HRIPIE欢迎您！ **日期:**

HRIPIE Problem Set

We share three problems which HRIPIE are currently working on:

1. 集团手机套餐优化
2. 下一代群体药理学软件
3. 中医药的“公理-推断”体系构建

the corresponding mathematical topics include:

1. Convex Optimization and Statistics
2. Numerical Linear Algebra, Statistics and ODE Parameter Estimation (Nonlinear Optimization)
3. Category Theory

1 集团手机套餐优化

准确预估耗用量，对于方案选型、费用节约有重要意义。今天向大家分享一个利用工业工程为企业开源节流的案例。

1.1 最优决策理论

我们用随机向量$\vec{X}$表示使用模式，其具体的取值$\vec{x}$为实际用量。若选择方案$s(\cdot)$，实际用量$\vec{x}$需支付$s(\vec{x})$元。给定诸多可选方案$s \in \mathcal{S}$，在能获得同等服务的前提下，我们预先选择期望成本 $\mathbb{E}_{\vec{X}}[s(\vec{X})]$最低的方案$s^*(\cdot)$，即最优方案

$$s^*(\cdot) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{E}_{\vec{X}}[s(\vec{X})]$$

1.2 案例——节省集团手机话费

我们用二维向量$\vec{x}$表示某位同事每月的通话时间、数据流量，比如

$$\vec{x} = (849 \text{ min}, 6.3 \text{ Gb})^T$$

如果那位同事选择了套餐方案$s_1(\cdot)$，该套餐基本费用为188元，含有500min通话和10Gb流量，套餐外费用为0.19元/min、9元/3Gb，那么其费用$s_1(\vec{x})$为

$$188 + (849 - 500) * 0.19 = 222.9 \text{ 元}$$

我们通过同事们过去一年的月度用量数据，拟合统计学模型，就能预判费用，并为每一位同事选择最省钱的套餐方案，为集团节流开支。

我们假设用量$\vec{x}$服从对数二维高斯分布：

$$\log \vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$$

其中$\vec{\mu}$ 和$\mathbf{\Sigma}$通过样本对数均值、样本对数协方差矩阵获得（MLE估计）。最后，从 $\mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$ 抽样并作指数还原

$$\vec{x}_1, \vec{x}_2, ..., \vec{x}_n \overset{i.i.d.}{\sim} \exp[\mathcal{N}(\vec{\mu}, \mathbf{\Sigma})]$$

大数定律保证样本均值收敛于期望值

$$\mathbb{E}_{\vec{X}}[s(\vec{X})] = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} s(\vec{x}_n)$$

对于计算期望值$\mathbb{E}_{\vec{X}}[s(\vec{X})]$，本质上是求积分，本问题适合数值近似。

我们通过Monte-Carlo均值逼近，比如当$N = 10000$时

$$\mathbb{E}_{\vec{X}}[s(\vec{X})] \approx \frac{1}{N} \sum_{n=1}^{N} s(\vec{x}_n)$$

其中数值最小的方案为最优决策。

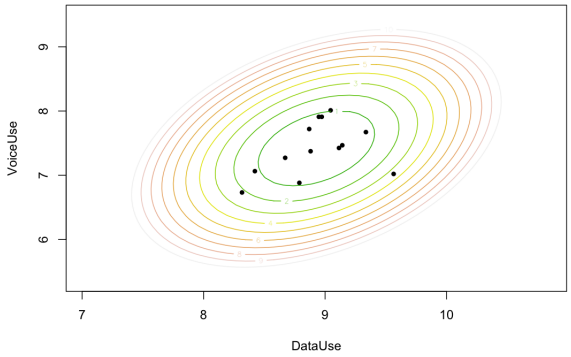
1.3 中山联通提供的可选方案

手机卡均为5G网络，兼容4G/3G/2G。

	套餐内		套餐外	
资费(元)	语音(min)	流量(Gb)	语音	流量
39	300	20	0.15元/min	5元/Gb
59	1500	20	0.15元/min	5元/Gb
69	2000	20	0.15元/min	5元/Gb
79	2500	20	0.15元/min	5元/Gb
89	3000	20	0.15元/min	5元/Gb

1.4 分析结果示例

号码	流量均值(Mb)	通话均值(min)
18507606665	7951.94	1801.91



套餐价格(元)	Monte-Carlo均值(元)	最佳决策
35.00	263.68	
59.00	128.68	
69.00	103.95	
79.00	95.67	✓
89.00	96.72	

1.5 问题

- 是否有更快更精确的计算

$$\mathbb{E}_{\vec{X}}[s(\vec{X})]$$

的方法?

