半参数预测模型在定量的分子结构与其活性之间关系中的应用

殷弘
（香港浸会大学数学系）

摘　　要

我们将要介绍的这个半参数预测模型（也叫 kriging 模型）是由一个参数模型和一个非参数随机过程联合构成的。它比单个的参数化模型更具有灵活性，同时又克服了非参数化模型处理高维数据存在的局限性。通过对一组实际数据的应用，我们发现它比单个的参数化模型具有更强的预测能力，值得在定量的分子结构与其活性之间的关系的研究中加以推广。

关键字 半参数， 回归，预测

1：引言

我们研究定量的分子结构与其活性之间的关系（QSAR），其目的是想在分子的活性与分子结构之间建立一个比较理想的统计回归模型：

$$\text{煦赴魂俶} = f(\text{煦赴賦凳}) = f(\text{煦赴鏡扮肄}) \qquad (1.1)$$

这样我们就可以通过此模型来预测未知某类化合物的物理化学的，生物学的以及毒物学的某种属性，模型中称为回归变量。而分子描述值是对分子结构的一种定量的描述，我们可以将其看成模型中的自变量。自从提出第一个分子描述值以来，现在有成千上百个分子描述值，这给模型建立带来了很多困难。比如说，如何选择变量？选好变量后建立什么样的模型等等？QSAR 研究中经常用到的参数化模型有普通的线性回归，主成份回归，偏最小二乘回归和邻回归。这些方法只是充分挖掘了自变量与回归变量之间的线性关系，对剩下的信息没有能力给出解释了。而本文将要介绍的半参数模型是由一个参数化模型和一个非参数化的随机过程组成的。其中非参数化的随机过程提高了整个模型的质量，现在我们将此方法介绍给大家。

2： Kriging 模型

Kriging 一词的意思是最优的空间预测，它是根据一个南非采矿工程师 Krige 的名字命名的，是他将随机过程模型首次运用在空间预测上的。详细内容读者可以参阅 Cressie (1993), Journel 和 Huijbregts (1978), Rivoirard (1994)。

假设我们采集到 m 个训练样本 $S=[s_1,s_2,\cdots,s_m]'$ 和 $Y=[y_1,y_2,\cdots y_m]'$，$s_i\in\Re^n, y_i\in\Re$。Kriging 方法用如下的模型来建立自变量与回归变量之间的关系（不含误差，含有误差的模型在后面介绍）：

$$y(s)=u(s)+z(s) \tag{2.1}$$

其中$u(s)$一个参数模型，它表现了回归变量$y(s)$的大部分信息，被称作平均结构。$z(s)$是一个均值为零的随机过程。常用的 kriging 模型假设$u(s)$是一个参数线性模型：

$$u(s)=f(s)\beta=\sum_{j=1}^{p}f_j(s)\beta_j\text{。} \tag{2.2}$$

$f(s)=[f_1(s),f_2(s),\cdots f_p(s)], f_j:\Re^n\to\Re$，$\{f_j(s)\}_1^p$ 就是任一一组基函数。$\beta=[\beta_1,\beta_2,\cdots,\beta_p]'$是需要估计的参数项量。除了假定随机过程$z(s)$的均值零以外，还要定义它的协方差：

$$E(z(s_i)z(s_j))=\sigma^2R_\theta(s_i,s_j)=\prod_{j=1}^{n}R_\theta^{(j)}(s_i-s_j)\text{，} \tag{2.3}$$

σ^2被称为过程方差，R_θ是调节样本点的相关函数。我们列举最常用的相关函数$(d_j=s_i-s_j,\theta_j>0)$：

Name	$R_\theta^{(j)}(d_j)$
EXP	$\exp(-\theta_j\mid d_j\mid)$
GUASS	$\exp(-\theta j\, d_j^2)$
LIN	$\max\{0,1-\theta_j\mid d_j\mid\}$
SPLINE	$1-3\xi_j^2+2\xi_j^3;\ \xi_j=\min\{1,\theta_j\mid d_j\mid\}$

