arXiv:2303.02114v1 [math.ST] 3 Mar 2023

通过凸规划选择多个自回归过程的滞后阶数并估计稳定参数

作者：Somnath Chakraborty、Johannes Lederer 和 Rainer von Sachs

[1]波鸿鲁尔大学数学系，44801 波鸿，德国，[a]somnath.chakraborty@ruhr-uni-bochum.de；[b]johannes.lederer@ruhr-uni-bochum.de

[2]鲁汶大学 LIDAM 统计学、生物统计学与精算科学研究所，1348 新鲁汶，比利时，[c]rainer.vonsachs@uclouvain.be

受多种应用的推动，高维时间序列已成为一个活跃的研究领域。特别是，针对滞后阶数已知的单个稳定自回归过程，近年来已经提出了多种方法及有限样本理论。相反，我们研究共享未知滞后阶数的多个稳定自回归过程。我们利用不同过程之间的信息，同时选择滞后阶数并估计参数。我们证明，所估计的过程是稳定的，并建立了预测误差的收敛速率；在我们的设定下，该速率可能优于已知速率。我们关于滞后阶数选择和稳定性的研究结果，对于单个自回归过程的情形同样具有意义。

1. 引言。如今，宏观经济学、金融、日常天气预测、脑成像等众多领域都在获取复杂数据，因此需要采用智能的建模方法，避免相对于可用样本量而言使用过多的模型参数。此外，这些数据通常具有高维特征——因为它们以面板形式或多元向量的形式共同出现。这些典型事实使得预测这些数据在（不久的）未来如何演变极具挑战性。要应对这一挑战，选择一种假设观测到的多元数据集中不同组成部分之间存在某种共同底层结构的数据生成模型，不仅会带来优势，也体现了一个观察结果：不同序列并非彼此独立——它们实际上可能由某种潜在（即不可观测的）机制驱动（例如领先经济指标或全球气候趋势等，通常以潜在因子模型来刻画）。此外，我们几乎总能观察到当前观测与过去观测之间存在序列相关；传统上，人们通过假设时间上存在某种弱依赖来对其建模（由此形成动态潜在因子模型，例如 Forni 等（2000））。

在此背景下，由于因子建模不一定能够实现逐分量预测，（参数）向量自回归（VAR）方法长期以来一直是对这类多元时间序列进行建模的一个极为重要的工具——其中特别体现为这样一种思想：共同的序列依赖性受到所有分量共有的最大滞后阶数存在这一条件的限制。然而，随着分量序列数量的增加，VAR 模型众所周知会趋于参数过多。由于这是一个高维参数估计问题，近年来应对这一问题的可能方法包括正则化方法，例如使用 LASSO 估计这些模型的参数（本质上采用某种正则化最小二乘法，参见例如 Nardi 和 Rinaldo（2011））。这与更为传统的方法形成对比（基于

MSC2020 学科分类：主要分类 62M10、62L99；次要分类 49M29。关键词和短语：正则化最小二乘、LASSO、自回归过程、层次组范数、对偶范数、稳定性、样本复杂度。

主要依赖用于滞后阶数选择的信息准则（例如 AIC、BIC 等），通过选择较低的滞后阶数来解决过参数化问题，其依据是假设存在短程依赖，并假设所有分量都适用同一个滞后阶数。在高维情境下，为了获得良好的预测表现，这些方法后来被 LASSO 超越——不过，直到最近，LASSO 仍未纳入滞后阶数选择的概念。直到 Nicholson 等人（2020）近期提出一类分层滞后结构，将滞后阶数选择的概念嵌入凸正则化项中。其关键建模工具是采用嵌套组的分组 LASSO，以确保滞后系数的稀疏模式符合 VAR 的有序结构。关于解决 VAR 过参数化问题的降维方法的文献详情，请参阅 Nicholson 等人（2020）上述研究的第2节。然而，这种方法明显的缺点是必须将观测到的多元时间序列的所有分量建模为具有相同的数据长度；这是经典 VAR 建模中的一项限制，无法规避。

受 Nicholson 等人（2020）所用方法的启发，本文在正则化 LASSO 框架下，提出一种分析多个（可能长度不同的）稳定自回归过程的方法，其中通过重叠分组范数实现正则化，从而在组层面诱导稀疏性。此外，我们证明，即使对这些过程的最大滞后阶数一无所知，所提出的框架仍能估计真实滞后阶数和 AR 模型系数。最后，我们证明，使用该方法所得 AR 系数拟合的模型是稳定的。由于我们关于统计保证的结果本质上具有非渐近性质——即使在只观测单条时间序列的情形下也颇具意义——我们首先回顾时间序列领域中关于这类非渐近结果的（稀疏）文献，然后再详细探讨我们的方法与 Nicholson 等人（2020）的方法之间的异同。

直到最近，时间序列分析领域的大多数研究都专注于推导预测变量的渐近行为。这严重限制了这些结果的适用性，尤其是在样本量与预测变量之比偏低的情况下。为克服这种依赖性带来的难题，常见做法是假设数据生成过程具有稳定性（从而满足平稳性）。例如，Negahban 和 Wainwright（2011）以及 Loh 和 Wainwright（2012）在推导小样本情形下的保证时都采用了稳定性假设；然而，他们的结果建立在一个条件之上：系数矩阵的范数受到严格约束（即系数矩阵的算子范数之和小于 1）。这一条件远强于由这些系数决定的过程稳定性。最近，Basu 和 Michailidis（2015）在理解小样本情形下时间依赖性和截面依赖性的影响方面取得了重大进展。该研究的基本假设是，数据适合用稳定的向量自回归过程建模；作者追踪了由过程稳定性所施加的谱域限制，并推导出了适用于具有高斯白噪声创新的高维向量自回归过程的非渐近预测保证。随后，多项后续研究（例如 Wong 等（2020）、Masini 等（2022）及其中的参考文献）将他们的结果扩展到创新项具有重尾分布的情形；此外，这些研究仅假设底层过程满足平稳性且二阶矩有限——这些条件比稳定性更弱。

在本引言余下部分，我们将进一步详细讨论我们的方法与 Nicholson 等人（2020）的方法之间的关系。从本质上说，后者的贡献在于发现，在存在关于真实未知滞后阶数上界 L 的先验信息时，Mairal 等人（2011）提出的方法（以及 Tseng（2009）对其所作的计算速度更快的改进）在算法层面上可用于此类情形，以估计维度为 M、阶数为L的向量自回归模型，以下简称 $\text{VAR}_M(L)$。我们的

工作现在可以看作是将一个共同的（即“对角向量”）自回归模型拟合到维度为 M 的多元时间序列上，也就是由 M 个观测到的（单变量）时间序列组成的面板，这些序列的长度通常并不相同 n_m，$1 \le m \le M$。我们要解决一个颇具挑战性的问题：如何仅凭观测数据模型中可用的信息，选择一个恰当的共同滞后阶数 L，从而能够通过惩罚型 LASSO 方法表述并求解我们的问题。我们推导了多元一步预测误差的非渐近界，并在稀疏性范式下估计自回归系数的集合 $\beta := (\beta_1^m, \ldots, \beta_L^m)_{1 \le m \le M}$。假设存在一个共同的真实未知滞后阶数 L_0，它生成了我们的 M 个时间序列，那么我们的算法与 Nicholson 等人（2020）（以及 Mairal 等人（2011））用于拟合共同模型的方法类似，基于对分层组 LASSO 方法的改进：我们首先根据对分组 LASSO 方法理论复杂度的深入分析，确定一个恰当的（最小）上界 $L \ge L_0$。该上界主要取决于最短观测时间序列分量的样本量 $n_{\min}$（当然也取决于 M）。有了这个恰当的 L，便可将我们的自回归问题纳入高维多元回归框架，并将现有的 LASSO 估计技术（采用重叠的、按层级构造的组）ˆ应用于该问题。我们为估计量 β（即上述学习算法的解）以及估计量 L_0（本质上取自 ˆβ 的支撑集）推导了统计保证。具体而言，我们给出了多元一步预测误差、 β 的估计误差、 β 支撑集的错误发现，以及拟合模型稳定性的非渐近界。对于后者，我们证明，阶数为 L_0 、系数估计值为 β 的拟合自回归模型满足真实模型的稳定性条件（通过我们提出的一个更明确的 ε-稳定性概念来衡量统计估计问题的难度）。

在下文中，我们将更明确地阐述本文贡献的确切性质，这些贡献源于我们在文献中发现的现有方法的局限性。

现有局限与我们的贡献。近期已有的非渐近时间序列分析研究未能充分回答以下一些问题。

L1 Nicholson 等人（2020）指出，基于带重叠组的 LASSO 的方法（如 Mairal 等人（2011）提出的方法）能够确定稳定的高维向量自回归过程的分量滞后阶数 L_m（包括 VAR 模型中第 $i-$ 个分量对第 $j-$ 个分量的依赖所对应的“交叉”滞后阶数 L_{ij}）。他们基于（对偶）凸规划的公式在计算上颇具吸引力，总体而言也很有意思，但无论在理论界限还是实际应用中，该方法都需要对 $L_{i,\ j}$ 设定一致上界。现有文献要么完全忽视这一问题，要么根据 AIC 或 BIC 等模型选择方法，或通过贝叶斯收缩来设定这些界限，而这些方法在此处是否适用尚不明确。

C1 我们沿用 Nicholson 等人（2020年）的方法，但为滞后阶数建立了一个合适的上界。我们对这一最小上界的选择保证了以下几点：

1. 经验预测风险较小（见定理 4.3 及定理之后的讨论）；2. 得出的调节参数不会过大（见推论 4.5）；3. 数据矩阵满足受限特征值性质（见定理 4.6 及其后的推论 4.7 ）。

从这个意义上说，我们针对当前设定得到的滞后阶数上界是最优的。

L2 现实应用通常涉及多个相互独立的单变量时间序列。这对应于采用对角系数矩阵进行向量自回归建模。然而，单纯

将现有的 VAR$_M$（L）建模方法迁移到实践中可能会很繁琐，尤其是在各个分量时间序列的可用样本数量不同时：显然，为了统一使用最小的单个样本量而截去部分样本，可能会导致理论保证和实际表现变差（见下文）。

C2 应用层次组范数使我们能够推导出预测保证，其依据是每个分量的平均样本数（而不是每个模型的最小样本数；若直接将其转换为 M 维 VAR 模型，情况便会如此）。更具体地说，若能准确获知真实滞后阶数 L_0 and$n_1 = L_0 + T_1$，$\cdots$，$n_m = L_0 + T_M$（相应的）各个分量的样本数 M，则可得出以下结果：转换为 VAR$_M$（L）模型将产生以下可证明的误差（见 Nicholson 等人（2020））；该结果表明，以高概率

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O\left(\sqrt{\frac{\log(M^2 L_0)}{n_{\min} M}}\right), \tag{1}$$

而我们所述算法产生的误差则为以下量级（见下文公式（33））

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O\left(\sqrt{\frac{\log(ML)}{D}}\right). \tag{2}$$

此处 $D = T_1 + \cdots + T_M$ 是“后样本”的总数（$T_m := n_m - L_0$）。此外，与相关的 VAR$_M$（L）模型转换不同，我们的策略使误差对反向特征多项式在单位圆盘上的行为的依赖程度更弱（可与 Basu 和 Michailidis（2015）比较）。

L3 回顾一下，若（单变量）自回归过程 $X_t = a_1 X_{t-1} + \cdots + a_L X_{t-L} + U_t$ 的“反向特征多项式” $1 - a_1 z - \cdots - a_L z^L$ 在闭单位圆盘上没有复根，则该过程是稳定的；众所周知，稳定性意味着平稳性（见下文第2节）。但拟合得到的自回归模型是否稳定呢？对于单变量自回归过程系数的 Yule-Walker估计，这个问题有肯定答案（不过这种估计效率较低）；然而，文献似乎尚未探讨由基于LASSO的方法估计出的参数重构而成的过程是否稳定或平稳。不过，既然这些算法的输入观测数据是由稳定过程生成的，那么有理由期待重构出的过程也是稳定的（因此，多步预测也同样可靠）。

C3 我们在定理 4.10 中证明：当可用作输入的样本由稳定过程生成时，根据我们的算法返回的参数重构出的过程是稳定的。正如 Basu 和 Michailidis（2015）所示，自回归过程的一种稳定性度量等价于对谱密度有界性的判定，而傅里叶域中的有界性在某种意义上对应于时间域中过程的“平滑性”。因此，直观而言，稳定自回归过程构成一个“平滑”的子类；任何旨在从这个平滑子类中学习过程参数的算法，都应当返回属于该子类的估计量。本文通过定理 4.10 保证了这一点。

需要在此指出，本文的结果表明，当基础过程稳定时，我们的重叠组 LASSO 方法能够产生一个稳定过程。本文的目标并不是选择最优的调节参数或某些绝对常数，而是证明这些参数/常数的存在，并为其提出合理的候选值。

本文结构。本文组织如下。第2节回顾现有文献中的相关定义，并设定记号。在第3节中，（1）我们明确模型，并将学习问题表述为一个带有重叠组范数的正则化组 LASSO 问题，其中数据矩阵是一个块对角矩阵——每个块由处理与相应分量时间序列对应的最小二乘问题的数据矩阵构成；（2）我们介绍基于组 LASSO 问题的学习算法。第4节包含主要的技术内容，尤其是对主要结果各项陈述的证明；这些主要结果已在第3节末尾给出。第4节分为四个小节：第4.1小节给出一个界定一步预测误差的预言机不等式（定理 4.2），以及模型有效噪声的高概率界（定理 4.3）。第4.2小节首先给出数据矩阵各块的受限特征值界（命题 4.6），随后整合分块结果，最终得到估计自回归系数误差的估计值（定理 4.9）。最后，结合这些小节中的结果，在第4.3小节中建立了估计的 AR 模型的稳定性（定理 4.10）。所有辅助结果的证明均推迟至一系列附录中。

2. 预备知识。

2.1. 一般记号。

• M_d（$\mathbb{F}$）表示元素属于域 $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$的 $d \times d$ 矩阵构成的环，对于$M \in M_d$（$\mathbb{F}$），我们用 $M^\top$ 表示转置；如果 $\mathbb{F} = \mathbb{C}$，则 $M^\star$表示共轭转置。

