（地震）数据分析的信息论观点

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

2021年1月21日

目 录

◆一、概述

◆广义的数据分析是当今时代各行各业的核心要务。

◆勘探地震领域，叠前数据采集、其它物理场的采集（近地表建模用）、井中地球物理数据的采集、岩石地球物理数据采集等等，以及地震数据分析构成了勘探地震领域的最重要工作。

◆数据分析的目的当然是为了决策。勘探地震数据分析的根本目的是勘探决策或钻井决策。

◆事实上，数据分析的各个阶段有各个阶段的目的。

◆一、概述

◆无论是广义的数据分析或是地震数据分析，首先要从数据中获得有意义的、可解释的信息，然后再提炼形成知识，最后用于决策。

目 录

◆二、数据中信息的赋存形式

◆什么是信息?

- ◆前已述及，这样一个广泛实用的词汇，至今也没有一个统一的定义。
- ◆狭义的信息：
  - ◆Rochester,1996: 信息是一个有组织的事实和数据集合。
    - ◆数据转化为信息；信息转变为知识；知识转变为智慧（决策）。
  - ◆Hartley,1928; Ursul,1971: 信息是被消除的不确定性。
- ◆广义的信息：
  - ◆我的理解：可被用于交流的、完整的知识。

◆二、数据中信息的赋存形式

◆数据的两种形式：

- ◆空间或/和时间有序的数据；
- ◆非空间或/和时间有序的数据。

◆勘探地震中的数据，一般地，被认为是空间或/和时间有序的数据。

◆最通常的数据表达形式，认为数据是多维（高维）随机向量。

◆一般地，认为数据中是包含信息的。

- ◆否者，这样的数据没有被分析和被处理的价值。

◆二、数据中信息的赋存形式

- ◆数据中，信息的赋存形式是什么样的?
  - ◆最基本的形式是各空间或/和时间点上对应的随机变量之间存在协变性，统计上，有相关性。
    - ◆所谓的相关性，形象地讲，就是“事件的发生存在内在关联性”。相互独立的事件的发生，不存在关联性。
  - ◆更为直观的表现：空间或/和时间有序的数据中存在结构信息。
    - ◆即便是非空间或/和时间有序的数据中内在地也存在“结构”信息。
  - ◆现代数据分析语境下，所谓的“结构”信息，常常用“模式”来代表。

目 录

◆三、数据中信息的提取思想与方法

◆所谓数据中信息的提取，本质上，就是对数据中蕴含的结构信息进行感知和表达，尤其是表达。

◆我认为：最基本的感知方法就是获取数据（随机过程）的各阶统计量。

- ◆尤其是一阶和二阶统计量。这已经获得了数据中蕴含的结构信息的大部分内容。
- ◆当前的数据分析，80%以上的研究和应用还是建立在二阶统计量的基础上的。

◆三、数据中信息的提取思想与方法

◆二阶统计量：

- ◆自相关、自功率谱；互相关、互功率谱。

◆自相关反映一个随机过程（随机数据）各空间或/和时间点上对应的随机变量之间的协变程度，完全不协变（独立），随机过程（随机数据）是白噪音，不包含任何信息量！此时的功率谱是各频率相等的，进一步说明了该随机数据不包含任何信息量。

- ◆从信息熵的定义，信息熵最大，不确定性最大！这是从另一个角度的理解！

◆三、数据中信息的提取思想与方法

◆二阶统计量：

- ◆自相关、功率谱；互相关、互功率谱。

◆自功率谱刻画随机过程（随机数据）中包含的特征成分及能量分布情况。

- ◆对数据中所含信号的特征表达的重要方式之一！

◆三、数据中信息的提取思想与方法

◆二阶统计量：

◆自相关、功率谱；互相关、互功率谱。

◆互相关反映两个随机变量（随机数据）不同空间或/和时间点上对应的随机变量之间的协变程度，完全不协变，两随机过程是独立的。互相关系数等于零。

◆互相关反映两个随机变量（随机数据）不同空间或/和时间点上对应的随机变量之间的同相部分能量减去反相部分能量的差，若同相部分能量和反相部分能量相等，互相关系数等于零。

◆三、数据中信息的提取思想与方法

◆二阶统计量：

◆自相关、功率谱；互相关、互功率谱。

◆互功率谱在频率域反映两个随机变量（随机数据）之间的协变程度。

$$P_{xy}(\omega)=\int_{-\infty}^{\infty}C_{xy}(\tau)e^{-j\omega\tau}d\tau \qquad P_{xy}(\omega)=\left|P_{xy}(\omega)\right|e^{j\phi(\omega)}$$

