一种基于广义KL距离和几何曲率的模型选择准则

杨 坚, 罗四维, 刘蕴辉
(北京交通大学计算机与信息技术学院计算机研究所, 北京 100044)

摘 要: 模型选择的目标就是识别产生给定数据的模型.通常模型的好坏由模型的泛化能力来度量,而泛化能力包含模型对给定数据的拟合度和模型自身复杂度两个方面.本文从信息几何的观点使用定义在流形上的广义KL距离来度量模型的拟合度;另一方面从微分几何的观点用曲率的概念来度量模型的内在复杂度;因此,拟合度和复杂度的表示都具有在参数变换下保持不变的特点.通过理论分析,我们证明了用于表示模型预测能力的未来残差与模型固有曲率的关系.由此提出一种新的基于广义KL距离和曲率的模型选择准则KLCIC.该准则不仅考虑了样本大小、参数个数和函数形式等影响复杂度的因素,而且具有非常清晰直观的几何意义.实验结果表明该方法的有效性.

关键词: 模型选择; 广义Kullback-Leibler距离; 曲率立体阵; 解轨迹; 未来残差
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2005) 12-2272-06

A New Model Selection Criterion Based on Kullback-Leibler Information Divergence and Geometric Curvature

YANG Jian, LUO Si-wei, LIU Yun-hui
(*Institute of Computer, School of Computer & Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: The goal of model selection is to identify the model that generated the data. Goodness of a model is measured using generalizability, which takes two opposite aspects: goodness of fit and model complexity into account. we use generalized KL-divergence defined on the manifold to measure the goodness of fit, and use the conception of curvature from the point of view of differential geometry to explore the intrinsic model complexity that is free of reparametrization; through theoretical analysis, we proved the future residual that is qualified to measure the generalizability can be expressed by using the intrinsic curvature array of model, from which we propose a new model selection criterion KL-divergence and curvature information criterion with very clear and intuitive geometric understanding of model selection. Experimental results reveal its validity.

Key words: model selection; generalized Kullback-Leibler divergence; curvature array; solution locus; expected residual

1 引言

从同一给定数据的几个可能的模型中选择一个好的模型并不是件容易的事.至今已经提出一些基于不同原则的模型选择方法,如Akaike信息准则(AIC)[1,2]、贝叶斯信息准则(BIC)[3]、最小描述长度(MDL)[4,5,6]、交叉验证(CV)[7,8]、贝叶斯模型选择(BMS)[9,10]、随机复杂度(SC)[11]、复杂度的信息理论度量(ICOMP)[12].AIC是通过真实模型与拟合模型间的Kullback-Leibler距离在大样本条件下导出的,其复杂度只考虑了参数个数;BIC来源于贝叶斯统计理论,它不仅考虑了参数个数,还包括样本个数;MDL源于算术编码理论,它既包括参数和样本个数,还通过Fisher信息矩阵反映了函数形式对复杂度的影响.其性能比AIC和BIC好,但积分项计算麻烦;CV是一种基于抽样的方法,它没有显示地给出复杂度表示,优点是容易实现,但小样本时不可靠;BMS在其积分式中隐含地考虑了复杂度问题,但需要事先给出参数的先验分布,且计算困难;SC和ICOMP通过引入参数的Hessian矩阵和协方差矩阵考虑了影响复杂度的各种因素,但它们不是在参数变换下不变的.其他的模型选择方法包括:从泛函分析方法出发、基于特定Hilbert空间的子空间信息准则(Subspace information criterion)[13],基于信息几何理论中e投影和m投影的最小最大Kullback-Leibler方法(Min-Max Kullback-Leibler)[14],以及用于图像和传感器数据三维建模的几何信息准则(Geometric information criterion)[15],它使用高维空间中的Mahalanobis几何投影.

通常,进行模型选择时至少有三个方面必须考虑:拟合度,复杂度和泛化能力.模型选择是个不适定问题,因为数据样本中所有可用信息不足以使我们的选择唯一.而使得问题更糟糕的就是在实际环境中数据往往不可避免的带有噪音,以至于我们无从知道被模型拟合的是噪音还是数据中内在的规律.一般来说,模型越复杂就越容易吸收随机噪音,因此增

收稿日期:2004-06-01;修回日期:2005-08-18
基金项目:国家自然科学基金(No.60373029),博士点基金(No.20020004020)

加对带有噪音的数据的拟合度未必增加它对数据内在规律的拟合度.事实上,总是可以通过增加复杂度来提高拟合度.然而,这样一来导致的过拟合和欠拟合都将引起相对低的泛化能力.这意味着一个其复杂度逼近真实模型复杂度的模型具有可能最好的泛化能力.描述泛化能力的一个形式化公式为[16]:

(模型泛化能力) = (- 模型拟合度) + (模型复杂度)

显然,复杂度与拟合度间相互制约的关系体现出所谓的'奥克姆剃刀'效应.模型选择的任务是选择能达到要求的最简单的模型.那么如何在拟合度和复杂度间达到一个最佳的平衡从而获得最佳的泛化能力呢? 常用的模型选择方法如 AIC, BIC, MDL, CV, BMS 等都是达到这一平衡的代表.以上方法都隐式或显式的考虑了影响复杂度的因素.但大多准则不能满足一个基本的要求:参数表示的不变性,即复杂度应与采用何种参数形式无关.这是任何一个有明确物理含义的复杂度度量准则必须满足的条件[17,18].MDL 虽然满足参数表示的不变性要求[17],但其思想来自算术编码理论,不能提供一个从模型本身内在特性理解模型选择的理论框架;其他几种准则大都属于启发式方法,仅仅从模型所含参数以及模型的函数表达形式考虑,因此缺乏对模型内在特性及模型复杂度、模型-数据拟合度的几何意义的清晰理解.