• $\mathbb{D}$ 是复闭单位圆盘 $\mathbb{D} := \{z \in \mathbb{C} : |z| \le 1\}$，其边界为 $\partial\mathbb{D} := \{z \in \mathbb{C} : |z| = 1\}$。

• 对于任意整数 $d > 0$，若 $x \in \mathbb{R}^d$，则 $\|x\|_2 = \sqrt{x^\top x}$，且 $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$；特别地，在标准同一化 $\mathbb{C} = \mathbb{R}^2$下，我们有 $\mathbb{S}^1 = \partial\mathbb{D}$。

• 对于整数 $n > 0$，我们将集合 $\{1, 2, \cdots, n\}$ 记作 $[n]$。

• 对于向量 $\boldsymbol{\beta}_m = (\boldsymbol{\beta}_{m,1}, \cdots, \boldsymbol{\beta}_{m,L})$ 和整数 $0 < L_0 \le L$，我们记

$$\hat{\boldsymbol{\beta}}_m(L_0) := (\hat{\boldsymbol{\beta}}_{m,1}, \cdots, \hat{\boldsymbol{\beta}}_{m,L_0}) . \tag{3}$$

2.2. 自回归过程的记号。

约定：我们用 $X, Y, Z, \ldots$ 表示随机变量，用 $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \ldots$ 表示随机向量。另一方面，我们用 $a, b, c, \ldots$ 表示实数（或复数）常数，用 $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \ldots$ 表示向量值常数。

记号：对于 d 维自回归过程 •

$$X_t := A_1 X_{t-1} + \cdots + A_L X_{t-L} + U_t , \tag{4}$$

以及 $z \in \mathbb{C}$，我们记作

$$\mathcal{A}_z := I - A_1 z - \cdots - A_L z^L . \tag{5}$$

• L 表示对式（4）中的过程预先确定的真实滞后阶数的“临时”上界，而 L_0 表示真实滞后阶数。

定义 2.1（弱平稳性）。若一个 d 维时间序列 $\{X_t\}_{t\in\mathbb{Z}}$ 满足以下条件，则称其为弱平稳：a） $\mathbb{E}[\|X_t\|_2^2] < \infty$ 对所有 $t \in \mathbb{Z}$, b） $\mathbb{E}[X_t] = \mu$ 均成立，对所有$t \in \mathbb{Z}$均成立，并且 c） $\mathbb{E}[X_t X_{t-h}^\top] = \Gamma(h)$ 对所有 t， $h \in \mathbb{Z}$均成立。

定义 2.2（强平稳性）。若一个 d 维时间序列 $\{X_t\}_{t\in\mathbb{Z}}$ 满足：对每个整数 $n>0$，以及所有整数 t_1，$\cdots$，t_n，h，向量 $(X_{t_1}$，$\cdots$，$X_{t_n})$ 和 $(X_{t_1+h}$，$\cdots$，$X_{t_n+h})$ 的分布相同，则称其为强平稳。

定义 2.3（自回归时间序列）。若存在 $d\times d$ 个矩阵 $A_1,\cdots,A_L$，使得 d 维时间序列 $\{X_t\}_{t\in\mathbb{Z}}$ 的滞后阶数至多为 $L>0$，则称其为自回归时间序列

$$X_t = A_1X_{t-1}+\cdots+A_LX_{t-L}+U_t\,. \tag{6}$$

对于某个随机白噪声过程 $\{U_t\}_{t\in\mathbb{Z}}$，该关系对所有 $t\in\mathbb{Z}$ 均成立。

与每个 d 维滞后阶数为 L 的自回归过程 $\{X_t\}_{t\in\mathbb{Z}}$（如方程 (6) 所示）相关联的，是相应的一阶过程 $\mathbf{X}_t=\mathbf{A}\mathbf{X}_{t-1}+\mathbf{U}_t$，其中

$$\mathbf{A}:=\begin{pmatrix}A_{1\to L},A_L\\ \mathbf{I}_{dL-d},\mathbf{0}\end{pmatrix},\qquad \mathbf{U}_t:=\begin{pmatrix}U_t\\ \mathbf{0}\end{pmatrix}, \tag{7}$$

且 $A_{1\to L}$ 是分块矩阵 $(A_1\cdots A_{L-1})$。

定义 2.4（稳定性）。若 d 维、滞后阶数为 L 的自回归过程 $\{X_t\}_{t\in\mathbb{Z}}$（如方程 6 所示）满足：对于 $|z|\le 1$，$\det(\mathbf{I}\ -\mathbf{A}z)\ \neq 0$，则称该过程稳定。等价地，若对于 $|z|\le 1$，$\det(\mathcal{A}_z)\ \neq 0$，则该过程稳定。

定义 2.5（反向特征多项式）。多项式 $\det(\mathcal{A}_z)$ 称为方程 6 中过程的反向特征多项式。

注意等式 $\det(\mathbf{I}\ -\mathbf{A}z)\ =\det(\mathcal{A}_z)$。特别地，当且仅当$\mathbf{A}$ 的每个特征值都位于开单位圆盘内时，方程 (6) 中的过程才是稳定的。

定义 2.6（ϵ-稳定性）。如果方程 (6) 中的稳定自回归过程满足以下条件，则称其为 ϵ-稳定的，其中 $\epsilon\in(0,1)$：

$$\epsilon\le\min_{|z|=1}|\det(\mathcal{A}_{\mathbf{z}})|\le\max_{|z|=1}|\det(\mathcal{A}_{\mathbf{z}})|\le\epsilon^{-1}\,. \tag{8}$$

备注。根据最大模原理，这等价于说

$$\epsilon\le\min_{|z|\le 1}|\det(\mathcal{A}_{\mathbf{z}})|\le\max_{|z|\le 1}|\det(\mathcal{A}_{\mathbf{z}})|\le\epsilon^{-1}\,.$$

这里的下界是对稳定性概念的一种便捷量化，而稳定性要求$\min_{|z|\le 1}|\det(\mathcal{A}_{\mathbf{z}})|>0$。请注意，我们还需要上界来推导统计保证。

关于自回归过程的一个众所周知的事实如下：详情请参见 Lütkepohl（2005年，命题 2.1）。

引理 2.7（稳定性蕴含弱平稳性）。稳定的自回归过程是弱平稳的。

3. 统计模型与估计量。本节介绍我们的统计模型和估计量，并提出一种算法，用于根据观测样本学习模型参数。

3.1. 统计模型。我们先介绍模型。假设我们观测到由 M 单变量自回归过程生成的时间样本，并且已知真实滞后阶数的统一上界为 L 。那么，我们可以将这些单变量滞后阶数至多为 M 的自回归过程

$$\begin{aligned} X_t^1 &= \beta_1^1 X_{t-1}^1 + \cdots + \beta_L^1 X_{t-L}^1 + U_t^1 ; \\ &\vdots \\ X_t^M &= \beta_1^M X_{t-1}^M + \cdots + \beta_L^M X_{t-L}^M + U_t^M . \end{aligned} \tag{9}$$

本文采用简化假设：所有 M 个分量过程的真实滞后阶数相同，即为 L_0，且 $L \geq L_0$ 是一般的； L_0 和 L 均为先验未知。此外，我们假设创新项均值为零且服从高斯白噪声；也就是说，对每个 $m \in [M]$，集合 $\{U_t^m\}_{t\in\mathbb{Z}}$ 由相互独立、均值为零的单变量高斯随机变量组成，其各坐标的标准差为 $\sigma_m \in (0,\infty)$。此外，我们假设对每个 $t \in \mathbb{Z}$，噪声变量 $U_t^1, \ldots, U_t^M$ 相互独立。我们通过 $\Theta_{ml} := \beta_l^m$ 将模型参数汇总到矩阵 $\Theta \in \mathbb{R}^{M\times L}$ 中，以便后文更轻松地指代各组参数。我们的目标是：1. 同时估计所有模型的参数；2. 评估滞后阶数 L_0，并假设它在所有 M 个过程间相同。

我们对最小的 β-系数的绝对值作如下假设，这在 LASSO 文献中通常称为 β-min 假设；例如，参见 Bunea（2008年）。我们指出，只有在进行 λ-阈值化后，为了得到定理 3.1 中的界，才需要这一假设；具体而言，不采用阈值化时，本文的分析不需要该假设。

A假设 1（β-min 假设）。存在一个绝对常数 $c_\beta > 0$，使得真实的自回归系数向量 $\boldsymbol{\beta}$ 满足

$$\boldsymbol{\beta}_j^m \neq 0 \;\Rightarrow\; \boldsymbol{\beta}_j^m \geq c_\beta . \tag{10}$$

宽泛地说，这一假设确保了非零系数能够首先被检测出来。现在我们来定义一个正则化项。令 n_j 表示来自

$$X_t^j = \beta_1^j X_{t-1}^j + \cdots + \beta_L^j X_{t-L}^j + U_t^j ,$$

和 $T_j := n_j - L$ 的样本总数。其主要思路如下。设 $L \geq L_0$ 是某个整数，并且对每个 $t \in \{-L+1, \ldots, 1, \ldots, T_m\}$ 和 $m \in \{1, \ldots, M\}$，我们都有一个关于 X_t^m 的观测值 x_t^m。对每个 $m \in [M]$，记 $\boldsymbol{\beta}_m := (\beta_1^m, \ldots..., \beta_L^m)^\top$，并令 $\boldsymbol{\beta} := (\boldsymbol{\beta}_1^\top, \ldots..., \boldsymbol{\beta}_m^\top)^\top$。我们定义$\mathcal{G}_1$，....., $\mathcal{G}_L \subset S_{ML} := \{1,2,\ldots,M\} \times \{1,2,\ldots,L\}$ 为

$$\mathcal{G}_l := \{1,2,\ldots,M\} \times \big\{l, l+1 \ldots, L\big\} \tag{11}$$

对所有 $l \in \{1,\ldots,L\}$ 均成立。各组是嵌套的： $\mathcal{G}_1 \supset \cdots \supset \mathcal{G}_L$。设 $\boldsymbol{\beta}_{\mathcal{G}_l} \in \mathbb{R}^{M\times(L-l+1)}$ 为 $\boldsymbol{\Theta}$ 的子矩阵，由索引大于或等于 l 的列组成。我们将组范数定义为

$$\|\boldsymbol{\beta}\|_{\mathcal{G}} := \sum_{l=1}^{L} \sqrt{M(L-l+1)}\, \|\boldsymbol{\beta}_{\mathcal{G}_l}\|_{\mathbb{F}} , \tag{12}$$

$$\text{where} \qquad \|\boldsymbol{\beta}_{\mathcal{G}_l}\|_{\mathbb{F}} := \sqrt{\sum_{m=1}^{M}\sum_{j=l}^{L} |\beta_j^m|^2} .$$

是 $\boldsymbol{\beta}_{\mathcal{G}_l}$ 的弗罗贝尼乌斯范数。为便于记号，我们也将组范数 $\|\boldsymbol{\beta}\|_{\mathcal{G}}$ 写作 $\mathcal{N}(\boldsymbol{\beta})$。

总体样本外样本量记为

$$D := T_1 + \cdots + T_M\,.$$

为了估计系数向量 $\boldsymbol{\beta}$，我们提出求解以下约束凸规划

$$\text{(13)}\qquad \text{最小化}\qquad \frac{1}{D}\sum_{m=1}^{M}\sum_{t=1}^{T_m}\bigl(x_t^m - \beta_1^m x_{t-1}^m - \cdots - \beta_L^m x_{t-L}^m\bigr)^2 + \lambda\|\boldsymbol{\beta}\|_{\mathcal{G}}\,,$$

并使用适当的调节参数 $\lambda > 0$。
目标函数可以写成简洁的形式。为此，我们定义向量 $\boldsymbol{y} \in \mathbb{R}^D$、矩阵 $X \in \mathbb{R}^{D\times(ML)}$ 和参数 $\boldsymbol{\beta} \in \mathbb{R}^{ML}$，如下：

$$\boldsymbol{y} := (x_1^1, \ldots, x_{T_1}^1, x_1^2, \ldots, x_{T_2}^2, \cdots, x_1^M, \ldots, x_{T_M}^M)^\top;$$

$$\text{(14)}\qquad X := \begin{pmatrix} x_{1-1}^1, \ldots, x_{1-L}^1 & & & & \\ \vdots & & & & \\ x_{T_1-1}^1, \ldots, x_{T_1-L}^1 & & & & \\ & x_{1-1}^2, \ldots, x_{1-L}^2 & & & \\ & \vdots & & & \\ & x_{T_2-1}^2, \ldots, x_{T_2-L}^2 & & & \\ & & \ddots & & \\ & & & x_{1-1}^M, \ldots, x_{1-L}^M & \\ & & & \vdots & \\ & & & x_{T_M-1}^M, \ldots, x_{T_M-L}^M \end{pmatrix};$$

$$\boldsymbol{\beta} := (\beta_1^1, \ldots, \beta_L^1, \beta_1^2, \ldots, \beta_L^2, \cdots, \cdots, \beta_1^M, \ldots, \beta_L^M)^\top.$$

显然，

$$\sum_{m=1}^{M}\sum_{t=1}^{T_m}\bigl(x_t^m - \beta_1^m x_{t-1}^m - \cdots - \beta_L^m x_{t-L}^m\bigr)^2 \;=\; \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2\,.$$

总之，上述估计程序等价于

$$\text{(15)}\qquad \widehat{\boldsymbol{\beta}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{ML}}\left\{\frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{\mathcal{G}}\right\}.$$

因此，该估计量可以表示为一种改进的组套索估计量，这意味着我们可以使用成熟的组套索算法来处理重叠组（Mairal 等人，2011年）。本质上，上述估计量将 Nicholson 等人（2020）中的逐元素估计量 HLag$^{\text{E}}$ 推广到了多个时间序列。请注意，普通的 LASSO 估计量，以及任何采用非重叠组的组 LASSO 估计量，都会通过将系数置零来实现稀疏性，却不考虑这样一个事实：当回归的输入仅包含真实滞后阶数 L_0 的上界时，索引介于 L_0+1 和 L之间的所有系数都应在索引小于或等于 L_0的系数之前置零；然而，通过链式分组 $\mathcal{G}_1 \supseteq \cdots \supseteq \mathcal{G}_L$ 得到的惩罚项恰好实现了这一点。