◆互功率谱的振幅谱称同相谱，二者同相部分的振幅频谱；相位谱称群延迟，各频率总体相位移的均值。群延迟为零，二者变成相干信号。相干系数等于1。

◆三、数据中信息的提取思想与方法

◆我的评价：

- ◆到目前为止，能深入理解二阶统计量，并能把它们恰当地用在数据分析中的各环节中，已经成为了数据分析的高手！

◆三、数据中信息的提取思想与方法

◆但是，仅仅停留在一阶和二阶统计量的利用上是不够的。

- ◆二阶统计量：刻画信号的背景部分；
- ◆高阶统计量：刻画信号的细节部分

◆可以毫不夸张地讲：凡是用功率谱或相关函数进行过分析与处理，而又未得到满意结果的任何问题都值得重新使用高阶统计量方法再分析一次。

◆三、数据中信息的提取思想与方法

◆二阶统计量是蕴含在随机过程（随机数据）中的潜变量，由实测数据获取这些潜变量，是一个标准的反问题（参数估计问题）。

◆参数估计问题的基本思想就是Bayes估计。

◆上述过程可以说是统计学家发展出来的数据分析的思想和做法。

◆三、数据中信息的提取思想与方法

◆从数据中提取信息，还有应用更为广泛的方法。这是数学分析学家进行数据分析的做法。

◆首先对实测数据进行建模：

$$\boldsymbol{u}(x,y,t)=\boldsymbol{s}(x,y,t)+\boldsymbol{\eta}_C(x,y,t)+\boldsymbol{\eta}_R(x,y,t)$$

◆其中，$\boldsymbol{u}(x,y,t)$ 是观测的数据；$\boldsymbol{s}(x,y,t)$ 是数据中包含的信号；$\boldsymbol{\eta}_C(x,y,t)$ 是相干噪声；$\boldsymbol{\eta}_R(x,y,t)$ 是随机噪声。

◆数据中的信号 $\boldsymbol{s}(x,y,t)$ 可以用近似函数来逼近。近似函数用一组基函数的线性组合构成。基函数代表了数据中信号的特征。

◆三、数据中信息的提取思想与方法

◆用基函数的线性叠加对数据中包含的信号进行最佳逼近，从而实现实测数据中信号的特征分析或（有用的）信息提取，很显然也是个标准的反问题。

◆ 基本的求解思想也是基于Bayes估计理论。

◆基函数与组合系数同时估计称字典学习问题；稀疏特征表达问题；非高斯或混合高斯噪音假设下的估计问题。当然，还可能有非线性、非高斯情形下的特征表达问题。

◆上述问题代表了当今数据中提取信息的研究热点和研究方向。

◆三、数据中信息的提取思想与方法

◆有监督学习算法、半监督学习算法、无监督学习算法更是目前数据分析的研究热点。

◆前已述及，由（上述）非学习类算法拓展到学习类算法是无法回避的。但是，只有合适的应用场景才能有效地使用学习类算法。

目 录

◆四、数据中信息提取方法的评价

◆到目前为止，从数据中提取信息的各种非学习类算法的思想已经非常清楚了。

◆基于概率统计的方法，就是抽取各阶统计量。尤其是二阶统计量。

- ◆基于自相关的PCA方法
- ◆基于互相关的CCA方法
- ◆基于高阶统计量的ICA方法
- ◆自相关矩阵SVD分解是提取信息核心做法。目前发展到了高维情形。

◆基于泛函分析与函数逼近论的方法，选取基函数（族），由它们代表信号的特征，基函数的线性组合构建信号逼近模型，Bayes理论下实现最佳逼近。

◆四、数据中信息提取方法的评价

◆构建（高维）自相关矩阵，进行SVD分解，得到数据中信号的特征。必须加窗适应信号的非平稳性。只能获取信号的线性特征。施加低秩约束，也是建立在信号的线性特征稀疏的基础上的，否则要么低秩效果不佳；要么对信号的逼近程度太低。引入ICA会改善低秩逼近特性，但不会提升过多。

◆四、数据中信息提取方法的评价

◆目前信号分析的主流方法，我个人的观点，主要还是选择基函数（族）或者基函数字典，在噪音高斯、混合Gauss或非Gauss假设下，进行稀疏反演，提取数据中信号的特征。

◆此时，提取信号特征的能力首先取决于我们构建基函数（族、字典）的能力。总体而言，受限制比较小。

◆当前，小波基（框架基）函数（族）算是最具代表性的方法。但是，小波基（框架基）函数（族）的构造，严重缺乏物理意义。S变换试图引入物理意义，但牺牲掉了什么？分辨率、计算有效性？

◆四、数据中信息提取方法的评价

◆图像可以视为是高维信号（至少大于二维！），图像处理更关注的是图像中模式的提取或模式识别，它是高级图像处理的基础。不能识别图像中的模式，就不可能很好地进行图像的分割、图像识别。模式识别的核心问题就是特征提取问题，与高维信号的特征表达高度一致。

