·专题一:双清论坛“大规模商务场景的统计管理理论”·

大规模商务场景的统计管理理论*

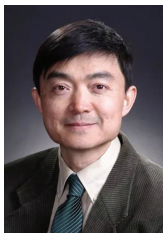
陈松蹊[1,7] 陈国青[2] 常晋源[3,4] 霍 红[5]
章 魏[5] 张新雨[4] 朱雪宁[6] 王汉生[7**]

1. 北京大学 数学科学学院,北京 100871
2. 清华大学 经济管理学院,北京 100084
3. 西南财经大学 数据科学与商业智能联合实验室,成都 610074
4. 中国科学院 数学与系统科学研究院,北京 100190
5. 国家自然科学基金委员会 管理科学部,北京 100085
6. 复旦大学 大数据学院,上海 200433
7. 北京大学 光华管理学院,北京 100871

[摘 要] 大规模商务场景是科学技术进步与商业实践发展的必然产物。大规模商务场景既覆盖了面向经济主战场的商务实践,也包括国家治理相关的重要领域,还关注数字孪生为核心的新一代数字管理技术。大规模商务场景的统计管理覆盖了管理学、经济学、计算机、环境治理、数学、统计学等多个交叉学科,为管理理论的创新提供了独特机遇。如何面向大规模商务场景,发展前沿统计方法,创新管理理论是政府部门、工业界和学术界共同关心的重要问题。基于国家自然科学基金委员会第 344 期“双清论坛”,本文从大规模商务场景出发,围绕复杂商务场景中的“数据分析方法”“统计计算与优化方法”以及“预测理论与管理决策”三方面进行了深入探讨。基于对相关概念的清晰界定和对国内外的重要文献进行系统梳理,总结了当前国内外研究现状与前沿,分析了发展趋势和方向,凝练了该领域未来 5 到 10 年的重大关键科学问题,探讨了前沿研究方向和科学基金资助战略。

[关键词] 大规模商务场景;统计管理理论;数据分析;统计计算与优化;预测理论

随着信息技术,特别是互联网、物联网、云计算、人工智能等新技术的蓬勃发展,人类社会已经进入了大数据时代。大数据作为一种新型生产资料以及具有重大价值的资产,正在深刻地改变着商业实践和国家治理的各个方面。首先,海量大数据的采集与分析,为数字经济时代新商业模式的诞生提供了物质基础。例如:金融科技、智慧零售、互联网营销、



陈松蹊 中国科学院院士,北京大学统计科学中心创始主任,斯坦福大学 2020—2022 年全球前 2%顶尖科学家。主要从事超高维大数据统计分析、统计地球物理、非参数统计方法等研究,在数学地球物理领域做出了前沿交叉成果,为精准度量污染排放和评估大气治理效果提供了科学方法。曾主持国家自然科学基金重大项目、重大研究计划集成项目和重点项目 5 项,国家重点研发专项 1 项。发表学术论文 120 余篇,担任 *Environmentrics* 和《中国科学(数学)》副主编。获 2018 年教育部自然科学奖一等奖。

王汉生 北京大学光华管理学院商务统计与经济计量系教授,国家杰出青年科学基金获得者,教育部长江学者奖励计划特聘教授,全国工业统计学教学研究会青年统计学家协会创始会长。发表学术论文 180 余篇,(合)著英文专著 1 本、中文教材 4 本。2022 年入选爱思唯尔中国高被引学者。

收稿日期:2023-12-12;修回日期:2024-02-02
* 本文根据国家自然科学基金委员会第 344 期“双清论坛”的讨论的内容整理。
** 通信作者,Email: hansheng@pku.edu.cn

短视频、自动驾驶、智能制造等。在直播电商领域，Lin 等[1]使用 1 450 场直播数据，构建面板向量自回归模型以探究主播情绪和观众行为之间的动态关系，发现主播的积极情绪可以增强观众与主播的互动行为，为直播公司提供了优化营销策略的重要依据。在新型零售终端领域，Liu 等[2]利用商品图像信息，构建并实现了一套智能无人售货柜数字孪生体，其在 6 756 件商品检测中的精确率达到了 93.7%，实现了“扫码开门，自主选购，无感结算”的全新购物模式，为新型零售行业提供了全新的思路和技术参考。除此以外，海量大数据的采集与分析也为发展新一代数字孪生技术提供了扎实的场景与数据基础，还为解决国家治理层面的重大管理问题提供了决策支持。

在 2020 年 9 月的科学家座谈会上，习近平总书记对科技创新作出坚持“四个面向”的战略部署：面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康。大规模商务场景已成为经济主战场的核心组成部分，展现出“大流量、大数据、大模型与大应用”的重要特征。此类场景孕育了一系列前沿的科学议题，成为管理科学、统计学、信息科学等学科领域研究的新焦点。对数字经济时代新商业模式、国家治理层面重大管理问题以及新一代数字孪生技术都有重要的科学意义。为回应国家的核心需求，更好地促进该方向的跨学科交叉研究，国家自然科学基金委员会（以下简称“自然科学基金委”）第 344 期双清论坛“大规模商务场景的统计管理理论”于 2023 年 9 月 14—15 日在成都成功举办。论坛围绕“复杂商务场景中的数据分析方法”“复杂商务场景中的统计计算与优化方法”及“复杂商务场景中的预测理论与管理决策”3 个议题，与会专家针对议题开展了深入研讨，并对未来 5 到 10 年自然科学基金委如何支持统计管理理论的相关研究、如何在顶层设计和队伍组织等方面发挥更大作用、如何促使我国数据科学家把握当前的重大机遇等提出了具体建议。

1 大规模商务场景的概念

本次论坛关注的核心是大规模商务场景。大规模商务场景首先关注的是来自经济主战场的商务实践，内容丰富，包括但不局限于：数智生活、供应链管理、智慧零售和金融科技等关键领域。

以直播电商为例。作为一种新型在线购物模式迅速崛起，它结合了实时视频广播与在线购物，通过直播展示产品并实时互动，让消费者亲见产品应用并直接与卖家交流，从而提升购物体验。这种模式产生了大量的结构化（如观众点赞、打赏、购买）以及非结构化（如视频弹幕）数据。Zhou 等[3]通过分析 317 309 条弹幕数据发现弹幕互动促进了观众打赏行为；Lin 等[1]使用 1 450 场直播视频数据探究主播情绪和观众行为的动态关系；Bharadwaj 等[4]在此基础上进一步得出主播情感交流直接促进观众购物行为的结论，以上发现均为直播公司提供了优化营销策略的重要依据。又如，零售行业的智能化对零售场景提出了新的需求。以智能无人售货柜为例，这种新型零售终端具有“扫码开门，自主选购，无感结算”的特点。为了更好地实现这一场景的落地应用，Zhang 等[5]构建了一套智能售货柜的标准数据集，这一数据集考虑了市面上常见的两种形态的智能无人售货柜，共计包含 37 098 份图像数据，为各类深度学习算法提供了参考。Liu 等[2]利用商品图像信息，构建并实现了一套智能无人售货柜数字孪生体，其在 6 756 件商品检测中的精确率达到了 93.7%。智能无人售货柜的解决方案，为新型零售行业提供了新思路。

此外，大规模商务场景相关的统计管理理论进展，不仅受益于纯商业相关的商业管理问题，也得益于其他交叉领域重要管理问题的研究。例如，大气污染治理，这是国家管理的重要问题，也是国家治理相关的重要场景。具体来说，以 PM2.5 为核心的空气质量评估、科学有效监控网络的建设、污染治理策略的制定及优化以及污染源的控制等，都是国家治理相关的重要场景。又如气候变化监控，在应对气候变化的管理场景中，需要收集和分析大量气候数据（例如，海洋温度和海平面高度等），然后根据分析结果制定气候变化的预测、预警和应对方案[6]。大气或者海洋治理领域的数据往往都具有一个强烈的时空（Spatial Temporal）特征，为此开发的一系列有效的统计学方法（如空间自回归模型），也完全适用于纯商业应用场景中的时空数据分析。

最后，大规模商务场景还包括以数字孪生为核心的新一代数字管理场景。通过虚实结合的数字孪生，能够模拟和优化复杂系统和过程。这项技术可以广泛适用于城市规划、智能制造和工程管理等场景，利用数字化转型有效提高管理的质量和效率。无论哪种大规模商务场景，最终都指向大规模复杂数据，这是技术进步的结果，更是现代化商业、管理、以及治理的前提条件。

2　大规模商务场景统计管理理论的主要研究内容及进展

大规模商务场景相关的统计管理理论有着丰富的落地场景以及深厚的理论研究基础。对相关的研究内容以及进展，将从三方面展开综述：(1) 复杂商务场景与数据分析方法，重点关注数据分析在各种复杂商务场景中的应用，并因此引出相关商业场景、分析方法以及综合应用需求；(2) 复杂商务场景中的统计计算与优化方法，主要关注统计计算与优化理论方向上的凝练和提高；(3) 复杂商务场景中的预测理论与管理决策，主要聚焦在综合应用方向中关于预测需求的重要响应。三部分内容各有侧重，且交叉融合、互相促进，如图 1 所示。

2.1　复杂商务场景与数据分析方法

复杂商务数据具有大规模、高维度、复杂相关、高噪音等显著特征。一方面，随着数字商务与大数据产业的快速发展，海量商务数据生成并被采集，形成了大规模与高维度的复杂数据，需要高效统计计算方法予以快速、实时分析；另一方面，由于金融、医疗、互联网、环境治理等场景的复杂性和多变性，时变数据、在线数据、网络数据、缺失数据、厚尾数据等复杂相关、高噪音数据被生成，为相关数据分析带来了巨大挑战。因此，应以高效计算、稳健分析为目标，结合统计计算与机器学习方法，进行复杂商务场景中的数据建模与理论研究，同时聚焦多源数据融合与隐私保护等实际要求，实现面向实际商务场景的具体应用。

