[1,2]
(1. School of Statistics, Renmin University of China, Beijing 100872, China;
2. School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China)

Abstract: The aim of this paper is to review some important aspects in the study of high dimensianl panel data. Differential types and methods in the high dimensioanl panel data are disscused, literaures about random effect and marginal effect high dimensional varianc－covariance matrix in mixed model are reviewed. The advances of multi indicators large dimensional data factor model are sumarized. Some unresolved key issues, the future development in the theory and application are commented and previewed.

Key words: high dimensional; panel data; dimensional reduction; variable selection

(责任编辑：郭诗梦)