模型选择可以看作从多个统计模型中选优的统计推断问题,而一个统计模型是一族概率分布的集合,并可看作由所有概率分布构成集合的子集,真实分布可能在模型中也可能在模型附近,为了进行统计推断,有必要知道统计模型在整个概率分布集合中占据哪一部分,以及统计模型的形状,这就是统计模型的几何问题,一些几何量也在推断中起了非常重要的作用,如两个分布之间的"距离"(或称差异量)、曲率等.由于几何性质具有内蕴性,它具有与坐标无关的参数不变性,反映的是模型内在的本质特性,因而研究统计模型的内在几何结构及性质对于设计或选择更有效的学习系统具有非常重要的作用,对于模型选择这样一个统计推断问题从几何角度去研究是可行而且是很有意义的,有助于研究者提出更有效更准确的选择准则.

2　关于拟合度

为了衡量模型表示的分布与真实分布的逼近程度,需要在整个概率分布空间中定义衡量两个分布之间"差异量"的度量,并且应该保证该度量与参数的表示形式是无关的.

2.1　拟合度中的不变差异度量

在信息几何理论中,含参数的分布族构成了一个微分流形,参数作为坐标将每个分布表示为一个点.信息几何的独特之处就是将分布看作为点,可以讨论由分布组成的几何体的长度、角度、曲率等等.信息几何定义了分布间的差异量(divergence)—"δ 差异量"作为线性空间中距离的推广,并保证了具有参数表示不变性,即这种定义的"距离"在输入空间、输出空间、参数空间的任何可逆变换下均保持不变.

对于一般的统计流形 S,相应的 δ-表示为

$$l_\delta(x;\theta) := F_\delta(p(x;\theta)) = \begin{cases} p^\delta/\delta, & \delta > 0 \\ \log p, & \delta = 0 \end{cases} \tag{1}$$

$\delta \in [0,1]$,特别的,$\delta = 0$ 时称为指数表示,$\delta = 1$ 时称为混合表示.关于统计上微分流形的理论,文献[19,20]已经证明了在 δ-表示下,统计流形的几何性质不依赖于随机变量的表示.对于正的有限测度集合(即任一测度的积分为有限的正数) S 中的两个测度 p 与 q 之间的 δ 差异量为:

$$D_\delta(p,q) = \frac{1}{\delta(1-\delta)}\int [\delta p + (1-\delta)q - p^\delta q^{1-\delta}], \delta \in (0,1)$$

因此对于积分为 1 的概率分布,其广义 Kullback-Leibler(KL)距离为:

$$D_\delta(p,q) = \frac{1}{\delta(1-\delta)}\left(1 - \int p^\delta q^{1-\delta}\right) \tag{2}$$

而通常的 KL 距离为:

$$KL(p,q) = \lim_{\delta\to 0} D_\delta(q,p) = \lim_{\delta\to 1} D_\delta(q.p)$$
$$= \int (p - q + p\ln\frac{p}{q}) = \int p\ln\frac{p}{q} \tag{3}$$

δ 差异量具有参数表示不变性,即 $D_\delta(p,q)$ 不依赖于所在空间的测度.δ 差异量可以作为线性理论中距离的推广,但它与距离不同,不满足对称性,因而文中用"差异量"表示.以往的模型选择准则使用对数似然函数来估计拟合度,本文从信息几何的角度,以一种更为直观更为自然的方式寻求模型-数据拟合程度的估计,并定义几何拟合度的概念.图 1 表示了样本空间 Z、参数空间 Θ、概率空间 P、模型空间 S 之间的映射关系.

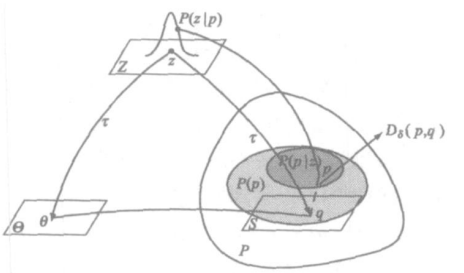

图 1　拟合度估计示意图

设真实分布为 $p \in P$,样本 $z \in Z$ 服从真实分布 p,映射 τ 将 z 映射为对 p 的一个估计 $q = \tau(z) \in S$,模型 S 中的每个点 q 对应的参数为 $\boldsymbol{\theta} \in \Theta$.服从分布 q 的模型对数据的拟合程度可以通过该分布与真实分布 p 之间的"距离"来度量,信息几何中对于 δ-平坦的概率统计模型,用式(2)中的 $D_\delta(p,q)$ 来表示.在这种拟合度衡量标准下选用的模型不仅对特定的观察数据集拟合程度好,还保证了能对同一分布产生的未来数据有较好的预测.

由于真实分布未知,我们用"距离" $D_\delta(p,q)$ 的贝叶斯后验均值作为对拟合度的度量.在这种考虑下,定义几何拟合度如下:

定义 1　设真实分布的先验概率为 $P(p) \in S$(如图 1 所示),则几何拟合度定义为:

$$GF = -E_\delta(q|z) = -\langle D_\delta(p,q)\rangle_z = -\int_p p(p|z) D_\delta(p,q) \tag{4}$$

定理 1　几何拟合度的计算由下面的定理给出:

$$E_\delta(q|z)\approx D_\delta(\hat{p},\hat{q}) \tag{5}$$

其中 $\hat{p}$(如图 2 所示)为 p 在 P 中的最优估计值,$\hat{q}$ 是 $\hat{p}$ 到 S 的 δ 投影.

