山东大学数学学院栾贻会老师及研究团队文章：高维变量下模型选择路径与模型置信集的构建Mathematics

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/32870.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

山东大学数学学院栾贻会老师及研究团队文章：高维变量下模型选择路径与模型置信集的构建Mathematics。



mathematics

MDPI

Article

Model Selection Path and Construction of Model Confidence Set under High-Dimensional Variables

Faguang Wen [1,2], Jiming Jiang [3] and Yihui Luan [1,*]

1 Frontiers Science Center for Nonlinear Expectations (Ministry of Education), Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China; 201711778@mail.sdu.edu.cn
2 School of Mathematics, Shandong University, Jinan 250100, China
3 Department of Statistics, University of California, Davis, CA 95616, USA; jimjiang@ucdavis.edu
* Correspondence: yhluan@sdu.edu.cn

一、文章亮点介绍：

文章给出的AMac方法对近期提出的模型置信集 (Model Confidence Set, MCS) 构造方法Mac进行了改进，其改善了Mac方法在小样本及大噪声下的有限样本表现。另外，针对高维变量引起备选模型集过大致使现有模型置信集构造方法计算失效的问题，文章提出了模型选择路径 (Model Selection Path, MSP) 的概念以高效缩减备选模型集，并在线性模型框架下给出了MSP的构造算法。试验表明MSP+现有MCS构造方法的两步法可在高维情形下高效构造满足给定置信水平的MCS。

二、文章介绍

1. 研究背景及目的

实际应用中，当研究问题有多个可选模型时，常用的做法是利用某一模型选择准则 (例如：BIC准则，AIC准则等) 从备选模型集中挑选出一个最佳的模型用于建模分析。尽管模型

选择在理论及方法论上取得了很多成果，然而在数据信息不足时，数据的微小扰动(增删一些数据) 会造成选择过程确定的最优模型不断变化，这便是模型选择的不确定性问题。在传统的数据分析中，数据分析人员往往忽略模型选择过程的不确定性，这往往会造成过于理想化的估计或预测结果。

考虑到单个模型结果存在较大不确定性，学者们更期望结合多个好模型的结果给出更有价值的结论。MCS是一个具有给定置信水平以包含最佳模型的多个备选模型的集合，反映了最佳模型被一些模型包含的概率大小，类似于参数估计的置信区间，可以给出模型选择不确定性的度量，同时综合MCS中模型的结果可以得出更加可靠的结论。

随着模型选择不确定性问题关注度的上升，近年来MCS的构造方法也引起了学者们的广泛研究。尽管，近年来先后出现了MCS、LRT、LMB、Mac等一系列优秀方法，但在MCS构造中依然存在如下关键问题：(1) 虽然Mac方法构造出的模型置信集较于其他方法有着模型集更小的优点，然而Mac在小样本和大噪声下的表现仍有待提高。(2) 已有方法在可选模型集较大时均因计算时间过长而难以实施，这在解释变量维数较高时极易发生。基于以上两个MCS构造中的关键问题，文章提出了AMac方法对Mac的有限样本表现进行改善，同时提出MSP的概念用于解决高维变量下MCS的计算难题。

2. 算法及结果展示

AMac算法图：

爱科学
iikx.com

Algorithm 1 Constructing MCS using AMac

1: Using the data $(\boldsymbol{X}, \boldsymbol{Y})$ and a criterion f, perform parameter estimation and ranking of the models to obtain ordered models $\mathcal{M}_f = \{\hat{M}_1^0, ..., \hat{M}_N^0\}$ and their corresponding parameters $(\hat{M}_i^0, \hat{\psi}_i^0), i = 1, ..., N$.

2: Choose a set of probability weights a_i for models $\hat{M}_i^0, i = 1, ..., N$, satisfying $\sum_{i=1}^{N} a_i = 1$.

3: Keep the data $\boldsymbol{X}$ unchanged, and generate new data $\boldsymbol{Y}_i^{[b]}, b = 1, ..., B, i = 1, ..., N$ under the models $(\hat{M}_i^0, \hat{\psi}_i^0), i = 1, ..., N$, respectively.