示例（正则化项）。我们考虑两个单变量滞后阶数至多为三的自回归过程；也就是说， $M=2$ 和 $L=3$。相应的组如下：

$$\begin{aligned}\mathcal{G}_1 &= \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}\,;\\ \mathcal{G}_2 &= \{(1,2),(1,3),(2,2),(2,3)\}\,;\\ \mathcal{G}_3 &= \{(1,3),(2,3)\}\,.\end{aligned}$$

因此， $\boldsymbol{\beta}\in\mathbb{R}^{2\times 3}$ 的组范数为

$$\begin{aligned}\|\boldsymbol{\beta}\|_{\mathcal{G}} = \sqrt{6}\sqrt{(\beta_1^1)^2+(\beta_1^2)^2+(\beta_2^1)^2+(\beta_2^2)^2+(\beta_3^1)^2+(\beta_3^2)^2}\\ +\ \sqrt{4}\sqrt{(\beta_2^1)^2+(\beta_2^2)^2+(\beta_3^1)^2+(\beta_3^2)^2}\\ +\ \sqrt{2}\sqrt{(\beta_3^1)^2+(\beta_3^2)^2}\,.\end{aligned}$$

请注意，当（正则化）LASSO 将某个组（例如上文中的第二组）设为$\mathbf{0}$时，它会自动将后续所有组也设为$\mathbf{0}$ 。更具体地说，当（正则化）LASSO 将 $\boldsymbol{\beta}_{\mathcal{G}_l}\mathbf{0}$设为零时，层次结构 $\mathcal{G}_l\supseteq\mathcal{G}_{l+1}\supseteq\cdots$意味着，对于所有 $r>0$， $\boldsymbol{\beta}_{\mathcal{G}_{l+r}}$的每个坐标都是 $\boldsymbol{\beta}_{\mathcal{G}_l}$的一个坐标，从而确保每个 $r\geq 0$都满足 $\boldsymbol{\beta}_{\mathcal{G}_{l+r}}=\mathbf{0}$。

下文中，我们使用以下记号。对于 $m\in[M]$ 和 $l\in[L]$，以及 $t\leq T_m$，我们记为

$$\begin{aligned}\boldsymbol{U}^{(m)} &:= (U_1^m,\cdots,U_{T_m}^m)^\top;\\ X^{(m,l)} &:= (X_{1-l}^m,\ldots,X_{T_m-l}^m)^\top;\\ X_t^{(m)} &:= (X_{t-1}^m,\ldots,X_{t-L}^m);\\ X^{(m)} &:= (X^{(m,1)},\ldots,X^{(m,L)})\,.\end{aligned}\tag{16}$$

此外，我们用 $X_{<j}^{(m)}$ 表示变量 $X_{j'}^m$ 中的任意一个，其中 $j'<j$。

3.2. 估计流程。学习多个长度可能不同且真实滞后阶数相同的时间序列的自回归系数，比通常的组 LASSO 问题更为复杂；在通常的组 LASSO 问题中，各组构成指标集的一个划分。从方法论角度看，当前面临的一些直接技术挑战包括：1. 在凸优化问题 (13) 的表述中应采用什么样的 L ； 2. 如何分解组范数的对偶范数（以便应用 Hölder 不等式，推导出如下文第 4.1 小节中的预言机预测保证）。从实践角度看，挑战在于如何将样本数量的差异纳入凸优化问题。

现在，我们概述一下估计流程（将在下文详述）：该流程以多个时间序列为输入，构造形式如式（15）的适当凸优化问题，并按照 Nicholson 等人（2020）的方法，通过随机近端梯度法求解该凸优化问题。本质上，我们首先检查数据集，找出样本数最少的分量所对应的样本（若有并列，则任意选择）。然后，我们利用该分量，按照式（38）的描述，确定初始输入滞后 L。接下来，我们求解式（13）中的正则化最小二乘问题，其中惩罚函数为组范数 $\mathcal{N}$，下文将进一步讨论（见式（12））。如 Nicholson 等人（2020）所述，该过程的收敛速率相对于计算步骤数呈二次关系。

算法 1：AR系数估计流程

输入：来自分量稳定过程的样本、置信参数 $A \geq 1$ 和 $\delta > 0$、稳定性参数 $\epsilon \in (0,1)$ˆ ˜。输出：估计滞后阶数 L_0 $\boldsymbol{\beta}$，以及估计的自回归系数向量。令 min 为各分量样本数的最小值。求解 L （见式（38））：

$$n_{\min} = L + 84Ae\zeta^{-2} L \log\left(\frac{ML}{\delta}\right).$$

2. 将 Nicholson 等人（2020）的学习算法作为子程序，用于求解 (15) 中的凸优化问题——其中 L 如上；将输出记为 ˆ $\boldsymbol{\beta}$。3. 如果对所有 m， $m' \in [M]$ 都满足 $\boldsymbol{\beta}_m = \boldsymbol{\beta}_{m'}$（等价地，分量时间序列来自相同AR过程），则返回 ˆ$L_0 := \max\{j : |(\hat{\boldsymbol{\beta}}_1)_j| > \lambda\}$ 和 $\tilde{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_0', \cdots, \hat{\boldsymbol{\beta}}_0')$，其中符号说明参见公式 (3)

$$\hat{\boldsymbol{\beta}}_0' := \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m(\hat{L}_0);$$

否则，返回

$$\hat{L}_0 := \max\{j : |\hat{\boldsymbol{\beta}}_j^m| > \lambda \text{ for some } m \in [M]\},$$

and $$\tilde{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_1(\hat{L}_0), \cdots, \hat{\boldsymbol{\beta}}_m(\hat{L}_0)).$$

以下是关于估计流程输出的理论性质的主定理。该定理基本上总结了第4节中的结果，并将在下文所示的 4.2.2 和 4.3 小节中进行详细讨论。

ˆ定理 3.1（主定理）。设 $\boldsymbol{\beta}$ 为上述式（15）中组 LASSO 的输出，其中

$$\lambda = 24(84Ae)^{\frac{1}{2}} \zeta^{-1} \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4}) \sqrt{\frac{L \log\left(\frac{ML}{\delta}\right)}{DM}},$$

其中 $\zeta = 6^{-3}\epsilon^4$，且 $L \geq L_0$ 满足

$$n_{\min} = L + 84Ae\zeta^{-2} L \log\left(\frac{ML}{\delta}\right).$$

假设 β-最小值条件（10）以 $c_\beta = \lambda$成立。若后样本的总数 D 满足

$$D \geq 3^9 \cdot (84Ae) C_\sharp^5 (1 + \epsilon^{-2} + \epsilon^{-4})^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4 (\epsilon^3\zeta)^{-2} M^a L_0^2 L^3 \log\left(\frac{ML}{\delta}\right) \log(2L),$$

且若 $T_{\min} \geq 84eA\zeta^{-2} L_0$ 对数 L，则以下结论以高概率成立：

1. 估计误差为

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \frac{81(84Ae)^{\frac{1}{2}} L L_0 \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4})}{\zeta \alpha \epsilon^2} \sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}};$$

2. 若 ˆ$S_\lambda := \{j : \ |\boldsymbol{\beta}_j| > \lambda\}$，则错误发现可由以下不等式界定：

$$|S_\lambda \setminus \operatorname{supp}(\boldsymbol{\beta})| \ \leq \frac{243(84Ae)^{\frac{1}{2}} L L_0^{\frac{3}{2}} \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4})}{\zeta \alpha \epsilon^2 \lambda} \sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}}.$$

ˆ3. 由算法 1（“AR系数估计流程”）返回的系数 $\boldsymbol{\beta}_0$ 拟合得到的 AR 模型具有稳定性，且该结论以高概率成立。

定理3.1的证明。在满足所述的 β-最小条件（10）的情况下，这可直接由定理（4.9）结合定理（4.10）得到。 □

在上述流程中，有必要考虑两个不同的稳定性参数 $A \geq 1$ 和 δ，因为无法将它们整合为单一参数——这主要是由于在我们的设置中，来自分量过程的样本数量不同。此外，我们分别提及两种情况（对于所有 m， $\boldsymbol{\beta}_m$ 是否相同），以特别强调：在第一种情况下，我们的算法所需的样本数量少于后一种情况（从而使样本复杂度降低一个 M 倍的因子）。

像上述子程序中使用的近端梯度法这样的优秀算法，可能会由于数值伪影而产生取值很小但非零的参数。也可以考虑其他类型的算法，但所需的计算步骤数可能会高得多。此外，如算法中所述，通过标准的 λ-阈值处理，也可以控制这些伪影以及更广泛意义上的统计假阳性。

4. 统计保证。本节介绍本文的主要理论结果。我们首先推导一步预测误差的界，首先将其表述为一个预言机不等式（见定理 4.2），该不等式取决于我们最小二乘惩罚方法的调节参数 λ。然后，我们通过控制 Lasso 优化问题的有效噪声来控制这个调节参数。综合这两点，最终可以得到一步预测误差更明确的收敛速率。本节第二部分讨论对估计的自回归系数的控制，包括对其支撑集错误发现的控制（定理 4.9）。最后，我们给出关于估计的 AR 模型稳定性的结果（定理 4.10）。

T首先，我们简要回顾上文定义的组 LASSO 范数 $\mathcal{N}$的对偶范数概念
定义的。
我们的底层空间为 $\mathbb{R}^{ML}$。注意，

$$\boldsymbol{\beta} \mapsto \mathcal{N}(\boldsymbol{\beta}) := \sum \sqrt{|\mathcal{G}_l|} \cdot \|\boldsymbol{\beta}_{\mathcal{G}_l}\|_2$$

对于任意覆盖 $[ML]$的 $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_l\}$，它都是一个范数。然而，在某个组内所有坐标均为零的点上，这个范数是奇异的。此外，所有较小规模组中的所有坐标都会为零；这一点对我们的分析很重要，因为我们基本上关注的是稀疏解。因此，我们不能借助微分技术来获得该范数的界，而需要采用对偶范数方法。

定义 4.1（对偶范数）。对于 $\mathbb{R}^{ML}$ 上的范数 $\mathcal{N}$， $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$ 的对偶范数 $\mathcal{N}_\star(\boldsymbol{\alpha})$ 是以下凸规划的最优解：

$$\begin{aligned} &\text{maximize } \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \\ &\text{subject to } \mathcal{N}(\boldsymbol{\beta}) \leq 1 . \end{aligned}$$

对偶范数用于概括 LASSO 的有效噪声。更具体地说，对偶范数以 $\mathcal{N}_\star(X^\top \boldsymbol{U})$ 的形式出现，在式 (15) 中称为 LASSO 的有效噪声。回想一下，有效噪声向量 $X^\top \boldsymbol{U}$ 可视为噪声向量 $\boldsymbol{U}$ 在 X 的列空间上的“投影”，因此， $\mathcal{N}_\star(X^\top \boldsymbol{U})$ 衡量了数据中存在的“真实”噪声。

在附录 D 中，我们得到了对偶范数的一个通用界，该界将用于获得本文的统计保证。

4.1. 预测误差。现在，我们给出已宣布的一步预测误差非渐近界的理论结果：首先通过一个预言机不等式进行表述（见定理 4.2），随后在下一小节中，通过控制 Lasso 优化问题的有效噪声来控制调节参数 λ。这使我们能够给出预测误差的具体速率。请注意，这种方法明确说明了如何选择 L（通过公式（24），随后通过公式（38）），即算法 1（“AR系数估计流程”）第 2 步的输入参数。

4.1.1. 预言机预测误差。问题 (15) 有解，因为对于时间序列的任意特定实现（因此，实际上固定了 $\boldsymbol{y}$ 和 X）以及任意 $\lambda > 0$，凸函数

$$f_\lambda(\boldsymbol{\beta}) = \frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{\mathcal{G}} \tag{17}$$

是连续的，并且

$$\lim_{\|\boldsymbol{\beta}\|_2 \to \infty} f_\lambda(\boldsymbol{\beta}) = \infty\,.$$

这也表明，我们可以将其视为定义在紧域上的凸规划，因为（至少在理论上）我们应该能够将最小化的定义域限制为半径适当的 ℓ^2-球，例如 $R_{r,X,\boldsymbol{y}} \in (0,\infty)$。令 $\boldsymbol{\beta}$ 为该凸规划的一个解。

现在将时间序列观测值写成矩阵形式，即 $\mathcal{X}$。我们关注“风险”的估计，等价地说，即样本内单步超前均方预测误差 $\mathbb{E}[\|\boldsymbol{y} - X\hat{}\boldsymbol{\beta}\|_2^2/D \mid \mathcal{X}]$。在下面的推导中，我们采用一种广为人知的方法来控制该误差，例如 Lederer（2021年，第 6 章）中所述。证明见附录 A。

定理 4.2（预测保证）。假设 $\lambda \geq \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$，其中 $\mathcal{N}(\boldsymbol{\alpha}) = \sum_{l=1}^{L}\sqrt{|\mathcal{G}_l|}\cdot\|\boldsymbol{\alpha}_{(\geq l)}\|_2$如命题 D.3 中所述。记 $\overline{\sigma}^2 = D^{-1}(T_1\sigma_1^2 + \cdots + T_M\sigma_M^2)$；则

$$\frac{1}{D}\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] \leq \overline{\sigma}^2 + \min_{\alpha\in\mathbb{R}^{ML}}\left(\frac{1}{D}\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + 2\lambda\mathcal{N}(\boldsymbol{\alpha})\right),$$

特别地，以下不等式成立：

$$\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] - \mathbb{E}\left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \mid \mathcal{X}\right] \leq \min_{\alpha\in\mathbb{R}^{ML}}\left(\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + 2D\lambda\mathcal{N}(\boldsymbol{\alpha})\right).$$