为了证明该定理,首先介绍下面的一个引理[20]:

引理 1 设 P 为一个 δ 平坦流形,设 $P(p)$ 为 P 中的一个先验概率,$\forall q\in P,\forall z\in Z$ 则

$$E_\delta(q|z)=E_\delta(\hat{p}|z)+D_\delta(\hat{p},q) \tag{6}$$

其中 $\hat{p}$(如图 2 所示)为 p 在 P 中的最优估计值,本文中取其 δ 后验均值.定理 1 的证明如下:

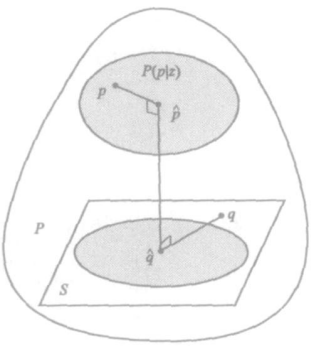

图 2 拟合度分解示意图

对于 δ 对偶平坦的统计流形,应用信息几何中的重要定理广义 Pythagoras 定理[19]将 $D_\delta(p,q)$ 进行如图 2 所示的分解并得到以下分解等式:

$$\begin{aligned}D_\delta(p,q)&=D_\delta(p,\hat{p})+D_\delta(\hat{p},q)\\&=D_\delta(p,\hat{p})+D_\delta(\hat{p},\hat{q})+D_\delta(\hat{q},q)\end{aligned} \tag{7}$$

其中 $\hat{q}$ 是 $\hat{p}$ 到 S 的 δ 投影,根据信息几何,$\hat{p},\hat{q}$ 具有对应的统计意义,$\hat{p}^\delta:=\langle p^\delta\rangle_2$ 为 $\hat{p}$ 的 δ 后验均值.

综合式(6)、(7)可得到下面的表达式:

$$\begin{aligned}E_\delta(q|z)&=E_\delta(\hat{p}|z)+D_\delta(\hat{p},q)\\&=\langle D_\delta(p,\hat{p})\rangle_z+D_\delta(\hat{p},\hat{q})+D_\delta(\hat{q},q)\end{aligned} \tag{8}$$

式(8)右边第一项是由于有限的样本数据而产生的均值分布与真实分布之间的偏差,科学的选取样本可以更好的逼近真实分布从而减少此项偏差,由于该项与选用的样本数据有关,与模型无关,因而在选择模型时可将此项略去不考虑.显然通过最小化式(8)我们可以得到最佳的拟合度:

$$\arg\operatorname*{Min}_q E_\delta(q|z)=\operatorname{Min}(D_\delta(\hat{p},\hat{q})+D_\delta(\hat{q},q)) \tag{9}$$

由于模型分布 q 未知,这里考虑其最好的一种情况,最小化 $E_\delta(q|z)$ 的模型分布 $q\in S$ 称为在样本为 z 时的最优估计,此时 $q=\hat{q}$,$D_\delta(\hat{q},q)=0$,则 $E_\delta(q|z)$ 取值为 $D_\delta(\hat{p},\hat{q})$,综上,有近似计算表示式:

$$E_\delta(q|z)\approx D_\delta(\hat{p},\hat{q}) \tag{10}$$

几何拟合度使用流形中两个分布之间的"距离"作为欧氏线性理论中距离概念的推广同样保证了参数表示不变性.

3 关于复杂度

如前所述,复杂度指的是模型拟合各种不同数据模式的能力.一般来说,模型越复杂就越容易吸收随机噪音,因此增加对带有噪音的数据的拟合度未必增加它对数据内在规律的拟合度.事实上,总是可以通过增加复杂度来提高拟合度.然而,这样一来导致的过拟合和欠拟合都将引起相对低的泛化能力.这意味着一个其复杂度逼近真实模型复杂度的模型具有可能最好的泛化能力.文献[17]通过引入几何复杂度对模型选择作了较深入的分析;文献[21]使用物理学中低温扩展方法揭示了贝叶斯方法的渐进性行为并且表明模型推断的贝叶斯方法同样是为了达到某种平衡.在该文中作者认为那些与数据样本大小成比例的项实质上是为了惩罚模型的曲率,而且是模型的几何特征对于统计推断的主要影响.正如微分几何表明的那样,具有清楚和直观几何意义的内蕴几何量——曲率可以用来衡量一个模型的非线性程度甚至可以做更多的事.本节从参数估计量邻域附近的解轨迹来讨论模型选择中的复杂度问题,并且考察曲率与泛化能力的关系.

3.1 立体阵与解轨迹

我们现在考虑一个随机系统.假设该系统由一个 p-维参数 $\boldsymbol{\theta}$ 参数化.定义 n 个观测值作为 n 个训练样本 $y_t=f(x_t,\theta)+\varepsilon,t=1,\cdots,n$,这里 ε 是随机噪音.我们的目的是找到一个合适的参数 $\boldsymbol{\theta}$,它能很好逼近数据中的规律.用向量重写上面等式为

$$\boldsymbol{Y}=f(\boldsymbol{X},\boldsymbol{\theta})+\boldsymbol{\varepsilon} \tag{10}$$

其中 $Y_{n\times}\in R,X_{n\times 1}\in R^k,\boldsymbol{\varepsilon}=(\boldsymbol{\varepsilon}_1,\cdots,\boldsymbol{\varepsilon}_n),\theta\in\Theta$,假设 Θ 是 p-维欧氏空间 R^p 的一个开子集,f 是关于 $\boldsymbol{\theta}$ 的 C^∞ 连续函数.$\boldsymbol{\varepsilon}\sim N(0,\sigma^2E)$,表示为 $iddN$.进一步描述之前,我们先给出有关立体阵的定义和运算[22,23].

定义 2 一个立体阵定义为 $\boldsymbol{X}_{tpq}=(\boldsymbol{X}_{tij}),t=1,\cdots,n;i=1,\cdots,n;j=1,\cdots,n$,其中 $\boldsymbol{t},\boldsymbol{p}$ 和 $\boldsymbol{q}$ 分别表示面、行和列.

定义 3 一个矩阵 $\boldsymbol{A}_{st}$ 和一个阵列 $\boldsymbol{X}_{tij}$ 的方括号乘积为 $\boldsymbol{Y}_{sij}=\sum_{t=1}^{n}\boldsymbol{A}_{st}\boldsymbol{X}_{tij}$,表示为 $\boldsymbol{Y}=[\boldsymbol{A}][\boldsymbol{X}]$,其维数关系为 $\boldsymbol{Y}_{m\times p\times q}=[\boldsymbol{A}_{m\times n}][\boldsymbol{X}_{n\times p\times q}]$.