- 如果每个人的套餐都可以个性化定制，资费区间为$[0, 89]$元，那么

$$s^*(\cdot) = \underset{s \in \mathcal{S}}{\operatorname{argmin}}\, \mathbb{E}_{\vec{X}}[s(\vec{X})]$$

是凸优化问题吗?

2 下一代群体药理学软件

2.1 Intro to Pharmacokinetics (PK)

Ceftazidime PK can be described by a two-compartment mamillary system:

$$\begin{cases} V_1 \cdot \dfrac{d}{dt}C_1(t) = Q \cdot [C_2(t) - C_1(t)] - \mathrm{CL} \cdot C_1(t) & V_1 \cdot C_1(0) = 1800 \text{ mg} \\ V_2 \cdot \dfrac{d}{dt}C_2(t) = Q \cdot [C_1(t) - C_2(t)] & V_2 \cdot C_2(0) = 0 \end{cases}$$

where $V_1 = 9.45$ L, $V_2 = 4.79$ L, $Q = 7.07$ L/hr, CL $= 6.73$ L/hr.

To solve $C_1(t)$ and $C_2(t)$, we can re-write:

$$\begin{cases} C_1' = -1.46 \cdot C_1 + 0.748 \cdot C_2 & C_1(0) = 190 \\ C_2' = +1.48 \cdot C_1 - 1.48 \cdot C_2 & C_2(0) = 0 \end{cases}$$

The standard solution that I was taught:

$$\downarrow \text{Laplace Transformation}$$

$$\begin{cases} s\bar{C_1}(s) - C_1(0) = -1.46 \cdot \bar{C_1}(s) + 0.748 \cdot \bar{C_2}(s) \\ s\bar{C_2}(s) - C_2(0) = +1.48 \cdot \bar{C_1}(s) - 1.48 \cdot \bar{C_2}(s) \end{cases}$$

which was

$$\begin{cases} \bar{C_1}(s) = \dfrac{190}{s + 1.46 - \dfrac{0.748 \times 1.48}{s + 1.48}} \\ \bar{C_2}(s) = \dfrac{1.48 \times 190}{(s + 1.48)(s + 1.46) - 0.748 \times 1.48} \end{cases}$$

$$\downarrow \text{Inverse Laplace Transformation}$$

$$\begin{cases} C_1(t) = +94.1e^{-2.52t} + 95.9e^{-0.418t} \\ C_2(t) = -134e^{-2.52t} + 134e^{-0.418t} \end{cases}$$

Now, we can solve it by matrix eigen-decomposition, which naturally supports linear PK systems of all kinds, even globally non-linear but locally linear PK systems.

We re-write the linear differential equations system as:

$$\begin{bmatrix} C_1' \\ C_2' \end{bmatrix} = \begin{bmatrix} -1.46 & 0.748 \\ 1.48 & -1.48 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \quad \begin{bmatrix} C_1(0) \\ C_2(0) \end{bmatrix} = \begin{bmatrix} 190 \\ 0 \end{bmatrix}$$

Given the eigen values λ_1 and λ_2 and their corresponding eigen vectors $\vec{\nu}_1$ and $\vec{\nu}_2$ of $\begin{bmatrix} -1.46 & 0.748 \\ 1.48 & -1.48 \end{bmatrix}$ that satisfy the initial condition, we can directly write solution as:

$$\vec{C} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} \vec{\nu}_1 & \vec{\nu}_2 \end{bmatrix} \begin{bmatrix} e^{\lambda_1 t} \\ e^{\lambda_2 t} \end{bmatrix} = \vec{\nu}_1 e^{\lambda_1 t} + \vec{\nu}_2 e^{\lambda_2 t}$$

the initial condition is $\begin{bmatrix} \vec{\nu}_1 & \vec{\nu}_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 190 \\ 0 \end{bmatrix}$, then we can solve eigen values and eigen vectors as:

$$\begin{cases} \lambda_1 = -2.52 & \vec{\nu}_1 = \begin{bmatrix} 94.1 \\ -134 \end{bmatrix} \\ \lambda_1 = -0.418 & \vec{\nu}_2 = \begin{bmatrix} 95.9 \\ 134 \end{bmatrix} \end{cases}$$

