统计数据分析的基本思想与方法

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06月15日

目 录

一、概述

- 大数据+高性能计算机+ML算法引领人类进入所谓的人工智能时代。
- 大数据的来源：各种各样的传感器。
- 大数据的传输：4G/5G、WIFI。
- 大数据的管理：数据标准、数据库。
- 大数据的存储：大型数据中心。
- 大数据是人工智能时代的核心资源。

一、概述

- 人工智能的本质是：
  - 利用高性能计算机和ML算法从大数据中挖掘出用于判断和决策的信息，基于挖掘出的信息做出决策错误概率最小的决策。

一、概述

◆人工智能的基础：从数据中挖掘出能用于决策的信息。

◆数据挖掘的思想基础：（概率与）数理统计

◆数理统计学的定义（中国大百科全书•数学）：数理统计学研究如何有效地收（采）集、整理和分析带有随机性的数据，以对所考查的问题做出**推断或预测，甚至为采取一定的决策和行动**提供依据和建议。

一、概述

- 可以认为：
  - 数理统计学为基于大数据的人工智能奠定了扎实的数学基础。

- 也可以不夸张地认为：
  - 基于大数据的人工智能奠基在数理统计学上。

一、概述

- 数据的随机性：
  - 无论是物理系统产生的数据或是非物理数据产生的数据都可以认为具有一定的随机性。

- 数据表现出随机性的原因：
  - 1、观测数据背后的影响因素很多，且各影响因素权重基本差不多。
  - 2、观测或收集数据的方式导致每次观测或收集的数据结果不同，且表现为随性。

一、概述

- 随机性数据的数学模型：
  - 随机变量：某种条件下按一定概率发生的事件（结果）称为随机变量。
  - 随机过程：时间或空间相关的随机变量族称为随机过程。
  - 描述随机变量和随机过程性的数学模型分别是概率分布函数（概率密度函数）和联合概率分布函数（概率密度函数）。

- 随机数据分析的主要任务：
  - 建立相应的（联合）概率分布函数（概率密度函数）
  - 估计（联合）概率分布函数（概率密度函数）中的参数

一、概述

- 随机数据分析的核心工作：
  - 模型推断和参数推断（参数估计）
  - 假设检验

- 我认为随机数据分析的本质依然是对（高维）随机数据（随机过程）进行建模。理论上，就是建立联合概率密度函数。
  - 高维随机过程的高斯分布函数就是一个典型。

目 录

二、统计数据分析的基本思想与方法

- **统计数据分析的核心问题是建立随机变量或随机过程的总体模型。**
  - 简言之，就是随机变量或随机过程满足的（联合）概率分布函数。

- **总体模型：**
  - 当刻画总体的随机向量 $\boldsymbol{X}$ 的分布族 $\{F_\theta, \theta \in \Theta\}$ 确定后，$\boldsymbol{X} \sim \{F_\theta, \theta \in \Theta\}$ 就形成了总体模型。

- **统计模型：**
  - 将采样数据 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 视为总体随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ 的一组独立观察值，其中 $X_1, X_2, \cdots, X_n \sim iidF_\theta$，称 $(X_1, X_2, \cdots, X_n)$ 为来自总体 $\boldsymbol{X}$ 的一个简单随机采样（即样本）。$\boldsymbol{X}$ 的取值空间 $\mathcal{X}$ 称为样本空间。$\boldsymbol{X}$ 和它相应的分布族 $\{P_\theta, \theta \in \Theta\}$ 称为统计模型。

二、统计数据分析的基本思想与方法

◆统计数据分析的中心工作是依据观测的随机数据估计统计模型 $\{P_\theta, \theta \in \Theta\}$ 中的参数 θ 或者估计统计模型 $\{P_\theta, \theta \in \Theta\}$ 的各阶统计量。

◆参数往往就是统计量。

◆由随机数据到统计模型是数据统计成为科学的关键。基于统计模型才能进行科学的统计推断。

◆概率论中，随机变量或随机过程的中心矩和原点矩给出了统计量的数学定义。

◆在统计模型 $\{P_\theta, \theta \in \Theta\}$ 的框架下，各阶统计量代表了随机数据中蕴含的信息。

二、统计数据分析的基本思想与方法

- **典型的概率分布模型：**
  - 二点分布是二项分布的特例（n=1）
  - 二项分布 $n, p(0<p<1)$，当n趋于无穷大，且 p 较小时变为泊松分布。当n趋于无穷大，且 p 较大时变为正态分布。

表 2.7.1 常用分布表

名称	概率分布	均值	方差	参数的范围
二点分布	$P(X=x)=p^x q^{1-x}$ $(x=0,1)$	p	pq	$0<p<1$ $q=1-p$
二项分布	$P(X=x)=C_n^x p^x q^{n-x}$ $(x=0,1,\cdots,n)$	np	npq	$0<p<1$ $q=1-p$ n 正整数
泊松分布	$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda}$ $(x=0,1,2,\cdots)$	λ	λ	$\lambda>0$
超几何分布	$P(X=x)=\frac{C_{N-M}^{n-x}C_M^x}{C_N^n}$ ① $(x=0,1,\cdots,\min(M,n))$	$\frac{nM}{N}$	$\frac{n(N-n)(N-M)M}{N^2(N-1)}$	n,M,N 正整数 $n\leqslant N$ $M\leqslant N$
负二项分布	$P(X=x)=C_{r+x-1}^{r-1}p^r q^x$ $(x=0,1,2,\cdots)$	$\frac{rq}{p}$	$\frac{rq}{p^2}$	$0<p<1$ $q=1-p$ r 正整数
均匀分布	$p(x)=\frac{1}{b-a}(a\leqslant x\leqslant b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$b>a$
指数分布	$p(x)=\lambda e^{-\lambda x}(\lambda>0,x>0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\lambda>0$
正态分布	$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	μ 任意 $\sigma>0$
伽玛分布	$p(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ $(x>0)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\alpha>0$ $\beta>0$
贝塔分布	$p(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}\cdot(1-x)^{\beta-1}$ $(0<x<1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$\alpha>0$ $\beta>0$
对数正态分布	$p(x)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\lg x-\mu)^2}{2\sigma^2}}$ $(x>0)$	$e^{\mu+\frac{1}{2}\sigma^2}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$	μ 任意 $\sigma>0$
威布尔分布	$p(x)=\frac{mx^{m-1}}{\eta^m}e^{-\left(\frac{x}{\eta}\right)^m}$ $(x>0)$	$\eta\Gamma\left(1+\frac{1}{m}\right)$	$\eta^2\left[\Gamma\left(1+\frac{2}{m}\right)-\Gamma^2\left(1+\frac{1}{m}\right)\right]$	$m>0$ $\eta>0$

