不确定性环境下维纳模型的随机变分贝叶斯学习

刘 切 [1] 李俊豪 [1] 王 浩 [1] 曾建学 [1] 柴 毅 [1]

摘 要 多重不确定性环境下的非线性系统辨识是一个开放问题. 贝叶斯学习在描述、处理不确定性方面具有显著优势, 已在线性系统辨识方面得到广泛应用, 但在非线性系统辨识的应用较少, 且面临概率估计复杂、计算量大等难题. 针对上述问题, 以典型维纳 (Wiener) 非线性过程为对象, 提出基于随机变分贝叶斯的非线性系统辨识方法. 首先对过程噪声、测量噪声以及参数不确定性进行概率描述; 然后利用随机变分贝叶斯方法对模型参数进行后验估计. 在估计过程中, 利用随机优化思想, 仅利用部分中间变量概率信息估计模型参数分布的自然梯度期望, 与利用所有中间变量概率信息估计模型参数比较, 显著降低了计算复杂性. 该方法是首次在系统辨识领域中的应用. 最后, 利用一个仿真实例和一个维纳模型的 Benchmark 问题, 证明了该方法在对大规模数据下非线性系统辨识的有效性.

关键词 非线性系统辨识, 随机优化, 变分贝叶斯, 维纳模型

引用格式 刘切, 李俊豪, 王浩, 曾建学, 柴毅. 不确定性环境下维纳模型的随机变分贝叶斯学习. 自动化学报, 2024, **50**(6): 1185−1198

DOI 10.16383/j.aas.c210925

Stochastic Variational Bayesian Learning of Wiener Model in the Presence of Uncertainty

LIU Qie[1] LI Jun-Hao[1] WANG Hao[1] ZENG Jian-Xue[1] CHAI Yi[1]

Abstract Nonlinear system identification in multiple uncertain environment is an open problem. Bayesian learning has significant advantages in describing and dealing with uncertainties and has been widely used in linear system identification. However, the use of Bayesian learning for nonlinear system identification has not been well studied, confronted with the complexity of the estimation of the probability and the high computational cost. Motivated by these problems, this paper proposes a nonlinear system identification method based on stochastic variational Bayesian for Wiener model, a typical nonlinear model. First, the process noise, measurement noise and parameter uncertainty are described in terms of probability distribution. Then, the posterior estimation of model parameters is carried out by using the stochastic variational Bayesian approach. In this framework, only a few intermediate variables are used to estimate the natural gradient of the lower bound function of the likelihood function based on the stochastic optimization idea. Compared with classical variational Bayesian approach, where the estimation of model parameters depends on the information of all the intermediate variables, the computational complexity is significantly reduced for the proposed method since it only depends on the information of a few intermediate variables. To the best of our knowledge, it is the first time to use the stochastic variational Bayesian to system identification. A numerical example and a Benchmark problem of Wiener model are used to show the effectiveness of this method in the nonlinear system identification in the presence of large-scale data.

Key words Nonlinear system identification, stochastic optimization, variational Bayesian, Wiener model

Citation Liu Qie, Li Jun-Hao, Wang Hao, Zeng Jian-Xue, Chai Yi. Stochastic variational Bayesian learning of Wiener model in the presence of uncertainty. *Acta Automatica Sinica*, 2024, **50**(6): 1185−1198

系统辨识是基于模型的控制系统设计基础, 是现代控制理论主要研究内容. 系统辨识主要目标是利用数学方法从输入输出数据中建立系统的动态模型. 过去几十年中, 国内外研究人员围绕系统辨识的实验设计、算法设计以及收敛性证明等做了大量工作[1−3], 特别是对于线性系统的辨识, 已经有很多成熟的解决方案. 随着系统规模、结构的增加以及对高精准控制的需求, 传统利用线性模型近似描述非线性过程的方法已经不能满足人们对辨识精度的要求. 非线性系统辨识日益成为辨识主要研究方向[4−5]. 由于非线性模型的复杂性、多样性以及模型自身和数据的不确定性, 使得非线性系统辨识异常复杂,

收稿日期 2021-09-27 录用日期 2022-03-01
Manuscript received September 27, 2021; accepted March 1, 2022
国家重点研发计划 (2021YFB1715000), 国家自然科学基金 (61903051, U2034209) 资助
Supported by National Key Research and Development Program of China (2021YFB1715000) and National Natural Science Foundation of China (61903051, U2034209)
本文责任编委 孙秋野
Recommended by Associate Editor SUN Qiu-Ye
1. 重庆大学自动化学院 重庆 400044
1. School of Automation, Chongqing University, Chongqing 400044

成为一个开放性问题[6]. 本文针对非线性系统辨识过程中数据不确定、模型不确定的问题, 采用贝叶斯学习方法, 提出基于随机变分贝叶斯的一类非线性系统辨识方法, 在提高模型辨识精度情况下, 显著减少了算法的计算量, 为非线性系统辨识提供了一种全新的思路.

