高维数据分析的本质问题思考

报告人：王华忠

波现象与智能反演成像研究组（WPI）

同济大学海洋与地球科学学院，上海

2023年07月13日

目 录

◆一、概述

◆大数据和人工智能时代，高维数据分析（多元数据分析）已经成为广泛的议题。

◆为什么会是这样呢?

◆物理世界和非物理世界，最基本的逻辑是：n个因素共同作用下，“系统”产生响应。

◆基于这样的激励+响应机制，构建一个模型，基于此模型，进行实际过程的预测，然后用于决策。从而解决实际问题。

◆这是一个深层次的、一般规律性的认识世界+改造世界的做法。

◆一、概述

◆n个因素共同作用下，“系统”产生响应。这个过程可以用一个n维函数表达（确定性的表达）；也可以用一个概率模型表达（不确定性表达）。

◆对这个过程建模，是高维数据分析的根本问题。

◆大数据时代，当然是用数据驱动的方式进行建模。

◆事实上，人类社会认识自然和社会规律的思维模式已经逐渐转向通过数据采集/收集、数据分析、建立模型、解决问题的模式上来。建立因果的函数关系的思维模式逐渐退居幕后。

◆一、概述

大数据时代的思维模式

目 录

◆二、高维数据的表现形式

◆勒内·笛卡尔1596年3月31日生于法国安德尔-卢瓦尔省的图赖讷(现笛卡尔，因笛卡儿得名)，1650年2月11日逝世于瑞典斯德哥尔摩，是世界著名的法国哲学家、数学家、物理学家。他对现代数学的发展做出了重要的贡献，因将几何坐标体系公式化而被认为是解析几何之父。他还是西方现代哲学思想的奠基人，是近代唯物论的开拓者且提出了“普遍怀疑”的主张。黑格尔称他为“现代哲学之父”。他的哲学思想深深影响了之后的几代欧洲人，开拓了所谓“欧陆理性主义”哲学。堪称17世纪的欧洲哲学界和科学界最有影响的巨匠之一，被誉为“近代科学的始祖”。 **“我思故我在”！**

◆二、高维数据的表现形式

◆笛卡尔引入坐标系支撑起一个抽象的空间是划时代的思想创举。据此构建起了抽象的“高维空间”概念。

◆在此思维模式下，很多问题变得形象且直观！

◆高维空间中到底有什么？

◆数学上，尤其从解析几何看，高维空间中有几何结构。

◆飘在高维空间中的几何结构到底应该是什么？

◆无非是空间曲面、空间曲线。

◆高维空间中的实体由空间曲面封闭形成。

◆二、高维数据的表现形式

◆从高维数据的角度看，飘在高维空间中的几何结构该如何理解?

- ◆相邻的高维数据取值大小相近。
- ◆更准确的讲：相邻的随机变量存在相关性。

- ◆由相邻的高维数据取值体现出来的相关性，指示了高维数据中包含某种信息。

◆二、高维数据的表现形式

◆高维空间的坐标向量就是高维数据的潜在影响因素形成的。

◆不要机械地从时间、空间坐标的角度理解高维空间的基向量。

◆二、高维数据的表现形式

- ◆总结性观点：
  - ◆高维数据的表现形式就是：
    - ◆飘在高维空间中的“几何结构”。
    - ◆几何上，可以认为“几何结构”是空间曲面、空间曲线。
    - ◆尤其在局部点处，“几何结构”可以是空间平面、空间直线。
    - ◆高维空间的基向量是（应该是）高维数据的特征向量（或潜在影响因素）构成。