二、统计数据分析的基本思想与方法

- 中心极限定理之一：棣莫弗---拉普拉斯定理
  - 设随机变量 $X_n, n=1,2,\cdots,n,\cdots$，服从参数为 $n, p(0<p<1)$ 的二项分布，则对任意 x，有：

$$\lim_{n\to\infty} p\left\{\frac{X_n-np}{\sqrt{np(1-p)}}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x)$$

- 随机变量符合什么样的概率分布本质上是背后的物理原因或客观原因决定的！

二、统计数据分析的基本思想与方法

◆**统计数据分析的本质**就是在统计模型 $\{P_\theta, \theta \in \Theta\}$ 的框架估计随机数据中蕴含的信息，就是估计统计模型 $\{P_\theta, \theta \in \Theta\}$ 中的参数 θ 或各阶统计量。

◆本质上，这就是要建立随机数据（高维随机过程）的统计模型，也就是联合概率密度函数。

◆Bayes统计学派和经典统计学派（频率学派）发展了两种代表性的统计估计思想。

二、统计数据分析的基本思想与方法

- Bayes统计学派和经典统计学派思想的差异：
  - Bayes统计学派把要估计的模型参数视为已知分布的随机变量。
  - 经典统计学派把要估计的模型参数视为确定的待估计变量。

- Bayes统计学派在统计数据分析中（包括在当前的ML/AI中）占统治地位。

二、统计数据分析的基本思想与方法

- Bayes统计思想的关键点：
  - 把要估计的模型参数视为随机变量 Θ，假设已知其分布 p_{Θ} 。
  - 假设已知观测向量 x 的先验条件分布 $p_{X|\Theta}$ 。
  - 由Bayes法则计算后验概率分布 $p_{X|\Theta}\left(\theta \middle| X=\mathbf{x}\right)$ 。
  - 按照估计准则进行参数估计。

二、统计数据分析的基本思想与方法

◆Bayes统计推断思想：

◆Bayes法则：

$$p_{\Theta|X}(\theta|\mathbf{x})=\frac{p_{\Theta}(\theta)\,p_{X|\Theta}(\mathbf{x}|\theta)}{\sum_{\theta'}p_{\Theta}(\theta')\,p_{X|\Theta}(\mathbf{x}|\theta')}$$

二、统计数据分析的基本思想与方法

- Bayes统计推断思想中的估计准则：
  - 后验条件期望作为参数估计结果
    - 后验条件方差作为参数估计结果的精度评价
  - MAP准则
  - 均方误差最小准则
  - 线性均方误差最小准则---Wiener（反）滤波
  - 最小二乘准则---参数估计（FWI/TOMO/LS_RTM、信号与图像建模）

退化到误差泛函取极小的优化问题时，统计估计理论与函数逼近理论就建立起了紧密的联系！

二、统计数据分析的基本思想与方法

- 统计数据分析中的假设检验问题：
  - 假设检验的目的是判断我们依据统计模型做出的判断是否是可以接受的，或者说是否是合适的。

- 假设检验问题是统计数据分析中的一个十分重要的问题，尤其在ML/AI算法中有重要地位。

二、统计数据分析的基本思想与方法

- 假设检验问题：
  - 设 $X \sim \{F_\theta, \theta \in \Theta\}$ 为总体模型，所谓假设检验问题是两个关于总体（参数）真值的互相对立判断 $(\theta \in \Theta_0, \text{or}, \theta \in \Theta_1)$ 的鉴定问题，其中 Θ_0 是 Θ 的一个真子集，$\Theta_1 = \Theta \setminus \Theta_0$ 为 Θ_0 的余集。判断 $\theta \in \Theta_0$ 称为零假设，记为 H_0；判断 $\theta \in \Theta_1$ 称为对立假设，记为 H_1。