该领域未来的发展趋势如下：在方法论与理论研究层面，围绕高维数据、时空数据、复杂相关数据等商务场景中常见数据的建模研究将获得国内外学术同行越来越多的关注，为大规模复杂数据分析提供坚实理论保证。此外，高效统计计算方法与机器学习相关算法成为研究主流，聚焦分布式计算、隐私保护、数据融合等具体商务需求，实现高效、稳健的商务分析。在实际应用方面，从场景驱动的角度出发，借助高效统计计算与机器学习算法，促进实际商务需求与监督管理需求的协同发展，实现理论、算法、场景的有效融合。主要涉及以下三个方面。

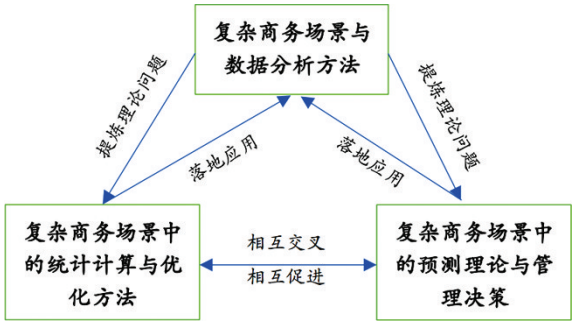


图 1　大规模商务场景统计管理理论研究内容与进展三个重要方面之间的关系

2.1.1　复杂商务数据相关的模型算法

复杂商务数据主要来源于大规模商务场景。这里的“商务场景”既包括典型的经济商业实践场景(如智慧银行、智能医保、数字经济等)，也包括国家治理相关的重要场景(如大气治理、气候变换监控等)，还包括以数字孪生为代表的新一代数字管理场景。相关数据具有复杂相关、高噪音、时变性强、异质性高的显著特征，也为其有效分析带来诸多难题。针对此现象，相关的数据分析与建模方法研究旨在通过统计、计量等模型，为各个职能部门分析、使用和理解生产运作过程中产生的商务数据，以提供可操作的商业决策模型、创造商业价值、并建立竞争优势。目前常见的相关数据建模方法可大致分为高维数据建模方法、时空数据建模方法、复杂数据的统计推断方法三类。

高维数据建模方法主要围绕大维随机矩阵数据与张量数据进行研究。大维随机矩阵理论是当前高维统计学研究的重要前沿[7]，为大维度系统的深入理解提供了重要支持。其重要的应用领域包括但不局限于：无线通讯、量化投资组合、金融因子分析及量子物理。大维随机矩阵理论的主要研究内容为特征根的分布和特征向量的分布[8]。目前在极限谱分布、中心极限定理、样本极值特征根的收敛性等方向已取得丰富的理论研究成果，并已被应用于多样的实际场景建模中。如 Zhen 等[9]设计了高维情况下的单样本及多样本检验统计量，最终可应用于分析正常个体、轻度认知障碍个体(即阿尔兹海默式个体)脑成像图差异，并确定差异的具体区域；Fan 等[10]通过调整后特征值阈值确定高维因子模型的因子个数，可应用于人脸图像分析。除大维随机矩阵外，张量数据建模也是高维数据建模研究的重要方向，但国内外对其分析尚属初步阶段。目前的张量数据分析集中于两大部分：在关系网络数据分析方面，如 SBM(Slacks-Based Measure)模型和混合成员 SBM 模型等概率图模型、网络向量自回归模型[11]等统计分析模型为张量数据的隐结构建模与统计推断提供了重要理论支撑；在特征矩阵数据分析中，学者们基于因子模型提出了多种分析方法，如

Wang 等[12]针对高维时序矩提出了矩阵因子模型，Chen 和 Fan[13]则利用主成分分析进行了矩阵截面因子模型的估计与统计推断，并应用于实际图像数据中。虽然在关系网络数据分析方面及特征矩阵数据方面已有了丰富的研究及分析方法，但对于更复杂的多源张量融合分析仍存在一定的空白，如何融合模型的低秩结构降维并进行低秩结构的融合与预测，仍是需要处理的难题。

时空数据建模方法可细化为实时数据建模与网络数据建模两大部分。实时数据建模旨在数字商务和智能系统的背景下，通过自适应采样与更新算法等技术，实现数据分析资源的节省化与计算资源的高效化。所谓自适应采样即每一时刻选择部分变量进行观测，以减轻大数据背景下数据大幅增长给硬件、能源、传输系统、还有存储计算的压力。目前有多篇文献基于自适应采样抽取的部分观测数据构建监控算法，实现稀疏异常的在线监控。另外，Guo 等[14]基于汤普森采样和贝叶斯因子模型，实现在线变换检测，提高了传统算法的有效性与及时性。此外，也有学者对实时数据的增量建模进行研究。在金融高频交易、传感器网络、航空航天等场景中往往会以极快的速度和极大的体量产生源源不断的数据流，给数据存储和模型计算带来极大的难题。因此需要研究增量建模方法，当有新数据产生时直接利用新数据对总结估计量进行更新，从而无需海量历史数据的存储，也不需从头计算，计算高效，适应大规模商务场景中电子商务的实际应用[15]。另一方面，新兴的网络型数据给已有的计量经济理论与方法带来了新的挑战，对于网络数据建模需要更深入的理论研究与应用尝试。目前针对网络数据，主要存在两大挑战：一是复杂数据的异质性挑战。目前的研究主要通过分组面板数据模型进行异质性网络数据建模：Su 等[16]提出了基于惩罚剖面似然 C-Lasso 方法，可以同时解决组识别和参数估计问题；Su 和 Ju[17]在 2016 年研究的基础上提出了惩罚主成分估计法，在未知分组的情况下，实现了组内斜率相同，组间斜率不同的模型估计。第二个挑战是大规模网络数据带来的计算效率问题，目前主要的解决方案是分布式计算，目前分布式计算研究大多聚焦于按行分布的计算[18]，且已有学者将此类分布式方法应用到网络数据建模中：如 Ren 等[19]将基于分而治之法的分布式最小二乘逼近应用于网络数据，解决了大型网络数据下的空间自回归模型估计问题。也有许多研究针对社群发现问题设计了分布式算法，如 Hui 等[20]研究了口袋交换网络移动轨迹的社区结构，并提出了三种分布式社群发现方法，在检测静态和时间社区方面具有巨大潜力。除分布式算法外，也可通过其他优化算法进行计算速度的进一步提高。如 Tan 等[21]提出了基于图像神经网络的小区交换方案(Ground-based Coud Image Database for Semantic Segmentation，GBCSS)，该方案比现有的启发式算法具有更小的计算复杂度，且具有显著的可扩展性和泛化能力。在应用方面，网络数据建模在社交网络分析、投资组合构建、金融风险评估等方面也有着广泛应用，助力风险管理、政策评估等多方位、多视角的管理决策。

稳健性统计推断方法旨在解决高噪音数据与缺失数据等受污染数据的统计分析。在高噪音数据研究中，有大量学者聚焦稳健降维问题。虽然传统的主成分降维方法能对数据进行有效降维，但基于 L2 损失函数的降维方法对于异常值不稳健，非常容易收到少数异常值的干扰。而在大数据时代背景下，数据异常值的比率和频率都明显增加，这要求我们发展能够对异常值免疫的降维方法。如 Fan 等[22]针对厚尾数据提出了高维稳健低秩矩阵估计方法，可以应用于稀疏线性模型，压缩感知，矩阵完成和多任务学习的四大场景；Chen 等[23]针对噪声，异常值和缺失数据，基于桥凸优化及非凸优化，提出了稳健主成分分析(Robust Principal Component Analysis，RPCA)方法；另有多位学者使用稳健的损失函数提出稳健因子模型，如 Chen 等[24]年提出了分位数因子模型。

尤其值得注意的是，大规模商务场景下的数据采集常常伴随着缺失，因此缺失数据研究也是重要的研究组成部分。其中一个重要的研究问题就是如何对缺失的矩阵数据进行补全。在这方面，Mao 等[25]通过使用额外的协变量数据，将目标矩阵在列空间上分解为协变量效应和低秩因子效应两部分，并提出了新型的惩罚估计方法。Liu 等[26]考虑了基于最小绝对偏差损失的矩阵补全问题，得到了更加稳健的中位数矩阵估计量，并提供了针对该非光滑损失的高效计算方法。复杂抽样调查中经常出现多元相应变量缺失的情况，Mao 等[27]在此场景下提出了一种允许同时使用行列特征来进行矩阵补全的方法，并采取增广逆概率加权方法来进一步估计感兴趣的未知参数。Wei 等[28]考虑了更高阶的张量补全问题，并提出了一种基于差分隐私的统一框架，在确保隐私保护的同时实现了张量补全的高准确性。

2.1.2　复杂商务数据相关的计算方法

在当前大数据时代的背景下，商务分析面临海量数据与隐私保护的新挑战，研究高效统计计算方法，实现计算高效、隐私友好的商务数据分析是科学研究的重要目标。因此，首先需要从内存节省、资源优化、计算实时的视角，研究高频数据、网络数据、图像数据等多种商务数据的分布式计算方法，实现大规模数据集训练过程的有效并行和局部估计量整合模型的精确创建。目前常见的分布式系统 Master-and-Worker 受到广泛关注，并应用于 Hadoop、Spark 等分布式环境中。其主要思路为：首先 Master 将任务进行分割并分配给各个 Worker；其次，所有 Worker 将自己本地计算机的局部结果进行计算并传给 Master；最后 Master 进行局部结果的最终整合。由于大型任务被分割成了若干小型任务，且各个 Worker 之间无沟通成本，因此有效提高了计算效率。目前已有大量学者针对 Master-and-Worker 框架下的分治法（Divide-and-Conquer）进行了丰富研究，如 Battey 等[29]将分治法与 Wald 检验、Rao 得分检验进行结合，得到高维分治法估计量，并将分治法低维的普通线性回归与广义线性回归推广至高维稀疏场景；Zhu 等[30]提出了分布式最小二乘近似方法，在主计算机上使用了加权最小二乘形式的损失函数得到最终结果，在线性回归、逻辑回归、Cox 模型上均可使用，并将该方法与自适应 Lasso 法结合进行收缩估计。

