# 在纵向数据分析中使用滑动平均Cholesky分解模型对协方差矩阵建模

**摘要**

我们提出一种新的回归模型来对协方差结构进行参数化。采用新颖的Cholesky因子, 这种分解中的项由移动平均序列的系数和对数化的方差组成，并且可以由协变量的线性函数建模得到。我们为了基于这个分解的均值-方差分析提出有效的最大似然估计，并得出系数估计值的渐近分布。此外，我们研究了局部搜索算法, 从计算上的角度来说比传统的所有子集选择都更为高效,它基于BIC准则的模型选择, 并且显示了其模型选择的一致性。 从而, Pan & MacKenzie (2003)提出的一个关键假设得到了验证。我们通过 CD4数据的轨迹分析以及仿真分析演示了该方法的有限样本表现。

关键词: BIC，纵向数据分析，最大似然估计，模型选择，改进的Cholesky分解，移动平均

**研究背景介绍**

纵向数据,来自于对相同个体的重复测量,非常之常见。在均值分析中，一系列的回归模型中被(Diggle et al., 2002)所研究. 最近,协方差结构的回归分析, 旨在提供一个可以描述重复测量之间的依赖结构特征的简约模型,得到了很多的关注。Pourahmadi (1999, 2000)首次提出了改进的Cholesky分解:对协方差矩阵的逆进行分解，这种分解一个引人注意的性质是，它为正定协方差矩阵提供了无约束的参数化。更为重要的是,这个分解中的项可以被解释为时间序列中的自回归参数和对数化的方差。回归模型可以近似于均值模型的方式被应用到这些项上，从而简单地描述了协方差矩阵的结构。相关讨论参考Pan & MacKenzie (2003), Ye & Pan (2006), Pourahmadi (2007)and Leng et al. (2010)的研究。

在本文中,我们引入一种新的Cholesky分解来分析个体间差异，直接分解协方差矩阵而非它的逆。这个分解中的项分别对应移动平均参数和对数化的方差。 如此，协方差建模变成了一个时间序列分析问题，为此，移动平均模型可以提供一种可选的, 同样有力并且简单的表示。在这个分解中，我们提出了一种新的回归模型来做均值-方差分析, 并证明它的最大似然估计是渐近正常并且完全有效的。此外, 最终得到的均值, 移动平均系数以及对数化的系数是渐进独立的。 这一结果可以得到一种计算策略，降低传统的基于BIC准则的全模型选择的复杂度。我们严格确立这个模型选择策略的一致性。我们的结果可以被用来证明Pan & MacKenzie (2003) 的一个猜想，也就是说他们的模型选择算法具有一致性。 Rothman et al. (2010) 用建议的分解和正则化相结合去分析高维协方差矩阵 当这个矩阵中存在一个带状结构。我们的模型则更为普通一些。

模型及估计方法

我们用表示第i个个体在时刻上的的次观测值. 我们假设响应变量满足正态分布. 通过设定和是个体特定的,我们的方法可以处理那些不定期的观测以及极不平衡的数据集。

为了参数化, Pourahmadi在1999年首次提出了的分解方法。下三角阵的对角线上的项全为1，下三角元素则为， 是自回归模型中的系数，。的对角线上的项为。

令,同样对角线上的项全为1,我们可以写出。中的项可以解释为下述模型中的滑动平均系数

 (1)  
此处且for。参数和均不受约束。

因为我们的分解和Pourahmadi (1999)的不同之处在于分解还是它的逆, 这有助于我们利用常见的协方差矩阵去评估这两种分解方法。如果协方差矩阵是一个的复合对称矩阵，其中是一个全一矩阵,那么根据Pourahmadi (1999)的分解有,而我们的分解则可得出。如果是序列的协方差阵，那么根据Pourahmadi的分解得到，而我们的分解得出。

为了尽可能地在参数降维的基础上参数化均值-方差结构，我们加入一个回归模型



这个想法来自于Pourahmadi (1999, 2000)和Pan & MacKenzie (2003)的研究. 这里是一个单调可微的联系函数,而,和分别为,和维的协变量,协变量and用于回归分析,而通常作为时间间隔的多项式或是随时间变化的协变量。接下来我们可以将上述三个回归模型统称为滑动平均模型,而把Pourahmadi (1999, 2000)提出的模型叫做自回归模型。对于高斯数据,我们将作为恒等函数。写出二次负对数似然函数



