研究论文

高斯过程及其在软测量建模中的应用

王华忠
(华东理工大学自动化研究所，上海 200237)

摘要：结合工业萘初馏塔关键质量指标估计问题，提出了采用高斯过程（GP）建立复杂工业过程软测量方法。将自动相关确定（ARD）原理与 GP 模型结合进行软测量模型辅助变量选择，通过建立 GP 软测量模型，同时得到关键质量指标估计值和相应的预测不确定度，有效解决了现有软测量建模方法不能给出估计值的测量不确定度的问题。研究表明，GP 软测量模型不仅能自动选择辅助变量，而且还具有较高的估计精度和较小的测量不确定度，能够更好地满足工业现场对测量可靠性的要求。

关键词：高斯过程；测量不确定度；软测量；建模
中图分类号：TQ 063；TP 183 **文献标识码**：A **文章编号**：0438－1157（2007）11－2840－06

Gaussian process and its application to soft-sensor modeling

WANG Huazhong
(*Research Institute of Automation*, *East China University of Science and Technology*, *Shanghai* 200237, *China*)

Abstract: With the estimation of key quality index in an industrial naphthalene distillation column, a novel soft-sensor modeling method based on Gaussian process (GP) was proposed for complex industrial processes. The principle of automatic relevance determination, implemented with GP model, was proposed to determine the secondary variables for the soft-sensor. To overcome the shortcomings existing in present methods, which can not determine the measurement uncertainty of soft-sensors, the GP based soft-sensor was developed to get both the prediction of key quality index and its measurement uncertainty simultaneously. Application studies showed that the GP soft sensor model not only determined the secondary variable automatically, but also possessed both high accuracy and small measurement uncertainty, which met the demands for reliable measurements in industrial application.

Key words: Gaussian process; measurement uncertainty; soft-sensor; modeling

引 言

现代工业生产对产品质量的要求越来越高，产品质量在某种程度上决定企业的生死存亡。但在过程工业中，许多重要的质量指标（反应器中反应物浓度，精馏塔产品组分等）却很难在线获得，软测量技术试图以过程可测辅助（直接）变量来估计直接质量指标（间接变量）。由于是采用过程模型来估计不可测变量，软测量模型也称作“软仪表”。

软测量技术起源于 20 世纪 70 年代Brosillow[1]提出的推断控制，它将工艺机理与控制理论有机地结合起来，在一定程度上解决了过程工业某些重要质量指标的在线控制问题。经过近 30 年的发展，软测量理论的研究从静态模型发展到动态模型，其

2006－11－29 收到初稿，2007－07－30 收到修改稿。
联系人：王华忠（1969—），男，博士。

Received date: 2006－11－29.
Corresponding author: Dr. WANG Huazhong. **E－mail**: hzwang@ecust.edu.cn

应用范围也从化工过程扩展到制造业等领域。软测量技术牵涉面很广，如数据预处理、辅助变量选择、软测量建模、模型校正等。正是由于其复杂性，现有的研究多侧重于该技术的一个或几个方面，而软测量建模方法是这类研究的重点。由于机理建模的复杂性，基于数据驱动（data-driven）的方法一直是最主要的软测量建模方法。在这些方法中，各种线性和非线性回归方法，典型的如 PLS[2] 和改进的非线性 PLS[3] 等一直备受重视。随着神经网络研究的兴起，各种基于神经网络的软测量方法[4]大量涌现，特别是 BP 神经网络和 RBF 神经网络。为了克服神经网络建模中存在的困难，将神经网络与进化[5]及其他优化算法以及模糊技术[6]相结合，产生了多种基于神经网络的软测量方法。而对于训练样本多、操作条件变化大的情况，采用多模型方法建立软测量模型可以提高模型预测精度和算法的实时性[7]。由于生产过程是处于动态中，为了克服静态软测量模型的局限性，提高软测量模型的长期适用性，动态软测量建模也引起了重视[8-9]。近年来，核函数方法，尤其是支持向量机[10]以其小样本处理能力强等优点，逐步成为软测量建模的重要方法。