对于非线性系统的辨识, 首要问题是选择合适的非线性模型对系统的动态过程进行描述. 一般而言, 并不存在一种通用的非线性模型能够描述所有的非线性过程, 而过于复杂的模型会显著增加后续参数估计的复杂度, 因此选择一个合适的非线性模型来描述非线性过程则至关重要. 常见的用于描述非线性过程的模型包括: 非线性状态空间模型[7]、非线性自回归滑动平均模型[8] 和模块化结构模型 (Block-oriented model) 等[9−10]. 在这些模型中, 模块化结构模型具有简单易实现等优点, 其中包括维纳 (Wiener) 结构、Hammerstein 结构以及 Wiener-Hammerstein 结构等[11]. 维纳模型是其中的一类基础模型, 已经成功用于描述 pH 中和过程[12]、蒸馏塔[13] 和通信系统等过程[14], 一些文献表明, 它可以用于几乎任何非线性系统[15]. 因此, 本文选择这一基础模型, 研究非线性过程的辨识.

维纳模型的结构示意图如图 1 所示, 其由动态线性部分和静态非线性部分组成. 如前所述, 数据不确定性带来的噪声处理是系统辨识的永恒主题. 目前大部分的维纳过程辨识集中在系统测量噪声 (如图 1 中 e_n 所示) 的处理, 而忽略过程噪声 (如图 1 中 w_n 所示) 对辨识的影响. 在实际过程中, 中间变量 x_n 也可能受到噪声的干扰; 同时, 在模型中增加过程噪声, 可以提高描述的准确性. 针对维纳模型, Hagenblad 等[15] 指出, 当测量噪声和过程噪声都存在时, 一些现有的方法无法准确估计出模型参数; 另一方面, 现有的大部分维纳系统辨识中, 均利用高斯模型描述测量噪声. 实际上, 在测量过程中由于传感器异常等原因, 数据可能受到较大扰动, 产生异常数据 (奇异点). 此时, 高斯模型不能准确描述数据奇异现象. 数据奇异现象下的系统辨识, 近年来受到广泛关注[16], 但在维纳系统的辨识过程中考虑并不多, 本文在后续内容中将充分讨论这一问题.

在对维纳模型的研究方法中, 预测误差方法 (Prediction error method, PEM)[17−19] 使用最为广泛, 它通过最小化预测误差拟合输入输出数据, 从而得到系统模型. 该方法原理简单, 是系统辨识的标准方法, 但是在模型噪声较大且出现奇异值的情况下, 该方法很难得到满意的参数估计效果. 极大似然估计 (Maximization likelihood estimation, MLE)[15] 是另一种系统辨识经典方法, 它通过最大化似然函数获得参数的无偏估计, 是处理强噪声情况下参数估计的有效手段. 文献 [15] 提出维纳模型的极大似然估计方法. 利用传统的 MLE 方法进行非线性系统辨识, 由于需要直接计算似然函数, 大量的指数运算和积分运算使得辨识计算量较大; 而在具有隐变量不能直接计算似然函数的情况下, 传统的 MLE 也不能用于参数估计. 在 MLE 方法不能使用的情况下, EM (Expectation-maximization) 算法通过直接计算隐变量 (除观测值外, 所有参数都可以看作隐变量) 的后验分布来极大化全概率似然函数, 从而达到参数估计的目的. 然而, 由于模型中的非线性环节, 很难直接计算隐变量的后验分布, 使得 EM 算法不能直接用于维纳系统的辨识. 针对此问题, Liu 等在文献 [20] 中提出基于变分贝叶斯期望最大化 (Variational Bayesian expectation-maximization, VBEM) 的维纳模型辨识方法. 该方法利用变分推断结合重要性采样技术近似求解隐变量的后验分布, 然后通过极大化全概率似然函数估计模型参数. 该方法是非线性系统辨识领域的首次应用, 极大提高了含奇异数据和过程噪声情况下的维纳模型辨识精度. 然而, 由于使用重要性采样技术且需要对每个隐变量都进行变分推断, 使得该方法计算量大, 不适合大规模数据下的系统辨识.