目 录

◆三、高维函数及分析方法

◆高维空间中的函数，就是多元自变量的函数（多元函数）。

◆数学分析明确指出：多元函数是n维笛卡尔空间中的空间曲面、空间曲线。

◆空间平面和空间直线是特例。

◆多元函数的微分学和积分学是高维函数的分析方法。

◆三、高维函数及分析方法

◆多元函数的微分学是高维函数分析的主体部分。

◆局部点处，多元函数的连续性、可微性、各阶偏导数、方向导数、梯度、偏微分、全微分、极值是高维函数分析的主要内容。

◆高维函数的积分学是奠基在高维函数的微分学的基础上的。

◆沿曲线的积分、曲面下体积的积分、封闭曲面的体积计算。这也是高维函数分析的重要内容。

◆三、高维函数及分析方法

◆进一步的，微分方程、积分方程、变分原理是高维函数分析的推广。一般地，这些问题中，函数的维度涉及空间三维+时间一维。

目 录

◆四、高维数据分析的目的及分析方法

◆与高维函数分析相比，高维数据分析的目的、思想与方法是很不相同的。

◆高维数据分析的目的：

◆寻找高维数据中存在的相关关系，对高维空间中的“几何结构”进行建模。基于模型进行预测，然后进行决策。从而解决相应的实际问题。

◆四、高维数据分析的目的及分析方法

◆高维数据分析的基本思想：

- ◆实际采集到的物理数据，背后的影响因素（潜在自变量）可能是“n维”的。
- ◆对实测数据进行特征表达，把实测数据升维到“n维”空间中。在这个“n维”空间中，“飘”一些空间几何结构（常被称为“流形”）。
- ◆这些空间几何结构或“流形”是相邻的采样数据取相近的数值体现出来的。从概率统计意义上，是随机采样数据的统计均值体现出这些空间几何结构。

◆四、高维数据分析的目的及分析方法

◆高维数据分析的基本思想：

- ◆把采样数据视为随机变量或随机过程。随机采样数据满足一定的概率分布函数。随机变量之间的相关性由二阶统计量描述，体现为高维空间的几何结构。
- ◆因此，高维数据分析的基本思想可以抽象为：
  - ◆视实测数据为随机过程，进行特征表达升维，形成“n维”空间，对“n维”空间中“飘”的几何结构进行广义线性表达，构建最佳逼近的表达式。基于该最佳逼近表达式，进行预测。解决实际问题。

◆四、高维数据分析的目的及分析方法

◆高维数据分析的基本方法：

- ◆方法1、视实测数据为随机过程，进行特征表达升维。人直接解释数据的特征，从而解决实际问题。
- ◆方法2、视实测数据为随机过程，进行特征表达升维，形成“n维”空间，对“n维”空间中“飘”的几何结构进行广义线性表达。基于标签数据，建立回归分析关系式。利用该关系式，对实际数据进行预测。解决实际问题。
  - ◆有监督学习算法。神经网络算法中，回归分析关系不是显式的，而是体现在“权系数” 中的。
- ◆方法3、视实测数据为随机过程，进行特征表达升维，形成“n维”空间，对“n维”空间中“飘”的几何结构，进行聚类分析。解决实际问题。
  - ◆无监督学习算法。

◆四、高维数据分析的目的及分析方法

◆高维数据分析的基本方法：

◆方法4、半监督学习算法的基本思想：

- ◆首先它应是辅助有监督学习提升学习效果的！根本道理在于：无标签样本的统计特征要与有标签样本保持一致。否则，半监督学习算法可能还会降低有监督学习算法的已有的学习效果。
- ◆因此，半监督学习算法中，如何选择无标签数据应是核心，如何设计半监督学习分类器还是其次。

◆四、高维数据分析的目的及分析方法

◆仿生人脑的神经网络算法的学习机制及高维数据分析：

- ◆感知学习+特征表达+监督学习分类是神经网络算法的思想基础。
- ◆首先，感知的思想是什么?
- ◆权系数向量 w 与输入数据向量 x 之间存在相关性是感知的基础（Hebb学习规则）！感知过程本质上就是特征表达的过程。把实测数据，即输入数据向量 x，进行升维表达（在“n维”空间中表达）的过程。
- ◆感知是通过内积、褶积实现的。内积是个匹配滤波器！把相似的、相关的东西过滤出来。
- ◆基于感知结果的分类，把一个非线性分类问题转化为“拟线性”的分类问题。