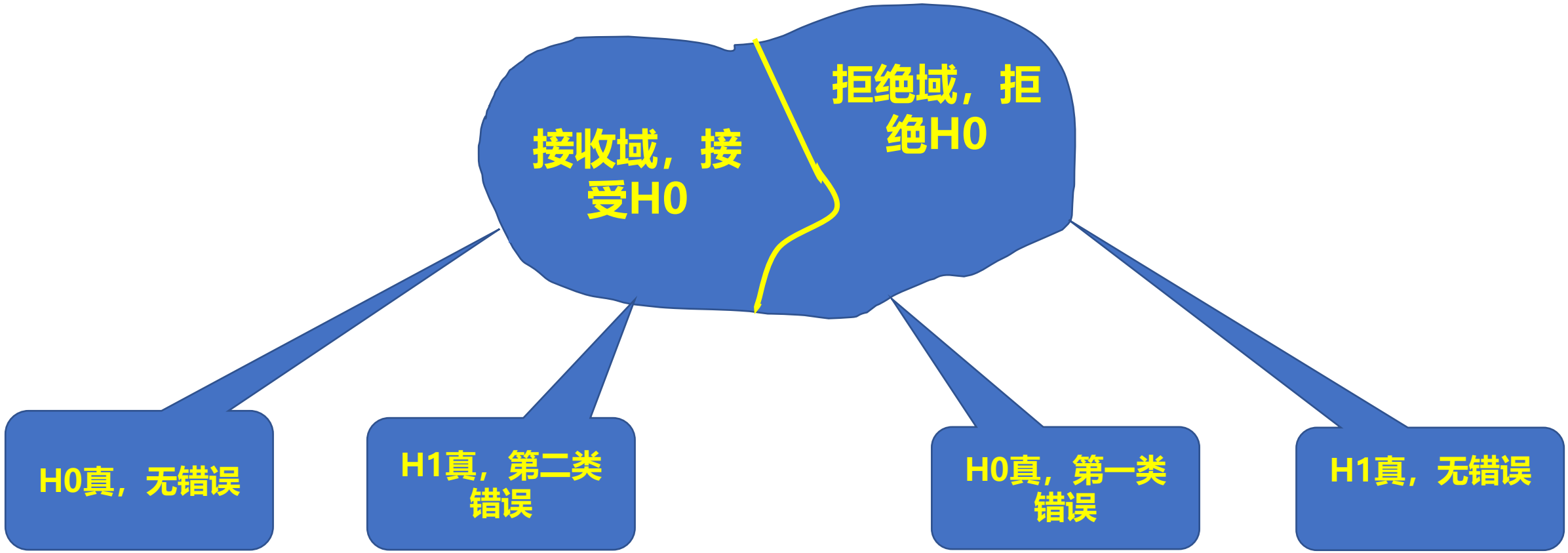
- 假设检验问题也可以认为是决策合理性的判决问题。

二、统计数据分析的基本思想与方法

- **假设检验中的功效函数及显著性水平：**
  - 假设(Θ_0,Θ_1)为总体模型$\boldsymbol{X}\sim\{F_\theta,\theta\in\Theta\}$的一个假设检验问题，$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$是总体的一个样本，$\mathcal{W}$为检验问题$(\Theta_0,\Theta_1)$的一个水平为$\alpha(\alpha\in(0,1))$的否定域。
  - （1）称定义在Θ上的函数$\beta_{\mathcal{W}}(\theta)\triangleq P_\theta(X\in\mathcal{W})$为$\mathcal{W}$的功效函数。
  - （2）若$\mathcal{W}$满足条件$\sup_{\theta\in\Theta_0}\beta_{\mathcal{W}}\le\alpha$，称$\mathcal{W}$为检验问题$(\Theta_0,\Theta_1)$的一个显著性水平为$\alpha(\alpha\in(0,1))$的否定域。

二、统计数据分析的基本思想与方法

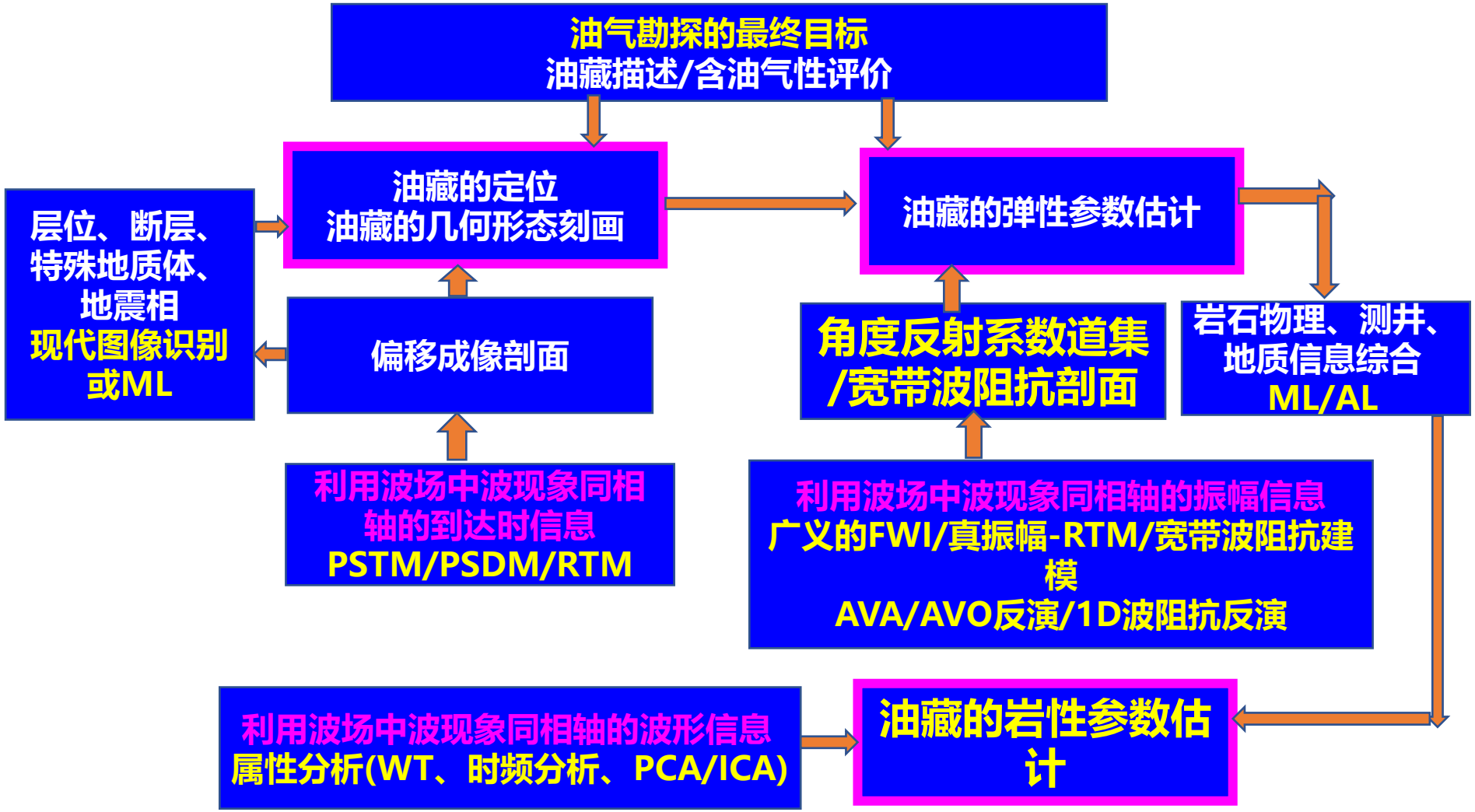
◆统计数据分析中的假设检验问题图示：



目 录

三、统计数据分析思想方法在勘探地震学中的应用

◆勘探地震中，地震数据分析的主要工作内容



三、统计数据分析思想方法在勘探地震学中的应用

- 解决地震数据分析中问题的统一思想框架：
  - 叠前地震数据（炮集），也包括其他的各种来源的数据，原则上，都可以视为随机数据。理论上，用随机变量或随机过程 x 来描述。
  - 随机过程 x 可以被描述为：$X=S+\eta$
    - 其中 s 被视为空间、时间或空间时间的函数。它是确定性的，是随机数据中蕴含的信息，是空间、时间或空间时间相关的，反映随机数据具有的内在结构。
    - η 是加性的随机变化部分，它规定了随机数据的统计模型。