4: Calculate the values of $f_M(\boldsymbol{X}, \boldsymbol{Y})$ for each candidate model in the candidate model set $\mathcal{M}$ using $(\boldsymbol{X}, \boldsymbol{Y}_i^{[b]})$ and sort them in ascending order:

$$f_{\hat{M}_{i,1}^{[b]}}(\boldsymbol{X}, Y_i^{[b]}) \leq f_{\hat{M}_{i,2}^{[b]}}(\boldsymbol{X}, Y_i^{[b]}) \leq \cdots \leq f_{\hat{M}_{i,N}^{[b]}}(\boldsymbol{X}, Y_i^{[b]}).$$

Resulting in an ordered sequence of models: $\hat{M}_{i,1}^{[b]}, \hat{M}_{i,2}^{[b]}, \ldots, \hat{M}_{i,N}^{[b]}$.

5: Repeat steps 3 and 4 for $i = 1, 2, \ldots, N$ and $b = 1, 2, \ldots, B$.

6: Calculate the empirical probabilities of $\hat{M}_{i,1}^* = \hat{M}_i^0, \hat{M}_{i,2}^* = \hat{M}_i^0, \ldots, \hat{M}_{i,N}^* = \hat{M}_i^0$:

$$P(\hat{M}_{i,l}^* = \hat{M}_i^0) = \frac{1}{B} \sum_{b=1}^{B} I_{(\hat{M}_{i,l}^{[b]} = \hat{M}_i^0)} \quad l, i = 1, ..., N.$$

7: According to Equation (6), calculate for any $l \in \{1, \cdots, N\}$:

$$\begin{aligned} P(\hat{M}_l^0 = M_{opt}) &\approx P(\hat{M}_l^* = M_{opt}) \\ &\approx P(\hat{M}_{1,l}^* = \hat{M}_1^0)a_1 + P(\hat{M}_{2,l}^* = \hat{M}_2^0)a_2 + \cdots + P(\hat{M}_{N,l}^* = \hat{M}_N^0)a_N. \end{aligned}$$

Obtain

$$P(\mathcal{M}_f) = \left\{ P(M_{opt} = \hat{M}_1^0), P(M_{opt} = \hat{M}_2^0), \cdots, P(M_{opt} = \hat{M}_N^0) \right\}.$$

8: Based on the resorting of candidate models in descending order of $P(\mathcal{M}_f)$, we obtain