当国内外大部分学者将研究的重点放在参数估计上时，如何对大规模数据在分布式框架下做统计学推断就变成了一个亟待解决的重要理论课题。我国学者在这方面做出了重要的贡献。Chen 和 Peng[31]考虑了一类具有一般展开形式的统计量（包含 U 统计量和 M 估计）的分布式推断问题，给出了这类统计量在线性项主导和退化两种情形下的渐近分布结果，进一步提出了分布式自助法和伪分布式自助法两种方法来估计和近似统计量的分布。Gu 和 Chen[32]考虑了分布式节点具有异质性参数的分布式推断问题，探讨了在异质性场景下分治法 M 估计和加权分布式 M 估计两种估计量的统计学推断理论，进一步考虑使用加权分布式 M 估计进行纠偏，放宽了经典理论中对于节点数目的限制条件。Chen 等[33]提出了一种分布式一阶牛顿型估计量，能够在不求解海森矩阵的逆矩阵情况下，对估计量进行统计推断。

另外，虽然分布式框架极大提高了计算速度，但也给数据隐私造成了潜在威胁，需要进一步研究分布式框架下，数据预处理、模型学习、知识提取、中间结果获取等数据全生命周期的去中心化与隐私保护算法，实现政府、企业、用户等多端商务分析的安全性与可信性。其中，联邦学习算法是较为受欢迎的隐私保护算法之一，其主要目的是解决“数据孤岛”问题，同时保留数据的隐私性。目前已有大量学者对此进行了研究，如 Chen 等[34]提出了一种满足差分隐私要求的去中心化联邦学习算法，这一算法极大地增强了网络安全性。Wu 等[35]基于网络梯度下降法，研究在异质性数据和弱平衡网络结构下，仍能保持全局统计效率的去中心化联邦学习算法，从而大幅提升了该网络的适用性。

除了高效计算的要求外，当前商务交易场景与数据来源也日趋复杂多样，跨国交易、大型电商、供应链管理、云计算等多方平台催生大量多源数据，如何将这些数据进行融合与协同学习也是关注的重点与难点。对此，一方面需要基于机器学习与深度学习方法，从多平台互通、多视角协同的角度出发，进一步挖掘商务数据价值，以助力多源数据的深度融合。在多源数据融合中，多视图聚类方法可以是高效融合的新范式，目前已有学者对其进行研究。如 Liu[36]提出了简单多核 K 聚类，采用 Mini-Max 优化来实现多视图聚类的新范式，同时给出了全局最优多视图聚类算法的理论保证。Wang 等[37]提出快速多视图锚—对应聚类框架，解决了在多视图场景下，由于锚点在特征维度上不一致，正确对应关系获取困难的问题。另一方面，对产生的多源数据，也要建立公平、高效、安全的数据定价理论，实现多源数据质量的可靠评估，为数字经济的进一步发展注入动能。数据定价理论主要包括两大方面：数据估值理论和数据供求关系理论。目前在数据要素估值方面，已有较多基于合作博弈模型的研究。如 Ghorbani 和 Zou[38]利用有监督机器学习算法，提出了 Data Shapley 方法，满足了公平数据估值的几个自然属性；Liu 等[39]基于不同数据供应商提供的不同样本与不同特征，提出了数据估值的 2D-Shapley 方法，定义了二维的合作博弈、公理化体系与效用函数，解决了交易场景的高复杂性问题。

除上述高效统计计算方法及多源融合算法外，另有一些前沿机器学习算法如半监督学习、深度学习及强化学习算法，在估计精度、稳健性、收敛速度、避免局部最优解方面具有突出优势。一些研究将其与经典统计学问题，如分位数回归、高维数据分析、

因果推断结合，优化了估计量的理论性质，并为大规模商务场景下数据的有效分析带来了有力支持。在半监督学习方向，一些统计学研究将海量无标签数据纳入回归，以充分利用无标签样本中的信息，所得估计量具有优异的统计学性质。Deng 等[40]聚焦高维半监督学习，得到了传统 Lasso 及 Dantzig 估计量无法达到的 Mini-Max 下界，并借助无标签数据对得分函数进行调整，所得半监督估计量估计精度至少不差于有监督估计量。Zhang 和 Bradic[41]讨论了半监督场景下，高维数据均值与方差的估计及统计推断问题，通过适当的偏差校正，显著提高了原始估计量的有效性，并将其应用于异质性处理效应分析中。深度学习凭借其在估计精度、解决维数灾难问题方面的出色表现，同样受到统计领域研究的广泛关注。在分位数回归方面，Tambwekar 等[42]将条件分位数概念拓展到二分类领域，并使用深度神经网络进行学习，得到了其非渐近误差速率，对标签噪音更加稳健。此外，Chernozhukov 等[43]提出了自动去偏机器学习方法，可应用于神经网络、随机森林、Lasso 等多种高维模型，并得到了稳健估计误差、收敛速度及渐近推断的基本条件。强化学习的核心思想是让智能体（Agent）在与未知环境（Environment）的交互中学习和采取行动，以最大限度地提高其获得的累积奖励。一些研究将强化学习与分位数回归结合，提高了算法的稳健性。在因果推断领域，Shi 等[44]讨论了商业场景中常用的 A/B 测试背景下的强化学习处理方法，可对长期的处理效应进行刻画，同时允许连续检测与在线更新。Ge 等[45]利用强化学习，解决了因果推断中处理发生在不同时间点的问题，并提出稳健和半参有效估计量，以对提出的几种因果效应进行推断。

2.1.3 复杂商务场景的综合应用

目前，大数据技术与人工智能快速兴起，金融、医疗、能源等领域内的重大应用场景加速涌现，促进了对复杂商务中场景驱动的应用创新的大量需求。场景驱动创新既是将现有技术应用于特定场景，创造更大价值的过程，也是基于需求愿景，进行多方要素整合，创造新技术、新渠道、新商业模式的过程。因此，从复杂商业场景出发，基于统计计算与机器学习方法进行具体的应用研究，实现场景与技术的循环促进与协同成长，是至关重要的课题。当前，有众多学者在大科技信贷、图文数据识别、智能医保、直播平台、大气治理等重要领域进行了商务场景重要应用问题的研究。

（1）以大科技信贷为例。大科技信贷是指大科技公司利用大科技生态系统和大数据风控模型两大工具提供信贷服务，创新信用风险管理框架。随着数字信用的发展，利用数字足迹累积形成大数据、并使用大数据与机器学习方法预测还款能力与还款意愿成为大科技信贷业务的两大技术支柱。但在此背景下，大科技信贷复杂场景也带来了内生与外生两大风险，需要针对性构架数据驱动的风险管理方案。首先，在隐私管理方面，需要利用联邦学习、多方安全计算、区块链等核心技术，实现跨链或跨平台互通，在互通过程中，进一步借助隐私计算实现数据的可用不可见，为金融数据有效共享提供路径。在外部风险方面，应契合我国当前市场环境，考虑政府监管与市场约束的不同监管模式下信贷市场效率[46]，针对性给出监管者决策。具体而言，已有学者通过运用机器学习的约束性算法，提出可以抵抗交易噪音的高维正定协方差矩阵估计方法，可在此基础上进行拓展，允许数据存在时变性，通过实时数据更新协方差估计，并结合估计量，提取重要特征，实现大科技信贷平台的智能监控。

（2）以图文数据识别为例。图文数据识别是一种技术，旨在识别和理解包含文本和图像信息的数据。这种技术可以分析和处理包含文字和图形的复杂数据，例如扫描文档、照片、网页截图等，在许多领域都有广泛的应用，包括文档管理、信息检索、自然语言处理、计算机视觉等。面对复杂商务场景中，由于时间和环境变换、采集手段差异等原因，极易产生分布差异极大的图文数据，为其识别造成困难。因此，有许多学者研究了对抗分布偏移的机器学习理论与方法，对此问题予以解决。如北京大学团队针对大规模商务数据的领域偏移问题，提出一种以商品实例为核心的多模态预训练范式，在文本检索图像任务达到 87.4％的精度[47]；复旦大学团队针对电商场景图文大数据，提出一种多模态知识图谱增强模型，在大规模文本—图像检索任务上达到 72.67％的准确率[48]。

（3）以智能医保为例。智能医保是指运用人工智能、大数据、物联网等智能技术和理念，对医保公共服务、医保经办管理、“三医”共享联动等各个方面，在体制机制、组织架构、方式流程、手段工具等方面实现全方位、系统性、智能化的重塑，实现服务便捷智慧、经办高效协同、治理智能精准、协作融合共享、支撑安全可靠的智慧化要求，逐步从以治病为中心向以健康为中心转变。学者利用领域泛化模型进

行应用研究。当前针对领域泛化的专用优化方式较少被关注，在模型优化路径有限、抗干扰能力较弱的问题方面，Zhang 等[49]提出了在训练阶段使用多任务和多优化路径、在测试阶段对样本进行多视角扰动并集成的思路，提高了泛化性与适应性；Wang 等[50]针对泛化过程中出现的新类，利用领域—类别交叉划分任务，元训练集—元测试集隐式对齐梯度，实现了可泛化的决策边界。上述领域泛化模型在智能医保方面均已实现应用，如在医疗文本方面，存在着千万级别的大规模医疗文本分类问题，为医保标准名精准映射造成了巨大难题。为此，学者提出了国际疾病分类（International Classification of Diseases，ICD）智能分析系统[51]，可达到 90％以上的模型预测精度。