三、统计数据分析思想方法在勘探地震学中的应用

- 解决地震数据分析中问题的统一思想框架：
  - 对于数理统计学家，随机数据的总体模型（统计模型）的建立是核心问题。主要工作是估计总体模型（统计模型）中的参数。
    - 矩估计、点估计、区间估计是常用的参数估计方法。
    - 线性与非线性回归分析也可以归结为总体模型（统计模型）中的参数估计问题。
    - Bayes估计与最大似然估计是最基本的思想与方法。
    - 假设检验可以用来评价估计或推断的合理性。

三、统计数据分析思想方法在勘探地震学中的应用

- 解决地震数据分析中问题的统一思想框架：
  - 对于数学分析学家，对随机数据中包含的函数 S 进行最佳的函数逼近（对函数 S 进行建模）是核心问题。
  - 数学分析学家仅仅关注对函数 S 的逼近，认为函数 S 是确定性的，无噪音的。首先提出逼近误差的度量（无穷范数、L2范数、L1范数，Lp范数等），这相当于定义了一个变分问题。然后，对函数 S 的逼近是在Hilbert空间的一个子空间上进行的，子空间是人为选定的一组基（框架）函数或基（框架）函数字典张成的。最后，在度量结果最小（逼近误差最小）的意义下，用一定的算法求解出对函数 S 的最佳逼近预测器。使用该最佳逼近预测器，就可以把 S 从实测的、含噪的随机数据中“最佳地”提取出来。

三、统计数据分析思想方法在勘探地震学中的应用

- 解决地震数据分析中问题的统一思想框架：
  - 数理统计学家和数学分析学家的思想与方法本质上是存在统一性的。数理统计学家估计的各阶统计量（尤其是二阶统计量）与数学分析学家获得的对数据中蕴含的函数 S 进行的最佳逼近的基函数存在本质上的一致性。
    - 譬如自相关函数的正交分解得到平稳、高斯数据的基函数；Fourier基函数族的线性组合可以最佳逼近平稳、高斯数据。
    - 理论上可以证明：自相关函数的正交分解得到的平稳、高斯数据的基函数就是Fourier基函数。

三、统计数据分析思想方法在勘探地震学中的应用

- 解决地震数据分析中问题的统一思想框架：
  - 数理统计学家建立了更抽象的数据分析的理论框架，但是总体模型（统计模型）是很难预先获得的。
    - 这限制了基于数理统计进行数据分析的能力。
  - 数学分析学家从函数逼近的角度构建的对数据中包含的函数 S 的“最佳”逼近能力是非常强大的。但是，不从概率统计的角度出发，难以从本质上讲清楚误差泛函构造的含义。

- 事实上，这两种不同的数据分析思想和方法已经高度融合了！

三、统计数据分析思想方法在勘探地震学中的应用

- 勘探地震中三种典型的数据分析问题模型：
  - 图像处理中ROF模型：$J(\boldsymbol{u}) = \frac{1}{2}\|A\boldsymbol{u}-\boldsymbol{g}\|^2 + \alpha TV(\boldsymbol{u})$ 。对于图像去噪问题，$A=\boldsymbol{I}$ 单位矩阵 。
  - 信号处理中去噪或反褶积模型：$J(\boldsymbol{\beta}) = \frac{1}{2}\|\Psi\boldsymbol{\beta}-d^{obs}\|^2 + \alpha|\boldsymbol{\beta}|_1$
  - 物理系统参数估计模型：$J(\boldsymbol{m}) = \frac{1}{2}\|A\boldsymbol{m}-\boldsymbol{d}^{obs}\|^2 + \alpha\Omega(\boldsymbol{m})$ 。$\Omega(\boldsymbol{m})$ 代表不同形式的正则化方式。

三、统计数据分析思想方法在勘探地震学中的应用

- 当前，地震数据分析中绝大部分的应用问题都可归结如下

- 正问题模型：
  - 2D褶积模型：$g(x,y)=\iint f(x',y')h(x-x',y-y')dx'dy'+\varepsilon(x,y)$
  - 高维褶积模型的抽象表示：$g(s)=\left[Hf(r)\right](s)+\varepsilon(s), r\in R, s\in S$

- 反问题模型：$J(f)=\left\|g-H(f)\right\|^2+\lambda\Omega(f)$

三、统计数据分析思想方法在勘探地震学中的应用

系统参数估计的Bayes理论框架

求取后验概率密度分布

$$\rho_{\mathbf{M}}(\mathbf{m}\,|\,\mathbf{d})=\frac{\rho(\mathbf{d},\mathbf{m})}{\rho(\mathbf{d})}=\frac{\rho(\mathbf{d}\,|\,\mathbf{m})\rho(\mathbf{m})}{\rho(\mathbf{d})}$$

求取估计结果的均值和方差

$$\hat{\mathbf{m}}=\int \mathbf{m}\rho_{\mathbf{M}}(\mathbf{m}\,|\,\mathbf{d})dm \quad C_M=\int(\mathbf{m}-\hat{\mathbf{m}})(\mathbf{m}-\hat{\mathbf{m}})^T\rho_{\mathbf{M}}(\mathbf{m}\,|\,\mathbf{d})d\mathbf{m}$$