$$\mathcal{M}_{ff} = \{\tilde{M}_1^0, \tilde{M}_2^0, \ldots, \tilde{M}_N^0\}.$$

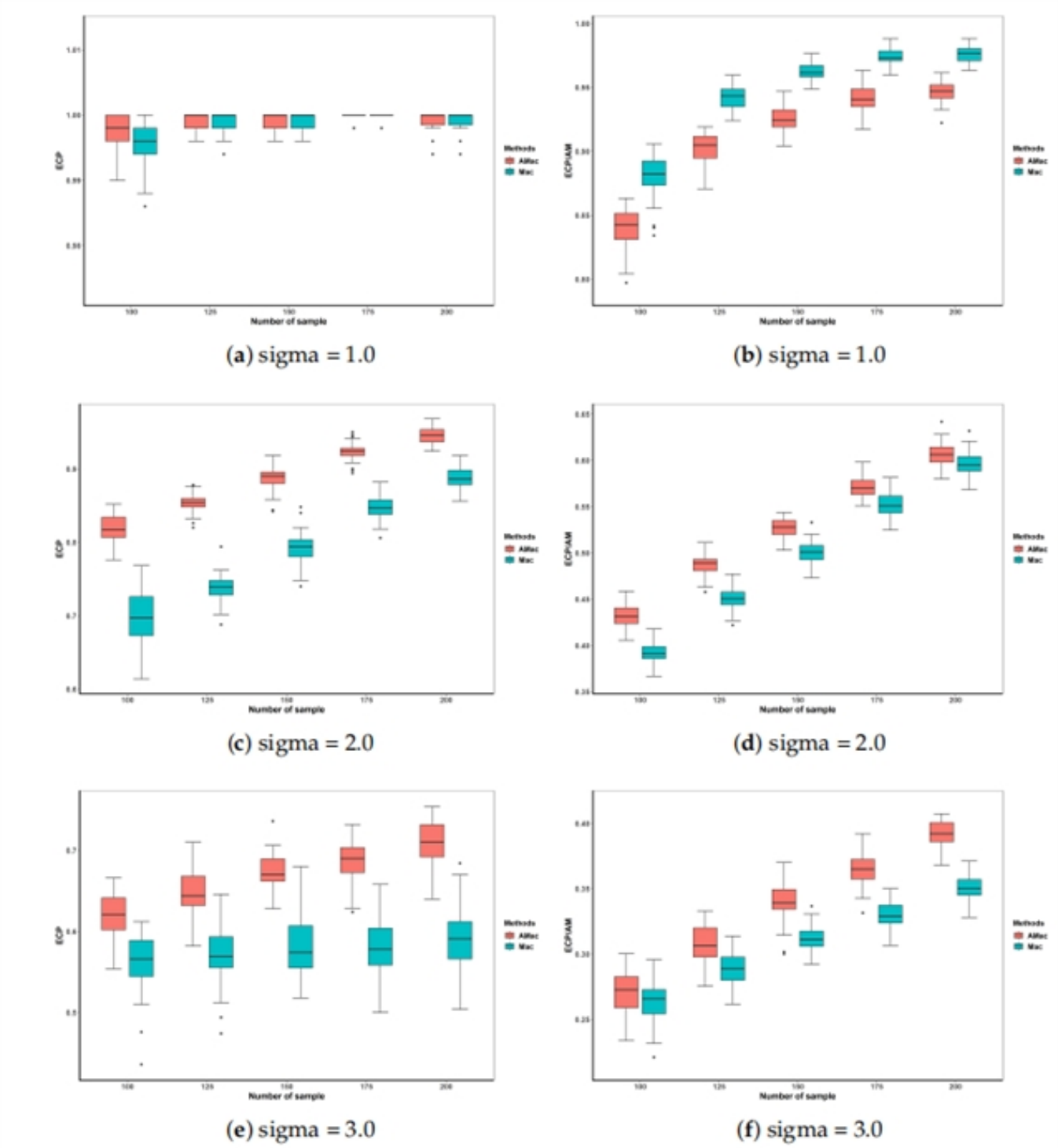
9: By calculating

$$\sum_{l=1}^{k} P(M_{opt} = \tilde{M}_l^0) \geqslant 1 - \alpha,$$

we determine the minimum value of k, and the $1 - \alpha$ confidence set for M_{opt} is $\mathcal{M}^* = \{\tilde{M}_1^0, \tilde{M}_2^0, \ldots, \tilde{M}_k^0\}$.

AMac与Mac结果对比图：拥有更高的覆盖率，小样本下所得MCS覆盖率更接近指定值，受噪声变化的影响更小。

爱科学
iikx.com



(a) sigma = 1.0 (b) sigma = 1.0

(c) sigma = 2.0 (d) sigma = 2.0

(e) sigma = 3.0 (f) sigma = 3.0

Figure 1. ECP and ECP/AM for AMac and Mac at a confidence level of $\alpha = 0.1$.

MSP算法图：

Algorithm 2 Alasso–lars (AL): the construction algorithm for MSP

1: Compute the least squares estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ under the full variable set, and calculate $\hat{w}_j = 1/|\hat{\beta}_j|, j = 1, ..., p$.
2: Define $x_j^* = x_j/\hat{w}_j, j = 1, 2, ..., p$.
3: Solve the lasso problem shown in Equation (8) using the lars–lasso algorithm:

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|. \quad (8)$$

4: Assuming that the lars–lasso algorithm takes a total of N steps to solve Equation (8), save the model M_i consisting of nonzero variables at each step and let $\mathcal{M}_r = \{M_i, i = 1, ..., N\} = \{M_l, l = 1, ..., N^*\}$ be the new set of candidate models. Here, $M_l, l = 1, ..., N^*$ are mutually distinct models.