其中。通过对的各个参数求偏导,它的极大似然估计等式可以写成



这里是连接函数的逆的衍生函数。并且我们使用记号，其中。

是一个的矩阵，第一行元素全为0,第j行 (j > 2) ;,,其中, 是由1组成的向量。参数和如通常在移动平均模型中一样被递归地定义。

由于,和的解满足方程(2), 这些参数可以通过其他参数迭代地求解.(2) 的quasi-Fisher得分算法可以直接产生这些参数的数值解。Hessian的期望值细节 将在补充材料中讨论. 更具体地,该算法的运作如下:

1. 初始化各参数为,和.令=0.
2. 用和计算.用下式更新

 （3）

1. 给定和, 用下式更新

 （4）

这里



是一个的矩阵，而是矩阵的行列元素。的矩阵=0。

1. 给定和, 用下式更新

 （5）

1. 设并且重复步骤2-5直到预先规定的收敛准则被满足。

此算法收敛到一个局部最优解，这严格地取决于初始值。的一个天然的起始值可以是与(3)中的同维的单位矩阵.接着我们初始化(4)中的，假设.

不难看出，这些初步估计是√n-一致的. 根据Section 3中的Theorem 1和补充材料中的的理论分析, 负对数似然函数在真实参数的小邻域内渐近凸。这确保了通过这种迭代算法得到的最终估计,用，和表示,渐进于全局最优解，比初始值更有效。在我们的数据分析和模拟研究中，通常在十次迭代中得到收敛。

渐近性质与模型选择

渐进性质

由于我们使用的是最大似然估计，因此得到的估计是有效的。为了正式地确立理论性质，我们施加了以下正则条件。

条件A1:协变量,和的维度，,是固定的;以及是有界的。

条件A2:真值在参数空间的内部，是的一个紧子集。

条件A3: 当, 收敛到正定矩阵。

条件A1 and A2在纵向数据分析中是标准的.最大似然方程的渐进性质 包括期望Hessian矩阵with的负,期望在协变量,和上是有条件的. 条件A3在回归分析中是标准的。我们可以正式地给极大似然估计得出以下渐进结果。

定理1. 若并且正则条件A1–A3成立,那么: (a)极大似然估计量对真值具有强一致性;(b) 如下渐进分布:



此处是块对角矩阵，其,and.

由此可得和都是渐进独立的。根据补充材料的(2)式,的块对角元素满足



由于和是的一致估计，渐近协方差矩阵里的是块对角矩阵的块组成部分的一致估计



模型选择

为均值-方差结构选择最优的模型，一个标准的方法是使用贝叶斯信息准则，BIC。我们讨论一种计算效率高的算法，给出了一致的模型。

为了便于表述，我们使用通用符号来表示任意候选模型，这里，包括作为相关预测值去给均值建模；和有类似定义。真模型用表示，例如，是由所有非零系数组成的向量。我们定义过度拟合模型的族为，而未拟合模型为。因此，对于任意我们可以得到，而对于任意我们可以得到，等等。用表示模型的大小，也就是，接下来，定义类似地去定义和根据这些记号，我们定义

 （7）

这里的是和模型的极大似然估计。Shao (1997)和Shi以及Tsai (2002)证明如果有限维的真模型存在并且预测值维度是固定的，（7）能够一致地识别真模型。 为了用它进行模型选择，我们运用所有子集选择，通过拟合个模型并且选择那个BIC值最小的模型。

受到定理1中和渐进独立性质的启发，我们在这里研究一个计算算法，可以大大降低所有子集选择的复杂度与Pan和Mackenzie (2003)类似，我们提出了寻求最优模型的搜索策略，利用下述搜索，包括通过saturating成对参数集得到的似然估计



这里分别代表均值，移动平均，对数更新的全部模型。因此，我们只需要应用所有子集选择为一组中特定的系数，通过使用对于其他两组系数的全模型。这个策略要求我们比对个模型的BIC值，在计算上比比较全部子集选择要少得多。该算法的选型一致性成立于下面的定理。

定理2：如果定理1中的条件成立，我们得到



Pan & MacKenzie (2003) 在Pourahmadi中选择多项顺序时使用了相似的算法，并给出了其成功的验证证据。他们推测他们的算法是选型一致的。由于在他们的模型中最大似然估计的渐近独立性，他们的猜想直接跟从定理2的证明。