根据测量原理，精密测量过程不仅应该得到一定的测量结果，而且应该给出该测量结果的精度参数（如测量不确定度）。对于直接测量变量，其不确定度较容易获得。但对于间接测量，特别是采用软测量这种方式，由于软测量模型通常较复杂，很难采用误差传递定律等方法来获得其估计值的精度参数。正因为如此，前面所述的各种软测量建模方法都不考虑软测量模型估计值的测量不确定度，这不能不说是现有软测量方法的显著不足之处。

高斯过程（GP）是随机变量的集合，集合中任意数量的随机变量组合服从联合高斯分布。高斯过程可以由均值函数和协方差函数唯一确定。高斯过程模型的主要优点体现在：它是一种非参数概率模型，不仅能对未知输入做输出预测，而且同时给出该预测的精度参数（即估计方差）；可以以先验概率的形式表示过程的先验知识，从而提高过程模型性能；与神经网络、支持向量机等方法相比，高斯过程模型参数明显减少，因而参数优化相对容易，且更易收敛。高斯过程既可以用于回归建模，也可以用于分类研究。近年来，高斯过程在机器学习等领域得到了成功的应用[11-12]，引起了控制界的关注，在内模控制[13]与软测量[14]中得到初步的应用。

基于高斯过程的上述特性，本文将其用于复杂工业过程建模。首先介绍了 GP 回归模型的基本原理和模型选择，然后以某厂工业萘初馏塔酚油含萘量软测量建模为例，研究了高斯过程软测量建模，结果表明高斯过程软测量模型在精度、可靠性等方面较好，可满足工业现场应用要求。

1　GP 原理分析

1.1　GP 模型

设给定训练样本

$$D=\{(\boldsymbol{x}_i,y_i)\}\qquad(i=1,\cdots,l)$$

其中，$\boldsymbol{x}_i\in R^n$，$y_i\in R$，l 为训练样本数，n 为输入向量维数。

对于新的测试样本 $\boldsymbol{x}$，GP 模型的预测值（即均值函数）为

$$\hat{y}=\mu(\boldsymbol{x})=\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{K}^{-1}\boldsymbol{y}\tag{1}$$

对于预测值的方差为

$$\sigma_{\hat{y}}^2(\boldsymbol{x})=k(\boldsymbol{x},\boldsymbol{x})-\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{K}^{-1}\boldsymbol{k}(\boldsymbol{x})\tag{2}$$

其中，$\boldsymbol{k}(\boldsymbol{x})=[C(\boldsymbol{x},\boldsymbol{x}_1),\cdots,C(\boldsymbol{x},\boldsymbol{x}_l)]^{\mathrm{T}}$ 为测试输入和训练样本输入值间的 $l\times1$ 维协方差向量，$\boldsymbol{K}$ 为 $l\times l$ 维训练样本间的协方差矩阵，其元素 $\boldsymbol{K}_{ij}=C(\boldsymbol{x}_i,\boldsymbol{x}_j)$，$k(\boldsymbol{x},\boldsymbol{x})$ 为测试输入和其自身的协方差，$\boldsymbol{y}$ 是 $l\times l$ 维输出向量。式（1）和式（2）表明，GP 可以根据均值函数、协方差函数和训练样本进行预测，并同时得到预测值的精度参数。而以往的多数方法不能对预测值的精度进行估计，这是高斯过程非常突出的一个优点。因为从统计预测角度来说，不仅希望得到估计值，而且还希望知道该估计值的精度参数。

协方差函数是高斯过程中最重要的组成之一，可以通过它对期望的函数特性做某些假设。比如，反映相似性（similarity）的概念，即相邻的输入产生相邻的输出。协方差函数应该满足对于任意的一个输入，相应的协方差矩阵 $\boldsymbol{K}$ 应该是对称正半定矩阵。常用的协方差函数包括平稳协方差函数和非平稳协方差函数。考虑到模型选择中对协方差函数连续性和可导性的要求，下列径向基函数是最常用的一类协方差函数

$$C(\boldsymbol{x}_i,\boldsymbol{x}_j)=v_0\exp\left[-\frac{1}{2}\sum_{t=1}^{n}w_t(x_i^t-x_j^t)^2\right]+v_1\delta_{ij}\tag{3}$$