此外，若 $\lambda \geq \frac{4}{D}\mathcal{N}_\star(X^\top\boldsymbol{U})$，则

$$\frac{1}{D}\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] \leq \overline{\sigma}^2 + \frac{\lambda}{2}\min\left\{3\mathcal{N}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), 3\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}})\right\}. \tag{18}$$

因此，

$$\frac{1}{D}\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] \leq \overline{\sigma}^2 + 2\lambda\sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell}\,. \tag{19}$$

该结果具有高维统计学中的标准 oracle 不等式形式（Lederer，2021年，第 6 章）。它表明，估计量使一步预测的均方风险最小化，其误差至多为一个复杂度项，该项与调节参数和模型复杂度成线性关系。因此，一旦我们能够通过有效噪声的上界控制调节参数 λ，上述界便给出了收敛速率的上界。

根据定理 4.2 和命题 D.3，为了找到满足 $\lambda \geq 4\mathcal{N}_\star(X^\top\boldsymbol{U})/D$ 的最小调节参数 λ，只需推导出关于$D^{-1}L^{-\frac{1}{2}}\|X^\top\boldsymbol{U}\|_\infty$的一个高概率上界（这恰好是关于 $X^\top U$ 的对偶范数的界）。

4.1.2. 控制有效噪声以界定调节参数。我们可以证明调节参数（等价地，有效噪声）的以下高概率界。同样，其证明见附录 B。请注意，这是一个重要的不等式；尽管相关论证广为人知，但我们将这些论证应用到了稀疏性强制正则化器由重叠组范数诱导的情形。

定理 4.3（有效噪声的界）。设 $C_\sharp := T_{\max}/T_{\min}$。对于任意满足以下条件的 $\eta > 0$

$$\eta \geq \frac{8C_\sharp \sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{M}, \tag{20}$$

对于任意 $\delta > 0$，若 $D = T_1 + \cdots + T_m$ 满足

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right), \tag{21}$$

则以下不等式成立：

$$\mathbb{P}\left[\frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \geq \frac{3\eta}{2\sqrt{L}}\right] \leq \delta. \tag{22}$$

这里，$c_0 > 0$ 是命题 E.1 中高斯集中不等式的绝对常数。

上述结果的含义是，若要使 LASSO 预言机不等式（定理 4.2）以高概率成立，只需将 λ 设为与 $(3\eta)/(2\sqrt{L})$ 一样大，但没有必要设得更大。事实上，公式（22）表明，$2/D\mathcal{N}_\star(X^TU)$ 大于 $3\eta/2\sqrt{L}$ 的概率很小；因此，我们可以假设 $2/D\mathcal{N}_\star(X^TU) < 3\eta/2\sqrt{L}$ 以概率 $1-\delta$ 成立。

T上述结果界定了有效噪声的尾部。对于任意 $A \geq 1$，我们现在令

$$\eta = C_0\sqrt{A}C_\epsilon C_\sharp^{\frac{3}{2}}\sigma_{\max}^2\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}},$$

对于某个绝对常数 $C_0 > 0$ 和仅取决于稳定性参数 ϵ 的参数 $C_\epsilon > 0$。以下我们取

$$\eta := 8(84Ae)^{\frac{1}{2}}\zeta^{-1}(1+\epsilon^{-2}+\epsilon^{-4})C_\sharp^{\frac{3}{2}}\sigma_{\max}^2\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}, \tag{23}$$

将 C_0 和 C_ϵ 的值——如下面引理4.4的证明中所示——代入公式（23）即可得到该结果；我们不尝试优化这些常数。

引理 4.4（关于 L 的数据依赖选择）。存在一个绝对常数 $C \in (0,\infty)$以及一个仅取决于稳定性参数 $\epsilon > 0$ 的参数 $C_\epsilon > 0$，使得如果 η 如公式（23）中所示，并且

$$T_{\min} \leq 84Ae\zeta^{-2}L\log\left(\frac{ML}{\delta}\right), \tag{24}$$

则不等式（20）成立。

这一结果表明，存在一个合适的 η，使我们的理论成立。如何在实践中找到这样的 η，需要在今后更具应用性的研究中加以探讨。

证明。令$C_0 := (12)^3\sqrt{84e}$，并且

$$C_\epsilon := \epsilon^{-4}(1+\epsilon^{-2}+\epsilon^{-4})\,.$$

注意

$$C_\sharp T_{\min} = T_{\max} \geq \frac{D}{M}\,. \tag{25}$$

现在，取 (23) 中的 η，则以下一系列不等式保证了不等式（20）成立：

$$\begin{aligned}
& (84Ae)^{\frac{1}{2}}\zeta^{-1}\sqrt{L\log\left(\tfrac{ML}{\delta}\right)} && \geq \sqrt{T_{\min}} \\
\Rightarrow \quad & (84Ae)^{\frac{1}{2}}\zeta^{-1}C_\sharp^{\frac{1}{2}}\sqrt{L\log\left(\tfrac{ML}{\delta}\right)} && \geq \sqrt{\frac{D}{M}} \\
\Rightarrow \quad & 8(84Ae)^{\frac{1}{2}}\zeta^{-1}(1+\epsilon^{-2}+\epsilon^{-4})C_\sharp^{\frac{3}{2}}\sigma_{\max}^2\sqrt{\tfrac{L\log\left(\frac{ML}{\delta}\right)}{DM}} && \geq \frac{8C_\sharp\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{M}\,.
\end{aligned}$$

□

不等式（24）为我们在公式（15）中构造惩罚最小二乘程序时选择 L 提供了理论依据。

上述证明中参数的显式形式至关重要，因为它既要满足上述（20）和（21），又要符合下文涉及受限特征值性质（如推论 4.7）的要求。需要注意的是，当前基于 LASSO 的文献往往忽略了将标准预言机不等式类结果（定理 4.2）的要求与受限特征值类结果的要求结合起来。

备注（关于通过公式（24）选择 L）。采用这样的上界 L，符合压缩感知文献中子采样高斯矩阵（即元素为独立同分布高斯变量的矩阵）广为人知的受限等距性质；更多详情请参见（例如）Candes 和 Tao（2006，第 1.E 节）及其中的参考文献。

对于如式 (21) 中的 D，我们得到以下界：

$$\mathbb{P}\left[\mathcal{N}_\star\left(\frac{2}{D}X^\top \boldsymbol{U}\right) > 12(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp{}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}\right] \leq \delta,$$

鉴于定理 4.3，相应地令

$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp{}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}} \tag{26}$$

以确保 (18) 成立。

备注。从上述内容可以看出，当 M, D 中任意一项增大时，λ 会减小（因为当 $x\to\infty$ 时，函数 $x^{-1}\log x \to 0$）。这是符合直觉的，因为较大的 M 要求 LASSO 不能将太多系数设为 0；而 D 越大，未正则化估计量对真实系数的近似效果越好。

结合定理 4.2，以上结果给出如下界：

推论 4.5（预测误差的具体速率）。假设

$$\eta = 8(84Ae)^{\frac{1}{2}}\zeta^{-1}(1+\epsilon^{-2}+\epsilon^{-4})C_\sharp^{\frac{3}{2}}\sigma_{\max}^2\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}},$$

如上文公式（23）所设，并且

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right).$$

则以下不等式以至少 $1-\delta$ 的概率成立：

$$\mathbb{E}\left[\frac{1}{D}\|\boldsymbol{y}-X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] - \overline{\sigma}^2$$

$$\leq \min_{\alpha\in\mathbb{R}^{ML}}\left(\frac{1}{D}\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + \frac{24}{\zeta}(84Ae)^{\frac{1}{2}}\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}\mathcal{N}(\boldsymbol{\alpha})\right).$$

4.2. 参数估计误差与错误发现。现在，我们将处理主定理 3.1 中关于估计误差和支撑集（通过控制错误发现）的断言 1 和 2。为此，我们需要处理受限特征值性质，因为该性质通常出现在 LASSO 分析中（参见 Basu 和 Michailidis（2015））。

4.2.1. 受限特征值性质的控制。到目前为止，我们所陈述的预测保证并未对每个组成部分的最小样本数施加限制。然而，对于估计和稳定性保证，我们需要要求这些样本达到最小数量。具体而言，我们将需要以下关于格拉姆矩阵 $X^\top X$ 的受限特征值性质的命题。附录 C 中给出的证明沿用了 Basu 和 Michailidis（2015）中的论证思路。由于公式（14）中数据矩阵的分块维度并不均一，同时系数 $\boldsymbol{\beta}$ 的各组成部分被赋予了不同的权重，我们只能期待得到如下所述的分块结果。

下面的命题给出了误差表达式的下界，其中 $\|X_m\boldsymbol{v}_m\|_2^2\,\hat{}\,\boldsymbol{v}_m$ 是 $\boldsymbol{\beta}-\boldsymbol{\beta}$ 的第 $m-$ 个分量。

命题 4.6（受限特征值的界）。对于任意置信参数$\delta>0$以及所有向量 $\boldsymbol{v}=(\boldsymbol{v}_1^\top,\ldots,\boldsymbol{v}_M^\top)^\top\in(\mathbb{R}^L)^M$，当 $\zeta:=\frac{\epsilon^4}{216}$ 和 s_m 为偶正整数时，以下不等式成立：

$$\mathbb{P}\left[\forall\, m\in[M]\inf_{\boldsymbol{v}_m\in\mathbb{R}^L}\left(\|X_m\boldsymbol{v}_m\|_2^2 - \frac{T_m\sigma_m^2\epsilon^2}{2}\left(\|\boldsymbol{v}_m\|_2^2 - \frac{2}{s_m}\|\boldsymbol{v}_m\|_1^2\right)\right)\geq 0\right]$$

$$\geq 1-2\sum_{m=1}^{M}\exp\left(-\frac{T_m\min\{\zeta,\zeta^2\}}{2}+s_m\min\{\log L,\log(21eL/s_m)\}\right). \tag{27}$$

命题 4.6 的证明见附录 C。

设定

$$s_m := 2\left\lfloor\frac{T_m\zeta^2}{8\log L}\right\rfloor, \tag{28}$$

我们立即得到命题 4.6 的以下推论：

推论 4.7（特定 s_m 的受限特征值上界）。若

$$T_{\min} \geq 84e\zeta^{-2}\log L\,, \tag{29}$$

其中 $\zeta = 6^{-3}\epsilon^4$，则有：

$$\begin{aligned}&\mathbb{P}\left[\forall m\in[M]\inf_{\boldsymbol{v}_m\in\mathbb{R}^L}\left(\|X_m\boldsymbol{v}_m\|_2^2-\frac{T_m\sigma_m^2\epsilon^2}{2}\|\boldsymbol{v}_m\|_2^2\left(1-\frac{8L\log L}{T_m\zeta^2}\right)\right)\geq 0\right]\\ &\geq\ 1-2\sum_{m=1}^{M}e^{-\frac{T_m\zeta^2}{4}}\,.\end{aligned} \tag{30}$$

证明。我们使用命题 4.6 中的 $\|\boldsymbol{v}_m\|_1^2 \leq L\|\boldsymbol{v}_m\|_2^2$。 □

4.2.2. 自回归系数的估计误差界及错误发现。本节给出了自回归系数估计误差的界，以及错误发现的界。

以下是下述定理证明中使用的紧凑记号（也用于定理 4.2 的证明）。

定义 4.8（记号）。

$$\mathcal{N}_{\leq L_0}(\boldsymbol{v}) := \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell}\,. \tag{31}$$

下述定理使用式 (15) 中的惩罚 LASSO 形式，给出了估计自回归系数 $\boldsymbol{\beta}$ 时的 ℓ^2 误差界，以及假阳性集合的大小界。

ˆT定理 4.9（估计误差和错误发现的界）。设 $\boldsymbol{\beta}$ 是 (15) 中组正则化 LASSO 的解，其中 λ 如上方程 (26) 所示，即

$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}$$

其中 $\zeta = 6^{-3}\epsilon^4$；若

$$\alpha := \min_{m\in[M]}\frac{T_m\sigma_m^2}{D}\,, \tag{32}$$

and

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right),$$

那么，对于任意置信参数 $\delta > 0$，不等式

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}} \tag{33}$$

以至少如下概率成立

$$(1-\delta)\left(1-2\sum_{m=1}^{M}L^{-\frac{T_m}{4\log L}}\right), \tag{34}$$

在$T_{\min}\geq 84eA\zeta^{-2}L_0$对数$L$条件下，给定 $\zeta=6^{-3}\epsilon^4$。

ˆ此外，引入记号 $S_\lambda:=\{j:\ \ |\boldsymbol{\beta}_j|>\lambda\}$，则错误发现由下述不等式界定，该不等式以与上文 (34) 中相同的概率成立：

$$|S_\lambda\setminus\operatorname{supp}(\boldsymbol{\beta})|\leq\frac{243(84Ae)^{\frac{1}{2}}LL_0^{\frac{3}{2}}\sigma_{\max}^2C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2\lambda}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}}. \tag{35}$$

备注。在上述内容中，平方根内的项渐近地减小，而外部的项可能发生变化，也可能保持不变（取决于每个分量的样本数量如何增加）。

定理 4.9 的证明。为简化记号，记

$$\boldsymbol{v}:=\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}.$$

根据上文的命题 4.6（由于 $T_{\min}\geq 84e\zeta^{-2}$ 对数 L，可应用该命题），有

$$\begin{aligned}\frac{1}{D}\|X\boldsymbol{v}\|_2^2&\geq\frac{1}{D}\sum_{m=1}^{M}\|X_m\boldsymbol{v}_m\|_2^2\\&\geq\frac{\epsilon^2}{2}\sum_{m=1}^{M}\frac{\sigma_m^2T_m}{D}\|\boldsymbol{v}_m\|_2^2\left(1-\frac{8L\log L}{T_m\zeta^2}\right)\\&>\frac{4\alpha\epsilon^2}{9}\|\boldsymbol{v}\|_2^2.\end{aligned}$$

对此，我们现在应用（见附录 A 中的方程 (45)）中的不等式

$$\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2\leq 3\lambda\mathcal{N}_{\leq L_0}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),$$