◆四、高维数据分析的目的及分析方法

◆仿生人脑的神经网络算法的学习机制及高维数据分析：

- ◆通过感知（内积、褶积）进行特征表达是神经网络算法的本质基础，与标签数据进行差异度量，然后反传播调整权系数。不断地用标签数据训练网络达到稳定的学习效果后，开始应用。
- ◆差异度量+反传播调整权系数的过程，尽管也很重要，但这是第二位的。

◆四、高维数据分析的目的及分析方法

◆仿生人脑的神经网络算法的学习机制及高维数据分析：

- ◆基于神经网络的高维数据分析的适用场景及局限性是什么？它能用于（大规模）参数估计中吗？
- ◆我的观点是：神经网络算法进行（大规模）参数估计逻辑上不通！
- ◆有些ML书上讲：参数估计是连续分类问题。但是，没有理论证明连续分类可以用目前的Bayes分类下发展的方法来解决。SVM用于连续分类似乎是不可思议的。
- ◆参数估计问题是基于因果关系（数学物理方程）进行的。线性系统参数估计的一整套反演解的概念讲得很清楚了（$Ax = b \Rightarrow \hat{x} = \left(A^T A\right)^{-1} A^T b^{obs}$）。标签数据之间存在的是相关关系，基于相关关系能进行参数估计吗？如果能，理论依据是什么？

◆四、高维数据分析的目的及分析方法

◆仿生人脑的神经网络算法的学习机制及高维数据分析：

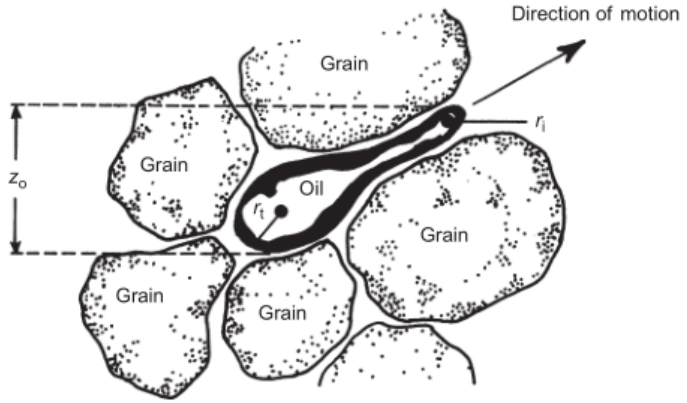
- ◆基于数据间存在的相关关系进行统计推断（统计决策）是没有问题的，目前大数据/AI做的就是这样的事情。

目 录

◆五、地震数据分析向何处去?

◆地震数据分析的一个核心问题---系统参数估计

- ◆波在地下岩石介质中的传播过程及其数学物理描述，是地震数据分析的基础。这种因果关系是地震数据分析的真正基础。
- ◆岩石的物理模型：骨架+孔隙+流体



- ◆岩石的数学模型（平均/等价模型）- 弹性模量与物性参数的关系：
  - ◆Gassman模型（1951）；Biot – Gassman模型；Hill 平均模型；Hudson 模型；.....
  - ◆一大推各种各样的数学模型？！为什么会这样?

◆五、地震数据分析向何处去?

◆地震数据分析的一个核心问题---系统参数估计

- ◆胡克定律
- ◆波动方程---已知激发与系统对输出的描述
  - ◆波在等效介质中传播的数学表达
- ◆对实测波场的预测
- ◆基于Bayes估计理论，构建反问题，进行参数估计。

勘探地震中，三种典型的正问题：
1、岩石物理模型---弹性模量与物性参数的关系；
2、各种形式的波动方程；
3、Zoeppritz方程及其各种简化形式。

◆这样的系统参数估计方法有什么质疑的地方？正问题的正确性！

◆五、地震数据分析向何处去?