后验概率密度最大化

$$\hat{\mathbf{m}}_{\mathrm{MAP}}=\arg\max_{\mathbf{m}}\rho_{\mathbf{M}}(\mathbf{m}\,|\,\mathbf{d})$$

非线性的FWI

引入高斯分布假设，转化为代价函数最小

$$\hat{\mathbf{m}}=\arg\min_{\mathbf{m}}\{S(\mathbf{m})\}=\arg\min_{\mathbf{m}}\left\{(\mathbf{m}-\mathbf{m}_{prior})^T\mathbf{C}_{\mathbf{M}}^{-1}(\mathbf{m}-\mathbf{m}_{prior})+(\mathbf{g}(\mathbf{m})-\mathbf{d}_{obs})^T\mathbf{C}_{\mathbf{D}}^{-1}(\mathbf{g}(\mathbf{m})-\mathbf{d}_{obs})\right\}$$

正问题的线性化

$$\mathbf{d}_{obs}=\mathbf{g}(\mathbf{m}_{\mathrm{B}}+\Delta\mathbf{m})=\mathbf{g}(\mathbf{m}_{\mathrm{B}})+\frac{\partial\mathbf{g}}{\partial\mathbf{m}}\Delta\mathbf{m}+\frac{1}{2}\frac{\partial^2\mathbf{g}}{\partial\mathbf{m}^2}\Delta\mathbf{m}^2+\cdots$$

$$\Delta\mathbf{d}_{obs}\approx\mathbf{g}(\mathbf{m}_{\mathrm{B}}+\Delta\mathbf{m})-\mathbf{g}(\mathbf{m}_{\mathrm{B}})=\frac{\partial\mathbf{g}}{\partial\mathbf{m}}\Delta\mathbf{m}$$

$$\Delta\mathbf{d}_{obs}=\mathrm{G}\Delta\mathbf{m}$$

线性化的FWI：层析成像；LS_PSMD

$$\hat{\mathbf{m}}=\arg\min_{\mathbf{m}}\left\{(\mathbf{m}-\mathbf{m}_{prior})^T\mathbf{C}_{\mathbf{M}}^{-1}(\mathbf{m}-\mathbf{m}_{prior})+(\mathrm{G}\Delta\mathbf{m}-\Delta\mathbf{d}_{obs})^T\mathbf{C}_{\mathbf{D}}^{-1}(\mathrm{G}\Delta\mathbf{m}-\Delta\mathbf{d}_{obs})\right\}$$

三、统计数据分析思想方法在勘探地震学中的应用

◆统计量与数据的内在关系分析（以正态分布为例）：

◆定义：称 n 维随机向量 $\xi=(X_1,X_2,\cdots,X_n)^T$ 服从 n 维正态分布，若 ξ 有如下概率密度函数

$$p(x)=\frac{1}{(2\pi)^{1/2}|C_{xx}|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T C_{xx}^{-1}(x-\mu)\right\}\xi$$

◆其中，$x=(x_1,x_2,\cdots,x_n)^T$，$\mu=(\mu_1,\mu_2,\cdots,\mu_n)^T$ 是 n 维向量，C_{xx} 是 $n\times n$ 正定对称矩阵。记为 $\xi\sim N(\mu,C_{xx})$。

三、统计数据分析思想方法在勘探地震学中的应用

◆统计量与数据的内在关系分析（以正态分布为例）：

◆定理：设 $\xi=(X_1,X_2,\cdots,X_n)^T\sim N(\boldsymbol{\mu},C_{xx})$，

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},|A|\neq 0 \quad \zeta\left(=(Y_1,Y_2,\cdots,Y_n)^T\right)=A\xi\left(=(X_1,X_2,\cdots,X_n)^T\right)$$

◆则，$\zeta=(Y_1,Y_2,\cdots,Y_n)^T\sim N(A\boldsymbol{\mu},AC_{xx}A^T)$。

◆自相关矩阵C_{xx} 存在正交分解 $C_{xx}=Q^T\Sigma Q$。则 $A=Q^T$。矩阵 A 的列向量形成一组正交基。K-L变换的理论基础。数据驱动的特征表达。特征就是基函数。可以扩展到基于数据驱动的高维数据的特征表达。

三、统计数据分析思想方法在勘探地震学中的应用

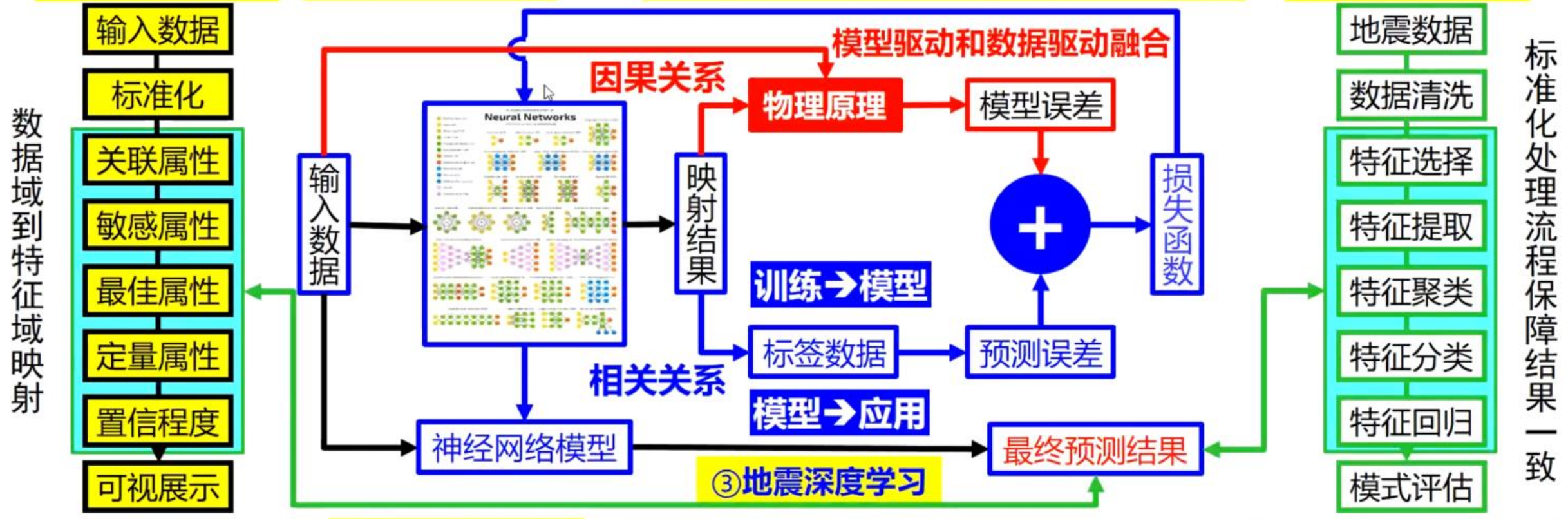
- 统计量与数据的内在关系分析（以正态分布为例）：
  - 二阶统计量---自相关矩阵 C_{xx} 蕴含了数据的（线性结构）特征。对应的功率谱从频率域反映了数据的特征。
  - 基于数据驱动的数据特征提取（挖掘）构成了AI/ML的核心任务。