其中，v_0 表示先验知识的总体度量，v_1 表示服从高斯分布的噪声的方差，δ_{ij} 是 Kronecker 算子。

1.2 GP 模型选择

虽然给定了协方差函数后，就可以根据给定训练样本对新的输入进行预测。但是由于协方差函数中的参数没有经过优化，这时 GP 模型的性能难于保证，因此必须进行模型选择。严格地说，模型选择包括协方差函数类型选择和 GP 模型相应的参数确定。由于同时确定这两者十分困难，目前最常用的方法就是先确定协方差函数类型，再优化模型参数。优化模型参数的主要方法有最大似然法、交叉验证和根据超参数的后验分布采用 Bayesian 推理方法[11]。对于上述 GP 过程，其超参数对数似然函数是

$$L(\boldsymbol{\Theta}) = -\frac{1}{2}\lg(|\boldsymbol{K}|) - \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{l}{2}\lg(2\pi) \quad (4)$$

其中，协方差函数的超参数 $\boldsymbol{\Theta}=[w_1, \cdots, w_n, v_0, v_1]$，$|\boldsymbol{K}|$ 表示 $\boldsymbol{K}$ 的行列式。

在优化过程中，要计算对数似然函数 L 对各个参数的导数

$$\frac{\partial L}{\partial \Theta_i} = -\frac{1}{2}Tr\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \Theta_i}\right) + \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \Theta_i}\boldsymbol{K}^{-1}\boldsymbol{y} = \frac{1}{2}Tr\left[(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} - \boldsymbol{K}^{-1})\frac{\partial \boldsymbol{K}}{\partial \Theta_i}\right] \quad (5)$$

其中，$\boldsymbol{\alpha}=\boldsymbol{K}^{-1}\boldsymbol{y}$。

从式（5）可以看出，在模型选择中，每一步迭代时都要计算 $l\times l$ 维矩阵 $\boldsymbol{K}$ 的逆矩阵，如果直接求逆，当 l 较大时（如大于 1000），该运算会十分耗时且数值解可能不稳定，内存消耗也很大。不过近年开发出的一些算法已经较好地解决了这个问题，文献［11］对这些方法进行了分析和比较。对软测量建模而言，由于有效样本数不会太多，算法实施并不困难。

2 基于 GP 的酚油含萘量软测量建模

2.1 工艺分析与数据采集

工业萘生产工艺流程主要由初馏和精馏二部分组成。初馏塔塔顶馏分为酚油和轻质杂茚馏分等，初馏塔塔底产品为脱酚萘洗油，用热油循环泵抽出，一部分经加热后返回塔底，以提供初馏塔蒸馏所需的热量，另一部分送入精馏塔作为进料。由于缺少在线分析仪表，以初馏塔塔顶温度作为塔顶组分质量控制指标，通过调节酚油的回流量来控制塔顶温度。实际操作中为了确保塔底萘洗油中不含或少含酚油等杂质，塔顶温度通常被控制得较高，从而导致萘酚油中萘含量过高，这样既不利于酚油提纯，也降低了精馏塔工业萘产量。因此，有必要建立酚油含萘量软测量模型，以提高控制水平。

首先从工艺机理出发，定性分析了影响酚油含萘量的 8 个因素，分别是初馏塔塔顶温度（Var1）、第 65 块塔板温度（Var2）、第 40 块塔板温度（Var3）、原料入塔温度（Var4）、初馏塔循环油温度（Var5）、初炉出口温度（Var6）、初馏塔进料量（Var7）、初馏塔向精馏塔投料量（Var8）。从 DCS 每隔 0.5 h 采集的数据中进行筛选和预处理，尽量使采集数据包含尽可能多的操作模式。最后共得到 350 组数据，其中 300 组数据用于建模，50 组数据用于校验。