将设计矩阵S进行标准化后，$-2\le d_j\le 2$。我们可以通过以下图形对相关函数的选择进行说明。

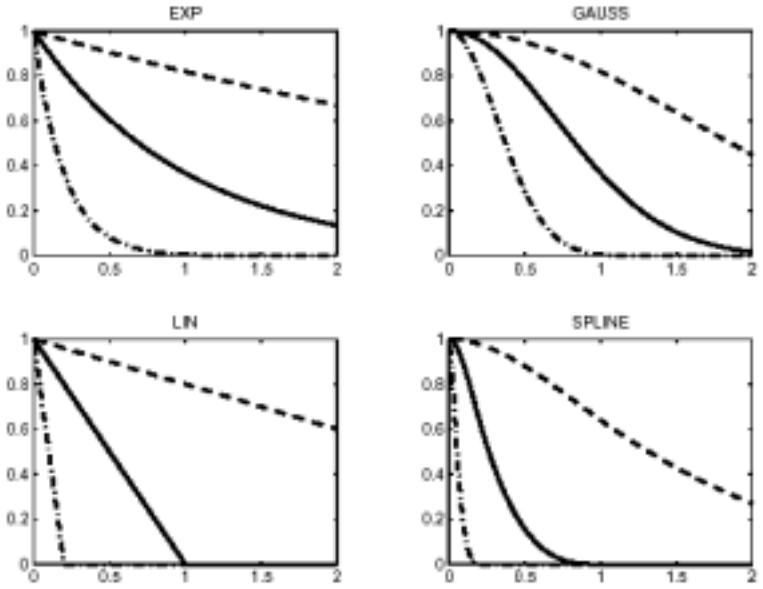

Figure 1.1 Correlation functions for $0 \le d_j \le 2$,
Dashed, full and dash-dotted line: $\theta_j = 0.2, 1, 5$

由以上图形可以看出，相关函数可以分为两类：一类是 Spline 和 Gauss，它们在原点处表现出曲线行为；另一类是 Lin 和 Exp，它们在原点处表现的是线性行为。我们可以将这些特点和你要考虑的实际数据的背景结合起来。如果实际数据满足的函数是连续可微的，Spline 和 Gauss 这样的相关函数就优选，反之， 如果函数在零点附近表现出线性行为，Lin 和 Exp 的效果就比 Spline 和 Gauss 要好（Isaaks 和 Srivastava, 1989）.

我们可以将最简单的 kriging 模型和单个的参数化线性模型作比较，除了线性参数β
外，kriging 模型还引进了参数(σ^2, θ_j)，这样无疑使得它更具有灵活性 (Sacks, et al., 1989)。

3. Kriging 模型参数的估计

当取定好一组基 $f(s) = [f_1(s), f_2(s), \cdots, f_p(s)]$，我们就得到一个 $m \times p$ 的扩展设计矩阵 F ， $F_{ij} = f_j(s_i), i = 1, \cdots m, j = 1, \cdots p$ ， 和一个 $m \times m$ 的相关矩阵 $R = R_\theta(s_i, s_j), i, j = 1, \cdots m$。通常我们用已知训练样本的响应值的线性组合来估计任一个给定样本 x 的响应值 $\hat{y}(x) = c'Y, c \in \Re^m$。很容易求出在线性无偏的条件下使得 $\varphi(x) = E[(\hat{y}(x) - y(x))^2]$ 达到最小的 c 的估计为：

$$\hat{c} = R^{-1}(r - F\tilde{\lambda}),\ \tilde{\lambda} = (F'R^{-1}F)^{-1}(F'R^{-1}r - f') \qquad (3.1)$$

其中定义 $r=[R_\theta(s_1,x),\cdots,R_\theta(s_{m,}x)]'$。 $\hat{y}(x)=c'Y$ 可以表达成：

$$\hat{y}(x)=(r-F\tilde{\lambda})'R^{-1}Y=f(x)\hat{\beta}+r'\gamma^{*} \qquad (3.2)$$

其中 $\hat{\beta}=(F'R^{-1}F)^{-1}F'R^{-1}Y$ 被称为模型（2.1）的广义最小二乘估计，$\gamma^{*}=R^{-1}(Y-F\hat{\beta})$。所以对每一个新的样本，我们只要求出向量 $f(x)$ 和 r，就可以估计新样本的响应值。