从而得到

$$\begin{aligned}\|\boldsymbol{v}\|_2^2&\leq\frac{27\lambda}{8\alpha\epsilon^2}\mathcal{N}_{\leq L_0}(\boldsymbol{v})\\&\leq\frac{27\lambda L_0\sqrt{ML}}{8\alpha\epsilon^2}\|\boldsymbol{v}\|_2\\\Rightarrow\qquad\|\boldsymbol{v}\|_2&\leq\frac{27\lambda L_0\sqrt{ML}}{8\alpha\epsilon^2}\end{aligned}$$

取 λ 如方程 (26) 所示，即

$$\lambda=24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}},$$

由此可得

$$\|\boldsymbol{v}\|_2\leq\frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}},$$

如所述。为得到错误发现的界，令 $S := \mathrm{supp}(\boldsymbol{\beta})$；错误发现的界可按如下方式得到：

$$
\begin{aligned}
|S_\lambda \setminus S| &= \sum_{j \in [ML] \setminus S} 1_{|\hat{\boldsymbol{\beta}}_j| > \lambda}(\hat{\boldsymbol{\beta}}) \\
&= \sum_{j \in [ML] \setminus S} 1_{|\hat{\boldsymbol{v}}_j| > \lambda}(\hat{\boldsymbol{v}}) \\
&\le \frac{1}{\lambda} \sum_{j \in [ML] \setminus S} |\hat{\boldsymbol{v}}_j| \\
\text{by remark (A)} \qquad &\le \frac{3}{\lambda} \sum_{j \in S} |\hat{\boldsymbol{v}}_j| \\
&\le \frac{3\sqrt{L_0}}{\lambda} \|\hat{\boldsymbol{v}}\|_2
\end{aligned}
$$

由不等式 (33) 可得界 (35)。 □

备注。对于方程 (26) 中的 λ 的显式取值，即，

$$
\lambda = 24(84Ae)^{\frac{1}{2}} \zeta^{-1} \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4}) \sqrt{\frac{L \log\left(\frac{ML}{\delta}\right)}{DM}}
$$

取 $\zeta = 6^{-3}\epsilon^4$ 时，不等式 (35) 变为

$$
|S_\lambda \setminus \mathrm{supp}(\boldsymbol{\beta})| \le \frac{81(ML)^{\frac{1}{2}} L_0^{\frac{3}{2}}}{8\alpha\epsilon^2}. \tag{36}
$$

4.3. 估计的 AR 模型的稳定性。最后，我们证明，按照 (15) 中的重叠组 LASSO 估计得到的系数（其中 λ 取 (26) 中的值）位于单变量滞后阶数为 L 的自回归过程的稳定区域内，即使后续样本数量“不太大”也是如此。更具体地，我们有以下定理。

ˆ定理 4.10（稳定性保证）。设 $A \ge 1$ 为置信参数。设 $\boldsymbol{\beta}$ 为上述 (15) 中组 LASSO 的输出，使用 D 个后续样本，其中

$$
D \ge 3^9 \cdot (84Ae) C_\sharp^5 (1 + \epsilon^{-2} + \epsilon^{-4})^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4 (\epsilon^3\zeta)^{-2} M^a L_0^2 L^3 \log\left(\frac{ML}{\delta}\right) \log(2L). \tag{37}
$$

若 $T_{\min} \ge 84eA\zeta^{-2}L_0$ 对数 L，且 L 满足

$$
n_{\min} = L + 84Ae\zeta^{-2} L \log\left(\frac{ML}{\delta}\right), \tag{38}
$$

满足 $\zeta = 6^{-3}\epsilon^4$ 时，由算法 1（“AR系数估计流程”）返回的系数 ˆ $\boldsymbol{\beta}_0$ 所拟合的 AR 模型是稳定的，其概率如式 (34) 所示，适用于以下情形：

1. 当所有时间序列都是同一个潜在随机过程的不同实现时， $a \ge 1$，并且

$$
\hat{\boldsymbol{\beta}}_0 := \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m ;
$$

2. 当所有时间序列都是不同潜在随机过程的实现时ˆ ˆ（等价地，所有 $\boldsymbol{\beta}_m$ 都各不相同），$a \geq 2$，并且 $\boldsymbol{\beta}_0 := \boldsymbol{\beta}$。

$T_{\min}$ 的下界与公式（24）中的上界在 $L \geq L_0$ 处一致。 $n_{\min}$ 的给定值选取了满足 $T_{\min}$上界和下界条件的 L 最小值。

定理 4.10 的证明。1. 我们从第一种情况开始。证明思路如下。 $\boldsymbol{\beta}$ 的所有分量 $\boldsymbol{\beta}_m$ 都是对所有 $m \in [M]$ 而言同一个底层 $\boldsymbol{\beta}_0 := \boldsymbol{\beta}_m$ 的近似。因此，根据上文命题 4.9 中的界以及平方函数的凸性，它们的均值必定与 $\boldsymbol{\beta}_0$ 相差不超过 ℓ^2，所需样本数为 $D = O(ML_0^2L^3\log(ML)\log(2L))$。此外，对系数进行 $0.5L^{-1}\epsilon$ 扰动仍能保持 ϵ-稳定过程的稳定性。

更明确地说，我们有

$$\begin{aligned}\left\|\left(\sum_{m=1}^{M}\frac{1}{M}\hat{\boldsymbol{\beta}}_m\right)-\boldsymbol{\beta}_0\right\|_2^2 &= \left\|\frac{1}{M}\sum_{m=1}^{M}(\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_m)\right\|_2^2\\ \text{convexity } \Rightarrow \qquad &\leq \frac{1}{M}\sum_{m=1}^{M}\|\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_m\|_2^2\\ &= \frac{1}{M}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2,\end{aligned}$$

根据定理 4.9，可得

$$\begin{aligned}\left\|\left(\sum_{m=1}^{M}\frac{1}{M}\hat{\boldsymbol{\beta}}_m\right)-\boldsymbol{\beta}\right\|_2 &\leq \frac{1}{\sqrt{M}}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\\ &\leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{MD}} \qquad (39)\end{aligned}$$

以高概率。注意,

$$\alpha = \min_{m\in[M]}\frac{T_m\sigma_m^2}{D} \geq \frac{T_{\min}\sigma_{\min}^2}{D} \geq \frac{T_{\min}\sigma_{\min}^2}{MT_{\max}} \geq \frac{\sigma_{\min}^2}{MC_\sharp}. \qquad (40)$$

When

$$D \geq 3^9\cdot(84Ae)C_\sharp^5(1+\epsilon^{-2}+\epsilon^{-4})^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4(\epsilon^3\zeta)^{-2}ML_0^2L^3\log\left(\frac{ML}{\delta}\right)\log(2L),$$

由此可得

$$\begin{aligned}\left\|\left(\sum_{m=1}^{M}\frac{1}{M}\hat{\boldsymbol{\beta}}_m\right)-\boldsymbol{\beta}\right\|_2 &\leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{MD}}\\ &\leq \frac{\epsilon\sigma_{\min}^2}{\sqrt{3}C_\sharp M\alpha}\sqrt{\frac{1}{L\log(2L)}}\\ (40)\ \Rightarrow \qquad &< \frac{\epsilon}{\sqrt{3}}\sqrt{\frac{1}{L\log(2L)}}.\end{aligned}$$

ˆ由于 $\epsilon \in (0,1)$，根据三角不等式即可推出稳定性。更明确地说，若 $\boldsymbol{\beta}_0 := M^{-1}(\hat{}\boldsymbol{\beta}_1 + \cdots + \hat{}\boldsymbol{\beta}_M)$，则对于每个 $z \in \mathbb{D}$，都有：

$$\begin{aligned}
|\boldsymbol{f}_{\hat{\boldsymbol{\beta}}_0}(z)| &= |1 - \hat{\boldsymbol{\beta}}_0 \cdot (z, z^2, \cdots, z^L)| \\
&= |1 - \boldsymbol{\beta}_0 \cdot (z, z^2, \cdots, z^L) - (\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \cdot (z, z^2, \cdots, z^L)| \\
&\geq |\boldsymbol{f}_{\boldsymbol{\beta}_0}(z)| - \|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|_2 \cdot \|(z, z^2, \cdots, z^L)\|_2 \\
&> \epsilon - \frac{\epsilon}{\sqrt{3}} \sqrt{\frac{1}{\log(2L)}} \\
&> 0 .
\end{aligned}$$

ˆ显然，这表明具有 $\boldsymbol{\beta}_0$ 系数的自回归过程是稳定的。

2. 第二种情况下的论证类似。由于

$$D \geq 3^9 \cdot (84Ae) C_\sharp^5 (1 + \epsilon^{-2} + \epsilon^{-4})^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4 (\epsilon^3 \zeta)^{-2} M^2 L_0^2 L^3 \log\left(\frac{ML}{\delta}\right) \log(2L) ,$$

根据定理 4.9，对于任意 $m \in [M]$，我们有

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\|_2 &\leq \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\
&\leq \frac{81(84Ae)^{\frac{1}{2}} L L_0 \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4})}{\zeta \alpha \epsilon^2} \sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}} \\
\text{(40)} \ \Rightarrow \qquad &< \frac{\epsilon}{\sqrt{3}} \sqrt{\frac{1}{L \log(2L)}} .
\end{aligned}$$

与第一种情况一样，这意味着每个分量都具有稳定性。

□

5. 算法方面。本节简要介绍算法的核心内容——即近端算子——以了解如何求解 LASSO，方法与 Nicholson 等人（2020年）所述相同。为了概述求解带有重叠分组范数惩罚的正则化 LASSO 的标准流程，我们先给出以下定义（例如，参见 Mairal 等人（2011年））。

定义 5.1（近端算子）。给定 $\mathbb{R}^{ML}$ 上的一个范数 $\mathcal{N}$ 和一个调节参数 λ，相应的近端算子 $\mathrm{Prox}_{\mathcal{N},\lambda}$ 对每个 $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$ 定义为以下凸优化问题的最优值：

$$\underset{\mathcal{N},\lambda}{\mathrm{Prox}}(\boldsymbol{\alpha}) := \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2 + \lambda \mathcal{N}(\boldsymbol{\beta}) \right\} .$$

根据 Zhao 等人（2009），近端算子 $\mathrm{Prox}_{\mathcal{N},\ \lambda}$——对于 (12) 中定义的 $\mathcal{N}$——是以下组合

$$\underset{\mathcal{G}_1,\lambda}{\mathrm{Prox}} \circ \cdots \circ \underset{\mathcal{G}_L,\lambda}{\mathrm{Prox}}$$

各个组的近端算子，并且可以通过归纳方式计算——从$\mathrm{Prox}_{\mathcal{G}_L,\ \lambda}$开始，这是众所周知的软阈值——在 $O(ML)$ 计算

步骤。根据 Combettes 和 Wajs（2005年，命题 3.1(iii)b），(15) 中正则化 LASSO 问题的解恰好是算子的不动点

$$\boldsymbol{\beta} \longmapsto \operatorname*{Prox}_{\mathcal{N},\lambda}\left(\boldsymbol{\beta} + \frac{2\lambda}{D}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right). \tag{41}$$

这里我们使用适当的 λ（如上文公式（26）所示）。正如 Nicholson 等人（2020年）所指出的，近端算子可通过对偶性求得。随后，Mairal 等人（2011年）的近端梯度法求出该近端算子的固定点（由公式（15）的凸性可知，该固定点存在且唯一）。有关此过程的伪代码，请参见 Nicholson 等人（2020年）；他们采用了加速版近端下降法，以实现二次收敛率。

6. 结论。我们建立了一套框架（见公式（15）及该公式之前的讨论），在这一框架下，可使用经层次组范数正则化的 LASSO，在不同长度但真实滞后阶数相同的多个 ϵ-稳定（定义（2.6））单变量自回归过程中，就一步预测误差给出统计保证（定理 4.2）。这里的结果不要求预先知道真实滞后阶数（或其任何上界）；事实上，我们表明，ˆ样本量本身就能提示应为组 LASSO 选用某个滞后阶数 L；而当ˆ样本量足够大时，L 将成为真实滞后阶数 L_0的上界。ˆ此外，这个 L 将达到理论保证成立所需的阶数。我们证明，采用适当调节参数 λ 构造的组 LASSO，能够以任意高的精度估计 AR 系数。我们还表明，估计所得系数集的支撑集与原始参数的支撑集基本一致（定理 4.9）。最后，我们证明，由组 LASSO 估计的系数所构成的拟合模型具有 ϵ-稳定性（定理 4.10）；在文献中，这一性质此前仅针对单变量自回归过程参数的 Yule-Walker 估计有所论述。

从理论角度来看，研究如何将组 LASSO 方法推广到具有多变量分量的多个解耦 AR 过程，将会很有意义。我们预计，这需要采用多种方法，其中包括：(1) 运用额外技术处理组范数；(2) 将稳定性问题与受限特征值问题结合起来；在多变量分量的情形下，这些工作的技术难度将大得多。在实践方面，最重要的问题是更准确地确定调节参数λ （见方程 (26)），这需要比例如引理4.4的证明中所用的常数/参数更精确的取值。更好地控制这些量，有助于对 L 作出更切合实际的ˆ估计，以便让组 LASSO 将其用作真实滞后阶数 L_0的上界。

基金资助。S.C. 和 J.L. 感谢德国研究联合会（DFG）通过资助项目编号 451920280 提供资助。S.C. 还部分获得了波鸿鲁尔大学研究生院通过德国学术交流中心（DAAD）的 STIBET 资金提供的“青年博士后研究助理奖学金”。

参考文献

Sumanta Basu 和 George Michailidis：稀疏高维时间序列模型中的正则化估计。《统计年刊》43(4)：1535–1567，2015年。Florentina Bunea：通过 ℓ_1 和 $\ell_1+\ell_2$ 惩罚在与逻辑回归模型中进行诚实的变量选择。《电子统计期刊》2，2008年。Stephen Boyd、Lieven Vandenberghe：《凸优化》。剑桥大学出版社，剑桥，2004年。Emmanuel Candes、Terence Tao：从随机投影中实现近乎最优的信号恢复：通用编码策略？《IEEE信息论汇刊》52，第12期，2006年