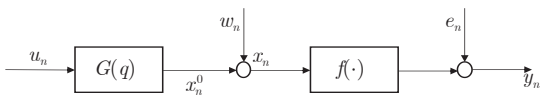


图 1 维纳模型结构示意图

Fig. 1 The structure of Wiener model

基于上述方法的局限性, 本文提出随机变分贝叶斯推断 (Stochastic variational Bayesian inference, SVBI) 的方法来解决存在过程噪声、奇异点及参数不确定情况下的维纳模型辨识问题. 区别于文献 [20] 提出的 VBEM 算法, 本文利用随机优化思想, 采用自然梯度下降的方法对模型参数进行更新. 因为使用随机梯度下降方法, 在迭代中只要知道梯度期望值即可保证梯度下降的收敛性, 因此, 在隐变量独立的假设下, 只需要部分隐变量信息即可对模型信息进行更新, 从而显著降低变分推理的计算量. 据作者所知, 这是本方法在系统辨识领域的首次应用. 本文认为该方法是基于贝叶斯学习的系统辨识的重要进展, 为非线性系统辨识提供了一个新的思路. 本文通过一个仿真实例详细分析该方法在辨识准确率和计算时间成本上的优势; 通过一个非线性电路辨识的 Benchmark 问题验证该方法在实际数据上的应用, 结果表明本文提出的方法在提高辨识准确度和计算效率方面均具有较大优势.

1　问题描述

1.1　系统模型

本文考虑的维纳模型如图 1 所示, 其结构如式 (1) 所示. 其中, u_n 为系统输入变量, y_n 为系统输出变量并受到测量噪声 e_n 的干扰, x_n 为中间不可测变量, 并受到过程噪声 w_n 的干扰, $f(x_n)$ 为系统的非线性部分.

$$\begin{cases} x_n^0 = G(q)u_n \\ x_n = x_n^0 + w_n \\ y_n = f(x_n) + e_n \end{cases} \tag{1}$$

其中, $G(q)$ 为输入传递函数, 表示为

$$G(q) = \frac{b_0 + b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b}}{1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}} \tag{2}$$

其中, n_a 和n_b 分别代表模型输出和输入的阶数, a_i, b_j $(i = 0, \cdots, n_a, j = 0, \cdots, n_b)$ 为系数.

为方便后面利用贝叶斯学习进行辨识, 首先将输入传递函数转换成一个有限脉冲响应模型 (Finity impluse response, FIR), 即

$$x_n^0 = (\theta_0 + \theta_1 z^{-1} + \cdots + \theta_L z^{-L})u_n \tag{3}$$

其中, z表示移位算子, L为 FIR 模型的阶数, 当 FIR 模型中的阶数足够高时, 其可以用来准确近似传递函数模型. FIR模型中的参数可以通过下式得到[21]:

$$\begin{cases} \theta_0 = b_0 \\ \theta_j = b_j - \sum\limits_{l=0}^{j-1} a_{j-l}\theta_l, \ j = 1, 2, \cdots, L \end{cases} \tag{4}$$

基于这一变换, 可以使用式 (3) 来描述系统的输入动态过程, 从而将对 $G(q)$ 的辨识转换为对参数 $\boldsymbol{\Theta} = [\theta_0, \theta_1, \theta_2, \cdots, \theta_L]^{\mathrm{T}}$ 的辨识.

典型地, 对于系统的非线性动态环节, 用一组非线性基函数的线性组合来表示

$$f(x_n) = \sum_{i=0}^{M} \lambda_i f_i(x_n) \tag{5}$$

其中, M 表示非线性基函数的数量, $f_i(\cdot)$ 是非线性基函数. 在给定非线性基函数条件下对系统非线性动态环节的辨识可以转换为对各非线性基函数系数 $\boldsymbol{\Lambda} = [\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_M]^{\mathrm{T}}$ 的辨识.

考虑到维纳模型中的测量噪声以及过程噪声, 假设两种噪声相互独立. 将过程噪声 w_n 假设为均值为 0、精度为 δ_w 的高斯分布; 为准确描述测量数据中的奇异值, 将测量噪声 e_n 假设为均值为 0、精度为 δ_e、自由度为 v 的学生 t 分布[22], 即

$$\begin{cases} p(w_n|\delta_w) = \mathrm{N}(w_n|0, \delta_w^{-1}) \\ p(e_n|\delta_e, v) = st(e_n|0, \delta_e^{-1}, v) \end{cases} \tag{6}$$

其中, st 表示学生 t 分布, 其可以表示为[23]

$$p(e_n|\delta_e, v) = \int \mathrm{N}\left(0, \frac{1}{\lambda_e r_n}\right) g(r_n|v)\mathrm{d}r_n \tag{7}$$

其中, r_n 为缩放比例因子并且服从伽马分布, 即

$$g(r_n|v) = \mathcal{G}\left(r_n|\frac{v}{2}, \frac{v}{2}\right) \tag{8}$$

其中, $\mathcal{G}$ 表示伽马分布, 假设参数 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Lambda}$ 的先验分布服从高斯–伽马分布, 参数 δ_w 和 δ_e 的先验分布服从伽马分布, 即

$$\begin{cases} p(\boldsymbol{\Theta}|\alpha) = \mathrm{N}(\boldsymbol{\Theta}|0, \alpha^{-1}\boldsymbol{I})\mathcal{G}(\alpha|a_0, b_0) \\ p(\boldsymbol{\Lambda}|\alpha) = \mathrm{N}(\boldsymbol{\Lambda}|0, \alpha^{-1}\boldsymbol{I})\mathcal{G}(\alpha|a_0, b_0) \\ p(\delta_w|a_0, b_0) = \mathcal{G}(\delta_w|a_0, b_0) \\ p(\delta_e|a_0 b_0) = \mathcal{G}(\delta_e|a_0, b_0) \end{cases} \tag{9}$$

其中, α^{-1} 为参数 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Lambda}$ 的协方差对角线元素; $\boldsymbol{I}$ 是与参数 $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Lambda}$ 具有相同维度的单位矩阵; a_0 和 b_0 表示系统的超参数, 为常量.

