析取克立格方法分析*

Disjunctive kriging analysis

余先川，王桂安，杨 萌，刘 敏

(北京师范大学信息科学与技术学院，北京 100875)

YU XianChuan, WANG GuiAn, YANG Meng and LIU Min

(College of Information Science & Technology, Beijing Normal University, Beijing 100875, China)

摘 要 非线性空间信息统计学主要包括对数正态克立格、指示克立格、析取克立格、条件期望等理论方法。通过源代码和运行结果对各种克立格方法在数据特点、时间复杂度等方面进行了比较。DK估计较普通克立格更有效。首先，DK考虑到了Z(*x*)的二维概率分布，充分应用有效信息，克服了线性克立格的前提条件太少，导致数值离散性大时，估计不够准确的缺点；其次，DK利用区域化变量的非线性组合估计，所以可以方便地得到变量的任意函数的估计结果，在局部可回采储量等方面比线性方法更容易、结果更准确。

关键词 非线性地质统计学；变异函数；普通克立格；析取克立格；厄尔米特多项式

地质统计学是以区域化变量理论为基础，以变异函数为重要工具，研究那些在空间分布上既有随机性又有结构性的自然科学（儒尔奈耳，1982；侯景儒等，1993；1998）。因此，凡是要研究空间数据分布的随机性和结构性，并对这批数据进行最优无偏内插估计就要应用地质统计学理论及相应的方法。

本文介绍了正态变换、变差函数计算以及厄尔米特多项式等基本过程。在研究中程序实现了析取克立格(DK)和普通克立格(OK)，并对它们进行了比较分析。

1 析取克立格

1.1 线性空间信息统计学的不足

线性克立格统计法有局限性，线性空间信息统计学仅限于有效数据的线性组合：$Z_k=\sum_{\alpha=1}^{n}\lambda_\alpha Z_\alpha$，其前提条件只限于随机函数的二阶矩，从而使线性空间信息统计学有以下不足（Rivoirard，1994）：

当已知随机函数V的数值离散性太大时，对于待估点 x_0 (或域V)的真值 $Z(x_0)$〔或 $Z_V(x_0)$〕的估值不精确；只能估计真值 $Z(x)$ 的值 $Z^*(x)$,不能估计 $Z(x)$ 的函数；线性估计量不能再现真值 $Z(x)$ 的空间变异性。因而，不能按照这些估计量的分布提出真值大于其边界值的比例（Rivoirard，1994；侯景儒，1994）。

为了从根本上解决线性空间信息统计学的不足，G.Matheron提出了NSIS（Theory and Methods of Nonlinear Spatial Information Statistics and Its Application in Geosciences）的基本构想：

（1）当对随机函数 $Z(x)$ 的性质有较多了解时，可以推断出它的一、二或k元分布。就可以建立一些非线性估计量。

*本文得到国家自然科学基金资助项目（40202030 和 40372129）
第一作者简介 余先川，男，1967 年生，教授，主要从事空间数据挖掘、遥感图像处理、矿产预测等研究。

（它比线性更准确）。例如，可以把一有效数据的n个函数之和 $f_\alpha(Z_\alpha)$ 作为估计量：$Z^{DK}=\sum_\alpha f_\alpha(Z_\alpha)$ 这就是下面就要详细研究的非线性估计方法——析取克立格。

（2）或更好的情况：把*n*个有效数据的唯一函数 $Z^E=f(Z_1,Z_2,...,Z_n)$ 作为估计量。

1.2 正态化

因为析取克立格方法估计的对象要满足正态分布，所以在进行估值前先要根据数据的分布特点选择不同的正态化方法进行正态化：

对数变换 $x_{ij}^{'}=\ln(x_{ij}+C)$

平方根变换 $x_{ij}^{'}=\sqrt{x_{ij}+C}$

反正弦变换 $x_{ij}^{'}=\sin^{-1}(\sqrt{x_{ij}/10^n})$（*n*为 x_{ij} 的位数）

反余弦变换 $x_{ij}^{'}=\cos^{-1}(\sqrt{x_{ij}/10^n})$（*n*为 x_{ij} 的位数）

若数据呈正偏（低值多，高值少）或称左偏，按左偏程度从大到小分别选择对数、平方根、反余弦变换。若数据呈负偏（低值少，高值多）或称右偏，则采用反正弦变换。

1.3 厄尔米特多项式（Hermite Polynomial）

析取克立格应用了一个特殊的并且非常有效的工具——厄尔米特多项式。

厄尔米特多项式是属于特殊函数的关于正态分布特征的多项式,厄尔米特多项式在DK中至为重要，因为它和正态概率密度函数有着密切的关系。厄尔米特多项式定义（Rivoirard，1994）：