◆四、数据中信息提取方法的评价

◆在信号复杂、噪音复杂的情况下，获取数据中包含的信号的特征是当前数据分析的最核心问题。

◆类似勘探地震数据分析中的地震成像。

◆从含噪数据中，提取信号的特征或者对信号进行有效的表达，在信号复杂、噪音复杂的情况下，远远没有得到很好的解决。

◆我相信，这是个高度非线性的反问题。彻底解决它，任重而道远。

◆目前，我们还没有形成解决这个强非线性问题的独特的观点，更不要提算法了。加窗之类的想法尽管会有效，但毕竟是老生常谈！Radon变换的引入会降低非线性性，也是很经典的做法了！

◆四、数据中信息提取方法的评价

◆对数据中的信号进行建模、进行特征表达的目的，是为了利用它们进行后续的、所谓的“判决”。譬如：

- ◆去噪音
- ◆数据规则化
- ◆数据压缩
- ◆同相轴追踪（初至拾取、层位拾取、断层识别）
- ◆模式识别
- ◆基于特征的划分或分类（近地表特征分区）
- ◆……

◆四、数据中信息提取方法的评价

◆到目前为止，从数据中提取信息并进行自主决策的各种学习类算法的思想，大的框架是清楚的。

◆但是，模拟人脑神经网络的人工深度神经网络算法，尽管在图像和语音处理中比较成功，但它解决非线性问题的能力还要思考与评价！

◆也许，其他的仿生算法，譬如遗传算法、粒子群算法、蚁群算法等，与合理构建的正问题结合，可以更好实现数据中信号特征的表达。噪音的非高斯性是必须要考虑的。稀疏性是非常值得引入的正则化要求！

目 录

目 录

◆五、空间数据中的信息提取与融合

◆勘探地震数据分析是独具特色的。最根本的原因是：地表观测的弹性波场是由波动方程来描述的。地震数据分析的核心问题-弹性参数估计（或称地震波成像）是建立在波动方程解对实测数据逼近的基础上的。

◆Bayes估计理论下的参数估计与前述数据分析并无本质不同。

◆事实上，到目前为止，勘探地震数据分析中的空间数据分析没有受到成像专家的充分重视！

◆大数据时代，这个问题变得无法回避！

◆五、空间数据中的信息提取与融合

◆什么是空间数据?

- ◆由空间位置坐标标识的数据体。

- ◆此类数据是大量的，远远比波场数据存在的场景更为广泛。
- ◆勘探地震中，井数据、近地表数据、速度场数据、深度域成像剖面、各种参数场数据、油田开发的各种相关数据等都是空间数据。

◆五、空间数据中的信息提取与融合

◆空间数据分析的本质是什么？

- ◆寻找同一类空间数据的相关性。
- ◆寻找不同类型空间数据的相关性。

本质上，空间数据分析的根本问题依然是抽取空间数据中蕴含的相关性！

◆空间数据分析的核心问题是什么？

- ◆相关性的表达
- ◆空间数据的相关性不仅仅体现在空间邻域内数据的相关性更强，也体现在属性相同的数据相关性更强。或者二者兼而有之。
  - ◆很多时候后者更重要，但数学书上没有体现，因为这是物理规律决定的！

◆五、空间数据中的信息提取与融合

◆空间数据分析的若干重要应用场景：

- ◆1、散乱数据插值
- ◆2、数据融合
  - ◆本质上应该是信息融合
- ◆3、数据同化

◆五、空间数据中的信息提取与融合

◆散乱数据插值

- ◆利用空间数据的统计相关性（方差）构建局部支撑的基函数，基函数的组合，形成对散乱数据的逼近函数。Bayes理论下，构建最佳逼近反问题。

- ◆最核心的问题：
  - ◆如何利用空间数据的统计相关性（方差）构建局部支撑的基函数?

◆五、空间数据中的信息提取与融合

◆【多（传感器）】信息融合

- ◆融合的定义：
  - ◆对由不同传感器获得的数据进行综合处理和分析，并进行协调、优化、整合，从中提取更多的信息或获得**新的有效的**信息，从而提高决策能力的技术和过程。
- ◆融合的作用：
  - ◆可以扩展对空间和时间信息检测的覆盖范围，提高和改善检测能力，降低信息的模糊性，增加决策的可信度和系统的可靠性。
    - ◆因为引入了冗余信息！