Patrick Combettes、Valérie Wajs：通过近端前向-后向分裂进行信号恢复。《多尺度建模与仿真》4，第4期，1168–1200，2005年。Mario Forni、Marc Hallin、Marco Lippi、Lucrezia Reichlin：《广义动态因子模型：识别与估计》。《经济学与统计学评论》82，第4期，540–554，2000年。Johannes Lederer：《高维统计学基础——含习题与 R 实验》。Springer 统计学丛书。Springer，尚，2021年。Po-Ling Loh 和 Martin J. Wainwright：含噪与缺失数据下的高维回归：具有非凸性的可证明保证。《统计年刊》40(3)：1637–1664，2012年。Helmut Lütkepohl：《多元时间序列分析新导论》。Springer 出版社，柏林，2005年。Julien Mairal、Rodolphe Jenatton、Guillaume Obozinski 和 Francis Bach：面向结构化稀疏性的凸优化与网络流优化。《机器学习研究杂志》12：2681–2720，2011年。Ricardo Masini、Marcelo Medeiros 和 Eduardo Mendes：弱相关创新项下高维向量自回归的正则化估计。《时间序列分析杂志》43：532–557，2011年。Yuval Nardi 和 Alessandro Rinald：通过 LASSO 方法进行自回归过程建模。《多元分析杂志》102，第3期，528–549，2011年。Sahand Negahban 和 Martin J. Wainwright：含噪声且具有高维尺度的（近似）低秩矩阵估计。《统计年刊》39(2)：1069–1097，2011年。William B. Nicholson、Ines Wilms、Jacob Bien 和 David S. Matteson：通过可解释的向量自回归进行高维预测。《机器学习研究杂志》21：论文第166号，52页，2020年。Paul Tseng：关于凸-凹优化的加速近端梯度法。2008年。可访问：http://www. math. washington. edu/˜tseng/papers/apgm. pdf。Roman Vershynin：《高维概率：数据科学中的应用导论》。Sara van de Geer 作序。剑桥统计与概率数学丛书，第47卷。剑桥大学出版社，剑桥，2018年。Kam Chung Wong、Zifan Li 和 Ambuj Tewari：针对 β-混合重尾时间序列的 Lasso 保证。《统计年刊》48(2)，2020年。Peng Zhao、Guilherme Rocha、Bin Yu：用于分组与层级变量选择的复合绝对惩罚族。《统计年刊》37，第6A期，3468–3497，2009年

附录 A：定理 4.2 证明

证明。将数据记为 $\mathcal{X}$。定义随机向量 $\boldsymbol{U} \in \mathbb{R}^D$ 如下：

$$\boldsymbol{U} := (U_1^1, \dots, U_{T_1}^1, U_1^2, \dots, U_{T_2}^2, \cdots, U_1^M, \dots, U_{T_M}^M)^\top .$$

我们注意到以下一系列不等式，它们都由期望的线性性得出：这里，我们用 $\boldsymbol{\beta}$ 表示模型的真实系数向量。于是有

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid X\right] \\
&= \mathbb{E}\left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \mid X\right] + \mathbb{E}\left[\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \mid X\right] \\
&\qquad\qquad + 2\mathbb{E}\left[\langle \boldsymbol{y} - X\boldsymbol{\beta}, X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}} \rangle \mid X\right] \\
&= \mathbb{E}\left[\|\boldsymbol{U}\|_2^2 \mid X\right] + \mathbb{E}\left[\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \mid X\right] \\
&\qquad\qquad + 2\mathbb{E}\left[\langle \boldsymbol{U}, X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}} \rangle \mid X\right] \\
&= T_1\sigma_1^2 + \cdots + T_M\sigma_M^2 + \mathbb{E}\left[\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \mid X\right] + 2\langle \mathbb{E}\left[\boldsymbol{U} \mid X\right], X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}} \rangle \\
&= T_1\sigma_1^2 + \cdots + T_M\sigma_M^2 + \|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 ,
\end{aligned}$$

ˆ其中，最后一个等式由 $\mathbb{E}[\boldsymbol{U} \mid X] = \boldsymbol{U}$ 和 $\boldsymbol{U} \perp (X\boldsymbol{\beta} - X\boldsymbol{\beta})$ 得出。因此，我们需要推导误差 $\|X(\boldsymbol{\beta} - \hat{}\boldsymbol{\beta})\|_2^2$ 的估计。我们注意到，

$$\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 = \sum_{m=1}^{M} \sum_{t=1}^{T_m} \left(\sum_{l=1}^{L} \left(\hat{\beta}_l^m - \boldsymbol{\beta}_l^m \right) x_{t-l}^m \right)^2 ,$$

ˆ其中 $\{\boldsymbol{\beta}_l^m\}$ 表示模型的真实参数。利用 $\boldsymbol{\beta}$ 是 $\frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{l=1}^{L} \mathcal{N}(\boldsymbol{\alpha})$在 $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$上的极小值点这一事实，可得

$$\begin{aligned}
&\tfrac{1}{D}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 + \lambda \mathcal{N}(\hat{\boldsymbol{\beta}}) \le \frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda \mathcal{N}(\boldsymbol{\alpha}) \\
\Rightarrow \quad &\tfrac{1}{D}\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \le \frac{1}{D}\|X(\boldsymbol{\beta} - \boldsymbol{\alpha})\|_2^2 + \frac{2}{D}\langle \boldsymbol{U}, X(\hat{\boldsymbol{\beta}} - \boldsymbol{\alpha})\rangle + \lambda \mathcal{N}(\boldsymbol{\alpha}) - \lambda \mathcal{N}(\hat{\boldsymbol{\beta}}) \\
&\le \frac{1}{D}\|X(\boldsymbol{\beta} - \boldsymbol{\alpha})\|_2^2 + \frac{2}{D}\langle X^\top \boldsymbol{U}, \hat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\rangle + \lambda \mathcal{N}(\boldsymbol{\alpha}) - \lambda \mathcal{N}(\hat{\boldsymbol{\beta}})
\end{aligned}$$

对所有 $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$均成立。由柯西-施瓦茨不等式可知

$$\begin{aligned}
\frac{1}{D}\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 &\le \frac{1}{D}\|X(\boldsymbol{\beta} - \boldsymbol{\alpha})\|_2^2 + \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})\mathcal{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\alpha}) + \lambda(\mathcal{N}(\boldsymbol{\alpha}) - \mathcal{N}(\hat{\boldsymbol{\beta}})) \\
&\le \frac{1}{D}\|X(\boldsymbol{\beta} - \boldsymbol{\alpha})\|_2^2 + \lambda \mathcal{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\alpha}) + \lambda(\mathcal{N}(\boldsymbol{\alpha}) - \mathcal{N}(\hat{\boldsymbol{\beta}}))
\end{aligned}$$

因为 $\lambda \ge \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$。最后，由三角不等式 $\mathcal{N}(\hat{}\boldsymbol{\beta} - \boldsymbol{\alpha}) \le \mathcal{N}(\hat{}\boldsymbol{\beta}) + \mathcal{N}(\boldsymbol{\alpha})$可证得定理的第一部分。

接下来，我们将为调节参数 λ选择一个略有不同的阈值，ˆ如定理陈述中所提出的。令 $\boldsymbol{v} := \boldsymbol{\beta} + \boldsymbol{\beta}$，并注意到真实滞后阶数 L_0满足

$$L_0 = \max\{\ell \in [L] : \|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} \ne 0\} .$$

我们有

$$\begin{aligned}
\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}}) &= \mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\boldsymbol{\beta} + \boldsymbol{v}) \\
&= \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} + \sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|} \cdot \|\cancelto{0}{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell} \\
&\quad - \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{\beta} + \boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|} \cdot \|\cancelto{0}{\boldsymbol{\beta}} + \boldsymbol{v}\|_{\mathcal{G}_\ell} \\
&= \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} - \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{\beta} + \boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{v}\|_{\mathcal{G}_\ell} \\
&\le \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{v}\|_{\mathcal{G}_\ell}, \qquad (42)
\end{aligned}$$

其中最后一步是由三角不等式得出的：$\|\boldsymbol{\beta} + \boldsymbol{v}\|_{\mathcal{G}_\ell} \ge \|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} - \|\boldsymbol{v}\|_{\mathcal{G}_\ell}$。假设 $\lambda \ge \frac{4}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$。利用 ˆ $\boldsymbol{\beta}$ 是目标函数 $\frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda \mathcal{N}(\boldsymbol{\alpha})$ 在 $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$ 上的最小值点这一事实，可得

$$\begin{aligned}
&\tfrac{1}{D}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 + \lambda \mathcal{N}(\hat{\boldsymbol{\beta}}) \le \frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \mathcal{N}(\boldsymbol{\beta}) \\
\Rightarrow \quad &\tfrac{1}{D}\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \le \frac{2}{D}\langle X^\top \boldsymbol{U}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle + \lambda \mathcal{N}(\boldsymbol{\beta}) - \lambda \mathcal{N}(\hat{\boldsymbol{\beta}}) .
\end{aligned}$$

根据柯西-施瓦茨不等式，可得

$$\begin{aligned}\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 &\le \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\lambda(\mathcal{N}(\boldsymbol{\beta})-\mathcal{N}(\hat{\boldsymbol{\beta}})) \\ &\le \frac{\lambda}{2}\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\lambda(\mathcal{N}(\boldsymbol{\beta})-\mathcal{N}(\hat{\boldsymbol{\beta}})) \qquad \because\ \lambda \ge \frac{4}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \\ &\le \frac{\lambda}{2}\min\left\{3\mathcal{N}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), 3\mathcal{N}(\boldsymbol{\beta})-\mathcal{N}(\hat{\boldsymbol{\beta}})\right\}. \end{aligned} \tag{43}$$

不等式

$$0 \le \frac{2}{\lambda D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 \le \mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+2(\mathcal{N}(\boldsymbol{\beta})-\mathcal{N}(\hat{\boldsymbol{\beta}}))\,,$$

and

$$\mathcal{N}(\boldsymbol{\beta})-\mathcal{N}(\hat{\boldsymbol{\beta}}) \le \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell}$$

得到如下结果:

$$\begin{aligned} 0 &\le \mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+2(\mathcal{N}(\boldsymbol{\beta})-\mathcal{N}(\hat{\boldsymbol{\beta}})) \\ &\le \mathcal{N}(\hat{\boldsymbol{v}})+2\sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} - 2\sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} \\ &= 3\sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell}\,. \end{aligned}$$

也就是说，我们有

$$\sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell} \le 3\sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell}\,. \tag{44}$$

根据方程（43），我们可得

$$\begin{aligned}\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 &\le 2\lambda\sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell} \\ &= 2\lambda\mathcal{N}_{\le L_0}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\,. \end{aligned} \tag{45}$$

□

注释（A）。如果我们不取 $\mathcal{N}$，而是在 $\mathbb{R}^{ML}$ 中取标准 ℓ^2（对应于 $\mathcal{N}$ 中最大的组），那么导出式 (44) 的论证可得

$$\begin{aligned}\sum_{\ell=L_0+1}^{L}|\hat{\boldsymbol{\beta}}_\ell^m| &= \sum_{\ell=L_0+1}^{L}|\boldsymbol{\beta}_\ell^m-\hat{\boldsymbol{\beta}}_\ell^m| \\ &\le 3\sum_{\ell=1}^{L_0}|\boldsymbol{\beta}_\ell^m-\hat{\boldsymbol{\beta}}_\ell^m|\,. \end{aligned}$$

附录 B：定理 4.3 的证明

证明。由于引理 D.3 推出

$$\begin{aligned}\mathcal{N}_\star\left(\frac{2}{D}X^\top \boldsymbol{U}\right) &\le \frac{1}{\sqrt{L}}\left\|\frac{2}{D}X^\top \boldsymbol{U}\right\|_\infty \\ &= \frac{1}{\sqrt{L}}\max_{m\in[M]}\max_{l\in[L]}\left|\frac{2}{D}\sum_{t=1}^{T_m} U_t^m X_{t-l}^m\right|,\end{aligned}$$

由并集界论证可知，只需求出内积的高概率上界（记号见 (16)）

$$\tau_{m,l} := 2\langle \boldsymbol{U}^m, X^{m,l}\rangle = 2\left(\sum_{t=1}^{T_m} U_t^m X_{t-l}^m\right). \tag{46}$$

现在我们介绍一些标准技巧，例如 Basu 和 Michailidis（2015）中使用的方法。由此可得

$$\tau_{m,l} = \left\|\boldsymbol{U}^{(m)} + X^{(m,l)}\right\|_2^2 - \left\|\boldsymbol{U}^{(m)}\right\|_2^2 - \left\|X^{(m,l)}\right\|_2^2.$$

因此，对任意 $\eta>0$，结合并集界论证以及高斯随机变量 U_j^m 与对每个 $j\in[T_m]$ 而言的变量 $X_{<j}^m$ 相互独立，可得

$$\begin{aligned}&\mathbb{P}\left[\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right] \\ &\le \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)})^\top(\boldsymbol{U}^{(m)}) - \mathrm{Var}(U^m)\right| > \frac{D}{2}\eta\right] + \mathbb{P}\left[\left|(X^{(m,l)})^\top(X^{(m,l)}) - \mathrm{Var}(X^m)\right| > \frac{D}{2}\eta\right] \\ &\qquad + \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)} + X^{(m,l)})^\top(\boldsymbol{U}^{(m)} + X^{(m,l)}) - \mathrm{Var}(X^m + U^m)\right| > \frac{D}{2}\eta\right].\end{aligned}$$

这里，$\mathrm{Var}(Y)$ 表示随机变量 Y 的协方差矩阵的迹。现在我们分别估计每一项，从以下项开始

$$p_{m,l,\eta} := \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)})^\top(\boldsymbol{U}^{(m)}) - \mathrm{Var}(U^{(m)})\right| > \frac{D}{2}\eta\right].$$