$$H_n(x)=\frac{1}{\sqrt{n!}g(x)}\frac{d^n g(x)}{dx^n} \qquad (2\text{-}1)$$

式中，$g(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$，为标准正态概率密度。

当样品对$[Y(x),Y(x+h)]$为二元标准正态分布，且相关性为(*h*)时，厄尔米特多项式的统计特征表示如下：

均值：$E[H_n(Y(x+h))\,|\,Y(x)]=\rho(h)^n H_n(Y(x))$ （2-2）

协方差：

$$\begin{aligned}&Cov[H_n(Y(x)),H_n(Y(x+h))]=E[H_p(Y(x))H_n(Y(x+h))]\\&=E[H_p(Y(x))E(H_n(Y(x+h))\,|\,Y(x))]=\rho(h)^n E[H_p(Y(x))H_n(Y(x))]\end{aligned} \qquad (2\text{-}3)$$

当*p*=n，$Cov[H_n(Y(x)),H_n(Y(x+h))]=\rho(h)^n$

p *n*，$Cov[H_p(Y(x)),H_n(Y(x+h))]=E[H_p(Y(x))H_n(Y(x+h))]=0$

随机函数 $Y(x)$ 的任意函数均可用厄尔米特多项式扩展：

$$f[Y(x)]=f_0+f_1H_1[Y(x)]+f_2H_2[Y(x)]+...=\sum_{n=0}^{\infty}f_nH_n[Y(x)] \qquad (2\text{-}4)$$

1.4　析取克立格基本算法

已知观测值$\{Z(x_\alpha)|\alpha=1,2,\mathrm{K}\ n\}$则未知值$Z(x_0)$的DK估计量为：$Z^*_{DK}(x_0)=\sum_{\alpha=1}^{n} f_\alpha[Z(x_\alpha)]$，式中$f_\alpha$是可测函数。DK不仅可以估计品味值本身（$f(x)=x$）还可以估计他的某个函数值，这个通过定义$f(x)$实现，例如求估计值大于某个边界值Z的概率，只要令$f(x)=\begin{cases}1 & z(x)>Z\\ 0 & z(x)\le Z\end{cases}$即可。在DK中$f_\alpha$有以下约束条件：

$$\sum_{i=1}^{n} E[f_\alpha[Z(x_\alpha)]\,|\,Z(x_\beta)=z(x_\beta)]=E[Z(x_0)\,|\,Z(x_\beta)=z(x_\beta)]，\ 其中\beta=1,2\mathrm{L}\ n$$

该式是线性克立格法的一个推广，它依赖于已知Z_β时，$Z(x_0)$和Z_α的条件概率分布，它是一个含义比一般线性估计更深广的条件。

DK中要求变量*Z*(*x*)满足正态分布，但实际上满足这个条件的情况很少，所以必须按上面的方法把*Z*(*x*)变换成正态变量是必要的：*Z*(*x*)　*Y*(*x*)

由（2-4）知 $[Y(x)]^{DK}=f_0+f_1H_1[Y(x)]+f_2H_2[Y(x)]+\ldots$

其中$f(x)=x$，　$f_k=<f(u),H_k(u)>=\int_{-\infty}^{\infty} f(u)H_k(u)g(u)du,\quad \forall k$

其中$g(x)=\dfrac{1}{\sqrt{2\pi}}e^{-x^2}$

$$H_k[Y(x)]=\sum_\alpha \lambda_{k\alpha}H_k(Y_\alpha)$$

α=1，2，……，*n* 表示用于估计的信息点的下标, k表示埃尔米特多项式的阶数，这里的$\lambda_{k\alpha}$满足方程组

$$\sum_\beta \lambda_{k\beta}Cov[H_k(Y_\alpha),H_k(Y_\beta)]=Cov[H_k(Y_\alpha),H_k(Y(x))]$$

即：$\sum_{\beta=1}^{n}\lambda_{k\beta}[\rho_{\alpha\beta}]^k=[\rho_{\alpha x}]^k$　　（对所有的　=1，2，……，*n*成立）

其中$\rho_{\alpha\beta}$表示信息点v_α和v_β间的相关性$\rho(v_a,v_\beta)$，　$\rho_{\alpha x}$表示信息点v_α和待估点v_x间的相关性$\rho(v_a,v_x)$

因为$H_k(Y(x))$的相关结构$[\rho(h)]^k$很快趋于一种纯块金结构，在任一未知点的克立格估计值也迅速趋于它的均值，即0。所以尽管系数f_k不可忽略，但只须对少数几个多项式进行克立格估计，就可以得到估计结果。通常它的个数少于12个。