理论上可以证明上面的广义最小二乘估计等于极大似然估计：假设随机过程 $z(s)=[z(s_1),z(s_2),\cdots,z(s_m)]$ 是高斯过程，那幺 $y(s)=[y(s_1),y(s_2),\cdots,y(s_m)]$ 也是一个高斯过程。此过程的对数似然函数为：

$$-\frac{1}{2}[n\ln\sigma^2+\ln|R|)+(Y-F\beta)'R^{-1}(Y-F\beta)/\sigma^2] \qquad (3.5)$$

在给定参数 $\theta=[\theta_{1,}\cdots\theta_n]$ 的情况下，对上式的 β 和 σ^2 求微分，得到它们的极大似然估计：

$$\hat{\beta}=(F'R^{-1}F)^{-1}F'R^{-1}Y,\qquad \hat{\sigma}^2=\frac{1}{m}(Y-F\hat{\beta})'R^{-1}(Y-F\hat{\beta})。 \qquad (3.6)$$

再将 $\hat{\beta}$ 和 $\hat{\sigma}^2$ 带入(3.1)式得：

$$-\frac{1}{2}(n\ln\hat{\sigma}^2+\ln|R|)。 \qquad (3.6)$$

我们要求 θ 的极大似然估计使得(3.3)达到极大（也可以说求 θ 使: $\hat{\sigma}^2|R|^{1/m}$ 极小），如果直接对此式求参数 θ 的极大似然估计是比较困难的，也消耗时间。我们采用 Welch, et al. (1992) 所提出的一个估计 θ 的算法来计算 $\hat{\theta}$。

4．具体的计算方法

以上只是理论上的求法。在实际应用中，由于样本的数量大，维数高，常常使得一些矩阵过于稀疏。直接计算会带来较大的误差。我们采用以下的求法得到参数的估计值.

初始化 $\theta^{(0)}=[\theta_1^{(0)},\cdots,\theta_n^{(0)}]$，循环下列步骤：

（1）将相关矩阵 R 进行 Cholesky 分解： $R=CC'$;

（2）对原有的设计矩阵 F 和回归变量 Y 进行正交转化： $\tilde{F}=C^{-1}F,\ Y=C^{-1}Y$ ；

（3）然后对 $\tilde{F}$ 做 QR 分解： $\tilde{F}=QG'$ ； Q 的列向量正交， G' 是上三角矩阵；

（4）计算参数 β 的估计： $\hat{\beta}=G^{-1}Q'\tilde{Y}$ ；

（5）计算参数σ^2的估计：$\hat{\sigma}^2 = \frac{1}{m}(\tilde{Y} - \tilde{F}\hat{\beta})'(\tilde{Y} - \tilde{F}\hat{\beta})$；

（6）计算目标函数：$\hat{\sigma}^2 |R|^{1/m}$。

只到找到目标函数的极小值。

4. 试验设计

Kriging 模型有一个很突出的性质：它是训练样本的插值函数。当$x = s_i$，即要预测的样本是某一个训练样本时：

$$\begin{aligned} \hat{y}(s_i) &= f(s_i)\hat{\beta} + r(s_i)'R^{-1}(Y - F\hat{\beta}) \\ &= f(s_i)\hat{\beta} + e_i'(Y - F\hat{\beta}) \\ &= f(s_i)\hat{\beta} + y_i - F_{i,:}\hat{\beta} \\ &= y_i \end{aligned} \tag{4.1}$$

这样一来 Kriging 模型对外插点的预测是很不准确的（Walter, et al., 1997）. 所以训练样本的选取很重要，要在它们所允许的范围内尽量散布开来。通过实验设计选取训练样本在某种程度上可以提高模型的质量，需要充满空间的试验设计的方法，如 Latin Hypercube Sampling 及其变形（Sacks, et al., 1989）和（Simpson, et al., 2001）或均匀设计(方开泰和马长兴,2001)。下面将介绍几种通过实验设计选取具有代表性样本的方法。