◆地震数据分析的一个另一个核心问题---Learning From Data

◆基于地震数据中体现出的相关关系进行地震数据分析。分析地震数据的组成成分，然后进行建模，然后进行分类/聚类，然后进行推断/判决。

◆学习的原则是什么?

◆随机数据满足缓平稳性假设，可以建模进行预测，对期望/标签继续预测。

◆ARMA/AR/MA模型

◆谐波叠加模型、平面波叠加模型

◆自适应滤波是机器学习的根本思想来源。

◆所谓缓平稳性假设，本质上是相邻的随机向量之间存在相关性。

◆高维空间中，“飘”着几何结构。

◆几何结构的建模表达方法---基函数的线性叠加。

◆五、地震数据分析向何处去?

◆地震数据分析的一个另一个核心问题---Learning From Data

- ◆能学到什么?
  - ◆感知数据中包含什么成分
  - ◆学到数据中的特征成分（基函数）
- ◆怎样学习的?
  - ◆感知的方法是用基函数进行相关，匹配。
  - ◆神经网络的（多层）感知机本质上就是在实现基函数的匹配。数据驱动的基函数的匹配。
  - ◆PCA/RPCA/SSA是数据驱动方式感知数据中的基函数的典型算法。

◆五、地震数据分析向何处去?

◆地震数据分析的一个另一个核心问题---Learning From Data

- ◆目前，数据驱动的方式进行地震数据分析广泛地用在数据预处理（去噪、规则化、Deblending）、地震图像分析（层位拾取、断层识别等）中。

- ◆数据驱动的方式，进行特征表达，进行分类/聚类，是合理的做法。但是，进行（大规模）参数估计，不是合理的逻辑选择。说不清楚反演解的性质。无法评判反演解的精度。构不成一个理论体系。

◆五、地震数据分析向何处去?

◆地震数据分析向何处去?

- ◆我的观点是：勘探地震中的参数估计是一个信息不足的反问题。高精度的参数估计，信息不足更为凸显。数据驱动的算法，用在有助于提高参数估计精度的信息提取上，是今后地震数据分析的重要方向。
- ◆Bayes参数估计理论及在勘探地震中的应用（即FWI），我们已经了解很多了。但是，Bayes推断/决策理论及在高维数据/图像分析中的应用（即机器学习算法及应用），我们还要持续强化学习和实践。目的是后者服务于前者，把广义的高精度地震波成像问题解决好，服务于油气地震勘探。

目 录

◆六、总结与讨论

◆20世纪中叶之前的物理科学，主要是寻找物理世界中的、能用数学物理关系表达的因果关系。解决对物理世界的解释问题。“大”科学家对认识物理世界的规律更感兴趣，认为这才是科学问题。

◆基于“大”科学家提出的物理定理（数学物理关系），解决物理世界中改造世界的问题，被认为是技术问题。“大”科学家不屑于做这样的事情。

◆但是，“大”科学家也无法在高维/“n维”空间中建立可行的数学物理关系，尤其是“n维”因变量对结果的影响都不怎么占优的情形下，更无法总结出可用的物理规律。

◆此时，大数据分析就要登场了。

◆六、总结与讨论

◆21世纪以来，技术科学的发展，促进了数据收集/采集技术的极大进步，超大规模算力的计算机系统提供了大数据分析所需的计算能力，于是，人们认识世界的方式发生了巨大的变化。

◆把数据视为高维随机向量，在高维空间中，由于高维随机向量的相关性，高维空间中呈现出“几何结构”。这便是高维空间中的“函数关系”。对此“函数关系”进行建模，就可以进行实际数据的预测，从而解释实际问题，解决实际问题。

◆这就是我理解的高维数据分析的本质。

谢谢
欢迎批评指正