$C_1(t)$ and $C_2(t)$:

$$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} \vec{\nu}_1 & \vec{\nu}_2 \end{bmatrix} \begin{bmatrix} e^{\lambda_1 t} \\ e^{\lambda_2 t} \end{bmatrix} = \begin{bmatrix} 94.1 \\ -134 \end{bmatrix} e^{-2.52t} + \begin{bmatrix} 95.9 \\ 134 \end{bmatrix} e^{-0.418t}$$

2.2 问题

- compute area under curve (AUC) of C_1 by

  a) $\text{AUC} = \dfrac{\text{Dose}}{\text{CL}}$

  b) integrate $\text{AUC} = \int_0^\infty C_1(t)dt = \int_0^\infty 94.1e^{-2.52t} + 95.9e^{-0.418t}dt$

- the eigen-decomposition was achieved by running `eigen` of matrix $\begin{bmatrix} -1.46 & 0.748 \\ 1.48 & -1.48 \end{bmatrix}$ to get the unitary eigen-vectors and eigen-values, then by solving a linear system of initial condition $\begin{bmatrix} \vec{\nu}_1 & \vec{\nu}_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 190 \\ 0 \end{bmatrix}$. Can you develop a faster eigen-decomposition algorithm that iterates with initial condition and therefore directly outputs $\vec{\nu}_1$ and $\vec{\nu}_2$?

2.3 定量药理学概论

2.3.1 定量药理学所代表的学科范式

定量药理学所代表的学科范式是：**用微分方程刻画世界的运行轨迹**。

为刻画某个自然现象，靠一个微分方程往往不够，而需要一个系统。这个系统既有很多微分方程，也有很多非微分方程。系统里有若干参数可供调节，就像控制面板上的旋钮和开关。

理论上，透过这样的系统，**参数空间**向**数据空间**所作的**映射**，便是世界的运动轨迹。考虑到我们观测的数据含有扰动，因此这种**映射与逆映射**的关系可以概括为**概率论与统计学**，或**模拟仿真与贝叶斯推断**[1]。

对于工程学来说，比如航空航天、车辆工程，因为系统是人造的，所以里面的微分方程和系统参数可认为是已知的，因此模拟仿真——从调试好的参数产生逼真的数据，是工程学的家常便饭。除了逆向工程，不太会有贝叶斯推断的需要。

但对于自然科学来说，比如生物、物理、化学，系统是黑箱。我们只能通过观察系统的输出数据，揣测“上帝的设计”，因此从数据推测参数的贝叶斯推断[2]是常见的科研场景。

通过数据空间和参数空间的来回捣腾，人类对世界的认知不断深化。值得一提的是，大自然的系统大多是非线性的，而非线性系统的参数估计是**非线性优化**，这是一门永无止境的艺术。因此，我们说**追求真理的过程是永无止境的**。

2.3.2 定量药理学的目标

定量药理学的目标是对药理作用定量，分为两个层面：

1. 对药物的有效成分定量（简称PK）；
2. 对药效定量（简称PD，也包括毒效）

一般认为，**有效成分是药效的驱动力**，因此PD通常指的是量效关系（简称ER）研究。有了PK和PD就可以进行联合分析[3]。

[1]本文采用贝叶斯派，我们把所有涉及频数派的内容认为是“频数派作为先验的贝叶斯派”，即在贝叶斯推断中，只需采用某种特定先验，也可以得到与频数派一样的结果，由此把频数派纳入贝叶斯派

[2]定量药理学中的贝叶斯推断：根据数据提出微分方程并估计系统参数，对象从个体水平到群体水平

[3]联合分析记为PK/PD：先由PK得到有效成分浓度随时间的变化$C(t)$，再由PD得到量效关系$R(C)$，最后进行函数复合，得到药效随时间的变化$R[C(t)]$

2.3.3 PK的研究范式

PK的研究方法通常为：

1. 测定药物从进入到排出个体的生物转化（biotransformation）全过程[4]；
2. 鉴定药物及其生物转化中间体的有效成分；
3. 测定体内**有效成分的时空浓度**；
4. 提出微分方程并估计参数