MSP构造MCS缩减时间图：由指数增长缩减为线性增长

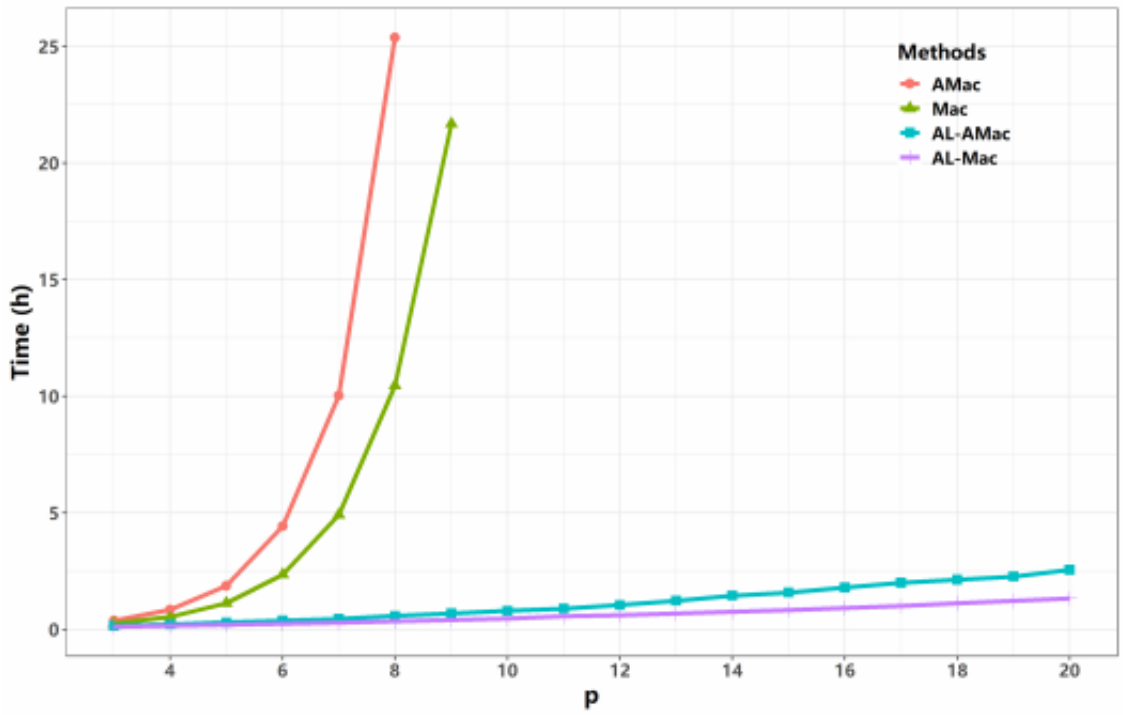
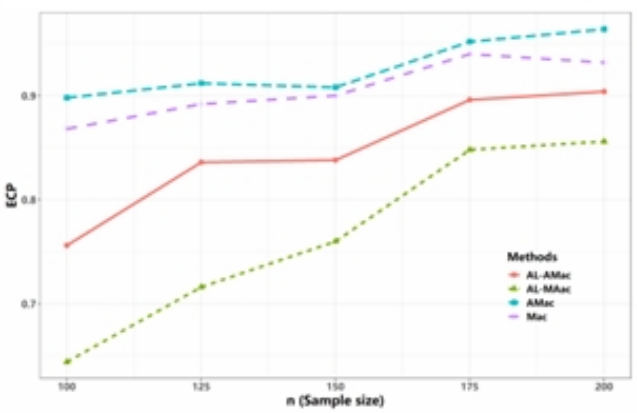
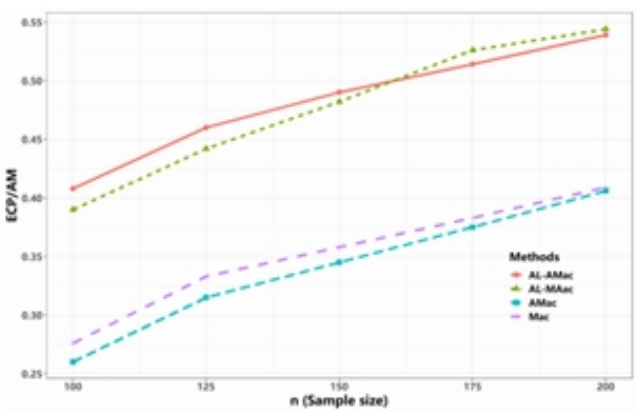


Figure 3. Running time of the four methods.