◆五、空间数据中的信息提取与融合

◆数据同化：

- ◆数据同化(data assimilation)是指**在考虑数据时空分布以及观测场和背景场误差**的基础上,在数值模型的动态运行过程中**融合新的观测数据**的方法。
- ◆它是在过程模型的动态框架内,通过数据同化算法不断融合时空上离散分布的不同来源和不同分辨率的直接或间接观测信息来自动调整模型轨迹,以改善动态模型状态的估计精度，提高模型预测能力。

◆五、空间数据中的信息提取与融合

◆数据同化的目的是利用不同来源、不同分辨率的数据来调整模型。也可以说调整系统状态输出。

◆建立一个模型，逼近两个同类数据，类似于数据同化！

◆基于不同来源、不同分辨率的数据，估计系统参数，调整系统状态（输出），也可认为是数据同化的另一条路。

◆总之，数据同化的目的是调整系统状态。

◆也许在勘探地震中，有独特的应用。譬如动态调整钻井轨迹。

◆数据融合和数据同化显然是两个非常不同的问题！

◆五、空间数据中的信息提取与融合

◆空间数据融合的基本思想：

- ◆1、存在数据融合和信息融合（特征融合）两个层次，数据融合是基础层次，更核心的是信息融合。
- ◆2、融合存在交集融合和并集融合两种概念。一般地，应该是并集融合。前者会减少决策信息。决策是需要冗余信息的。
- ◆3、融合的根本点是提取二者或多者间的互相关信息。

◆五、空间数据中的信息提取与融合

◆空间数据融合的基本方法：

- ◆1、加权平均方法；
- ◆2、尺度（特征）分解，同尺度加权平均。一般在变换域进行。
  - ◆Fourier变换，PCA、小波变换、Hilbert-Huang变换

◆五、空间数据中的信息提取与融合

◆空间数据融合结果的评价方法：

- ◆1、信息熵；
- ◆2、交叉熵；
- ◆3、相关熵/联合熵；
- ◆4、互信息。
  - ◆各种信息熵计算中的概率密度函数，由灰度直方图近似。

◆融合结果的熵增加，说明融合后结果包含的信息比融合前结果更丰富。

◆五、空间数据中的信息提取与融合

◆空间数据分析，在勘探地震中，应逐渐加强。目的是把各种与参数场相关的信息都融合到弹性参数估计得到的参数模型中，提高参数模型的精度、分辨率。

◆最终目的是提升储层描画的精度、提升钻井决策的有效性。

目 录

◆六、总结与讨论

◆（地震）数据中包含的信息是随机过程（实测数据）中空间/时间点的随机变量间的统计相关性体现出来的。空间/时间点的随机变量间不存在统计相关性，也就不存在信号，也就没有了信息。

◆更确切地讲，信息由信号的特征表达出来。

◆（地震）数据分析（信号/图像分析）的中心任务就是把数据中包含的信号的特征挖掘出来，用于后续的“判决”中。

◆六、总结与讨论

◆是时候把地震数据分析与一般性数据分析完全融合在一起了，尽管地震数据分析的正问题是由波动方程描述的（这是地震数据分析的最大特点了）。

◆无论什么类型的数据分析，最核心的事情就是从含噪数据中提取信号的特征或者对信号进行表达/逼近。复杂噪音和复杂信号情况下，这是个强非线性问题。

◆勘探地震中，我们关注的是参数估计这个强非线性问题，基于勘探地震的数据特点、观测系统特点和介质分布特点，提出了我们一列观点（CWI）。

◆六、总结与讨论

◆广义线性模型、（混合）高斯或非高斯噪声、字典基与稀疏表达代表了当前数据分析中信号特征提取的研究热点。

◆但是，复杂噪音和复杂信号情况下，信号分析/图像分析问题依然得不到很好的解决。问题的关键就是此时信号的表达/逼近是个强非线性问题。正如FWI是地震数据分析中强非线性问题。

◆我认为（至少目前认为）：强非线性问题没有理想解法。

◆六、总结与讨论

◆我认为：无论何种来源的非线性问题的求解，FWI/信号分析/图像分析/ML，减缓非线性性的根本还是在于正问题的构造。

◆二阶统计量描述的是信号的缓变部分或主要部分；高阶统计量描述的是信号的快变部分。分尺度描述信号的思想已经体现出来。

◆Hilbert-Huang变换也是分尺度描述信号的典型方法。

◆B样条函数及B样条与2带小波变换结合应该是描述信号的、值得探讨的方法。

◆总之，还是要从描述信号/图像的正问题入手，才能更好地解决信号/图像的特征表达问题。

◆六、总结与讨论

◆WPI正在做的所有研究工作，都与解一个强非线性问题有关。希望所有人的具体工作，都抽象在此思想框架下。聚焦在根本问题上，从而找到创新的解决办法。

◆只有创新才是科研的唯一目的！

谢谢

欢迎批评指正