经典PK理论是房室模型。房室模型参数少，参数估计的稳定性强，并且只需要血药浓度。近来，生理药理学（简称PBPK）理论得到发展，其由菲克定律解释血液循环系统与各器官间的药物传输，能刻画器官水平的药物浓度。但随着模型参数增多，即便有各器官的药物浓度数据，也常常不能稳定地估计参数。因此药理学家改用实验手段测量部分系统参数，减少需要估计的参数数量，提升参数估计的稳定性。

2.3.4 PD的研究范式

PD的研究取决于研究者的观测维度、临床终点的选择，其理论百花齐放的特点常常令初学者感到无所适从。

经典PD理论，力求建立简单的量效关系，其系统参数少，参数估计的稳定性强。但往往缺乏物理的第一性原理，高度依赖个人经验。

当前，系统药理学（简称QSP）希望从亚细胞水平建立量效关系，即便只能做似是而非的模拟仿真，也往往十分复杂，令人不知所云。

2.3.5 定量药理学的未来

定量药理学的目标是对药理作用定量。对药物作用机理的探究，无疑会从宏观走向微观。

我们知道，成功的新理论应当兼容成功的旧理论，并把旧理论作为新理论的特例——比如牛顿力学是广义相对论的特例。

从PK理论的角度看，如何将房室模型作为PBPK的先验，进而使PBPK的参数估计更为稳定，是很好的问题。此前最接近的讨论是器官合并（lumping），即探讨怎样把多个类似的器官合并成一个大器官，从而简化PBPK。

从PD理论的角度看，QSP与经典PD理论存在鸿沟。显然，当前QSP理论试图跳过经典PD理论，直接把微观描述作为起点，这样的做法是否科学尚存疑。

定量药理学的本质困难在于非线性优化。当尺度从宏观迈向微观，系统的复杂度骤然升高，非线性优化就近乎不可能。如果我们能以宏观描述作为起点，利用经典PKPD理论参数数量少、参数估计稳定性强的优

[4]至少包括主要的代谢（metabolism）、排泄（excretion）过程

点，向微观世界延展下去，保证宏观和微观描述的等价性、一致性，从而对复杂系统实现层级化描述，那么我相信就能克服这样的本质困难。

2.4 问题

- 以下是一个PK模型参数拟合的结果，且不论模型的细节

Parameter	Initial Value	Final Estimate	SE (CV%)	Confidence interval (95%)
A	232.0	164.1	4.973	[147.8 , 180.4]
ar	11.40	23.24	28.90	[9.809 , 36.68]
E	45.00	118.1	6.316	[103.2 , 133.0]
er	0.6700	1.296	9.004	[1.063 , 1.530]
C	18.00	14.40	11.55	[11.08 , 17.73]
cr	0.2000E-01	0.2498E-01	24.29	[0.1284E-01, 0.3712E-01]
A2	8137.	8137.	0.7090	[8022. , 8252.]
ar2	2.800	2.830	3.766	[2.617 , 3.043]
E2	665.0	665.0	11.38	[513.7 , 816.4]
er2	0.3000E-01	0.3225E-01	22.84	[0.1752E-01, 0.4698E-01]
C2	469.0	469.4	13.74	[340.4 , 598.4]
cr2	0.2000E-02	0.2101E-02	20.57	[0.1236E-02, 0.2965E-02]

  为什么此处Confidence Interval是以Final Estimate为中心的对称区间?

- 定量药理学的数据十分宝贵，工业软件通常只在本地运行。请评价这样一种数据加密的方法：把原始数据全都乘以某个随机值α加密后发送到远程端进行参数拟合；拟合完成的结果返回本地后，除以或乘以α（取决于参数的单位），得到解密的参数拟合值。

3 中医药的“公理-推断”体系构建

3.1 问题

- 请解释一个你熟悉的公理体系（e.g. Peano axioms to construct $\mathbb{N}$; Euclid's axioms of plane geometry）
- See figure 1, a weighted graph representing cost of traveling between x, y, z. What is minimal cost of traveling from y to z?

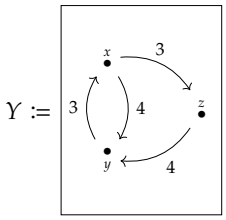


Figure 1: a weighted graph

a short tutorial of solving minimal-cost problem by enriching path with $\mathbf{Cost} = ([0, \infty], \geq, 0, +))$ will be given for further discussion on generalized matrix multiplication