MSP构造MCS效果展示图：构造出的MCS拥有更小的模型集，且能够达到指定的覆盖率。

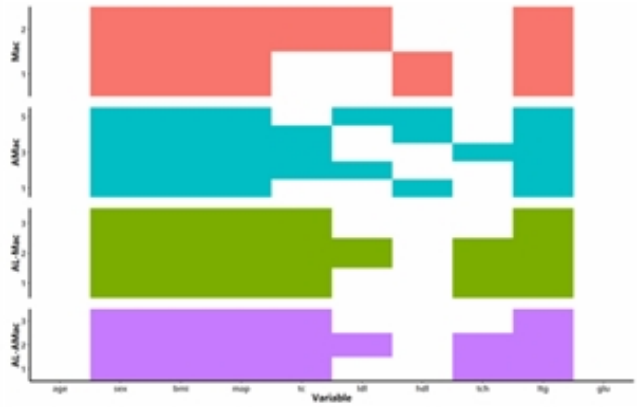


(a) sigma = 2.0

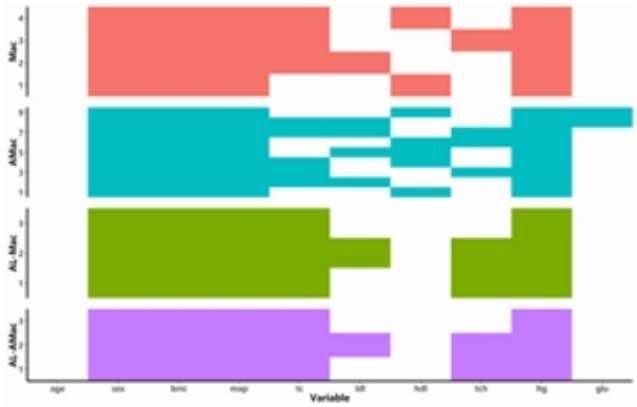


(b) sigma = 2.0

AMac及MSP在Diabetes数据集上效果展示：



(a) 90% confidence level ($\alpha = 0.1$)



(b) 95% confidence level ($\alpha = 0.05$)

3. 总结与讨论

论文围绕模型置信集构造问题，首先提出了一种新的模型置信集构造算法AMac，其次给出了MSP的定义及线性模型下MSP构造算法以解决高维变量下模型置信集构造中的计算难题。通过理论证明及模拟试验可知：(1) AMac较Mac方法有着更高的经验覆盖率，AMac构造的MCS在数据信息较少或数据波动较大时具有更理想的效果，同时AMac较Mac对于噪声具有更好的稳定性。(2) MSP能够很好的同现有MCS构造算法结合，并给出较为理想的MCS构造结果。

论文在线性回归模型下给出了MSP构造算法以解决高维变量下的MCS构造问题。但是，通过MSP来缩减模型集的想法却不仅仅针对线性回归模型，任何可以快速得到模型选择路径的模型集均可以采用本文的思想以解决高维问题，例如，对于广义线性模型，添加Adaptive lasso惩罚项的对数似然函数，其参数估计依然具有变量选择的一致性，并且可以通过局部二次近似算法得到其解路径，因此对于广义线性模型也可以构造MSP解决高维变量问题，然而如何通过LQA算法高效构造出MSP还有待进一步研究。

论文AMac方法采用的Bootstrap方法是固定模型的做法，即模型参数事先由估计给出，然后将估计参数带入模型通过不断产生误差项以构造新的数据进行模拟。另外Bootstrap还可以采用扰动数据的形式，即保持解释变量X不变，通过在原始数据Y上添加扰动来产生新数据进行模拟。能否

将两种形式的Bootstrap方法进行有效结合以给出更好的模拟概率还有待进一步研究。

三、论文作者

文法广

山东大学数学学院

研究方向：变量选择

Jiming Jiang (蒋继明)