由于参数的相互独立性, 其联合先验分布可以表示为

$$\begin{aligned} &p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \delta_w, \delta_e, \alpha) = \\ &\quad p(\boldsymbol{\Theta}|\alpha)p(\boldsymbol{\Lambda}|\alpha)p(\delta_w|a_0, b_0)p(\delta_e|a_0 b_0) \end{aligned} \tag{10}$$

令 $u_{1:N} = \{u_1, u_2, \cdots, u_N\}$ 为系统输入数据, $y_{1:N} = \{y_1, y_2, \cdots, y_N\}$ 为系统观测数据. 本文利用 MLE 方法进行系统辨识, 对参数的估计转化为如下的优化问题:

$$\left(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Lambda}}, \hat{\delta}_w, \hat{\delta}_e, \hat{\alpha}, \hat{v}\right) = \arg \max_{\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \delta_w, \delta_e, \alpha, v} p(y_{1:N}) \tag{11}$$

其中, $\boldsymbol{\Theta}$ 为动态线性环节的参数, $\boldsymbol{\Lambda}$ 为静态非线性环节的参数, δ_w 为过程噪声 w_n 的精度, δ_e 为过程噪声 e_n 的精度, v 为过程噪声 e_n 的自由度, 且所有参数均属于不同形式的指数族分布 (式 (8) 和式 (9)). 由于系统的非线性以及隐变量的引入, 很难直接计算观测数据的似然函数, 鉴于此, 本文利用变分贝叶斯方法求解上述优化问题.

1.2　VBEM 方法

VBEM 方法用于对系统参数和隐变量的真实后验分布进行估计, 得到变分后验分布. 令 $x_{1:N} = \{x_1, x_2, \cdots, x_N\}$, $r_{1:N} = \{r_1, r_2, \cdots, r_N\}$. 在本文所考虑到的维纳模型中, 记 $\mathcal{Y} = \{y_{1:N}\}$ 为观测数据, $\mathcal{X} = \{x_{1:N}, r_{1:N}\}$ 为局部隐变量, $\mathcal{Z} = \{\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \delta_w, \delta_e, \alpha\}$ 为全局隐变量, $\mathcal{V} = \{v\}$ 为结构参数集. 在给

定结构参数的情况下, 得到观测数据的概率密度函数 $p(\mathcal{Y}|\mathcal{V})$, 根据贝叶斯的规则, 有

$$p(\mathcal{Y}|\mathcal{V}) = \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}|\mathcal{V})}{p(\mathcal{X}, \mathcal{Z}|\mathcal{Y}, \mathcal{V})}$$

其似然函数的对数形式为

$$\ln p(\mathcal{Y}|\mathcal{V}) = \ln \int q(\mathcal{X}, \mathcal{Z}|\mathcal{V}) p(\mathcal{Y}|\mathcal{V}) \mathrm{d}\mathcal{X}\mathrm{d}\mathcal{Z} = \int q(\mathcal{X}, \mathcal{Z}|\mathcal{V}) \ln p(\mathcal{Y}|\mathcal{V}) \mathrm{d}\mathcal{X}\mathrm{d}\mathcal{Z} = \mathcal{L}(q) + KL(q||p)$$

其中, $q(\mathcal{X}, \mathcal{Z}|\mathcal{V})$ 为任意概率密度函数, $\mathcal{L}(q)$ 为 $\ln p(\mathcal{Y}|\mathcal{V})$ 的下界函数, 表示为

$$\mathcal{L}(q) = \int q(\mathcal{X}, \mathcal{Z}|\mathcal{V}) \ln \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}|\mathcal{V})}{q(\mathcal{X}, \mathcal{Z}|\mathcal{V})} \mathrm{d}\mathcal{X}\mathrm{d}\mathcal{Z}$$

$KL(q||p)$ 为 $q(\mathcal{X}, \mathcal{Z}|\mathcal{V})$ 和 $p(\mathcal{X}, \mathcal{Z}|\mathcal{Y}, \mathcal{V})$ 之间的 KL (Kullback-Leibler) 散度, 表示为

$$KL(q||p) = \int q(\mathcal{X}, \mathcal{Z}|\mathcal{V}) \ln \frac{q(\mathcal{X}, \mathcal{Z}|\mathcal{V})}{p(\mathcal{X}, \mathcal{Z}|\mathcal{Y}, \mathcal{V})} \mathrm{d}\mathcal{X}\mathrm{d}\mathcal{Z}$$