定义 4 $\boldsymbol{Y}=\boldsymbol{AX}$ 表示矩阵 $\boldsymbol{A}$ 与立体阵 $\boldsymbol{X}$ 后两个指标的矩阵乘法.即若记 $\boldsymbol{Y}=(\boldsymbol{Y}_{tij}),\boldsymbol{A}=(\boldsymbol{A}_{ie})$,则定义 $\boldsymbol{Y}_{tij}=\sum_{e=1}^{p}\boldsymbol{A}_{ie}\boldsymbol{X}_{tej}$,其维数关系为 $\boldsymbol{Y}_{n\times r\times q}=\boldsymbol{A}_{r\times p}\boldsymbol{X}_{n\times p\times q}$.

定义 5 设 $\boldsymbol{X}$ 为 $n\times p\times q$ 阶立体阵.X 的迹定义为一个 n 维向量,记为 $\operatorname{tr}[\boldsymbol{X}]$,其形式为 $\operatorname{tr}[\boldsymbol{X}]=(\operatorname{tr}X_1,\cdots,\operatorname{tr}X_n)'$.

为了便于今后讨论,我们以定义的形式引入以下记号:

定义 6 $f(\boldsymbol{X},\boldsymbol{\theta})=f(\boldsymbol{\theta})=(f_t),t=1,\cdots,n.\boldsymbol{e}(\boldsymbol{\theta})=\boldsymbol{Y}-f(\boldsymbol{\theta})$,若 $\boldsymbol{\theta}$ 为真参数,则 $\boldsymbol{e}(\boldsymbol{\theta})=\boldsymbol{\varepsilon}$

$$\boldsymbol{V}(\boldsymbol{\theta})=\left(\frac{\partial f_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\right),\boldsymbol{W}(\boldsymbol{\theta})=\left(\frac{\partial^2 f_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\right),$$

$$t=1,\cdots,n;i,j=1,\cdots,p$$

$$\boldsymbol{S}(\boldsymbol{\theta})=\|\boldsymbol{Y}-f(\boldsymbol{\theta})\|^2=(\boldsymbol{Y}-f(\boldsymbol{\theta}))'(\boldsymbol{Y}-f(\boldsymbol{\theta}))$$

其中撇号表示向量或矩阵的转置,$\boldsymbol{V}(\boldsymbol{\theta})$ 为 $n\times p$ 阶矩阵,$\boldsymbol{W}(\boldsymbol{\theta})$ 为 $n\times p\times q$ 阶立体阵.设 $\boldsymbol{\eta}=f(\boldsymbol{X},\boldsymbol{\theta})$,它可以看成从参数空间 Θ 到样本空间 R^n 的映射.在 R^n 中该映射定义了一个由所有向量 $\boldsymbol{\eta}=f(\boldsymbol{\theta})$ 的端点组成的流形,我们把该流形称为模型的解轨迹,表示为 π,π 的几何性质对于分析模型的统计特征很重要.如果模型是线性的,那么 π 是一个由 $\boldsymbol{X}$ 的列向量生成的超平面.对于一个非线性模型,切平面可以用来逼近 π,但假设噪音水平相对于 π 充分光滑来讲很小.直观上,这意味着噪音水平比起 π 的曲率半径来说相对较小.

对于模型(10),如果存在一个估计量 $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ 满足 $S(\hat{\boldsymbol{\theta}})=\inf_{\theta\in\Theta}S(\boldsymbol{\theta})$,称 $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ 为 $\boldsymbol{\theta}$ 的最小二乘估计量 LSE.$S(\boldsymbol{\theta})$ 表示

$\boldsymbol{Y}$ 与 $f(\boldsymbol{\theta})$ 的距离,因此 $f(\hat{\boldsymbol{\theta}})$ 就是样本空间中解轨迹 π 上离 $\boldsymbol{Y}$ 最近的点,见图 3. 容易看到 $f(\hat{\boldsymbol{\theta}})$ 是 π 上最接近 $\boldsymbol{Y}$ 的点,并且 $\boldsymbol{Y}$ 与 π 间的最短距离为 $\|\hat{\boldsymbol{e}}\|$. 在式(10)的模型中,若 $f(\boldsymbol{X},\boldsymbol{\theta})$ 在 Θ 上关于 $\boldsymbol{\theta}$ 存在一阶连续偏导数,且 $\boldsymbol{\theta}$ 的 LSE 估计量 $\hat{\boldsymbol{\theta}}$ 存在,则残差向量 $\hat{\boldsymbol{e}} = \boldsymbol{Y} - f(\hat{\boldsymbol{\theta}})$ 在点 $\hat{\boldsymbol{\theta}}$ 处与该切空间正交. LSE 的存在性由 Jennrich 定理[24]保证. 如果假设 $\boldsymbol{\varepsilon}$ 为 $iddN$,那么 $\boldsymbol{\theta}$ 的 LSE 就是 $\boldsymbol{\theta}$ 的最大似然估计量 MLE.

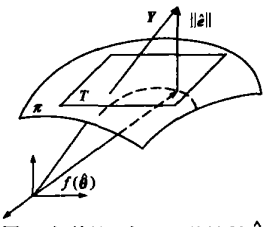

图 3　解轨迹 π 与 LSE 估计量 $\hat{\boldsymbol{\theta}}$

3.2　曲率度量与曲率立体阵

Efron[25]首先注意到统计模型的曲率在统计推断中的作用. Amari[19]做了进一步的工作,详细讨论了指数曲率和混合曲率的统计作用. 现在考虑参数空间 Θ 中一条过 $\boldsymbol{\theta}_0$ 以 $\boldsymbol{h}$ 为固定方向的直线 $l_h: \boldsymbol{\theta}(b) = \boldsymbol{\theta}_0 + b\boldsymbol{h}$,其中 b 为实参数. l_h 通过 $\boldsymbol{\eta} = f(\boldsymbol{\theta})$ 映射到样本空间 R^n 中解轨迹 π 上的一条曲线 $c_h: \boldsymbol{\eta} = \boldsymbol{\eta}_h(b) = f(\boldsymbol{\theta}_0) + b\boldsymbol{h}$. 这条曲线叫提升线. 由于

$$\frac{d\boldsymbol{\eta}_t}{db} = \sum_{i=1}^{p}\frac{\partial f_t}{\partial \boldsymbol{\theta}_i}\frac{d\boldsymbol{\theta}_i}{db} = \sum_{i=1}^{p} V_{ti}\boldsymbol{h}_i,$$

$$\frac{d^2\boldsymbol{\eta}_t}{db^2} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\partial^2 f_t}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\frac{d\boldsymbol{\theta}_i}{db}\frac{d\boldsymbol{\theta}_j}{db} = \sum_{i=1}^{p}\sum_{j=1}^{q} W_{tij}\boldsymbol{h}_i\boldsymbol{h}_j,$$