加利福尼亚大学戴维斯分校统计学系

研究方向：统计学

栾贻会

山东大学数学学院

研究方向：时间序列分析

原文出自Mathematics 期刊：https://www.mdpi.com/2227-7390/12/5/664

四、Mathematics 期刊及期刊活动介绍

1. 期刊介绍

主编：Francisco Chiclana, School of Computer Science and Informatics, De Montfort University, UK

Mathematics (ISSN 2227-7390) 是一个专注于数学领域的国际开放获取同行评审期刊，涵盖纯数学和应用数学的各个领域。自2013年创刊以来，Mathematics 期刊发展迅速，其于2015年被ESCI数据库收录，于2017年被Scopus数据库收录，2018年被SCIE数据库收录，2021年起，Mathematics 在JCR‘Mathematics’学科类别中稳居Q1分区。期刊最新影响因子为2.3，JCR位列Q1，排名21/490；最新CiteScore为4.0，在General Mathematics类别位列Q1。

Mathematics 涵盖纯数学和应用数学的各个领域。重点发表代数与逻辑、几何与拓扑、数学分析、统计与运筹学、应用数学，包括数学与计算机科学、控制理论与力学、数学生物学、数学物理、金融数学等数学在其他各学科应用的文章。期刊致力于为研究者提供一个高质量高效率的学术交流平台。目前投稿至首次决定的时间中位数为18.3天，从接受稿件到正式发表，时间中位数为1.9天。

期刊在国内100多位编委
(学者名单详见：https://www.mdpi.com/journal/mathematics/editors?search=china) 及其他全球700多位学编的支持下，秉承着开放获取的理念，确保发表的研究成果可被全球学者免费获取，加速科研成果的传播和影响力。

2. 期刊奖项及会议

・女性数学家奖 (Women Mathematician Award)

旨在表彰女性数学家的杰出成就，激励更多女性投身数学研究和相关事业，共同营造数学科学领域性别平等、学术多元的良好氛围。

奖项奖金：2000瑞士法郎；

免费发文：可在2025年12月31日前经同行评审后在Mathematics 期刊免费发表一篇论文 (文章类型：article/review)；

荣誉证书。

提名截止时间：2025年5月31日。

详见：https://www.mdpi.com/journal/mathematics/awards/2818

・博士论文奖 (PhD Thesis Award)

旨在奖励即将毕业或即将参加毕业答辩的博士研究生。

奖项奖金：500瑞士法郎；

免费发文：可在2025年7月31日前经同行评审后在Mathematics 期刊免费发表一篇论文 (文章类型：article/review)；

荣誉证书。

提名截止时间：2025年5月31日。

详见：https://www.mdpi.com/journal/mathematics/awards/2800

・第二届国际数学与应用在线会议 (IOCMA2026)

由MDPI期刊Mathematics 主办的第二届国际数学与应用在线会议 (The 2nd International Online Conference on Mathematics and Applications, IOCMA2026) 将于2026年6月10日至12日在线举行，本次活动免费对全球数学及应用数学领域所有学者开放。会议将开放口头报告和海报展示，并设立多项学术奖项，以鼓励优秀研究成果。获奖者将有机会获得200瑞士法郎的奖金及在Mathematics 期刊上优惠发表论文的机会。欢迎全球数学及相关领域的研究人员、专家学者踊跃投稿并参与会议。

摘要投稿截止日期：2026年2月10日；

注册截止日期：2026年6月5日。

详见：https://sciforum.net/event/IOCMA2026

欢迎广大数学领域的学者投稿与合作，共同推动数学研究的发展与创新。

2023 Impact Factor 2.3 2023 CiteScore 4.0 Time to First Decision 18.3 Days Acceptance to Publication 1.9 Days 特别声明：本文转载仅仅是出于传播信息的需要，并不意味着代表本网站观点或证实其内容的真实性；如其他媒体、网站或个人从本网站转载使用，须保留本网站注明的“来源”，并自负版权等法律责任；作者如果不希望被转载或者联系转载稿费等事宜，请与我们接洽。
来源：MDPI

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发