已知 $KL(q||p) \geq 0$, 且仅当 $q(\mathcal{X}, \mathcal{Z}|\mathcal{V})$ 等于真实后验分布 $p(\mathcal{X}, \mathcal{Z}|\mathcal{Y}, \mathcal{V})$ 时, 有 $KL(q||p) = 0$, 此时变分下界 $\mathcal{L}(q)$ 有最大值. 因此, VBEM 方法通过使用变分方法极大化下界函数 $\mathcal{L}(q)$ 来求解后验分布 $q(\mathcal{X}, \mathcal{Z}|\mathcal{V})$. 然而, 在非线性情况下, 由于真实后验分布往往是不可解析且不能直接计算, VBEM 方法通过迭代更新让变分后验分布 $q(\mathcal{X}, \mathcal{Z}|\mathcal{V})$ 逐渐接近于真实后验分布, 达到极大化下界函数 $\mathcal{L}(q)$ 的目的. 根据平均场理论, 在各隐变量互相独立的基础上, 有

$$q(\mathcal{X}, \mathcal{Z}|\mathcal{V}) = q(\mathcal{X}|\mathcal{V}) q(\mathcal{Z}|\mathcal{V}) \tag{12}$$

VBEM 算法是在贝叶斯定理的基础上, 对 EM 算法的重要发展, 与 EM 算法类似, 也包括 E 步和 M 步[20]. 对于第 k 次迭代, VB-E 步, 通过固定第 $k-1$ 次迭代得到的结构参数集 $\mathcal{V}^{k-1} = \{v\}$, 对隐变量 $\mathcal{X} = \{x_{1:N}, r_{1:N}\}$ 和 $\mathcal{Z} = \{\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \delta_w, \delta_e, \alpha\}$ 的后验分布进行更新, 最大化下界函数 $\mathcal{L}(q)$; VB-M 步, 基于 E 步中估计得到的隐变量, 对第 k 次迭代的结构参数集进行更新, 最大化下界函数 $\mathcal{L}(q)$. E 步和 M 步不断迭代直至算法收敛, 此时获得的隐变量的变分后验分布 $q(\mathcal{X}, \mathcal{Z}|\mathcal{V})$ 可以近似等效为其真实后验分布 $p(\mathcal{X}, \mathcal{Z}|\mathcal{Y}, \mathcal{V})$.

具体地, 对于第 k 次迭代, 在 VB-E 步骤中, 根据变分方法[17], 针对某一特定变量 (其他变量固定), 下界 $\mathcal{L}(q)$ 在如下更新方式下取得极大值:

通过关于 $q(\mathcal{X})$ 和 $q(\mathcal{Z})$ 的完全分解并逐步更新实现最大化, 隐变量的变分后验分布的更新形式为

$$q(\mathcal{X}_{j,n}) = \frac{\exp\left[\mathrm{E}_{q(\mathcal{Z})q(\mathcal{X}_{-j})q(\mathcal{X}_{j,-n})} \ln p(\cdot)\right]}{\int \exp\left[\mathrm{E}_{q(\mathcal{Z})q(\mathcal{X}_{-j})q(\mathcal{X}_{j,-n})} \ln p(\cdot)\right] \mathrm{d}\mathcal{X}_{j,n}} \tag{13}$$

$$q(\mathcal{Z}_m) = \frac{\exp\left[\mathrm{E}_{q(\mathcal{X})q(\mathcal{Z}_{-m})} \ln p(\cdot)\right]}{\int \exp\left[\mathrm{E}_{q(\mathcal{X})q(\mathcal{Z}_{-m})} \ln p(\cdot)\right] \mathrm{d}\mathcal{Z}_m} \tag{14}$$

其中, $\mathcal{X}_j$ 表示第 j 类局部隐变量, $j \in [1, 2]$, $\mathcal{X}_{-j} = \{\mathcal{X} \setminus \mathcal{X}_j\}$, $\mathcal{X}_{j,-n} = \{\mathcal{X}_j \setminus \mathcal{X}_{j,n}\}$, $\mathcal{Z}_m$ 表示第 m 个全局隐变量, $\mathcal{Z}_{-m} = \{\mathcal{Z} \setminus \mathcal{Z}_m\}$, $\mathrm{E}_{q(\cdot)}(\cdot)$ 表示关于 $q(\cdot)$ 的期望运算, $\ln p(\cdot)$ 表示 $\ln p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}|\mathcal{V}^{k-1})$.