（4）以大气治理为例。大气治理是一种综合性的环境管理和政策实践，旨在改善大气环境质量、减少空气污染、保护人类健康以及维护生态平衡。它涉及到监测和评估大气污染情况、制定和实施相关政策、开展环境保护技术研究和应用、促进清洁能源的发展等多方面工作。大气治理的目标通常包括降低空气污染物排放、改善大气质量、减少人类对空气污染的暴露以及保护生态系统的可持续发展。随着人们对大气污染愈发重视，许多统计学者为大气治理提供了统计学角度的真知灼见，为大气治理决策提供了重要的参考依据。例如，为了针对城市的本地排放采取有效的空气质量管理措施，Zhu 等[52]通过数据选择算法测量了三种空气污染物的增长，并使用面板数据回归模型量化从 2013 年 3 月至 2019 年 2 月的三个中国北方城市的本地排放。为了深入了解气象因素对空气污染的影响，Huang 等[53]研究了地表气象变量和边界层高度（Boundary Layer Height，BLH）对六种主要空气污染物的相对重要性，根据前向变量选择算法选择变量的顺序，发现在中国北方六个主要城市中，主要污染物相对重要性的顺序存在较强的一致性，这意味着该地区空气污染的气象过程存在规律性。Tong 等[54]基于检测 PM10 浓度的时空变化点，提出了一种基于地面监测网络的空气质量数据的尘暴检测和跟踪程序。该方法通过具有高时空分辨率和更好的天气适应性的方法，对现有基于遥感的方法提供了补充方法。Zhang 等[55]研究在生成过程中空气污染物浓度的时空趋势和气象影响，而排除了复杂的风驱动传播效应干扰，其模型还对没有监测站的地点进行了空气污染物浓度的插值，并提供了空气稳定期间空气污染浓度的地图，可以用于识别空气污染物容易积聚的地点。Chen 等[6]详细回顾了气象指纹分析中最优指纹法的统计学理论，严格明确了最优指纹法成立的统计学理论条件。

2.2　复杂商务场景中的统计计算与优化方法

大型商务场景进行统计分析常常面临海量高维且复杂的数据，并需进行快速的统计计算、优化及仿真。这些数据展现出的多模态、高维度、大样本、不完全、非线性、相依性及异质性等复杂特征，需要统计计算与优化方法做出重大转变，包括但不限于从中心化转向去中心化、从固定样本数据转向流数据、从无隐私保护的计算方式转向考虑隐私保护的方式、从完全观测数据迈向部分观测数据、从全量数据计算进化到小批次随机采样计算，以及从处理小数据和小模型的策略转向应对大数据和大模型的策略[56,57]。这些重大转变都引发了一系列亟需解决的基础科学问题。

复杂商务场景中的统计计算与优化方法可以分为三个关键层面：（1）针对统计学理论支撑下的大数据计算，主要研究大数据计算方法和统计学理论的交叉融合：一方面，为大数据计算方法提供统计学理论支撑；另一方面，基于统计学理论提出新的大数据计算方法，实现更加高效率的大数据计算。（2）针对统计计算的优化方法，主要研究优化方法与统计学理论的交叉融合：一方面，为前沿的高维复杂统计学模型发展有统计学理论支撑的优化方法；另一方面，基于统计学理论提出新的优化算法，为大数据优化提供新的视角。（3）针对统计学模型的仿真与决策，主要关注有统计学理论支撑的仿真方法，以及决策问题，尤其关注在全数据特征和网络特征的框架下，如何做仿真与决策。

2.2.1　统计学理论支撑下的大数据计算

大规模商务环境经常面临海量复杂数据，对大数据统计计算和隐私保护都提出了新的需求。在大数据计算方面，传统的基于分布式系统的计算方法对计算资源有着很高的要求。然而，现实的研究者经常受到计算资源的限制。在计算资源约束的情况下，如何完成大数据计算就成了一个重要问题。在隐私保护方面，为了适应日益增高的数据隐私保护需求，数据拥有者常常需要对原始数据增加人为噪音，以便于为敏感的原始数提供必要的隐私保护。但也因此带来了信息的损失和后续统计分析效率的降低，并衍生了新的研究课题。

（1）现代统计分析常涉及大规模数据集，而计

算资源非常有限。因此，如何在有限的计算资源下做有效的统计分析已成为一个重要的问题。子抽样方法是一种简洁而有效的解决方案。该方法通过对大数据集进行随机、多次、重复的子抽样，并对所获得的子抽样样本进行统计计算，最终将这些估计量整合为最终估计结果。为了实现这一目标，需要研究合适的子抽样设计方案，验证最终统计量的一致性，并发展相应的统计推断方法。这些工作对于在有限计算资源下处理大规模数据集具有重要意义。过去的研究已经提出了多种子采样方法。现有方法的核心思想是设计新的采样策略，以便在小样本量的情况下实现卓越的统计效率。例如，Wang 等[58]研究了一种基于杠杆分数选择最优子样本的问题，并提出了 A-最优准则。Yu 等[59]开发了一种最优泊松子采样方法。尽管这些方法很有用，但存在两个局限性。首先，为了适应不同分析目的，必须仔细设计具体的采样策略；其次，这些方法的计算成本很高，大多数情况下采样成本与全样本大小成正比。因此，在实际应用中，需要权衡计算成本和统计效率，选择最适合特定问题的子采样方法。最适合特定问题的子采样方法设计应该足够简单，可以在大多数实际计算机系统上轻松实现，并具有更广泛的适用性，且采用可以显著降低估计值的偏差的 Jackknife 技术用于纠偏。

(2) 在大规模商务数据分析中，数据时常会分布在不同的计算节点上，例如，不同地区医院的数据库，用户个人的移动设备等。为了估计与业务相关的关键参数，分布式计算的框架被广泛使用。以中心化的分布式计算系统为例，中心服务器通过收集来自各节点的信息来进行模型优化和参数估计。然而，出于对数据隐私的考虑，本地节点可能并不愿意向其他节点或中心服务器上传本地的原始数据，这对分布式系统下的统计分析提出了挑战。为了解决这一挑战，一个简单可行的思路是使用分布式梯度下降算法进行模型优化，本地节点仅需要向中心服务器传输梯度而非原始数据，从而一定程度降低了隐私泄露的风险。近年来，Chen 等[60]发现即使传输梯度，依然存在隐私泄露的风险，这使得人们对隐私保护的要求进一步提高。在这样的背景下，差分隐私[56]受到了学术界越来越多的关注。具体到分布式梯度下降算法中，差分隐私技术要求在传输的梯度信息上添加独立噪音，这使得算法的隐私保护能力进一步提升。然而，这一方法在带来隐私保护能力提升的同时，也引入了额外的噪音误差，这使得算法的收敛性质受到影响，进一步使得模型参数估计存在误差。因此，需要重新探究使用差分隐私技术后分布式梯度下降算法的收敛性质并提出改进方案。

(3) 对去中心化参数估计的统计学理论而言，主要研究去中心化联邦学习(Federated Learning)。联邦学习是一种新颖的分布式计算方法，旨在解决原始分布式计算中存在的数据安全问题。该方法允许多个终端在不传输原始数据的情况下合作训练模型，其关键思想是：本地终端仅向中央服务器传输参数估计，而不共享原始数据，由此极大保护了隐私。传统联邦学习方法属于中心化方法，存在过于依赖中央服务器的缺陷。因此，去中心化联邦学习受到越来越多的关注。同传统联邦学习方法相比，去中心化联邦学习不再依靠中央服务器，而是各个终端通过合理设计的通讯网络连接，彼此传输信息来实现模型更新。去中心化联邦学习的核心算法主要包含两步：第一步，每个终端首先将邻居终端传输的参数估计进行平均，随后基于本地数据进行一步梯度下降。这样的方法不依赖中央服务器，因此更加稳健、对带宽的要求更低。去中心化联邦学习的算法步骤清晰，但是其最终所产生的估计量的数值收敛性和统计学性质等重要理论问题亟待研究。

2.2.2 面向统计计算的优化方法

面向大规模商务场景的大数据分析，统计和优化是两个重要的内容。一方面，大量统计学估计量的产生就是对特定目标函数(如似然函数)的优化结果。另一方面，先进的优化策略往往从统计学理论中汲取新的思想和方法。因此，这两种方法的交叉融合是大势所趋。在实际工作中，人们常常面临的挑战有大数据大模型计算与计算资源有限的矛盾、大规模在线数据的实时建模与计算需求与传统计算方法需要全样本多次迭代计算的矛盾、以及大规模统计优化与不完全数据之间的矛盾[61, 62]。与此相关的主要科学问题如下。

(1) 面对大规模数据的统计计算问题，数据量往往超出计算机的内存(或显存)容量。这意味着无法一次性全部加载完整的数据集到计算机内存(或显存)之中，从而导致传统基于全样本的统计计算方法难以直接应用。为了解决这个问题，一个可行的策略是先将原始数据集随机划分成若干个小批次数据，然后依次将这些小批次数据加载到计算机内存(或显存)中进行优化迭代。实际上，这种基于随机划分的小批次随机梯度下降算法已经被广泛应用于

大规模优化问题[57]，并已形成了标准的软件框架（如 TensorFlow 和 PyTorch）。在使用这类随机梯度下降算法时，通常需要设定一些关键的调节参数（如学习率和动量参数）。这些调节参数不仅影响算法的数值收敛速率，还会影响最终得到的估计量的统计学性质[63]。为了深入理解各调节参数对算法结果的影响，有必要从理论上研究相关估计值的统计性质与调节参数之间的关系。这些理论结果可以为实践者提供指导，帮助其设置合适的调节参数，在计算效率和统计效率之间取得更好的平衡。总体而言，系统研究基于随机划分的随机梯度下降类算法理论基础，不仅有助于改进实际应用效果，还能够推动统计方法在大规模数据分析等领域的进一步发展与广泛应用。