注意 $f(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 的前二阶导数分别为 $\boldsymbol{V}(\boldsymbol{\theta}) = (V_{ti})$, $\boldsymbol{W}(\boldsymbol{\theta}) = (\boldsymbol{W}_{tij})$,因此提升线 c_h 沿 $\boldsymbol{h}$ 方向的前二阶导数分别为

$$\dot{\boldsymbol{\eta}}_h = \frac{\partial \boldsymbol{\eta}_h}{\partial b} = \boldsymbol{V}\boldsymbol{h},\ \ddot{\boldsymbol{\eta}}_h = \frac{\partial^2 \boldsymbol{\eta}_h}{\partial b^2} = \boldsymbol{h}'\boldsymbol{W}\boldsymbol{h} \tag{11}$$

当 $b = 0$ 时,上式对应于 $\boldsymbol{\theta}_0$,以后一般取 $\boldsymbol{\theta}_0$ 为真参数或 $\hat{\boldsymbol{\theta}}_0$. $\ddot{\boldsymbol{\eta}}_h$ 可分解为三个分量,垂直于切平面的法分量 $\ddot{\boldsymbol{\eta}}_h^N$ 与在切平面上平行于切方向 $\dot{\boldsymbol{\eta}}_h$ 及垂直于切方向的分量 $\ddot{\boldsymbol{\eta}}_h^P$ 和 $\ddot{\boldsymbol{\eta}}_h^C$(这两者构成了切分量 $\ddot{\boldsymbol{\eta}}_h^T$). 分量 $\ddot{\boldsymbol{\eta}}_h^N$ 是由于解轨迹沿着法方向的弯曲而引起的,$\ddot{\boldsymbol{\eta}}_h^T$ 则是由于切平面上沿着 $\boldsymbol{h}$ 方向及垂直方向的不均匀性引起的. Bates 和 Watts[22]从微分几何观点定义了如下的关于模型的固有曲率 K_h^N 和参数效应曲率 K_h^T.

定义 7　式(10)的模型沿着 $\boldsymbol{h}$ 方向在 $\boldsymbol{\theta}_0$ 处的固有曲率 K_h^N 和参数效应曲率 K_h^T 分别为:

$$K_h^N = \frac{\|\ddot{\boldsymbol{\eta}}_h^N\|}{\|\dot{\boldsymbol{\eta}}_h\|^2} = \frac{\|(\boldsymbol{h}'\boldsymbol{W}\boldsymbol{h})^N\|}{\boldsymbol{h}'\boldsymbol{V}'\boldsymbol{V}\boldsymbol{h}}$$

$$K_h^T = \frac{\|\ddot{\boldsymbol{\eta}}_h^T\|}{\|\dot{\boldsymbol{\eta}}_h\|^2} = \frac{\|(\boldsymbol{h}'\boldsymbol{W}\boldsymbol{h})^T\|}{\boldsymbol{h}'\boldsymbol{V}'\boldsymbol{V}\boldsymbol{h}} \tag{12}$$

其中 $(\boldsymbol{h}'\boldsymbol{W}\boldsymbol{h})^N$ 和 $(\boldsymbol{h}'\boldsymbol{W}\boldsymbol{h})^T$ 分别为向量 $(\boldsymbol{h}'\boldsymbol{W}\boldsymbol{h})$ 的法分量和切分量. $\boldsymbol{h}$ 表示方向向量,通常可取单位向量.

以上定义与微分几何中的曲率是等价的. K_h^N 与参数的选择无关,它由模型本身固有的性质决定; K_h^T 则不仅由模型本身决定,而且强烈的依赖于参数的选择. 基于以上的几何曲率,文献[22]引入曲率立体阵避免了复杂的计算并使用 QR 分解避免对方向的依赖,方法是用一组标准正交基向量替换由 $\boldsymbol{V}(\boldsymbol{\theta})$ 的列向量生成的切空间的基

$$\boldsymbol{V}_{np} = (\boldsymbol{Q},\boldsymbol{N})\begin{pmatrix}\boldsymbol{R}\\ 0\end{pmatrix} = \boldsymbol{Q}_{np}\boldsymbol{R}_{pp} \tag{13}$$

其中 $\boldsymbol{Q}$ 和 $\boldsymbol{N}$ 的列向量分别是切空间和法空间的正交向量. 在这个新的变换下, $\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{h}$ 分别变为 $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{L}$、$\boldsymbol{U} = \boldsymbol{L}'\boldsymbol{W}\boldsymbol{L}$,($\boldsymbol{L} = \boldsymbol{R}^{-1}$)和 $\boldsymbol{d} = \boldsymbol{R}\boldsymbol{h}$. 因此得到相应的固有曲率立体阵和参数效应曲率立体阵.

定义 8　式(10)的模型的固有曲率立体阵和参数效应曲率立体阵分别为:

$$\boldsymbol{I}_{(n-p)pp} = [\boldsymbol{N}'][\boldsymbol{U}],\ \boldsymbol{P}_{ppp} = [\boldsymbol{Q}'][\boldsymbol{U}] \tag{14}$$

3.3　预测问题

为了分析 LSE 渐进特性,需要一些约束. 关于所需的正则条件及定理参看[24,26~28]. 给定一个非线性系统见式(10),我们考虑给定新数据 $\boldsymbol{X}_0$ 时 $\boldsymbol{Y}_0$ 的预测;一般 $\boldsymbol{Y}_0 = f(\boldsymbol{X}_0, \hat{\boldsymbol{\theta}})$ 作为 $\boldsymbol{Y}_0$ 的预测值. 下面的定理表明残差向量 $\hat{\boldsymbol{e}} = \boldsymbol{Y} - f(\hat{\boldsymbol{\theta}})$ 和拟合误差向量 $\Delta f = f(\hat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}_0)$ 可以用与参数化无关的曲率立体阵 $\boldsymbol{I}$ 表示.