在 VB-M 步骤中, 下界 $\mathcal{L}(q)$ 表示为

$$\int q(\mathcal{X}, \mathcal{Z}|\mathcal{V}^{k-1}) \ln \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}|\mathcal{V}^k)}{q(\mathcal{X}, \mathcal{Z}|\mathcal{V}^{k-1})} \mathrm{d}\mathcal{X}\mathrm{d}\mathcal{Z}$$

其中, $q(\mathcal{X}, \mathcal{Z}|\mathcal{V}^{k-1})$ 表示 VB-E 步中获得的变分后验分布, 此时下界 $\mathcal{L}(q)$ 通过调整 $\mathcal{V}^k$ 来最大化, 即

$$\mathcal{V}^k = \arg\max_{\mathcal{V}^k} \mathcal{L}(q)$$

下界 $\mathcal{L}(q)$ 中的联合概率分布为 $p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}|\mathcal{V})$, 在后文推理中和迭代中需要多次使用, 根据链式法则, 对于本文所考虑到的维纳模型, 其全概率似然函数可以表示为

$$p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}|\mathcal{V}) = p(y_{1:N}, x_{1:N}, r_{1:N}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \delta_w, \delta_e, \alpha|v) = \prod_{n=1}^{N} p(y_n|x_n, r_n, \boldsymbol{\Lambda}, \delta_e, v) \prod_{n=1}^{N} p(x_n|\boldsymbol{\Theta}, \delta_w) \times \prod_{n=1}^{N} p(r_n|v) p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}|\alpha) p(\delta_w) p(\delta_e) p(\alpha)$$

上述方法中, 在对模型参数 (全局隐变量) 的后验分布进行更新时, 需要所有局部变量的后验分布信息 (式 (14)). 由于非线性的存在, 对局部隐变量的后验分布更新需要采用随机采样方法, 这使得对全局变量的更新需要很大的计算量, 限制了该方法在大规模样本下的应用. 鉴于此, 本文在下文中提出使用随机变分推理方法对模型参数进行更新, 将显著降低辨识的计算量.

2 维纳模型的随机变分贝叶斯学习

2.1 随机变分贝叶斯推理

如前所述, 如果利用传统 VBEM 算法对提出的模型进行参数估计, 面临计算量大的问题. 针对此问题, 本文提出利用随机优化的方法求解提出的辨识问题. 为此, 首先重新改写变分推理中的下界函数 (目标函数) $\mathcal{L}(q)$, 即

$$\mathcal{L}(q)=\int q(\mathcal{X},\mathcal{Z}|\mathcal{V})\ln\frac{p(\mathcal{Z}|\mathcal{Y},\mathcal{X},\mathcal{V})}{q(\mathcal{X},\mathcal{Z}|\mathcal{V})}\mathrm{d}\mathcal{X}\mathrm{d}\mathcal{Z}+const$$

其中, $const$ 表示常数项. 易知 $\mathcal{L}(q)$ 在 $q(\mathcal{Z}|\mathcal{V})=p(\mathcal{Z}|\mathcal{Y},\mathcal{X},\mathcal{V})$ 时取得最大值, 因此 $q(\mathcal{Z}|\mathcal{V})$ 与全条件分布 $p(\mathcal{Z}|\mathcal{Y},\mathcal{X},\mathcal{V})$ 具有相同的分布形式.

考虑全局隐变量集 $\mathcal{Z}=\{\boldsymbol{\Theta},\boldsymbol{\Lambda},\delta_w,\delta_e,\alpha\}$, 前文所描述的各全局隐变量的先验分布均为指数族分布, 不失一般性, 可以表示为

$$p(\mathcal{Z}_m)=h(\mathcal{Z}_m)\exp\left\{\Omega^{\mathrm{T}}t(\mathcal{Z}_m)-a_g(\Omega)\right\}$$

其中, $\mathcal{Z}_m$ 表示第 m 个全局隐变量, $h(\cdot)$ 为基础度量值, Ω 为自然参数, $a_g(\cdot)$ 为对数归一化因子, $t(\cdot)$ 为充分统计量[23].

根据过程噪声、测量噪声的概率分布信息 (均为指数族分布), 可以得到

$$\begin{aligned}p(\mathcal{Y},\mathcal{X}|\mathcal{Z}_m)&=p(\mathcal{Y}|\mathcal{X},\mathcal{Z}_m)p(\mathcal{X}|\mathcal{Z}_m)=\\&h(\mathcal{Y},\mathcal{X})\exp\left\{\mathcal{Z}_m^{\mathrm{T}}t(\mathcal{Y},\mathcal{X})-Na_l(\mathcal{Z}_m)\right\}=\\&h(\mathcal{Y},\mathcal{X})\exp\left\{[t(\mathcal{Y},\mathcal{X}),N][\mathcal{Z}_m^{\mathrm{T}},-a_l(\mathcal{Z}_m)]^{\mathrm{T}}\right\}\end{aligned}$$