（2）对大规模流数据而言，需要研究面向大规模流数据的统计计算方法及其理论。在处理大规模流数据时，一个核心挑战是实时性。与传统的静态数据集不同，流数据以高速度不断产生，这要求统计计算方法能够在数据到达时立即进行处理。此时面向大规模数据的子抽样和分布式方法均不再适用。Schifano 等[62]首先在线性模型中引入了在线估计和统计推断方法，获得了和传统全样本估计方程方法可比的统计学效率。通过改进经典的随机梯度下降算法，在加快计算速度的基础上获得了和全样本极大似然估计相同的估计效率。受到这些研究的启发，人们提出了更多的基于流数据的统计计算方法，并对其统计学性质也需要进行系统深入研究。例如，以随机优化领域的随机梯度下降方法为基础，Zhu 等[64]通过鞅差序列的随机误差假定，建立了基于流数据的渐近协方差估计理论，并将其应用于高维线性回归的高效统计推断。类似的理论性质也可以被拓展到其他重要场景，如梯度缺失、在线协方差估计[64]、以及序贯决策问题[65]等。

（3）对面向不完全数据的统计优化方法，需要研究在部分标注数据情况下的统计学优化算法，并为此发展相应的半监督统计学习理论框架。例如：在电子病历系统中，确诊病人是否患类风湿关节炎或者多发性硬化症需要较高的诊断成本。因此，确定病人是否患有上述疾病比起收集病人的基本信息更加困难。这就产生了一类非常有趣的数据，其中因变量只能获得部分观测，但自变量却能被完全观测。如果给予这样的数据做统计学习，就成了一个重要的理论问题。此类问题在文献中被称为半监督学习问题。半监督学习的场景在现实中常常出现，如抽样调查、文本分析和图像分类。例如，分类问题，一种常见的解决方法是高斯混合模型。对于高斯混合模型，可利用经典的 EM（Expectation-Maximum）算法求解。EM 算法的数值收敛性也已经得到了充分的研究。Wu[66]证明了在一定条件下，EM 算法会数值收敛到极大似然估计量。Xu 和 Jordan[67]研究了 EM 算法在高斯混合模型下的收敛速度，并将 EM 算法与梯度上升算法进行了比较。上述研究都是基于标签平衡的情况。在现实中可能面临某种标签出现频率极低的情况，如欺诈检测，CT 图像疾病识别等，这类标签往往被称为稀疏事件[68]。在稀疏事件的存在下，传统计算方法（如 EM 算法）往往会表现出非常不同于常规的规律。例如，面对稀疏事件数据，学者发现 EM 算法的收敛速度会非常慢。因此，发展面向稀疏事件统计学和机器学习算法理论是重要的研究课题。

2.2.3　基于统计学模型的仿真与决策

基于统计学模型的仿真与决策对于发展以数字孪生技术为代表的新一代数字管理工具意义重大。数字孪生是一种通过数字模型在计算机中对实际物体或系统进行仿真的新一代数字管理工具。这种数字模型是实体的虚拟表示，包括实体的结构、行为、性能等多个方面。数字孪生可以通过虚拟环境对实体状态和变化的实时反映，实现监测、优化、预测的目的。数字孪生在智能制造、工业升级和数字化转型方面已经发挥了关键作用。而统计学模型为仿真与决策提供了扎实的理论支撑，也提出了新的挑战。为此，需要对以下科学问题做系统性研究。

（1）对现实商务场景中的全数据特征和网络关系进行统计分析，将其抽象为复杂系统是进行仿真建模和决策优化的基础，有利于决策者掌握其构成要素之间的相互作用和变化关系。然而，由于大规模商务场景中涉及的多元主体规模庞大，使得关联网络呈现异质性和稀疏性等特点，需要研究如何表征全数据特征和关联关系，准确刻画各节点统计学性质和全局网络结构。特别是在如今的商务场景中，大规模网络数据普遍存在，为了揭示不同节点对之间的潜在链接、精确量化与潜在联系识别相关的统计不确定性，Fan 等[10]提出了一种大型网络成员关系统计推断方法，基于经验特征向量构建 Hotelling 型统计量，可以在大规模网络上更加高效的执行，并且能够处理节点属性信息混合的情况。另外，相关理论方法在诸多现实问题中得到了应用。例如，在智能制造领域，Mykoniatis 和 Harris[69]提

出了一种数据驱动的混合模拟建模方法，以实现数字孪生中的生产控制系统测试与验证。在网络安全领域，Zhang 等[70]研究了非线性复杂网络系统的攻击隔离和攻击定位问题，通过构造两个残差，并给出残差区间估计的方式，提出了相应的人工智能方案。针对能源市场，Chen 等[65]通过分析构建的多元价格运动演化网络拓扑特征，探讨了碳燃料能源市场系统中多元价格运动的动态演化规律，为碳交易市场的市场监管者开辟了一个全新视角。

（2）对于复杂系统仿真而言，高精度的仿真模型（如数字孪生模型）意味着仿真成本和仿真时间更高。而成本低、耗时少的仿真模型精度又低。如何平衡仿真效率和仿真精度的矛盾，如何优化仿真资源分配，在有限的计算资源下实现仿真效率的最大化，是本研究需要突破的科学难题。现代仿真技术作为强有力的建模工具被广泛用于复杂系统的评估与分析过程中，如医疗保健系统、工业生产系统等。系统仿真模型的随机特性，使得决策者需要多次的系统仿真复制才能获取平稳的系统估计，然而仿真模型的时间与经济成本随着系统复杂性的增加而急剧提升。因此，如何优化有限仿真资源的分配以最大限度提高系统仿真效率，十分具有现实意义。近些年来，涌现出诸多基于单精度仿真模型的仿真优化算法，例如：最优计量分配（Optimal Computing Budget Allocation）算法、无区别区间（Indifferent-zone）算法、期望提升（Expected Improvement）算法等。然而，随着经济与社会生产力的急剧提升，复杂系统趋于高度随机化，仅借助于单精度仿真模型已难以同时保证系统的评估效率与分析精度，这极大地增加了系统管理的难度。此外，仿真模型的精度与其运行速度还呈现出显著的负相关关系。因此，深入挖掘不同精度仿真模型之间的相互协作机制，并建立实现系统仿真效率与精度间有效权衡的多精度仿真优化方法，是实现复杂系统有效分析与评估的关键问题之一。为此，Xu 等[71]提出了两阶段多精度仿真优化算法，该算法分为有序变换（Ordinal Transformation）和最优采样（Optimal Sampling）两步以实现仿真预算的有效分配。

（3）现实的商务场景是不断演进和变化的，如何基于仿真模型对大规模复杂商务系统进行高质量评估，需要构建自适应数据关联与统计模型参数调整机制，对大规模商务场景下的决策行为形成和演化进行分析，建立动态环境下复杂商务系统的评估与优化决策理论与方法。在瞬息万变的信息化时代下，复杂系统的内外场景趋于高度动态多变化。动态场景下的系统管理与决策问题本质上是一个序贯决策过程，决策者需要不断借助仿真模型对系统进行动态评估分析来不断重新调整管理策略，以应对系统场景的实时变化，这对复杂系统的高质量管理与决策提出了严峻考验。为此，需要研究合适的系统仿真动态优化与决策方法，以精确捕捉系统的动态特性，并实现系统有效的动态决策与评估。既往研究已经提出了多种连续仿真方法，用以实现复杂动态系统的优化与管理，但这些研究鲜有考虑优化仿真的预算分配问题。现有研究以数据驱动为核心方法对复杂系统动态性进行分析，并基于此建立了仿真建模与优化算法。此外，为了进一步考虑系统决策与内外场景变化的相互影响，Peng 等[72]等研究，在贝叶斯框架下，将系统仿真决策过程描述成为一个随机控制问题，并基于近似动态规划方法，提出了一种渐近最优仿真优化方法，以实现系统仿真的动态优化与决策。特别地，复杂商务系统作为一类重要的复杂系统，一方面继承了复杂系统高度动态多变化的显著特征；另一方面，又有别于大多数复杂系统，拥有着较为丰富的数据资源。因此，有必要充分挖掘系统多源数据的有价值信息，综合考虑商务场景变化与系统决策的相互影响，并深入分析大规模商务场景下的决策行为形成和演化规律，以探究基于自适应多源数据关联与仿真统计模型参数调整机制，这对实现动态环境下复杂商务系统的高质量评估与分析具有十分重要的意义。

2.3 复杂商务场景中的预测理论与管理决策

复杂商务数据预测分析是商务智能与商务经济发展的重大理论基础。预测是判断决策优化场景出现的前提条件，其目标是为科学决策提供依据，进而提高决策的效益和效率。以电子商务、互联网金融、移动支付为场景的大规模复杂商务的预测与决策问题中通常会衍生出新型、复杂的数据，如：多源异构数据、部分无标签数据、分布漂移数据、混频数据、多模态数据等。针对这些新型且复杂的数据，传统的统计建模可能无法捕捉有效信息，亟需开发相应的统计预测方法，设计高效、鲁棒的决策算法。

复杂商务场景中的预测理论与管理决策可以分为两个关键层面：首先是管理决策驱动的预测研究。建立和应用预测模型来支持和指导实际管理决策，以改进决策质量、提高效率和应对不确定性。其次是复杂数据场景下的预测研究。主要研究如何识别新型复杂数据和传统数据的差异，从而有效地处理

多样性的数据类型，开发合适的预测模型。

2.3.1　管理决策驱动的预测研究

数据—预测—决策是当前管理科学领域进行科学决策的主要模式。精确的预测分析是进行科学决策的必由之路。从错综复杂的数据中提炼客观规律，开发提高预测精度的有力工具，为规避各级部门管理决策中的风险提供科学依据。然而，在实际决策制定中，传统的统计方法经常将预测误差作为优化目标来求解参数，在管理决策问题中，预测目标并不等于决策目标，单纯将预测误差作为目标可能会导致决策失误。比如，投资公司需要预测股票价格的波动情况来指导投资决策，实际往往将股票价格波动作为预测的目标，但决策目标可能是最大化投资回报或最小化风险，此时公司需要考虑到市场因素、公司基本面、行业趋势等因素。预测目标与决策目标之间存在差异，为了实现有效决策，需要根据实际决策目标来调整预测目标。此外，以库存管理代表的预测中的不确定性是一个挑战。如何在预测模型中有效地考虑和管理不确定性，以降低不确定性带来的风险，是一个重要问题。另一方面，管理决策中存在非对称损失的预测问题，比如，供应链中断的成本可能远大于过度供应的成本，这也是决策中需要考虑的问题。