统计数据分析思想方法在勘探地震学中的应用

中国石油

□ 数据驱动的非线性智能地震技术，深度学习融入行业知识，全流程重塑地震解决方案

①地震迭代映射　④地震人工智慧　$LOSS(\varepsilon)=\alpha D(f)+\beta M(f)$，$\alpha+\beta=1$　②地震数据挖掘



提供更好的属性数据　⑤地震自然智能　提供更好的解决方案　提供更好的地震特征

AI地震生态：平台层(系统管理)→数据层(数据管理)→计算层(集成算法和框架)→应用层(行业AI引擎)

中国石油勘探开发研究院 郑晓东　zxd@petrochina.com.cn

目 录

三、统计数据分析思想方法在机器学习中的应用

◆ML的目的是构建一个LM（Learning Machine）。其中的核心问题是：（随机）数据信息挖掘+基于所挖掘信息的决策。

◆勘探地震中的参数估计可以认为是（随机）数据信息挖掘。含油气性的判断可以认为是决策问题。

◆AI/ML的特点（侧重点）在于决策，最佳的决策，代价最小的决策。但是其中的数据挖掘或信息提取是核心！

三、统计数据分析思想方法在机器学习中的应用

◆Bayes 推断（Bayesian Inference）可以认为是ML的基本理论基础。

◆绝大部分、甚至所有的，ML算法都是奠基在Bayes 推断思想上的！

◆Bayes 推断基本理论：$$P(x)=\int P(x,\theta)d\theta=\int P(x|\theta)\hat{P}(\theta)d\theta=\int P(x|\theta)P(\theta|\mathcal{D})d\theta$$

$$=\int P(x|\theta)\frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}d\theta$$

◆其中，x 代表决策行动。$P(x)$ 是具体决策行动发生的概率。$\mathcal{D}$ 是样本数据。θ 是 $\mathcal{D}$ 所满足的统计模型中的参数。$P(\theta)$ 是参数 θ 的先验概率分布。

三、统计数据分析思想方法在机器学习中的应用

◆理论上讲，决策行动只能发生在$P(x)$比较大的地方。

◆很显然，$\hat{x}=\max_{x\in\Omega_x} P(x)$ 是一个很好的决策原则。$\hat{x}=E(x)=\int xP(x)dx$ 也是很好的决策原则。

◆但是，$P(x)=\int P(x|\theta)\frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}d\theta$ 的具体构建和计算是个困难的问题。

三、统计数据分析思想方法在机器学习中的应用

- 上述描述构成了ML中Bayes判决的基本思想。

- ML中的所有算法，都是围绕着$P(x)$展开的。
  - Monte Carlo方法
  - 概率图模型（Probabilistic Graphical Model）方法
  - 粒子滤波（Particle Filtering）方法
    - 这些是与$P(x)$的计算有直接关系的算法。

  - ANN/DL本质上也是在Bayes判决的思想下发展出的算法，尽管看起来它是在模仿人脑，但数学算法上，还是在具体实现Bayes判决。

三、统计数据分析思想方法在机器学习中的应用

- 从数学上看，完整地解决了决策变量 x 的统计建模

$$P(x)=\int P(x|\theta)\frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}d\theta$$

- 等于解决了统计决策问题。

- $P(\theta|\mathcal{D})$ 的问题等价于回归问题，进一步可退化为特征表达问题。
  - 这些是数据挖掘要解决的核心问题。

- 基于 $P(x|\theta)$，要解决的是分类问题（或判决问题）。

三、统计数据分析思想方法在机器学习中的应用

◆回归分析（Regression）构成了ML算法的基本；Bayes推断（判决）（统计分类/Classification）才是ML的目的。

◆更具体地，数据挖掘构成了ML算法的基本，分类与聚类才是ML的目的。

◆数据挖掘与数据的特征表达在某种意义上是等价的。高维数据的（稀疏）特征表达是当前统计数据分析的最高发展阶段。（稀疏）特征表达又是当前（高维）函数逼近的核心问题。

三、统计数据分析思想方法在机器学习中的应用

◆LASSO问题（模型驱动），Dimensionality Reduction问题（数据驱动）是高维数据统计学的研究议题。

- ◆Statistics for high-dimensional data侧重理论研究，LASSO方法是核心。
- ◆R_PCA、ICA、Low_Rank Matrix Factorization更侧重直接估计高维随机数据中的稀疏结构特征。