定理 2　如果满足所需的正则条件,那么有

$$\begin{aligned}\hat{\boldsymbol{e}} &= N\boldsymbol{\lambda} - \boldsymbol{Q}[\boldsymbol{\lambda}']\cdot[\boldsymbol{I}]\boldsymbol{\tau} - \frac{1}{2}\boldsymbol{N}(\boldsymbol{\tau}'\boldsymbol{I}\boldsymbol{\tau}) + o(n^{-1})\\ \Delta f &= \boldsymbol{Q}\boldsymbol{\tau} + \boldsymbol{Q}[\boldsymbol{\lambda}']\cdot[\boldsymbol{I}]\boldsymbol{\tau} + \frac{1}{2}\boldsymbol{N}(\boldsymbol{\tau}'\boldsymbol{I}\boldsymbol{\tau}) + o(n^{-1})\end{aligned} \tag{15}$$

$\boldsymbol{\tau} = \boldsymbol{Q}'\boldsymbol{\varepsilon}$, $\boldsymbol{\lambda} = \boldsymbol{N}'\boldsymbol{\varepsilon}$ 和 $\boldsymbol{I}$ 在 $\boldsymbol{\theta}_0$ 处计算. 以上两个等式都与参数化无关,因为它们仅依赖于固有曲率立体阵 $\boldsymbol{I}$.

定理 3　如果满足所需的正则条件,那么等式(15)的期望和方差为

$$E[\Delta f] = \frac{\sigma^2}{2}N\mathrm{tr}[\boldsymbol{I}] + o(n^{-1}) \tag{16}$$

$$\mathrm{Var}(\Delta f) \approx \sigma^2\boldsymbol{P}_T + \sigma^4\boldsymbol{Q}\boldsymbol{V}_I\boldsymbol{Q}' + \frac{1}{2}\sigma^4\boldsymbol{N}\boldsymbol{V}_I^*\boldsymbol{N}' \tag{17}$$

其中 $\boldsymbol{V}_I = \sum_{t=1}^{n-p} I_t^2$, I_t 是 $\boldsymbol{I}$ 的第 t 面; $\boldsymbol{V}_I^* = \sum_{k=1}^{p}\sum_{l=1}^{p} I_{kl}I'_{kl}$, I_{kl} 是 $\boldsymbol{I}$ 在 (k,l) 处的一个 $(n-p)$-维向量; $\boldsymbol{P}_T = \boldsymbol{Q}\boldsymbol{Q}'$ 是切空间的投影矩阵; tr 表示矩阵的迹, $\boldsymbol{I}$ 在 $\boldsymbol{\theta}_o$ 处计算. 同样以上两个等式都与参数化无关,因为它们仅依赖于固有曲率立体阵 $\boldsymbol{I}$.

由于当前残差 $\|\boldsymbol{Y} - f(\hat{\boldsymbol{\theta}})\|^2$ 不足以衡量模型的好坏. 为了得到好的泛化能力,自然应当最小化对未来数据的未来残差. 定义的未来期望残差形式[15]为:

$$R_0 = E_0[E[\|\boldsymbol{Y}_0 - f(\hat{\boldsymbol{\theta}})\|]] \tag{18}$$

如果 R_0 越小,则模型越好. 这里, $E_0[\cdot]$ 和 $E[\cdot]$ 分别表示关于未来数据和当前数据的期望;假设 $\boldsymbol{Y}_0$ 和 $\boldsymbol{Y}$ 相互独立且服从同一分布. 由于 $f(\hat{\boldsymbol{\theta}})$ 由当前数据 $\boldsymbol{Y}$ 决定,且 $\boldsymbol{Y}$ 独立于未来数据 $\boldsymbol{Y}_0$. 因此,等式(18)简化为:

$$\begin{aligned}R_0 &= E_0[E[\|[\boldsymbol{Y}_0 - f(\boldsymbol{\theta}_0)] - [f(\hat{\boldsymbol{\theta}}) - f(\hat{\boldsymbol{\theta}}_0)]\|^2]]\\ &= E_0[\|\boldsymbol{Y}_0 - f(\boldsymbol{\theta}_0)\|^2] + E[\|f(\hat{\boldsymbol{\theta}}) - f(\hat{\boldsymbol{\theta}}_0)\|^2]\end{aligned}$$

又由于 $\boldsymbol{Y}_0$ 和 $\boldsymbol{Y}$ 服从同一分布,我们有 $E_0[\boldsymbol{Y} - f(\boldsymbol{\theta}_0)\|^2] = E[\|\boldsymbol{Y} - f(\boldsymbol{\theta}_0)\|^2]$. 故上式重写为:

$$R_0 = E[\|\boldsymbol{Y} - f(\boldsymbol{\theta}_0)\|^2] + E[\|f(\hat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}_0)\|^2] \tag{19}$$

注意到

$$\boldsymbol{Y} - f(\hat{\boldsymbol{\theta}}) = \boldsymbol{Y} - f(\boldsymbol{\theta}_0) - [f(\hat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}_0)] = \boldsymbol{\varepsilon} - \Delta f$$

从而有

$$R_0 = E[\|\boldsymbol{\varepsilon}\|^2] + E[\|\Delta f\|^2] \tag{20}$$

如果取 R_0 为未来残差,那么当给定一个 LSE 时,由它反映的模型的复杂度是什么呢?下面的定理给出了 R_0 与曲率的关系.