(1) 复杂商务场景中涉及了大量预测问题。给定一个预测问题，研究者通常会建立二次损失或绝对值损失作为优化目标来求解参数，然后将求得的参数带入到决策环境中，最终得到决策。经典的预测教科书同样将预测本身视为目的，没有考虑到预测的目的是什么(即管理决策，包括预算、资源调度与分配等)。一个很好的预测结果对应了很低的预测误差，但并不代表会得到很好的决策，即预测目标不等于决策目标。例如，制造型企业需要预测未来产量以制定生产计划决策，若以预测产量和实际产量十分接近为目标，不考虑材料供应情况、市场需求变化情况等因素，最终生产计划可能无法满足市场需求，导致资源配置不合理。Goltsod 等[73]综合考虑了预测和库存管理这两个相互关联的问题，通过库存表现来判断预测表现，从而对库存管理中用到的需求参数进行适当的估计和更新，以整体观点来解决库存管理的最终目标。然而，Goltsod 等[73]只是考虑了库存管理问题，需要推广到更加一般的管理决策问题中。此外，也需要考虑如何公式化决策目标。当决策目标表达式非凸非线性时，应该如何设计高效的迭代算法求解参数。有必要深入研究迭代算法的收敛性质、参数估计量的渐近性质，如相合性和渐近正态性。总体而言，对上述问题的研究可以帮助人们理解预测目标和决策目标的不同，为实际商务场景中决策提出提供理论支持，最终得到最优决策。

(2) 以库存模型为代表的商务管理预测问题是供应链中一个重要课题，主要研究零售商如何管理商品的库存来使得利益最大化。由于市场的激烈竞争，价格差异优势逐渐减小，在提供同等产品和服务下，降低库存成本成了业界关注的焦点，过高的库存会导致资金占用和库存成本增加，过低的库存可能会导致供应不足，影响销售。产品的未来需求是影响库存管理的一个很重要因素，对需求进行预测来设置合理的库存水平，以提高企业运营效率。因此，确定需求分布是库存管理的重要问题，现有库存管理教科书通常假定未来需求分布是已知的，但这并不符合实际。当需求分布未知时，可以采用鲁棒优化、样本经验分布代替需求分布、分位数回归，这些对确定需求分布中的未知参数进行考虑。现实中往往很难从有限数据中选择出一个“正确”的需求分布。不确定性在其他管理预测问题中普遍存在，例如，预测未来商品销量时，要考虑商品特征的选择、预测模型的形式等，需要采取适当的措施以最大程度地减少不确定性带来的风险。针对需求模型的不确定性，模型平均可以有效避免对需求模型的错误选择导致的昂贵代价。

(3) 商务管理中存在着非对称损失的预测问题。许多预测问题中最常用的损失函数是均方误差和平均绝对误差，这都是对称的。对称损失函数的使用前提是假设正、负偏差是同等重要的。然而，在某些情况下，正、负偏差带来的代价是不一样的。比如，供应链中断的成本可能远大于过度供应的成本。因此，在一些管理预测问题中需要一个允许非对称的灵活损失函数。同样，在宏观经济预测中，乐观的国内生产总值(Gross Domestic Product，GDP)增长预测可能导致名义赤字的预期水平较低，从而为所需的财政调整留下一些余地。Ray 等[74]指出非对称定价或非对称价格调整是指价格上涨比下跌更容易的现象，研究提出了批发价格非对称定价的新理论及其实证支持，同时还讨论了非对称定价、渠道、价格调整成本以及公共政策等方面的影响。Chen 等[75]使用非对称损失研究了好消息和坏消息对价格信息的影响，进而利用价格信息为公司的投资决策提供有用信息。

2.3.2 复杂数据场景下的预测研究

大规模复杂商务场景带来了更多新型复杂的数据，这些数据往往具有多模态、多源异构、大部分数据无标签、分布漂移等特点。在管理预测研究中，存在以下三个挑战：开发有效的方法将多源数据整合成有意义的信息以支持决策制定；开发半监督学习方法从无标签数据中提取有用信息；开发稳健学习方法处理随着时间或其他因素发生分布漂移的数据。

（1）在管理决策问题中经常存在复杂多源数据。比如，在信用评级中需要融合来自不同平台的数据（消费者的消费数据、资产情况和理财数据、企业的资产结构借贷行为数据）用于构建信用评估系统。多源数据的难点在于需要克服不同源的数据模式不一致性，确保数据的一致性和可用性；需要开发迁移学习的方法等来提取有效信息。针对多源数据，Li 等[76]使用迁移学习的思想，确定传统任务和新任务相似性，将传统银行业务数据迁移到新业务数据中，重建训练集来训练违约风险预测模型，凸显了迁移学习在金融风险领域的商业价值，为管理者提供了决策依据。预测性业务流程监控是业务流程管理的重要任务之一，Chen 等[77]提出了一种基于 BERT（Bidirectional Encoder Representations from Transformers）和迁移学习的多任务预测方法，可以更快地应用于多种不同的预测任务，且具有突出的性能。多源数据的融合关键是如何处理这些源域的差异，需要开发有效的迁移学习技术和因果推断技术准确地把握不同域之间的差异和因果关系。为了更好地利用不同源域信息，需要建立一个模型融合的新型算法，从最优模型平均预测出发，开发关于模型参数的统计推断方法及感兴趣参数的统计学推断。

（2）如前所述，实际工作中常常碰到部分无标签数据，也被称为半监督数据。如何利用部分无标签数据来提升整体的预测能力是关键问题[78]。早期关于半监督数据开发的半监督方法是针对分类问题的预测，比如，Zheng 等[79]针对半监督数据开发了在线评论质量挖掘系统，以区分客服有用和无用的评论，进而提升供应商的营销决策。相比于分类问题，回归也有广阔的应用背景。最近，关于半监督回归问题发展迅速，Chakrabortty 和 Cai 等[78]将线性模型作为工作模型，假设协变量和响应变量之间具有线性相关性，进而通过协变量预测响应变量。实际中，协变量和响应变量之间关系往往是非线性的，并且具有不确定性。因此，在建立半监督方法时应该考虑这种未知性和不确定性。未来针对管理预测下的半监督数据可以尝试如下三个探索方向：1）如何验证无标签数据是否会有利于决策；2）如何利用无标签数据来提升决策能力；3）如何衡量无标签数据提升能力的大小。

（3）面向复杂数据场景的应用，预测与决策扮演着重要的角色。目前，预测模型的精准度主要取决于训练数据和测试数据，是独立同分布的。随着业务涉及领域的增加以及数据快速的更新迭代，训练数据和测试数据独立同分布的假设不再广泛适用。一方面，因为数据变化的速度快于模型更新的速度，且测试数据的分布变化情况未知；另一方面，因为不准确预测导致的错误决策成本增加。这类问题需要获取不同测试数据分布下的不变因果结构，将因果关系纳入到管理预测问题中，让预测模型不依靠变量之间不稳定的相关性来进行建模，这样即使在未知的测试数据分布上也能有一个稳健的结果。未来的研究可以围绕三个方向展开：1）为什么稳健学习方法能对未知测试数据分布取得良好结果；2）在什么样的测试数据条件下会比按照独立同分布假设建立的模型表现好；3）在稳健学习的建模过程中，去掉了变量之间的相关性过度损失了数据中的信息，在不降低模型泛化性能的前提下，如何尽可能的利用这些数据信息，提高预测的精准度。

3 未来 5～10 年大规模商务场景的统计管理理论发展目标及资助重点

3.1 发展目标

当前大数据时代特有的数据类型、大规模的数据体量以及新的商务场景给传统的统计管理理论与方法带来了前所未有的挑战。我们迫切需要建立新的方法论体系，采用更加主动、全面的视角，面向未来可能发生的场景和情境进行积极的分析和预测，并将这些前瞻性分析和预测应用于发展大规模商务场景下的统计管理理论[80,81]。这能积极促进管理学的研究范式创新，同时有效回应习近平总书记关于“四个面向”中“面向经济主战场”的重要要求。大规模商务场景下统计管理理论的建立涉及诸多学科知识和技术，需要管理科学、统计学与计算机科学等多学科的深度交叉融合[82]。因此，在未来五到十年中，需要进一步凝练该研究领域的内涵、发展目标、核心科学问题、关键技术问题和典型应用场景；夯实学科理论基础、创新方法、培养具有前沿统计管理理

论的人才，以应对当前大数据时代对统计管理人才的需求与技术需求。

3.2　资助重点

本次“双清论坛”的与会专家经过深入研讨，凝练了大规模商务场景下统计管理理论的若干重大关键科学问题，并建议未来 5～10 年应着重围绕以下领域开展原创性研究。

（1）基于多源融合与隐私保护的复杂商务数据建模与应用创新

背靠大数据产业与人工智能技术，数字商务与数字经济建设已成为国家发展的重要战略，是经济主战场的最前沿。大数据产业下各行各业的多源数据是激活经济要素潜能的关键支撑，是加快经济社会发展中质量变革、效率变革的重要引擎。如果能够充分利用互联网电商、短视频平台、流媒体等多方媒介产生的大数据，通过统计建模方法与机器学习算法实现数据共享机制，完成数据多源融合，将会最大化数据要素潜力，助推商务发展，反哺数据质量，形成大数据与商务场景互相融合促进的良性循环。