$H_n(Y(x))$的克立格方差是：$\sigma^2_{Kn}=1-\sum_\alpha \lambda_{n\alpha}[\rho_{\alpha x}]^n$

$f[Y(x)]$的DK估计方差是：$\mathrm{var}(f[Y(x)]-f[Y(x)]^{DK})=\sum_1^\infty (f_n)^2\sigma^2_{Kn}$

2 实验分析与结论

实验所用的原始数据是一套来自某锡锌多金属矿床的三维钻孔数据（侯景儒等，1998），该套数据包括63个钻孔，2种金属元素（Sn、Zn）。从数据处理结果来看，各种克立格方法估计出的品位值大致相同，而且与其相邻的已知点的品位值相近，这表明了几种克立格方法的有效性和程序的正确性。

各种克立格方法比较结果如表1所示：

表1 几种克立格的比较

方 法	特 点	时间复杂度	空间复杂度
普通克立格	数据须满足二阶平稳，当已知区域化变量数值离散性太大时，估计不够精确	最小$O(n^2)$	较小$O(n^2)$
泛克立格	允许数据是不平稳的，可以存在漂移，适用范围较普通克立格更广	较小$O((n+10)^2)$	较大$O((n+10)^2)$
协同克立格	考虑数据多元性即其他元素对待估元素的影响，结果更准确	较大$O((2n)^2)$	最大$O((2n)^2)$
析取克立格	要求数据正态分布，增加假设条件，结果更精确；不仅可以估计未知点值本身而且可以估计其函数值	最大$O(12n^2)$	较小$O(n^2)$

通过源代码分析得出时间复杂度的阶数相同都为$O(n^2)$，看不出太多差别，但从现有几套数据的运行结果看来：一般，普通克立格的运行速度最快，但对输入数据要求比较高。而泛克立格对输入数据放宽了要求，但也影响了运行速度。协同克立格考虑了多种元素的影响，结果相对更准确一些，但同时增加了运算负担。析取克立格的运行速度比上述几种克立格都要慢，但由于它是非线性估值方法，克服了线性空间新系统计学的许多不足，运用更广泛，在很多实际情况下结果会更精确。

通过阈值检查比较可知运行结果的准确性。例如，对某一批数据，分别用普通克立格（OK）和析取克立格（DK）进行插值，得到的结果进行以下统计分析：

表2 阈值统计分析表

阈 值	0	0.05	0.1	0.15	0.2	0.3
已知数据	0.206	0.248	0.312	0.382	0.460	0.623
OK运行结果	0.105	0.107	0.118	0.197	0.280	0.338
DK运行结果	0.179	0.187	0.230	0.270	0.340	0.431

表2中，第一行是根据数据范围取得的几个阈值，第2行至第4行是对应各阈值数据的平均值，这样如果对应列与已知数据的各列越接近，结果越准确。

除了估计未知点的值之外，该系统还可以计算估计方差，这样我们可以通过某点的估计方差来判断估计的误差大小，方差越小，估计越准确即该点值出现估计偏差的风险越小。在实际矿山开采工作中，比较含量和方差的等值线图就可以衡量其风险性，取得最大经济效益。

从实验结果，可以看出DK估计较普通克立格更有效。首先，DK考虑到了$Z(x)$的二维概率分布，充分应用有效信息，克服了线性克立格的前提条件太少，导致数值离散性大时，估计不够准确的缺点；其次，DK利用区域化变量的非线性组合估计，所以可以方便地得到变量的任意函数的估计结果，在局部可回采储量的等方面比线性方法更容易结果更准确。析取克立格充分考虑了数据的空间变异性，证明了非线性方法的有效性。

参 考 文 献

侯景儒, 尹镇南, 李维明. 1990. 地质统计学的理论与方法. 北京: 地质出版社.

侯景儒. 1993. 矿床统计预测及地质统计学的理论与应用. 北京: 冶金工业出版社.

侯景儒, 黄竞先, 吴雨沛, 等. 1994. 非参数及多元地质统计学的理论分析及其应用. 北京: 冶金工业出版社.

侯景儒, 等. 1998. 实用地质统计学. 北京: 地质出版社.

儒尔奈耳 A G, 著. 侯景儒, 黄竞先, 等, 译. 1982. 矿业地质统计学. 北京: 冶金工业出版社.

Rivoirard J. 1994. Introduction to disjunctive kriging and non-linear geostatistics. Oxford: Clarendon Press.