4.1 矩形网格

假设我们感兴趣的样本的自变量的区间为：$l_j \le x^{(j)} \le u_j, j = 1,\cdots,n$。矩形网格选取的样本点定义为：

$$x_i^{(j)} = l_i + k_i^{(j)} \frac{u_j - l_j}{q_j}, \quad k_i^{(j)} = 0,1,\cdots,q_j \tag{4.2}$$

其中$\{q_j\}$是整数。如果所有的$\{q_j\}$都等于q，则选取的样本点的总数为$(q+1)^n$。

4.2 拉丁超立方体抽样

假设我们要在n维向量空间里抽取m个样本。拉丁超立方体抽样的步骤是：

（1） 将每一维分成互不重迭的m个区间，使得每个区间有相同的概率 （通常考虑一个均匀分布，这样区间的长度相同）。

（2） 在每一维里的每一个区间中随机的抽取一个点；

（3） 再从每一维里随机抽出（2）中选取的点，将它们组成向量。

5. Kriging 模型在 QSAR 中的应用

这组数据来中南大学化学计量学研究小组。他们想在一组拓扑指数（medv13r 和 medv44 共 7 个）和保留指数之间建立一个较好的模型来预测未来样本的保留指数，采集到的样本是 207 个饱和烷烃。我们用矩形网格设计从这 207 个样本中抽取了 128 个样本作为训练样本，用原始的自变量做为 Kriging 模型的基函数，选取 Gauss 相关函数。得到的训练结果如下：

Kriging results

$\hat{\beta}$ =[-0.0011, 0.0875, 0.0452, -0.0262, 0.8694, 0.1504, 0.0123, 0.1382]

$\hat{\theta}$ =[14.4334, 14.6081, 30.0000, 5.1153, 25.5323, 10.0596, 5.6594]；

$\hat{\sigma}^2 = 58.4152$

$rmse(train,128) \approx 0$ $rmse = 4.03$ (test，79)

Linear results

Rmse (train,128)=7.61 rmse (test, 79)=5.08

Improved=(5.08-4.03)/5.08=20%

我们再将建立好的 Kriging 模型用于剩下的 79 个检验样本，求出它们的 $rmse = 4.03$ (test data)；如果我们用相同的 Kriging 模型中的基函数，建立一个参数线性回归模型，训练样本和检验样本的均方根误差分别是 $rmse = 7.61$（training data），$rmse = 5.08$ (test data)，参数 $\hat{\beta}$ =[-0.0000，0.1551，0.1305，-0.0085，0.8697，0.1808，-0.0481，0.0048]。如果用矩形网格设计抽出 192 个训练样本建立 Kriging 模型和参数线性回归模型，它们的拟合和预测结果分别是：

Kriging model: $rmse(train,192) = 0$ $rmse(test,15) = 2.82$

Linear model: $rmse(train,192) = 6.86$ $rmse(test,15) = 3.58$

Improved=(3.58-2.82)/3.58=21%

从以上对实际数据的分析结果来看，Kriging 模型不仅在对训练样本的拟合上还是对检验样本的预测上，都比线性模型要好,训练样本的误差为零，预测误差提高了近 20%。

6. 总结

我们只是初步尝试做了 Kriging 模型和参数线性模型的比较，其实还有很多研究工作可以继续做，比如在选择基函数上，这样 Kriging 模型还可以和广义线形模型做比较。接下来的工作我们想比较 Kriging 模型和主成份回归，还有偏最小二乘回归的结果，这对 QSAR 研究无疑将会起到主要的帮助。

参考文献

Cressie, N. (1993) Statistics for Spatial Data. Revised Edition. Wiley, New York.

Journel, A. G., Huijbregts, C. J. (1978). Mining Geostatistics. Academic Press, London.

Rivoiraral, J. (1994). Introduction to Disjunctive Kriging and Non-linear Geostatistics. Oxford University Press, Oxford.

Isaaks, E. H. and Srivastava, R. M. (1989). An Introduction to Applied Geostatistics. Oxford University Press, New York, USA.

Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989). Design and Analysis of Computer Experiments. Statist. Sci. 4 (4) (1989).

Walter, E. and Pronzato, L. (1997). Identification of Parametric Models from Experimental Data. Springer, Heidelberg.

方开泰，马长兴，正交与均匀试验设计，科学出版社,北京，2001