2.2 基于 GP 和自动相关确定原理的辅助变量选择

在建立软测量模型时，重要的一步是选择辅助变量。实际操作中，由于工艺机理比较复杂，辅助变量的确定比较困难。为了防止漏掉一些对模型预测有贡献的变量，通常会选较多的变量作辅助变量。但这样一方面会导致模型结构复杂，降低其性能；另一方面会导致模型参数过多，参数优化比较困难。这里采用自动相关确定原理（automatic relevance determination，ARD）并结合 GP 模型，自动确定软测量模型的辅助变量。ARD 首先用于神经网络的贝叶斯学习训练[15]，其作用是自动确定神经网络的多个输入变量中，哪些与预测目标是相关的，并自动确定与目标的相关程度。对于式（3）形式的协方差函数，对每个输入变量 x_i，都有相应的参数 w_i 与之对应。若 w_i 大，则表示变量 x_i 对方差的贡献大，其对预测的贡献也大，即 x_i 是软测量模型重要的输入变量。反之，若 w_i 小，则变量 x_i 就不是重要的输入变量，可以不作为辅助变量。采用 GP 模型来实施 ARD 十分简便，一旦 GP 模型建立，w_i 的大小就自动确定，随即就可以根据 w_i 的数值确定软测量模型的辅助变量。需要指出的是，在使用 ARD 确定辅助变量时，所有的输入数据必须标准化，否则不能根据 w_i 的大小来判断输入对输出（或预测）的贡献度。

针对采集的初馏塔软测量建模训练和测试样本，采用 GP 方法建立软测量模型，协方差函数如式（3）所示。软测量模型超参数初始化数值为

$$\boldsymbol{\Theta}=[w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, v_0, v_1] = [1,1,1,1,1,1,1,1,1,0.001]$$

表 1　辅助变量 ARD 分析结果

Table 1　Results of ARD for secondary variables

w_1(Var1)	w_2(Var2)	w_3(Var3)	w_4(Var4)	w_5(Var5)	w_6(Var6)	w_7(Var7)	w_8(Var8)
1.735	0.629	0.331	0.506	0.274	0.016	0.126	0.0023

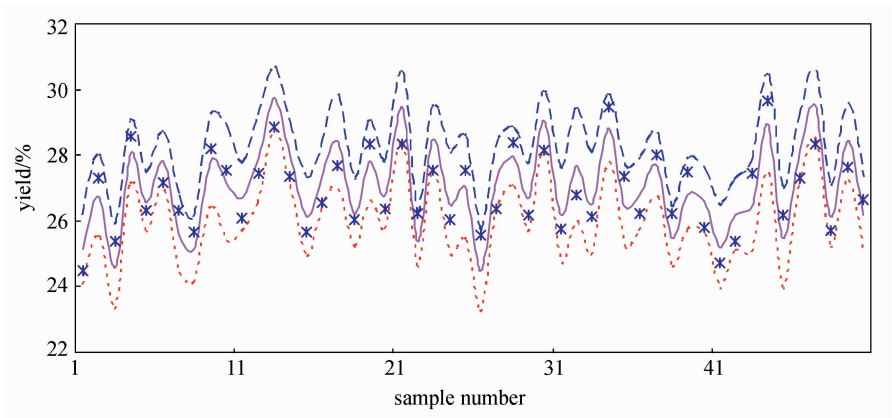

图 1　GP 软测量模型预测值、分析值和置信限度（95%置信概率）

Fig. 1　Curves of GP-PLS soft-sensor prediction, analysis and 95% confidence band

* analysis value; —— GP-PLS soft-sensor prediction value ($\hat{y}_i$); - - - $\hat{y}_i+2\sigma_i$; · · · · $\hat{y}_i-2\sigma_i$

即假设每个辅助变量对输出（萘酚油含萘量）预测的作用是一样的。用最大似然法优化 GP 软测量模型参数，得到与输入对应的模型参数如表 1 所示。从表 1 可以看出，w_1 数值最大，即初馏塔塔顶温度（Var1）与过程输出萘酚油含萘量关联最大，这一点与工艺知识是一致的。而 w_6 和 w_8 数值较小，即初炉出口温度（Var6）和初馏塔向精馏塔投料量（Var8）这两个变量对萘酚油含萘量预测作用不大，可以不作为辅助变量。若希望在初始化参数中体现过程的先验知识，则可以用不同的数值来初始化超参数，并且即使这样的参数有一定的差错，如给 w_6 赋予较大数值，而给 w_1 赋予较小数值，GP 模型最终仍然能够自动确定 w_1 较大的数值，w_6 较小的数值。但不合适的初始化参数会影响参数优化时间和算法的收敛性，因此建议在没有比较准确的先验知识时，超参数初始化时赋相同的数值。