定理 3 如果满足所需的正则条件,并假设 ε 为 $iidN$,那么有

$$R_0 \approx (n+p)\sigma^2 + \frac{3}{2}\sigma^4\left[\sum_{k=1}^{p}\sum_{l=1}^{p}\| I_{kl}\|^2 + \frac{1}{6}\|\mathrm{tr}[I]\|^2\right] \quad (21)$$

证明 首先有

$$E\|*\|^2 = \|E(*)\|^2 + \mathrm{tr}|\mathrm{Var}(*)|,\ *\text{代表向量} \quad (22)$$

由等式(16)有:$\|E(\Delta f)\|^2 = \frac{1}{4}\sigma^4\|\mathrm{tr}[I]\|^2$

再由等式(17)有:

$$\mathrm{tr}|\mathrm{Var}(\Delta f)| \approx \sigma^2\mathrm{tr}(P_T) + \sigma^4\mathrm{tr}(V_l Q' Q) + \frac{1}{2}\sigma^4\mathrm{tr}(V_l^* N' N)$$

$$\approx p\sigma^2 + \sigma^4\mathrm{tr}(V_l) + \frac{1}{2}\sigma^4\mathrm{tr}(V_l^*)$$

从等式 $V_l = \sum_{i=1}^{n-p} I_i^2$,得到

$$\mathrm{tr}(V_l) = \sum_{i=1}^{n-p}\mathrm{tr}(I_i^2) = \sum_{i=1}^{n-p}\sum_{k=1}^{p}\sum_{l=1}^{p} I_{ikl}^2 = \sum_{k=1}^{p}\sum_{l=1}^{p}\| I_{kl}\|^2$$

再由等式 $V_l^* = \sum_{k=1}^{p}\sum_{l=1}^{p} I_{kl} I'_{kl}$ 得到

$$\mathrm{tr}(V_l^*) = \sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{tr}(I_{kl}^* I_{kl}) = \sum_{k=1}^{p}\sum_{l=1}^{p}\| I_{kl}\|^2$$

将以上等式代入式(22)后得到

$$E\|\Delta f\|^2 \approx p\sigma^2 + \frac{3}{2}\sigma^4\sum_{k=1}^{p}\sum_{l=1}^{p}\| I_{kl}\|^2 + \frac{1}{4}\sigma^4\|\mathrm{tr}[I]\|^2$$

同样的计算可得:$E\|\varepsilon\|^2 = n\sigma^2$

由以上等式可得式(21).显然式(21)与参数化无关,因为它只依赖于固有曲率立体阵 I.当数据样本给定时,固有曲率立体阵真实反映了模型的内在特征,我们称为固有复杂度并用它作为一个共同的标准来衡量和比较模型的复杂度.

4 KLCIC 模型选择准则

由定理 3 和式(8)可以得到下面的模型选择准则:

定理 4 在所有给定的模型中,选择使下式最小的模型

$$KLCIC = -\ln E(q|z) + R_0 \quad (23)$$

我们称为 KL 距离和曲率信息准则(KLCIC).其中 R_0 在 $\hat{\theta}_0$ 处计算.当噪音水平比起 π 的曲率半径来说相对较小时,给定 n 个数据样本时由 KLCIC 选择的模型要求参数个数 p 尽可能少,并且选择具有尽可能小固有曲率的模型,这意味着所选择的模型在情况允许的条件下尽可能倾向于逼近线性模型.这并非说模型越简单越好,简单到何种程度最终还得由给定的数据决定.所以我们需要做的是对于所选模型,在它具有上式第一项所描述的最佳拟合度的条件下,尽可能减小超出真实模型复杂度的附加复杂度.此外,在某种意义上 I 反映了模型的结构,可以看成代表网络或模型的固有复杂度.

5 比较实验与分析

为了进行比较实验,我们给出两种常用的模型选择准则:

$$BIC = -\ln f(x|\hat{\theta}) + p\ln(n)$$

$$MDL = -\ln f(x|\hat{\theta}) + \frac{p}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int\sqrt{\det I(\theta)}\,d\theta$$

其中 $\hat{\theta}$ 是模型参数的 MLE,p 是模型中参数个数,n 是样本大小,$I(\theta)$是 Fisher 信息矩阵.以上各式中,第一项为拟合度,第二和第三项为复杂度.两个实验模型(M_2 比 M_1 复杂)为:

$$M_1: y = (1+\exp(a-bx))^{-1}$$

$$M_2: y = c + (1-c)(1-\exp(-ax^b))$$

其中 $a,b>0, 0<c<1$.实验方案如下:每个样本都只在 $x=(0.001,1,2,4,7)$5 个点上取值.每个模型分别生成 20,40,60 个样本并加入均值为零、方差为 1 的抽样噪音.用于生成样本数据的参数分别为:$M_1(a=2,b=3)$,$M_2(a=0.3,b=0.5,c=0.1)$.以上三种模型选择准则加上 KLCIC 分别应用于每一样本集中每一样本,即对每个样本进行模型选择.较小的值对应的模型为要选择的模型.选择准则的比较是基于它们多大程度恢复产生样本数据的真实模型的能力,即各准则选择真实模型的样本数占总样本数的百分比.一个理想的选择方法应该 100% 地识别出真实模型.当然,推断的质量强烈地依赖于数据特征(比如样本集大小、实验设计方法、噪音类型)和模型本身.实验结果如表 1 所示.

表 1 模型选择准则的恢复能力(%)的比较

样本容量	拟合模型	模型选择准则					
		BIC		MDL		KLCIC	
		产生数据的模型		产生数据的模型		产生数据的模型	
		$M1$	$M2$	$M1$	$M2$	$M1$	$M2$
20	$M1$	65.0	30.0	85.0	10.0	85.0	10.0
	$M2$	35.0	70.0	15.0	90.0	15.0	90.0
40	$M1$	75.0	20.0	90.0	2.5	87.5	5.0
	$M2$	25.0	80	10.0	97.5	12.5	95.0
60	$M1$	83.3	13.3	98.3	0.00	96.7	0.00
	$M2$	16.7	86.7	1.7	100	3.3	100

由表 1 中数据可见,三种模型选择方法在不同的样本容量下均能选出正确的模型.随着样本容量的增加,三种方法恢复正确模型的能力也随之增加.在三种方法中,BIC 方法的性能最差,这是由于它没有考虑模型的函数形式对复杂度的影响.MDL 和 KLCIC 都在参数变换下保持不变,并且都考虑了影响复杂度的样本容量、参数个数及函数形式.当样本容量较小时,MDL 比 KLCIC 的性能好一些;当样本容量较大时,两者性能基本相同,尽管 MDL 方法是目前公认的最有效的方法.但是 KLCIC 方法从另一角度揭示了模型选择的内在本质,它使我们能够更直观、更清晰地理解模型选择的几何意义.计算 KLCIC 中曲率立体阵有通用的计算机程序,可参考文献[26]或其中文译本文献[29].