根据现有假设，向量 $\boldsymbol{U}^{(m)}$ 是 T_m 维零均值高斯分布，协方差矩阵为 $\sigma_m^2\mathbb{I}_{T_m}$；由命题 E.1 中的不等式可得

$$p_{m,l,\eta} \le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2}\min\{1,\frac{D\eta}{8\sigma_m^2 T_m}\}}. \tag{47}$$

接下来，我们考虑

$$q_{m,l,\eta} := \mathbb{P}\left[\left|(\boldsymbol{X}^{(m,l)})^\top(\boldsymbol{X}^{(m,l)}) - \mathrm{Var}(X^{(m,l)})\right| > \frac{D}{2}\eta\right].$$

回顾一下，随机向量 $X^{(m,l)}$ 服从零均值高斯分布，其协方差矩阵为下文公式 (59) 中提到的 $\Gamma^{(m)}$。因此

$$\mathrm{Var}(X^{(m,l)}) = \mathrm{tr}(\Gamma^{(m)}) = T_m\mathbb{E}[(X_0^m)^2].$$

由引理 E.2 可得 $\|\Gamma^{(m)}\|_{\mathrm{op}} \le \epsilon^{-2}\sigma_m^2$；因此，命题 E.1 得出

$$q_{m,l,\eta} \le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2\epsilon^{-2}}\min\{1,\frac{D\eta}{8\sigma_m^2 T_m\epsilon^{-2}}\}}. \tag{48}$$

最后，我们考虑

$$r_{m,l,\eta} := \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)} + X^{(m,l)})^\top(\boldsymbol{U}^{(m)} + X^{(m,l)}) - \mathrm{Var}(X^m + U^m)\right| > \frac{D}{2}\eta\right].$$

注意， $\boldsymbol{U}^{(m)} + X^{(m,l)}$ 是一个均值为零、协方差矩阵为对称矩阵 $\tilde{\Gamma}^{(m)}$ 的高斯变量，其(t,s)项（对于 $t \geq s$）为

$$\begin{aligned}\tilde{\Gamma}^{(m)}_{t,s} &:= \mathbb{E}[(X^m_{t-l} + U^m_t)(X^m_{s-l} + U^m_s)] \\ &= \mathbb{E}[(X^m_{t-l}X^m_{s-l}] + \mathbb{E}[X^m_{s-l}U^m_t] + \mathbb{E}[X^m_{t-l}U^m_s] + \mathbb{E}[U^m_t U^m_s] \\ &= \mathbb{E}[(X^m_{t-l}X^m_{s-l}] + \mathbb{E}[X^m_{t-l}U^m_s] + \sigma_m^2 \mathbf{1}_{t=s}\,.\end{aligned}$$

结合引理 E.3，可得

$$\begin{aligned}\|\tilde{\Gamma}^{(m)}\|_{\mathrm{op}} &\leq \epsilon^{-2}\sigma_m^2 + \epsilon^{-4}\sigma_m^2 + \sigma_m^2 \\ &= \sigma_m^2(1 + \epsilon^{-2} + \epsilon^{-4})\,,\end{aligned}$$

根据命题 E.1，可得

$$r_{m,l,\eta} \leq 2e^{-\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}\min\{1, \frac{D\eta}{8\sigma_m^2 T_m(1+\epsilon^{-2}+\epsilon^{-4})}\}}. \tag{49}$$

现在假设

$$\eta \geq \frac{8C_\sharp\sigma^2_{\max}(1 + \epsilon^{-2} + \epsilon^{-4})}{M};$$

那么，不等式 (47)、(48) 和 (49) 可简化如下:

$$\begin{aligned} p_{m,l,\eta} &\leq 2e^{-\frac{c_0 D\eta}{8\sigma_m^2}}; \\ q_{m,l,\eta} &\leq 2e^{-\frac{c_0 D\eta}{8\sigma_m^2\epsilon^{-2}}}, \\ r_{m,l,\eta} &\leq 2e^{-\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}}. \end{aligned} \tag{50}$$

其中，最下面的不等式 (50) 右侧的值最大。注意，并集界意味着

$$\begin{aligned}\mathbb{P}\left[\sup_{m,l}\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right] &\leq \sum_{m,l}\mathbb{P}\left[\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right] \\ &\leq \sum_{l\in[L]}\sum_{m\in[M]}\mathbb{P}\left[\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right] \\ (50)\ \Rightarrow \qquad &\leq 6\sum_{l\in[L]}\sum_{m\in[M]} e^{-\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}}.\end{aligned}$$

因此，对于任意 $\delta > 0$，为了满足不等式

$$\mathbb{P}\left[\mathcal{N}_\star\left(\frac{2}{D}X^\top\boldsymbol{U}\right) > \frac{3\eta}{2\sqrt{L}}\right] \leq \delta\,,$$

只需满足

$$\exp\left(-\frac{c_0 D\eta}{8\sigma_m(1 + \epsilon^{-2} + \epsilon^{-4})}\right) \leq \frac{\delta}{ML} \tag{51}$$

对每个 $m \in [M]$ 都成立。等价地，只需对每个 $m \in [M]$ 满足以下条件:

$$\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})} \geq \log\left(\frac{ML}{\delta}\right)$$

$$\text{equivalently,} \qquad D \geq \frac{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right).$$

这等价于

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right). \tag{52}$$

由上述论证可得

$$\mathbb{P}\left[\frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \geq \frac{3\eta}{2\sqrt{L}}\right] \;\leq\; \delta \tag{53}$$

在.....成立的条件下,

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right).$$

□

附录 C：命题 4.6 的证明

我们首先观察 $M=1$ 情形下的情况：这相当于将 X 限制在左上方块

$$X_1 := \begin{pmatrix} x_{1-1}^1 & \cdots & x_{1-L}^1 \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ x_{T_1-1}^1 & \cdots & x_{T_1-L}^1 \end{pmatrix}.$$

引理 C.1（分块集中不等式）。对于任意整数 $s_1 > 0$，记作$\mathbb{K}(s_1) := \mathbb{B}_0(s_1) \cap \mathbb{B}_1(1)$。则

$$\mathbb{P}\left[\sup_{\boldsymbol{v}_1 \in \mathbb{K}(s_1)} \left|\boldsymbol{v}_1^\top (X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \geq \frac{T_1^2\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{2\mathfrak{m}(f)}\right]$$

$$\leq 2\exp\left(-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2} + s_1\min\{\log L,\ \log(21eL/s_1)\}\right). \tag{54}$$

引理 C.1 的证明。设 $\boldsymbol{v}_1 \in \mathbb{R}^L$ 是一个固定的单位范数向量。那么， $\boldsymbol{u}_1 := X_1\boldsymbol{v}_1$ 是 $\mathbb{R}^{T_1}$ 中的随机零均值高斯向量，其协方差矩阵为 $\Sigma_{\boldsymbol{v}_1}^{X_1} := \mathbb{E}\left[X_1\boldsymbol{v}_1\boldsymbol{v}_1^\top X_1^\top\right]$。有

$$\begin{aligned}(\Sigma_{\boldsymbol{v}_1}^{X_1})_{r,s} &= \sum_{j,l=1}^{L}\mathbb{E}\left[X_{r-j}^1\boldsymbol{v}_{1,j}\boldsymbol{v}_{1,l}X_{s-l}^1\right] \\ &= \sum_{j,l=1}^{L}\boldsymbol{v}_{1,j}\mathbb{E}\left[X_{r-j}^1X_{s-l}^1\right]\boldsymbol{v}_{1,l} \\ &= \boldsymbol{v}_1^\top\Gamma_{r,s}\boldsymbol{v}_1\,,\end{aligned}$$

其中，$\Gamma_{r,\ s}$ 是向量 $X_r^{(1)}$ 和 $X_s^{(1)}$ 的协方差。注意，

$$\begin{aligned}\mathrm{tr}(\Sigma_{\boldsymbol{v}_1}^{X_1}) &= \mathbb{E}\left[\mathrm{tr}(X_1\boldsymbol{v}_1\boldsymbol{v}_1^\top X_1^\top)\right]\\ &= T_1\mathbb{E}\left[(X_1\boldsymbol{v}_1)_{11}^2\right]\\ &= T_1\boldsymbol{v}_1^\top\Gamma^{(1)}\boldsymbol{v}_1\,,\end{aligned}$$

其中，$\Gamma^{(1)}$ 是 $\{X_t^1\}$的自协方差矩阵。此外，根据引理 E.2，有如下不等式

$$\|\Sigma_{\boldsymbol{v}_1}^{X_1}\|_{\mathrm{op}} \le \frac{\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\,.$$

根据命题 E.1 中的不等式，有

$$\mathbb{P}\left[\left|\boldsymbol{u}_1^T\boldsymbol{u}_1 - \mathrm{tr}(\Sigma_{\boldsymbol{v}_1}^{X_1})\right| \ge 4T_1\zeta_1\|\Sigma_{\boldsymbol{v}_1}^{X_1}\|_{\mathrm{op}}\right] \le 2e^{-\frac{T_1}{2}\min\{\zeta_1,\zeta_1^2\}}$$

对任意 $\zeta_1>0$均成立。特别地，这意味着对于任意固定的 $\boldsymbol{v}_1\in\mathbb{R}^L$，该不等式 y

$$\mathbb{P}\left[\left|\boldsymbol{v}_1^\top(X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \ge \frac{4T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\right] \le 2e^{-\frac{T_1}{2}\min\{\zeta_1,\zeta_1^2\}} \tag{55}$$

h对任意 $\zeta_1>0$均成立。将 $G := X_1^\top X_1 - T_1\Gamma^{(1)}$代入引理 E.4，可得 at

$$\begin{aligned}&\mathbb{P}\left[\sup_{\boldsymbol{v}_1\in\mathbb{K}(s_1)}\left|\boldsymbol{v}_1^\top(X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \ge \frac{4T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\right]\\ \le& 2\exp\left(-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2} + s_1\min\{\log L,\ \log(21eL/s_1)\}\right)\end{aligned} \tag{56}$$

对任意整数 $s_1\ge 1$和任意 $\zeta_1>0$均成立。 □

命题 4.6 的证明。

证明。当 $s_1>0$ 为偶数时，由引理 E.5 可得

$$\begin{aligned}&\mathbb{P}\left[\sup_{\boldsymbol{v}_1\in\mathbb{R}^L}\left|\boldsymbol{v}_1(X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \le \frac{108T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\left(\|\boldsymbol{v}_1\|_2^2 + \frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\right]\\ \ge\ & 1 - 2\exp\left(-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2} + s_1\min\{\log L,\ \log(21eL/s_0)\}\right)\,.\end{aligned}$$

应用引理 E.2 可知，对于

$$\delta := 2e^{-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2} + s_1\min\{\log L,\ \log(21eL/s_1)\}}\,,$$

以下不等式序列的结论对所有 $\boldsymbol{v}_1\in\mathbb{R}^L$ 均以至少 $1-\delta$ 的概率成立：

$$\begin{aligned}\|X_1\boldsymbol{v}_1\|_2^2 &\ge T_1\lambda_{\min}(\Gamma^{(1)})\|\boldsymbol{v}_1\|_2^2 - \frac{108T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\left(\|\boldsymbol{v}_1\|_2^2 + \frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\\ &\ge \frac{T_1\Lambda_{\min}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{M}(f)}\|\boldsymbol{v}_1\|_2^2 - \frac{108T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\left(\|\boldsymbol{v}_1\|_2^2 + \frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\\ &= T_1\left(\frac{\Lambda_{\min}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{M}(f_1)} - \frac{108\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f_1)}\right)\|\boldsymbol{v}_1\|_2^2 - \frac{216T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{s_1\mathfrak{m}(f_1)}\|\boldsymbol{v}_1\|_1^2\,.\end{aligned}$$

我们专门讨论单变量 ϵ-稳定自回归时间序列的情形，其中 $\Sigma_{\boldsymbol{U}_1} = \sigma_1^2$，并且（根据 ϵ-稳定性）

$$\epsilon^2 \le \mathfrak{m}(f_1) \le \mathfrak{M}(f_1) \le \epsilon^{-2}.$$

令 $\zeta_1 := 6^{-3}\epsilon^4$，我们得到

$$\mathbb{P}\left[\inf_{\boldsymbol{v}_1 \in \mathbb{R}^L} \left(\|X_1 \boldsymbol{v}_1\|_2^2 - \frac{T_1}{2}\sigma_1^2\epsilon^2 \left(\|\boldsymbol{v}_1\|_2^2 - \frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2 \right) \right) \ge 0 \right]$$

$$\ge 1 - 2e^{-\frac{T_1}{2}\zeta_1^2 + s_0 \min\{\log L,\ \log(21eL/s_0)\}}. \tag{57}$$

最后，我们考虑 $X^\top X$ 的所有块；上述偏差不等式给出

$$\mathbb{P}\left[\forall_{m\in[M]} \inf_{\boldsymbol{v}_m \in \mathbb{R}^L} \left(\|X_m \boldsymbol{v}_m\|_2^2 - \frac{T_m\sigma^2\epsilon^2}{2} \left(\|\boldsymbol{v}_m\|_2^2 - \frac{2}{s_m}\|\boldsymbol{v}_m\|_1^2 \right) \right) \ge 0 \right]$$

$$\ge 1 - 2\sum_{m=1}^{M} e^{-\frac{T_m\zeta^2}{2} + s_m \min\{\log L,\ \log(21eL/s_m)\}}. \tag{58}$$

这就证明了该命题。 □

附录 D：关于对偶范数的补充

由于 $\mathcal{N}_\star(\mathbf{0}) = 0$，在以下对偶范数的计算中，我们将假设 $\boldsymbol{\alpha} \neq \mathbf{0}$。这也保证了 $\mathcal{N}_\star(\boldsymbol{\alpha}) > 0$。注意，$\mathcal{N}_\star(\boldsymbol{\alpha}) = \mathcal{N}_\star(|\boldsymbol{\alpha}|)$，其中 $|\boldsymbol{\alpha}| = (|\alpha_1|, \cdots, |\alpha_L|)$；这是因为 $\mathcal{N}$ 在对其自变量的坐标进行任意符号变换时保持不变。因此，对于 $\mathcal{N}_\star(\boldsymbol{\alpha}) = \boldsymbol{\alpha} \cdot \boldsymbol{\beta}$，不失一般性，我们可以假设 $|\boldsymbol{\alpha}| = \boldsymbol{\alpha}$ 和 $|\boldsymbol{\beta}| = \boldsymbol{\beta}$。