在根据 w_i 的数值确定辅助变量时，采用主元贡献度与工艺分析相结合。主元贡献度是主元分析中确定主元数的一种方法。将 w_i 从大到小排列，若 $\sum_{i=1}^{p} w_i \Big/ \sum_{j=i}^{n} w_j \geqslant 90\%$，则 p 为辅助变量数目。最终确定辅助变量数为 6。

2.3　基于 GP 的萘酚油含萘量软测量模型的建立

当通过 ARD 确定了软测量模型辅助变量后，从训练和测试集合中删除不相关输入，得到新的训练和测试数据集合，再采用上述方法重新建立基于 GP 的萘酚油含萘量软测量模型，并同时得到预测模型方差。软测量模型预测值、分析值及 95%置信概率对应的 2 倍标准差置信区间曲线如图 1 所示，软测量模型预测标准差如图 2 所示。

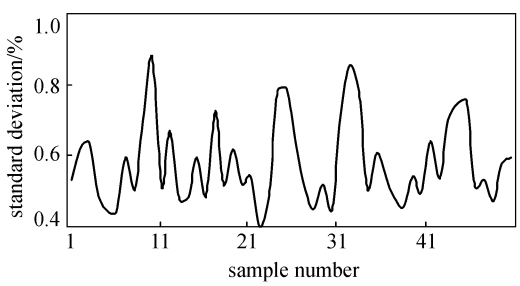

图 2　GP 软测量模型预测标准差

Fig. 2　Standard deviation of GP soft-sensor predictive model

采用预测均方根误差（RMSE）、打靶率（HR）和平均对数密度误差（LD）作模型性能比较标准。平均对数密度误差可以衡量模型精度与不确定性，该指标在以往的文献中几乎没有使用，这是因为那些方法很难给出软测量模型的不确定性。

RMSE 的定义为

$$\mathrm{RMSE}=\sqrt{\frac{1}{l}\sum_{i=1}^{l}(\hat{y}_i-y_i)^2} \tag{6}$$

HR 的定义为

$$\mathrm{HR}=\frac{r}{l}\times 100\% \tag{7}$$

LD 的定义为

$$\mathrm{LD}=\frac{1}{2l}\sum_{i=1}^{l}\left[\lg(2\pi)+\lg(\sigma_i^2)+\frac{(\hat{y}_i-y_i)^2}{\sigma_i^2}\right] \tag{8}$$

其中，$\hat{y}_i$为软测量模型预测输出值，y_i为酚油含萘量人工分析值，r 为软测量模型预测误差小于某一数值（本文为 0.5%）的数据个数，σ_i^2是高斯软测量模型方差。

在评价模型性能时，RMSE 越小，HR 越大，LD 越小，则模型的精度越高，在给定的置信概率下，模型的不确定度越小。对于测试样本，GP 软测量模型的性能参数分别为：RMSE＝0.691%，HR＝82%，LD＝1.18。从图 1 可以看出，测试样本全部处于由 $\hat{y}_i\pm 2\sigma_i$构成的置信区间内，这说明预测模型的精度较高，且不确定度较小，能够满足工业现场应用要求。从图 2 标准差曲线可以看出，总体上模型的标准差比较小。有部分测试样本的标准差较大，这主要是因为与该样本对应的操作区域附近训练数据较少。

2.4 与 BPNN 和 SVR 软测量模型比较

为了进一步比较 GP 软测量模型的性能，又针对该工业过程，采用前向神经网络（BPNN）和支持向量回归机（SVR）来建立软测量模型，训练和测试样本同上。BPNN 具有 1 个隐含层，节点数为 10，采用带动量因子的 BP 算法训练神经网络。为了防止 BPNN 模型过拟合，训练过程引入“early stoping”技术，即训练中一旦验证均方误差随训练误差的减小反而增大时，停止训练过程，并返回对应验证均方误差最小时的网络结构。ε 损失函数 SVR 采用序贯最小优化（SMO）算法，选用高斯核函数，核宽度为 0.47，正则化参数为 3.5，不灵敏度 ε＝0.0026。最终得到的软测量模型性能指标如表 2 所示，从表 2 可以看出，GP 模型性能与 SVR 和 BPNN 较接近，但 SVR 模型选择比较困难，BPNN 网络结构确定及训练过程比较复杂，训练时间长，且两者都不能给出估计的精度参数，此外，这两种方法都不能进行辅助变量的选择。因此，综合考虑，采用 GP 建立软测量模型仍然是较好的选择。