6 结论

本文从信息几何的观点使用定义在流形上的广义 KL 距离来度量模型的拟合度;另一方面从微分几何的观点用曲率的概念来度量模型的内在复杂度;因此,拟合度和复杂度的表示都具有在参数变换下保持不变的特点.我们给出了用于表示模型预测能力的未来期望残差与模型固有曲率的关系.由

此提出了一种新的基于广义 KL 距离和曲率的模型选择准则 KLCIC. 该准则不仅考虑了样本大小、参数个数和函数形式等影响复杂度的因素,而且具有非常清晰直观的几何意义. 实验结果表明了该方法的有效性.

参考文献:

[1] Akaike H. Information theory and an extension of the maximum likelihood principle[A]. B N Petrox, F Caski. Second International Symposium on Information Theory[C]. Budapest, 1973. 267 – 281.

[2] Bozdogan H. Akaike information criterion and recent developments in information complexity[J]. Journal of Mathematical Psychology, 2000, 44(1): 62 – 91.

[3] Schwarz G. Estimating the dimension of a model[J]. Annals of Statics, 1978, 7(2): 461 – 464.

[4] Rissanen J. A universal prior for integers and estimation by minimum description length[J]. Annals of Statistics, 1983, 11(2): 416 – 431.

[5] Rissanen J. Fisher information and stochastic complexity[J]. IEEE Trans Information theory, 1996, 42(1): 40 – 47.

[6] Hansen M H, et al. Model selection and the principle of minimum description length[J]. Journal of the American Statistical Association, 2001, 96(454): 746 – 774.

[7] Stone M. Cross-validation choice and assessment of statistical predictions[J]. Journal of the Royal Statistical Society, 1974, Series B, 36 (2): 461 – 464.

[8] Browne M W. Cross-validation methods[J]. Journal of Mathematical Psychology, 2000, 44(1): 108 – 132.

[9] Kass R E, et al. Bayes factor[J]. Journal of the American Statistical Association, 1995, 90(430): 773 – 795.

[10] Wasserman L. Bayesian model selection and model averaging[J]. Journal of Mathematical Psychology, 2000, 44(1): 92 – 107.

[11] Rissanen J. Stochastic complexity and the MDL principle[J]. Econometric Reviews, 1987, 6(3): 85 – 102.

[12] Bozdogan H. On the information-based measure of covariance complexity and its application to the multivariate linear models[J]. Commun Statist-Theory Meth, 1990, 19(1): 221 – 278.

[13] Sugiyama M, et al. Subspace information criterion for model selection [J]. Neural Computation, 2001, 13(8): 1863 – 1889.

[14] Johnson J K. Min-MaxKullback-Leibler model selection[CP/OL]. http://www.mit.edu/people/jasonj/termpaper-6.291-may02.pdf, 2002.

[15] Kanatani K. Geometric information criterion for model selection[J]. International Journal of Computer Vision, 1998, 26(3): 171 – 189.

[16] Mayung I J. Model comparison method[J]. Methods in Enzymology, 2003, 383: 351 – 366.

[17] Mayung I J, et al. Counting probability distributions: Differential geometry and model selection[A]. Proc. of National Academy of Science[C]. USA, 2000. 11170 – 11175.

[18] Bojan K, et al. Analysis of different model selection criteria[DB/OL]. http://citeseer.nj.nec.com/kverh99analysis.html. 1999.

[19] Amari S I. Differential-Geometrical Methods in Statistics[M]. Lecture Notes in Statistics, Berlin: Springer-Verlag, 1985.

[20] Zhu H Y, et al. Information geometric measurements of generalization [DB/OL]. Http://www.ncrg.aston.ac.uk/cgi-bin/tr avail.pl? tr-number = NCRG/95/005, 1995.

[21] Balasubramanian V. Statistical inference, Occcam's razor and statistical mechanics on space of probability distributions [J]. Neural Computation, 1997, 9(2): 349 – 368.

[22] Bates D, et al. Relative curvature measures of nonlinearity[J]. J R Statist Society, 1980, B42(1): 1 – 25.

[23] Tsia C. Contributions to the testing and analysis of nolinear models [D]. USA: Univ of Minnesota, 1983.

[24] Jennrich R I. Asymptotic properties of nonlinear least square estimators [J]. Ann. Math. Statist, 1969, 40(3): 633 – 643.

[25] Efron B. Defining the curvature of a statistical problem (with application to second efficiency)[J]. Ann Statist, 1975, 3(3): 1189 – 1242.

[26] Ratkowsky D A. Nonlinear Regression Modeling[M]. New York, USA: Marcel Dekker Inc, 1983.

[27] Wei B C. A unified methods for the moments of least squares estimator in nonlinear regression[A]. Proc of Sino-American Symposium on Statistics[C]. Beijing, 1987. 463 – 466.

[28] 韦博成. 近代非线性回归分析[M]. 南京: 东南大学出版社, 1989.

[29] 洪再吉, 等译. 非线性回归模型统一的实用方法[M]. 南京: 南京大学出版社, 1986.

作者简介:

杨　坚　男, 1970 年出生于贵州遵义, 目前为北京交通大学博士研究生, 主要研究方向为模型选择和神经网络.

E-mail: yj_swendy@citiz.net.

罗四维　男, 1943 年出生于北京, 北京交通大学教授、博士生导师, 目前主要研究方向为并行处理和神经网络.