因此，建议资助重点为：1）多源数据融合的复杂商务数据建模与应用研究：加强对网络数据、图像数据、音频数据、自然语言数据、高频数据和流数据等多种商务场景中的高价值数据的建模研究，提出系统的统计学分析框架，建立相关的统计学理论体系；2）隐私保护视角下的复杂商务数据建模与应用探索：加强基于联邦学习和差分隐私实现的安全数据互通的统计算法研究。

（2）复杂商务场景驱动的统计计算与优化方法

大型商务场景催生了海量数据，各种复杂模型也应运而生。这些模型包括但不局限于：超高维模型、网络数据模型、以深度学习强化学习为代表的前沿机器学习模型等。由于实际工作中所具有的计算资源永远是有限的，如何在有限的计算资源下完成具有统计学理论支撑的大数据复杂模型的计算，已成为一个重要的研究方向。另一方面，如何发展具有统计学理论支撑的新一代仿真技术与方法对支撑大规模商务场景下的管理学理论研究具有重要意义。

因此，建议资助重点为：1）面向大规模数据的计算方法与统计学理论研究。包括但不局限于：分布式计算方法、子抽样方法、流数据计算方法、不完备数据建模、以及主动学习等领域。2）去中心化的统计计算与优化。重点关注网络结构对算法收敛性以及最终统计量统计学性质的影响，还同时要兼顾典型超参数（学习率等）的重要作用，建立相关的统计学理论。3）面向大规模商务场景的全数据特征和网络关系进行统计分析，并建立高精度的仿真模型（数字孪生模型等）。

（3）面向复杂商务场景的预测理论与管理决策

在当今复杂商业场景下，预测理论和管理决策成为了日益重要的研究内容。预测问题是决策制定过程中不可或缺的一环。由于商务环境日益改变、风险不断升级、竞争愈发激烈，传统的预测方法已经不能满足当今商务决策的需求，因此需要发展更加智能和精确的预测模型。例如，考虑预测目标和决策目标的一致性、预测问题中的不确定性、非对称损失问题等。与此同时，新的商务场景带来了新型数据的挑战，多源异构、部分无标签数据、分布漂移数据等新型数据增加了数据处理和分析的难度。识别新型数据的特征、开发相应的预测方法，对决策制定具有重要的现实意义。

因此，建议资助重点为：1）管理决策驱动的预测研究：建立以决策目标为导向的预测框架，加强对预测问题中不确定性的研究，以及如何权衡不同类型错误的非对称损失预测问题。2）复杂数据场景下的预测研究：加强对多源异构数据、部分无标签数据、分布漂移数据等新型数据的建模研究，借助模型平均、迁移学习、半监督学习、稳健学习等技术的结合来提炼有效信息，进而实现精确预测和决策支持。

4　结　语

大规模商务场景中蕴含着很多新的管理学问题，为管理学科的发展提出了新的要求与新的方向。将大规模商务场景数据从资源转变成为生产力的关键在于对其进行有效建模与合理分析。在这个过程中，统计管理理论与方法扮演着不可或缺的重要角色。挖掘大规模商务场景数据的内在价值已成为当前的国家战略。如何挖掘与分析大规模商务场景中的数据资源并将其转变为有效信息辅助管理决策，一直是学术界关注的热点问题。我们应当充分发挥管理科学在管理统计、数据挖掘等领域的优势和特点，融合数学、信息科学等多学科领域的知识，提出一套新的统计管理理论和方法，全面提升各行业的商务分析能力和核心竞争力，从而有力保障我国经济高质量稳步发展。根据相关研究热点与趋势，本文凝练出了该领域未来五到十年的重大关键科学问题，探讨了前沿研究方向和科学基金资助战略。今后需要进一步凝练研究方向，规划和推进跨学科攻

关团队的培养，深入开展统计管理基础理论研究，促进相关研究成果的应用与推广。

参 考 文 献

[1] Lin Y, Yao D, Chen XY. Happiness begets money: emotion and engagement in live streaming. Journal of Marketing Research, 2021, 58(3): 417—438.

[2] Liu CX, Da ZY, Liang YZ, et al. Product recognition for unmanned vending machines. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(2): 1584—1597.

[3] Zhou JL, Zhou J, Ding Y, et al. The magic of danmaku: a social interaction perspective of gift sending on live streaming platforms. Electronic Commerce Research and Applications, 2019, 34: 100815.

[4] Bharadwaj N, Ballings M, Naik PA, et al. A new livestream retail analytics framework to assess the sales impact of emotional displays. Journal of Marketing, 2022, 86(1): 27—47.

[5] Zhang HJ, Li DH, Ji YZ, et al. Toward new retail: a benchmark dataset for smart unmanned vending machines. IEEE Transactions on Industrial Informatics, 2020, 16(12): 7722—7731.

[6] Chen HY, Chen SX, Mu M. A statistical review on the optimal fingerprinting approach in climate change studies. Climate Dynamics, 2024, 62(2): 1439—1446.

[7] Bai ZD, Silverstein JW. Spectral Analysis of Large Dimensional Random Matrices. New York, NY: Springer New York, 2010.

[8] Deshpande A, Mehta A, Vincent T, et al. Quantum computational advantage via high-dimensional Gaussian boson sampling. Science Advances, 2022, 8(1): eabi7894.

[9] Zheng SR, Cheng GH, Guo JH, et al. Test for high dimensional correlation matrices. The Annals of Statistics, 2019, 47(5): 2887—2921.

[10] Fan JQ, Fan YY, Han X, et al. Simple: statistical inference on membership profiles in large networks. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2022, 84(2): 630—653.

[11] Zhu XN, Pan R, Li GD, et al. Network vector autoregression. The Annals of Statistics, 2017, 45(3): 1096—1123.

[12] Wang D, Liu XL, Chen R. Factor models for matrix-valued high-dimensional time series. Journal of Econometrics, 2019, 208(1): 231—248.

[13] Chen EY, Fan JQ. Statistical inference for high-dimensional matrix-variate factor models. Journal of the American Statistical Association, 2023, 118(542): 1038—1055.

[14] Guo J, Yan H, Zhang C. A Bayesian partially observable online change detection approach with Thompson sampling. Technometrics, 2023, 65(2): 179—191.

[15] Wang KN, Wang HW, Li SM. Renewable quantile regression for streaming datasets. Knowledge-Based Systems, 2022, 235: 107675.

[16] Su LJ, Shi ZT, Phillips PCB. Identifying latent structures in panel data. Econometrica, 2016, 84(6): 2215—2264.

[17] Su LJ, Ju GS. Identifying latent grouped patterns in panel data models with interactive fixed effects. Journal of Econometrics, 2018, 206(2): 554—573.

[18] Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. Journal of the American Statistical Association, 2019, 114(526): 668—681.

[19] Ren YM, Li Z, Zhu XN, et al. Distributed estimation and inference for spatial autoregression model with large scale networks. Journal of Econometrics, 2024, 238(2): 105629.

[20] Hui P, Yoneki E, Chan SY, et al. Distributed community detection in delay tolerant networksProceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture. Kyoto Japan. ACM, 2007.

[21] Tan K, Bremner D, Le Kernec J, et al. Graph neural network-based cell switching for energy optimization in ultra-dense heterogeneous networks. Scientific Reports, 2022, 12(1): 21581.

[22] Fan JQ, Wang WC, Zhu ZW. A shrinkage principle for heavy-tailed data: high-dimensional robust low-rank matrix recovery. The Annals of Statistics, 2021, 49(3): 1239—1266.

[23] Chen YX, Fan JQ, Ma C, et al. Bridging convex and nonconvex optimization in robust pca: noise, outliers, and missing data. The Annals of Statistics, 2021, 49(5): 2948—2971.

[24] Chen L, Dolado JJ, Gonzalo J. Quantile factor models. Econometrica, 2021, 89(2): 875—910.

[25] Mao XJ, Chen SX, Wong RKW. Matrix completion with covariate information. Journal of the American Statistical Association, 2019, 114(525): 198—210.

[26] Liu W, Mao X, Wong RK. Median matrix completion: from embarrassment to optimality. International Conference on Machine Learning. International Conference on Machine Learning, 2020, 119: 6294—6304.

[27] Mao XJ, Wang ZL, Yang S. Matrix completion under complex survey sampling. Annals of the Institute of Statistical Mathematics, 2023, 75(3): 463—492.

[28] Wei Z, Li ZP, Mao XJ, et al. Applying differential privacy to tensor completion ICASSP 2022- 2022// IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore. IEEE, 2022: 3923—3927.

[29] Battey H, Fan JQ, Liu H, et al. Distributed testing and estimation under sparse high dimensional models. The Annals of Statistics, 2018, 46(3): 1352—1382.

[30] Zhu XN, Li F, Wang HS. Least-square approximation for a distributed system. Journal of Computational and Graphical Statistics, 2021, 30(4): 1004—1018.

[31] Chen SX, Peng LH. Distributed statistical inference for massive data. The Annals of Statistics, 2021, 49(5): 2851—2869.

[32] Gu J, Chen SX. Weighted distributed estimation under heterogeneity. 2022: arXiv: 2209. 06482. http://arxiv.org/abs/2209.06482

[33] Chen X, Liu WD, Zhang YC. First-order newton-type estimator for distributed estimation and inference. Journal of the American Statistical Association, 2022, 117(540): 1858—1874.

[34] Chen SZ, Yu DX, Zou YF, et al. Decentralized wireless federated learning with differential privacy. IEEE Transactions on Industrial Informatics, 2022, 18(9): 6273—6282.

[35] Wu SY, Huang DY, Wang HS. Network gradient descent algorithm for decentralized federated learning. Journal of Business & Economic Statistics, 2023, 41(3): 806—818.

[36] Liu XW. Hyperparameter-free localized simple multiple kernel K-means with global optimum. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023: 45(7) 8566—8576.