三、统计数据分析思想方法在机器学习中的应用

◆据此可以看出，ML/AI的侧重点在于统计判决，决策变量 x 的统计建模是中心问题。

◆这是一个理论问题。对ML的具体实践起到指导作用。但不可以生搬硬套。

◆Bayes估计理论与地震波成像也是这种关系。

◆ML/AI情形下的Bayes决策更抽象，$P(\theta|\mathcal{D})$ 和 $P(x|\theta)$ 的内涵更不容易把握。

目 录

四、结论与讨论

◆在大数据和人工智能阶段，把地震数据分析奠基在概率与统计数据分析的基础上是不可回避的。

◆概率与统计数据分析奠定了（随机）数据分析的更为抽象与扎实的思想基础。它完美地阐释了（随机）数据分析的本质。

◆（随机）数据分析的总体模型（统计模型）是最核心的概念。总体模型（统计模型）中所包含的参数的估计问题是（随机）数据分析的根本问题。估计结果的假设检验是另一个重要问题。

◆尤其在智能数据分析时代，推断或决策效果的评价是很重要的。

四、结论与讨论

- Bayes统计（估计）思想是占据最核心位置的思想。
  - 据此发展出的估计方法（MAP估计、均方误差最小估计、最小二乘估计等）构成了数据分析中的核心方法，广泛应用在信号分析、图像分析与识别、地震波反演成像（FWI/TOMO/LS_RTM）中。
- 函数逼近理论提供了强大的数据分析工具。小波变换、框架字典、压缩感知与稀疏表达代表了当前函数逼近理论的最高发展成果。
- 高维随机数据的统计稀疏表达是当今统计数据分析的代表性思想与方法。
- 基于此，数学分析学家和概率统计学家又走在了一起！

四、结论与讨论

◆但是，即便理论上发展到如此阶段，针对复杂的信号（图像或波场）、复杂的噪音情形下，从随机观测数据中有效地提取信号或对信号进行特征表达也是极为困难的事情。很难保证后续的判断或决策的可靠性。

◆对随机数据分析理论与方法的掌握是一个方面，对实际数据的深入理解是另一个方面，两个方面都很强，才能较好地解决随机数据分析中的问题，做出最佳的判决，获得最好的收益。

附、统计数据分析思想方法在勘探地震学中的应用点

- 信号分析、图像分析、空间数据分析中的应用问题：
  - 1、信号分析，理论上，继续探索复杂情形下的信号建模问题。这是信号分析的根本问题。
    - 高维情形下的模式组合，不是一维基函数组合，表达高维信号要引起关注。
    - 突破二阶统计量是必须的。
  - 2、信号分析，应用上，解决复杂噪音压制、数据压缩、压缩感知随机采样、解混叠/盲源分解、数据规则化、连续源数据的处理、展宽频带、子波估计等等问题。
    - 测量波场传播方向；测量波场之间的差异也是很值得探索研究的。
  - 3、图像分析、理论上、对高维图像的建模是核心问题。
    - PDE、变分法、高维小波变换、Bayes方法的综合应用没有达到信号处理的水平，主要是实践不够。同样地、图像建模是个核心问题！

附、统计数据分析思想方法在勘探地震学中的应用点

- 信号分析、图像分析、空间数据分析中的应用问题：
  - 4、图像分析，应用上，分为三个Level的问题。
    - 基本level中，图像反褶积展宽频带的问题没有得到很好的解决，PSF的构造，PSF反褶积的问题性等缺乏研究。
    - 中等Level中，图像增强，尤其是基于统计方法（相关方法）的而不是用微分方法的图像增强，做得不够。图像属性提取，做得更是不够，这限制了我们发展更好的层位追踪、断层识别等方法的精度。
      - 图像属性的提取是一个十分核心的事情！图像属性是图像处理中很多方法技术的基础信息。
    - 高级Level中，图像智能识别，我认为，可以人工神经网络，也可以用PDE、变分法、高维小波变换、Bayes方法综合的做法设计智能算法解决特定的问题。

附、统计数据分析思想方法在勘探地震学中的应用点

- 信号分析、图像分析、空间数据分析中的应用问题：
  - 5、空间数据分析，理论上，这也是对空间数据的建模问题。或散乱数据的逼近问题（函数逼近的特例！）。
    - Kriging方法、径向基函数方法都是经典的方法。
    - 对于统计规律确定的一个随机过程，它的规则采样和散乱采样如何影响对它的建模过程与结果?
  - 6、空间数据分析，应用上，首先解决散乱的空间数据的融合插值问题，为各种建模问题提供基础手段。希望从空间散乱数据变化规律进行预测的角度提出新的解决散乱数据插值的新方法。希望分析空间变化的带限子波统计特征，消除震源激发和检波值接收引起的振幅和相位畸变。