通过选择合适的先验分布, 在似然函数与先验分布均为指数族分布形式情况下, 可以使全条件分布与先验分布共轭, 即

$$\begin{aligned}p(\mathcal{Z}_m|\mathcal{Y},\mathcal{X},\Omega)=&\\&h(\mathcal{Z}_m)\exp\Big\{\eta_g^{\mathrm{T}}(\mathcal{Y},\mathcal{X},\Omega)t(\mathcal{Z}_m)-\\&a_g\big(\eta_g(\mathcal{Y},\mathcal{X},\Omega)\big)\Big\}\end{aligned}$$

其中

$$\begin{aligned}\eta_g^{\mathrm{T}}(\mathcal{Y},\mathcal{X},\Omega)&=\Omega^{\mathrm{T}}+[t(\mathcal{Y},\mathcal{X}),N]=\\&\Omega^{\mathrm{T}}+\sum_{i=1}^{N}\Big[t(\mathcal{Y}_i,\mathcal{X}_i),1\Big]\end{aligned}$$

根据此结论, 各隐变量后验分布的形式均已知, 将对隐变量后验分布的推理转换成后验分布参数的推理. 在各隐变量相互独立的假设基础上, 各隐变量的变分后验分布由相应的参数来控制, 于是式 (12) 写为

$$\begin{aligned}q(\mathcal{X},\mathcal{Z}|\mathcal{V})=&\\&\prod_{j=1}^{2}\prod_{n=1}^{N}q(\mathcal{X}_{j,n}|\phi_{j,n},\mathcal{V})\prod_{m=1}^{M}q(\mathcal{Z}_m|\beta_m,\mathcal{V})\end{aligned}$$

其中, 局部变分参数 $\phi_{j,n}$ 控制第 j 类的第 n 个局部隐变量的分布, 全局变分参数 β_m 控制第 m 个全局隐变量的分布. 在上述假设下, $q(\mathcal{Z}_m|\beta_m,\mathcal{V})$ 具有如下形式:

$$\begin{aligned}q(\mathcal{Z}_m|\beta_m,\mathcal{V})=&\\&h(\mathcal{Z}_m)\exp\left\{\beta_m^{\mathrm{T}}t(\mathcal{Z}_m)-a_g(\beta_m)\right\}=\\&h(\mathcal{Z}_m)\exp\Big\{\eta_g^{\mathrm{T}}(\mathcal{Y},\mathcal{X},\Omega)t(\mathcal{Z}_m)-\\&a_g\big(\eta_g(\mathcal{Y},\mathcal{X},\Omega)\big)\Big\}\end{aligned}\tag{15}$$

则下界函数 $\mathcal{L}(q)$ 关于 β_m 的表达式为

$$\begin{aligned}\mathcal{L}(\beta_m)=&\mathrm{E}_q\left[\ln p(\mathcal{Z}_m|\mathcal{Y},\mathcal{X},\Omega)\right]-\\&\mathrm{E}_q\left[q(\mathcal{Z}_m|\beta_m,\mathcal{V})\right]+const=\\&\mathrm{E}_q^{\mathrm{T}}[\eta_g(\mathcal{Y},\mathcal{X},\Omega)]\nabla_{\beta_m}a_g(\beta_m)-\\&\beta_m^{\mathrm{T}}\nabla_{\beta_m}a_g(\beta_m)+a_g(\beta_m)+const\end{aligned}\tag{16}$$

其中, $const$ 包含其他与 $q(\mathcal{Z}_m|\beta_m,\mathcal{V})$ 无关的项.

至此, 本文将下界函数写成关于隐变量后验分布参数的表达式. 此时, 可以通过更新参数 β_m 求解后验分布. 为减少更新过程的计算量, 本文采用随机优化方法对 $\mathcal{L}(\beta_m)$ 进行更新.

对于下界函数 $\mathcal{L}(\beta_m)$, 由于参数的变化而引起的概率分布的变化量并不适合用欧氏距离来表征, 使用经典梯度下降方法收敛速度很慢. 针对上述问题, 本文使用 $\mathcal{L}(\beta_m)$ 的自然梯度信息对其进行最大化. $\mathcal{L}(\beta_m)$ 的自然梯度定义为[24−25]

$$\hat{\nabla}_{\beta_m}\mathcal{L}(\beta_m)=G^{-1}(\beta_m)\nabla_{\beta_m}\mathcal{L}(\beta_m)\tag{17}$$

其中, $G(\beta_m)$ 为 $q(\mathcal{Z}_m)$ 的 Fisher 信息矩阵, 定义为[25]

$$\begin{aligned}G(\beta_m)=&\mathrm{E}_q\Big[\big(\nabla_{\beta_m}\ln q(\mathcal{Z}_m|\beta_m,\mathcal{V})\big)\times\\&\big(\nabla_{\beta_m}\ln q(\mathcal{Z}_m|\beta_m,\mathcal{V})\big)^{\mathrm{T}}\Big]=\\&\mathrm{E}_q\Big[\big(t(\beta_m)-\mathrm{E}_q\left[t(\beta_m)\right]\big)\times\\&\big(t(\beta_m)-\mathrm{E}_q\left[t(\beta_m)\right]\big)^{\mathrm{T}}\Big]=\\&\nabla_{\beta_m}^2a_g(\beta_m)\end{aligned}$$