表 2 不同软测量模型性能比较

Table 2 Comparison of different soft-sensors

Model	RMSE		HR/%	
	Train	Test	Train	Test
GP	0.00574	0.00691	85	82
SVR	0.00536	0.00667	89	84
BPNN	0.00568	0.00681	86	82

3 结 论

将高斯过程用于复杂工业过程软测量建模，不仅能够有效地处理工业过程复杂的特性，而且能解决软测量模型辅助变量选择和模型预测不确定性等问题。针对工业萘初馏塔酚油含萘量软测量建模的应用表明，高斯过程在模型精度和预测不确定性等方面具有较好综合性能，是一种很好的新型软测量建模方法。由于软测量模型是基于工业过程历史数据，当操作过程偏离了建模数据对应的工况时，软测量模型的性能将显著下降，这在 GP 软测量模型预测方差上能得到明显反映，因此如何根据预测方差提供的信息，以及不可测变量化验分析值，进一步研究软测量模型的在线校正，或对模型输出进行校正，是值得进一步研究的。

References

[1] Brosillow C B. Inferential control of process. *AIChE J.*, 1978, **24** (3): 485-509

[2] Kresta J V, Martin T E, MacGregor J F. Development of inferential process models using PLS. *Computers and Chemical Engineering*, 1994, **18** (7): 597-611

[3] Wang Huazhong (王华忠), Yu Jinshou (俞金寿). Soft sensor modeling of acrylonitrile yield based on hybrid SVR-PLS approach. *Control and Decision* (控制与决策), 2005, **20** (5): 549-552

[4] Willis M J, Montague G A, Massimo D C, *et al*. Artificial neural networks in process estimation and control. *Automatica*, 1992, **28** (6): 1181-1187

[5] Liu Ruilan, Su Hongye, *et al*. Fuzzy neural network model of 4-CBA concentration for industrial PTA oxidation process. *Chinese Journal of Chemical Engineering*, 2004, **12** (2): 234-239

[6] Yan Xuefeng (颜学峰), Yu Juan (余娟), Qian Feng (钱锋). An evolution algorithm with select-best and prepotency operator and parameter estimation of 4-CBA model. *Journal of Chemical Engineering of Chinese Universities* (高校化学工程学报), 2005, **19** (2): 238-243

[7] Luo Jianxu (罗健旭), Shao Huihe (邵惠鹤). Developing dynamic soft sensors using multiple neural networks. *Journal of Chemical Industry and Engineering* (*China*) (化工学报), 2003, **54** (12): 1770-1773

[8] Ma Yong (马勇), Huang Dexian (黄德先), Jin Yihui (金以慧). Discussion about dynamic soft sensing modeling. *Journal of Chemical Industry and Engineering* (*China*) (化工学报), 2005, **56** (8): 1516-1519

[9] Fortuna L, Graziani S, Xibilia M G. Soft sensors for product quality monitoring in debutanizer distillation columns. *Control Engineering Practice*, 2005, **13** (4): 499-508

[10] Desai K, Badhe Y, Tamble S S, *et al*. Soft-sensor development for fed-batch bioreactors using support vector regression. *Biochemical Engineering Journal*, 2006, **27** (3): 225-239

[11] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning. Cambridge: MIT Press, 2006

[12] Bermak A, Belhouari S B. Bayesian learning using Gaussian process for gas identification. *IEEE Transactions on Instrumentation and Measurement*, 2006, **55** (3): 787-792

[13] Gregor G, Gordon L. Internal model control based on a Gaussian process prior model//Proceedings of the American Control Conference. Denver, Colorado, 2003: 4981-4986

[14] Xiong Zhihua (熊志化), Zhang Jicheng (张继承), Shao Huihe (邵惠鹤). GP-based soft sensor modeling. *Journal of System Simulation* (系统仿真学报), 2005, **17** (4): 793-794, 800

[15] Neal R M. Bayesian Learning for Neural Networks. New York: Springer-Verlag Press, 1996