[37] Wang SW, Liu XW, Zhu XZ, et al. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. IEEE Transactions on Image Processing, 2022, 31: 556—568.

[38] Ghorbani A, Zou J. Data shapley: equitable valuation of data for machine learning. International Conference on Machine Learning, 2019, 97:2242—2251.

[39] Liu Z, Just HA, Chang X, et al. 2D-Shapley: a framework for fragmented data valuation// Proceedings of the 40th International Conference on Machine Learning, 2023, 899: 21730—21755.

[40] Deng SY, Ning Y, Zhao JW, et al. Optimal and safe estimation for high-dimensional semi-supervised learning. Journal of the American Statistical Association, 2024: 1—12.

[41] Zhang YQ, Bradic J. High-dimensional semi-supervised learning: in search of optimal inference of the mean. Biometrika, 2022, 109(2): 387—403.

[42] Tambwekar A, Maiya A, Dhavala S, et al. Estimation and applications of quantiles in deep binary classification. IEEE Transactions on Artificial Intelligence, 2022, 3(2): 275—286.

[43] Chernozhukov V, Newey WK, Singh R. Automatic debiased machine learning of causal and structural effects. Econometrica, 2022, 90(3): 967—1027.

[44] Shi CC, Wang XY, Luo SK, et al. Dynamic causal effects evaluation in A/B testing with a reinforcement learning framework. Journal of the American Statistical Association, 2023, 118(543): 2059—2071.

[45] Ge L, Wang JT, Shi CC, et al. A reinforcement learning framework for dynamic mediation analysis. 2023: arXiv: 2301.13348. http://arxiv.org/abs/2301.13348.

[46] Chen HQ, Gu M, Ni B. How price limit affects the market efficiency in a short-sale constrained market? evidence from a quasi-natural experiment. SSRN Electronic Journal, 2023, 73: 22—39.

[47] Jin Y, Li YZ, Yuan ZH, et al. Learning instance-level representation for large-scale multi-modal pretraining in E-commerce// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 11060—11069.

[48] Wang XD, Wang CY, Li L, et al. FashionKLIP: enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge// GraphProceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada. Stroudsburg: Association for Computational Linguistics, 2023: 149—158.

[49] Zhang J, Qi L, Shi YH, et al. MVDG: A unified multi-view framework for domain generalization// European Conference on Computer Vision. Cham: Springer, 2022: 161—177.

[50] Wang PF, Zhang ZX, Lei Z, et al. Sharpness-aware gradient matching for domain generalization 2023// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 3769—3778.

[51] Hong Y, Zeng ML. International classification of diseases (ICD). Knowledge Organization, 2022, 49(7): 496—528.

[52] Zhu YR, Liang YS, Chen SX. Assessing local emission for air pollution via data experiments. Atmospheric Environment, 2021, 252: 118323.

[53] Huang YX, Guo B, Sun HX, et al. Relative importance of meteorological variables on air quality and role of boundary layer height. Atmospheric Environment, 2021, 267: 118737.

[54] Tong PF, Chen SX, Tang CY. Detecting and evaluating dust-events in North China with ground air quality data. Earth and Space Science, 2022, 9(1).

[55] Zhang Y, Chen SX, Bao L. Air pollution estimation under air stagnation—a case study of Beijing. Environmetrics, 2023, 34(6): e2819.

[56] Dong JS, Roth A, Su WJ. Gaussian differential privacy. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2022, 84(1): 3—37.

[57] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need// Proceedings of the 31st International Conference on Neural Information Processing Systems. California: Curran Associates, 2017: 6000—6010.

[58] Wang HY, Zhu R, Ma P. Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association, 2018, 113(522): 829—844.

[59] Yu J, Wang HY, Ai MY, et al. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. Journal of the American Statistical Association, 2022, 117(537): 265—276.

[60] Chen JY, Lin X, Shi ZQ, et al. Link prediction adversarial attack via iterative gradient attack. IEEE Transactions on Computational Social Systems, 2020, 7(4): 1081—1094.

[61] Gronsbell JL, Cai TX. Semi-supervised approaches to efficient evaluation of model prediction performance. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2018, 80(3): 579—594.

[62] Schifano ED, Wu J, Wang C, et al. Online updating of statistical inference in the big data setting. Technometrics: a Journal of Statistics for the Physical, Chemical, and Engineering Sciences, 2016, 58(3): 393—403.

[63] Chen X, Lee JD, Tong XT, et al. Statistical inference for model parameters in stochastic gradient descent. The Annals of Statistics, 2020, 48(1): 251-273

[64] Zhu WR, Chen X, Wu WB. Online covariance matrix estimation in stochastic gradient descent. Journal of the American Statistical Association, 2023, 118 (541): 393—404.

[65] Chen WD, Xiong S, Chen QY. Characterizing the dynamic evolutionary behavior of multivariate price movement fluctuation in the carbon-fuel energy markets system from complex network perspective. Energy, 2022, 239: 121896.

[66] Jeff Wu CF. On the convergence properties of the EM algorithm. The Annals of Statistics, 1983, 11 (1): 95—103.

[67] Xu L, Jordan MI. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation, 1996, 8(1): 129—151.

[68] Wang, H. Logistic regression for massive data with rare events. International Conference on Machine Learning, 2020, 119:9829—9836.

[69] Mykoniatis K, Harris GA. A digital twin emulator of a modular production system using a data-driven hybrid modeling and simulation approach. Journal of Intelligent Manufacturing, 2021, 32(7): 1899—1911.

[70] Zhang XM, Zhu FL, Zhang JC, et al. Attack isolation and location for a complex network cyber-physical system via zonotope theory. Neurocomputing, 2022, 469: 239—250.

[71] Xu J, Zhang S, Huang E, et al. Mo2tos: multi-fidelity optimization with ordinal transformation and optimal sampling. Asia-Pacific Journal of Operational Research, 2016, 33(3), 1650017.

[72] Peng YJ, Chong EKP, Chen CH, et al. Ranking and selection as stochastic control. IEEE Transactions on Automatic Control, 2018, 63(8): 2359—2373.

[73] Goltsos TE, Syntetos AA, Glock CH, et al. Inventory-forecasting: mind the gap. European Journal of Operational Research, 2022, 299(2): 397—419.

[74] Ray S, Chen HA, Bergen ME, et al. Asymmetric wholesale pricing: theory and evidence. Marketing Science, 2006, 25 (2): 131—154.

[75] Chen Q, Huang ZQ, Jiang X, et al. Asymmetric reporting timeliness and informational feedback. Management Science, 2021, 67(8): 5194—5208.

[76] Li W, Ding S, Chen Y, et al. Transfer learning-based default prediction model for consumer credit in China. The Journal of Supercomputing, 2019, 75(2): 862—884.

[77] Chen H, Fang XW, Fang H. Multi-task prediction method of business process based on BERT and Transfer Learning. Knowledge-Based Systems, 2022, 254: 109603.

[78] Chakrabortty A, Cai TX. Efficient and adaptive linear regression in semi-supervised settings. The Annals of Statistics, 2018, 46(4): 1541—1572.

[79] Zheng XL, Zhu S, Lin ZX. Capturing the essence of word-of-mouth for social commerce: assessing the quality of online e-commerce reviews by a semi-supervised approach. Decision Support Systems, 2013, 56: 211—222.

[80] 陈国青，任明，卫强，等. 数智赋能:信息系统研究的新跃迁. 管理世界，2022，38(1)：180—196.

[81] 陈国青，吴刚，顾远东，等. 管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向. 管理科学学报，2018，21(7)：1—10.

[82] 陈松蹊，毛晓军，王聪.大数据情景下的数据完备化:挑战与对策. 管理世界，2022，38(1)：196—207.

Statistical Management Theory for Business Applications with Massive Scale

Song-Xi Chen[1, 7]　Guoqing Chen[2]　Jinyuan Chang[3, 4]　Hong Huo[5]
Wei Zhang[5]　Xinyu Zhang[4]　Xuening Zhu[6]　Hansheng Wang[7 *]

1. *School of Mathematical Science*, *Peking University*, *Beijing* 100871
2. *School of Economics and Management*, *Tsinghua University*, *Beijing* 100084
3. *Joint Laboratory of Data Science and Business Intelligence*, *Southwestern University of Finance and Economics*, *Chengdu* 610074
4. *Academy of Mathematics and Systems Science*, *Chinese Academy of Sciences*, *Beijing* 100190
5. *Department of Management Sciences*, *National Natural Science Foundation of China*, *Beijing* 100085
6. *School of Data Science*, *Fudan University*, *Shanghai* 200433
7. *Guanghua School of Management*, *Peking University*, *Beijing* 100871

Abstract　The large-scale business scenario is an inevitable outcome of the progress in science and technology and the development of business practices. It not only encompasses business practices geared towards the economic forefront but also includes crucial areas related to national governance, with a focus on the new generation of digital management technology centered around digital twins. Spanning multiple interdisciplinary fields such as management, economics, computer science, environmental governance, mathematics, and statistics, the large-scale business scenario presents a unique opportunity for innovation in management theory. The development of cutting-edge statistical methods and the innovation of data-driven management theories tailored to the large-scale business scenario are important concerns shared by government agencies, industry, and the academic community. Based on the 344th issue of the "Shuang Qing Forum", this paper initiates its exploration from the perspective of the large-scale business scenario, delving into three aspects within complex business scenarios: data analysis methods, statistical computation and optimization methods, and prediction theory and management decision-making. Through a clear definition of relevant concepts and a systematic review of important literature both domestically and internationally, the article summarizes the current research status and frontiers, analyzes development trends and directions, distills significant key scientific issues for the next 5 to 10 years in this field, and discusses cutting-edge research directions and strategies for scientific fund support.

Keywords　large-scale business scenarios; statistical management theory; data analysis; statistical computation and optimization; predictive theory

（责任编辑　张　强）

* Corresponding Author, Email: hansheng@pku.edu.cn