附、统计数据分析思想方法在勘探地震学中的应用点

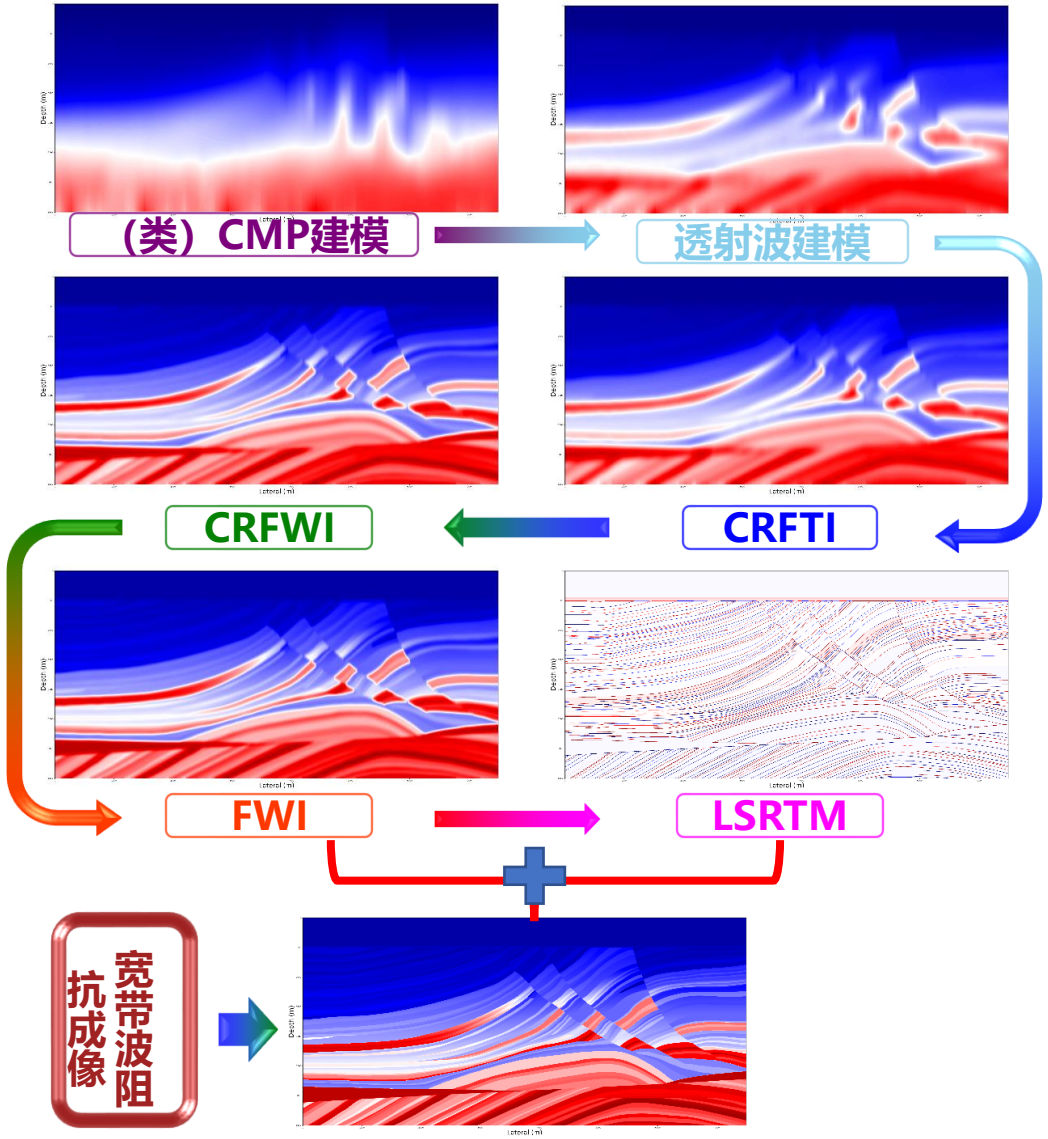
- 地震反演成像中的应用问题：
  - 1、地震波+电磁波波场特征分析与地表岩性区域分划；
  - 2、针对初至波的CMP道集速度扫描及初始速度建模；
  - 3、高维多属性+进化优化算法+Markov最佳演化的初至识别
  - 4、初至波到达时Beam层析；到达时波动理论层析。
    - 各向同性速度估计与建模
    - 各向异性速度估计与建模
    - Q值估计与建模
  - 5、基于CMP道集与类CMP道集（PSTM及PSDM CIG转化来的CMP道集）自动与智能速度建模

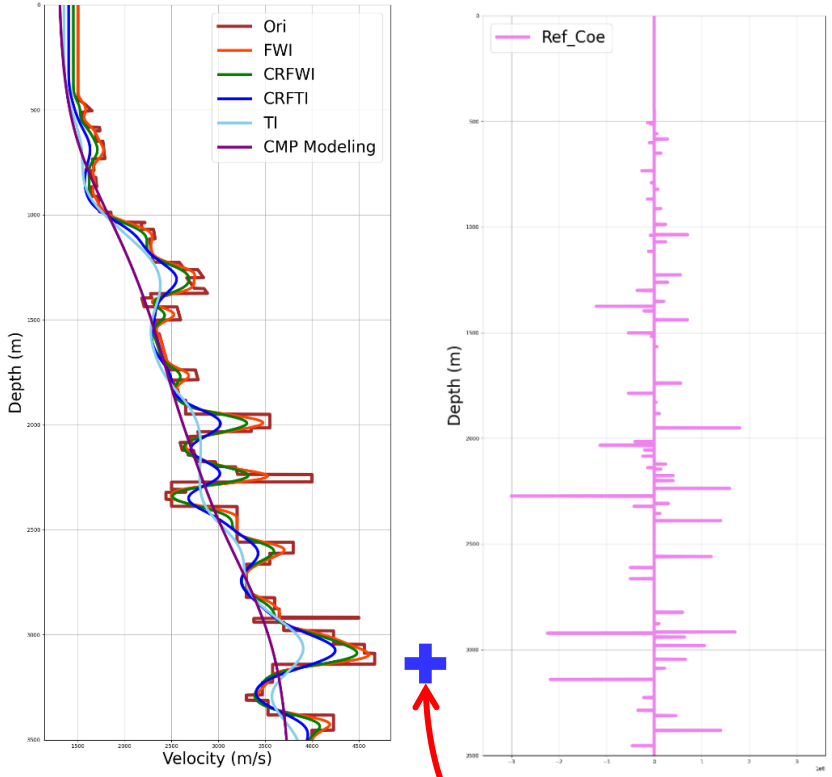
附、统计数据分析思想方法在勘探地震学中的应用点

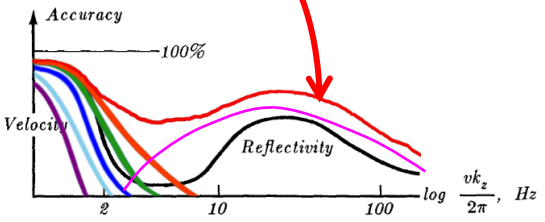
- 地震反演成像中的应用问题：
  - 6、完善CIG道集射线层析速度建模
  - 7、特征反射波波动理论层析速度估计与建模
    - 最后过渡到反射波FWI建模。
  - 8、最小二乘保真成像方法
    - 保真的方位角度反射系数估计
  - 9、自由表面多次波成像
    - 最后发展到层间多次波成像
  - 10，最终统一到宽带波阻抗成像
    - 提供给油藏描述阶段

附、统计数据分析思想方法在勘探地震学中的应用点

◆凸的地震波成像技术组合---CWI+宽带波阻抗成像







（Jon F. Claerbout,1984）

谢谢
欢迎批评指正