层次组范数 $\mathcal{N}(\boldsymbol{\beta})$ 的优点在于，在通过正则化最小二乘回归进行稀疏恢复时，它会将任意一组参数置为零；由于组结构存在层次性，层次顺序中位于更下方的所有组中的变量也会自动置为零。因此，这些范数非常适合用于确定自回归过程中的真实滞后阶数。

为了在分析中使用这种组范数，我们需要先了解该范数的一些基本性质。注意，该范数在组层面上类似于 ℓ^1，而在每个组内则是 ℓ^2范数；由此可以预期，该范数的对偶范数在组层面上应当“类似于” ℓ^∞-范数，而在组内则应为 ℓ^2。我们将证明，这一直觉相当准确：实际的对偶范数可以由这种混合 $\ell^\infty,\ _2$范数给出上界。

引理 D.1（对偶范数在边界上取得）。

$$\mathcal{N}_\star(\boldsymbol{\alpha}) = \max_{\mathcal{N}(\boldsymbol{\beta})=1} \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle.$$

引理 D.1 的证明。注意，由于 $\boldsymbol{\beta} \mapsto \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$ 的连续性、子集 $C_{\boldsymbol{\beta}} := \{\boldsymbol{\beta} :\ \mathcal{N}(\boldsymbol{\beta}) \le 1\}$ 的紧致性，以及 $\mathcal{N}_\star(\boldsymbol{\alpha}) \neq 0$，存在非零的 $\boldsymbol{\beta}_0 \in C_{\boldsymbol{\beta}}$，使得 $\mathcal{N}_\star(\boldsymbol{\alpha}) = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_0 \rangle$。假设（若可能）$\boldsymbol{\beta}_0$ 满足 $\mathcal{N}(\boldsymbol{\beta}_0) < 1$。在 $(0, +\infty)$ 中，

函数 $c \mapsto \mathcal{N}(c\boldsymbol{\beta}_0)$ 是严格递增的（连续）函数；因此，存在 $c>1$，使得 $c\mathcal{N}(\boldsymbol{\beta}_0)=\mathcal{N}(c\boldsymbol{\beta}_0) \leq 1$；于是有

$$\begin{aligned}\max\{\langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle : \mathcal{N}(\boldsymbol{\beta}) \leq 1\} &\geq \langle \boldsymbol{\alpha}, c\boldsymbol{\beta_0}\rangle \\ &= c \cdot \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_0\rangle \\ &> \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_0\rangle \\ &= \mathcal{N}_\star(\boldsymbol{\alpha}),\end{aligned}$$

这就产生了矛盾。 □

为了便于计算 $\mathbb{R}^{ML}$ 中的对偶范数，我们先考虑 $M=1$ 的特殊情形；在命题 D.2 之后，我们将把这一结果推广到一般情形。现在，我们定义

$$\mathcal{N}^1(\boldsymbol{\alpha}) := \sum_{l=1}^{L} \sqrt{(L-l+1)\sum_{j=l}^{L} \alpha_j^2}\,.$$

我们将 $\boldsymbol{\alpha}_{\geq j} = (\mathbf{0},\ \alpha_j,\ \alpha_{j+1},\ \cdots,\ \alpha_L)$ 记作 $j \in [L]$。从现在起，我们假设

$$\prod_{j\in[L]} \alpha_j \neq 0\,.$$

我们有如下命题。

命题 D.2（L^∞ 范数界定对偶范数，$M=1$ 情形）。对于 $\boldsymbol{\alpha} \in \mathbb{R}^L$，以下不等式成立：

$$\mathcal{N}_\star^1(\boldsymbol{\alpha}) \leq L^{-\frac{1}{2}} \|\boldsymbol{\alpha}\|_\infty\,.$$

命题 D.2 的证明。我们应用以下初等不等式

$$\begin{aligned}(\sqrt{L-l+1}\|\beta_{\geq l}\|_2 - \sqrt{L-l+1}|\beta_l|)^2 &\geq (L-l+1)\|\beta_{\geq l}\|_2^2 - (L-l+1)|\beta_l|^2 \\ &= (L-l+1)\sum_{j>l}\beta_j^2 \\ &= \left(\sqrt{(L-l+1)\sum_{j>l}\beta_j^2}\right)^2 \\ \text{Cauchy-Schwarz } \Rightarrow \qquad &\geq \left(\sum_{j>l}\beta_j\right)^2,\end{aligned}$$

推导

$$\begin{aligned}\mathcal{N}^1(\boldsymbol{\beta}) &= \sum_{l=1}^{L}\sqrt{L-l+1}\|\beta_{\geq l}\|_2 \\ &\geq \sum_{l=1}^{L}\left(\sqrt{L-l+1}|\beta_l| + \sum_{j>l}|\beta_j|\right) \\ &= \sum_{l=1}^{L}\left(\sqrt{L-l+1}+l-1\right)|\beta_l| \\ &\geq \sqrt{L}|\beta_l|\,.\end{aligned}$$

由于 $\mathcal{N}_\star^1(\boldsymbol{\alpha}) \le \|\boldsymbol{\alpha}\|_\infty \mathcal{N}^1(\mathbf{1})$，其中 $\mathbf{1} \in \mathbb{R}^L$ 是全 1 向量，因此只需证明 $\mathcal{N}_\star^1(\mathbf{1}) \le L^{-\frac{1}{2}}$。由于

$$\mathcal{N}_\star^1(\mathbf{1}) = \max_{\mathcal{N}^1(\boldsymbol{\beta})=1} \sum_{j\in[L]} \beta_j \,,$$

and

$$\begin{aligned} 1 &= \mathcal{N}^1(\boldsymbol{\beta}) \\ &= \sum_{j\in[L]} \sqrt{L-l+1} \|\beta_{\ge j}\|_2 \\ &\ge \sqrt{L} \sum_{j\in[L]} |\beta_j| \,, \end{aligned}$$

该结论得证。 □

结合上述针对 $M=1$ 情形的命题，我们可以一般性地提出以下命题。

命题 D.3（L^∞ 范数界定对偶范数，一般情形）。假设 $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$，并且

$$\mathcal{N}(\boldsymbol{\alpha}) = \sum_{l=1}^{L} \sqrt{M(L-l+1)} \|\boldsymbol{\alpha}_{(\ge l)}\|_2 \,,$$

$$\text{其中} \qquad \|\boldsymbol{\alpha}_{(\ge l)}\|_2 := \sqrt{\sum_{j=l}^{L} \sum_{m=1}^{M} \alpha_{(m-1)+j}^2} \,.$$

则 $\mathcal{N}_\star(\boldsymbol{\alpha}) \le L^{-\frac{1}{2}} \|\boldsymbol{\alpha}\|_\infty$。

命题 D.3 的证明。再次，只需证明当 $\boldsymbol{\alpha}$ 为全 1 向量时，$\mathcal{N}(\boldsymbol{\alpha}) \le L^{-\frac{1}{2}}$。

现在，$\mathcal{N}_\star(\boldsymbol{\alpha})$ 是该问题的最优值

$$\begin{aligned} \text{maximize} \quad & \sum_{j\in[ML]} \beta_j \\ \text{subject to} \quad & \mathcal{N}(\boldsymbol{\beta}) = 1 \,. \end{aligned}$$

对于 $l \in [L]$，记 $\boldsymbol{\beta}_{(l)} := (\beta_l, \beta_{L+l}, \cdots, \beta_{(M-1)L+l})^\top$ 和 $x_l := \sqrt{M} \|\boldsymbol{\beta}_{(l)}\|_2$；根据柯西-施瓦茨不等式，有

$$\sum_{j\in[ML]} \beta_j \le \sum_{l=1}^{L} x_l \,.$$

此外，$\mathcal{N}(\boldsymbol{\beta}) = \mathcal{N}^1(\boldsymbol{x})$，其中 $\boldsymbol{x} := (x_1, \cdots, x_L)$；因此，根据 D.2 可得该引理。 □

附录 E：证明中用到的一些已知结果

以下关于高斯随机向量范数的尾部界是众所周知的；详情请参见（Vershynin，2018年，定理6.2.1）。

命题 E.1（Wright-Hansen 不等式）。存在一个绝对常数 $c_0 > 0$，使得以下陈述成立。若 $\boldsymbol{q} \sim N(\mathbf{0}, \Sigma)$ 属于 $\mathbb{R}^d$，其中 Σ 为对称正定矩阵，则对任意 $\tau > 0$，以下结论成立：

$$\mathbb{P}\left[\frac{1}{d}\left|\boldsymbol{q}^T\boldsymbol{q} - tr(\Sigma)\right| \geq 4\tau\|\Sigma\|_{op}\right] \leq 2e^{-c_0 d\min\{\tau^2,\tau\}}.$$

我们需要收集关于方程 (9) 中时间序列模型的一些信息。令 $T := \text{最小值}_{m\in[M]} T_m$，并回顾一下我们记作 $D = T_1 + \cdots + T_M$。令 $\Gamma^{(m,l)}$ 为随机向量 $X^{(m,l)}$ 的协方差矩阵；这是一个对称矩阵。对于满足 $r \leq s$ 的任意 $r, s \in \{1-l, \ldots, T_m - l\}$，时间序列的平稳性给出

$$\Gamma^{(m,l)}_{r,s} = \mathbb{E}[X^{(m,l)}_{r-l} X^{(m,l)}_{s-l}] = \mathbb{E}[X^m_0 X^m_{s-r}]. \tag{59}$$

因此，我们可以将 $\Gamma^{(m)} := \Gamma^{(m,l)}$ 记作“自协方差”矩阵。

现在，进一步地，对于任意两个整数 $T > L > 0$ 和 $_{T,\ L}\Gamma$ 令其表示该 $T_1 L \times T_1 L$ 矩阵，其$(r,\ s)$个分块是向量$(X_{r-1},\ X_{r-2},\ \ldots,\ X_{r-L})$与$(X_{s-1},\ X_{s-2},\ \ldots,\ X_{s-L})$的协方差。

下面的引理给出了这些分块协方差矩阵的特征值界。

引理 E.2（Toeplitz 矩阵的特征值界）。令 $f(z)$ 表示稳定单变量 AR 过程 $\{X_t\}$的逆特征多项式。令 $\mathfrak{m}(f) = \inf_{|z|\leq 1}|f(z)|^2$，且令 $\mathfrak{M}(f) = \sup_{|z|\leq 1}|f(z)|^2$。以下不等式成立：

$$\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mathfrak{M}(f)} \leq \Lambda_{\min}({}_{T,L}\Gamma) \leq \Lambda_{\max}({}_{T,L}\Gamma) \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mathfrak{m}(f)}.$$

引理 E.2 的证明。若考虑稳定的维-L VAR 过程 $Y_t := (X_{t-1},\ \ldots,\ X_{t-L})$，则该结论可直接由（Basu 和 Michailidis，2015，命题 2.3）得到。 □

以下界限见 Basu 和 Michailidis（2015，命题 2.4b）。

引理 E.3（Basu-Michailidis）。设 $\Delta^{(m)}$ 表示 T_m维矩阵，且满足$\Delta^{(m)}_{t,\ s} := \mathbb{E}[X^m_{t-l} U^m_s]$。以下不等式成立：

$$\Lambda_{\max}(\Delta^{(m)}) \leq \frac{\Lambda_{\max}(\Sigma_{\epsilon_m})\mathfrak{M}(f_m)}{\mathfrak{m}(f)}.$$

以下内容见（Basu 和 Michailidis，2015，引理 F.2）。回顾一下，$\mathbb{M}_L(\mathbb{R})$表示所有 $L \times L$ 实矩阵构成的代数，而 $\mathbb{S}^{L-1}$ 是 $\mathbb{R}^L$中的单位球面。

引理 E.4（Basu-Michailidis）。假设 $G \in \mathbb{M}_L(\mathbb{R})$ 是一个随机对称矩阵，对每个 $\boldsymbol{u} \in \mathbb{S}^{L-1}$、$T_1 \in \mathbb{N}$和 $\eta > 0$，以下不等式均成立：

$$\mathbb{P}\left[\left|\boldsymbol{u}^T G \boldsymbol{u}\right| \geq C\eta\right] \leq 2e^{-cT_1\min\{\eta,\eta^2\}},$$

其中 $C,\ c > 0$ 是与 η 和 $\boldsymbol{u}$ 相互独立的参数。于是，对任意整数 $s_0 > 0$，以下不等式成立：

$$\mathbb{P}\left[\sup_{\|\boldsymbol{u}\|_0 \leq s_0,\ \|\boldsymbol{u}\|_2 \leq 1}\left|\boldsymbol{u}^T G \boldsymbol{u}\right| \geq C\eta\right] \leq 2e^{-cT_1\min\{\eta,\eta^2\} + s_0\min\{\log L, \log(21eL/s_0)\}}. \tag{60}$$

以下结果出自（Loh 和 Wainwright，2012，引理 12）。回顾一下，对于任意整数 $s_0 > 0$，我们记

$$\mathbb{K}(s_0) = \{\boldsymbol{v} \in \mathbb{R}^L : \|\boldsymbol{v}\|_2 \leq 1,\ \|\boldsymbol{v}\|_0 \leq s_0\}\,.$$

LEMMA E.5（Loh-Wainwright）。假设 $G \in \mathbb{M}_L(\mathbb{R})$ 是一个对称矩阵，满足以下条件：

$$\sup_{\|\boldsymbol{u}\|_0 \leq s_0,\ \|\boldsymbol{u}\|_2 \leq 1} \left|\boldsymbol{u}^T G \boldsymbol{u}\right| \leq \delta\,.$$

Then

$$\sup_{\boldsymbol{u} \in \mathbb{R}^L} \left|\boldsymbol{u}^T G \boldsymbol{u}\right| \leq 27\delta \left(\|\boldsymbol{u}\|_2^2 + \frac{2}{s_0}\|\boldsymbol{u}\|_1^2\right)\,. \tag{61}$$