$\nabla_{\beta_m}\mathcal{L}(\beta_m)$ 为 $\mathcal{L}(\beta_m)$ 关于 β_m 的经典梯度, 根据式 (16), 计算为[23]

$$\begin{aligned}\nabla_{\beta_m}\mathcal{L}(\beta_m)=&\\&\nabla_{\beta_m}^2a_g(\beta_m)(\mathrm{E}_q\left[\eta_g^{\mathrm{T}}(\mathcal{Y},\mathcal{X},\Omega)\right]-\beta_m)\end{aligned}$$

于是, 下界函数 $\mathcal{L}(\beta_m)$ 关于 β_m 自然梯度为

$$\begin{aligned}\hat{\nabla}_{\beta_m}\mathcal{L}(\beta_m)=G^{-1}(\beta_m)\nabla_{\beta_m}\mathcal{L}(\beta_m)=&\\\mathrm{E}_q\left[\eta_g(\mathcal{Y},\mathcal{X},\Omega)\right]-\beta_m=&\\\left[\Omega^{\mathrm{T}}+\sum_{i=1}^{N}\Big[\mathrm{E}_q[t(\mathcal{Y}_i,\mathcal{X}_i)],1\Big]\right]^{\mathrm{T}}-\beta_m&\end{aligned}\tag{18}$$

根据梯度下降法, 在第 k 次迭代时, 全局变分参数

β_m 的更新为[26]

$$\beta_m^{(k)} = \beta_m^{(k-1)} + \rho_k \hat{\nabla}_{\beta_m} \mathcal{L}(\beta_m^{(k-1)}) \tag{19}$$

其中, ρ_k 为步长, 满足 $\sum \rho_k = \infty$, $\sum \rho_k^2 < \infty$ [26]. 为提高算法的速度, 本文取 $\rho_k = (k+\tau)^{-\gamma} \leq 1$, k 表示第 k 个迭代时刻, γ 表示遗忘率, 用来控制旧信息遗忘的速率, 延迟因子为 $\tau \geq 0$.

在本文的全局参数变分推理中, 下界函数的自然梯度包含一个多项式求和 (如式 (18) 所示), 该多项式每一项与一个数据点对应. 全局参数真实梯度信息的计算量与采样点数量成正比, 那么在大规模数据情况下, 其自然梯度的计算量将显著增加. 针对上述问题, 随机优化的思想是利用部分数据点信息计算梯度的期望, 而不直接计算真实梯度, 从而显著降低计算复杂度. 根据随机优化的思想[27], 在利用式 (19) 最大化 $\mathcal{L}(\beta_m)$ 时, 其收敛条件为

$$\hat{\nabla}_{\beta_m} \mathcal{L}(\beta_m) = \mathrm{E}_q \left[\hat{\nabla}_{\beta_m} \bar{\mathcal{L}}(\beta_m) \right] \tag{20}$$

换言之, 只要计算下界函数自然梯度的期望即可完成参数的更新. 针对式 (3) 描述的维纳过程, 假设样本之间互相独立, 且数据样本服从 $1 \sim N$ 上的均匀分布[28]. 根据式 (18), $\mathcal{L}(\beta_m)$ 关于 β_m 的自然梯度期望为

$$\mathrm{E}_q \left[\hat{\nabla}_{\beta_m} \bar{\mathcal{L}}(\beta_m^{(k-1)}) \right] = \Omega + \frac{N}{Z} \sum_{I_z=1}^{Z} \left[\mathrm{E}_q[t(\mathcal{Y}_{I_z}, \mathcal{X}_{I_z})], 1 \right]^{\mathrm{T}} - \beta_m^{(k-1)} \tag{21}$$

其中, $I_z \sim \mathrm{U}(1, N)$, 通过均匀采样得到, Z 表示采样次数. 在此情况下, 根据文献 [26], 全局变分参数 β_m 的更新方式为

$$\beta_m^{(k)} = \beta_m^{(k-1)} + \rho_k \mathrm{E}_q \left[\hat{\nabla}_{\beta_m} \bar{\mathcal{L}}(\beta_m^{(k-1)}) \right] \tag{22}$$

从式 (22) 中可以看出, 利用随机优化方法, 全局变分参数的更新只与若干个局部变量的信息有关, 极端情况下, 当 $Z = 1$ 时, 全局变分参数的更新的计算量是原来的 $1/N$, 将极大地降低全局隐变量更新的计算量. 该方法有效地克服了非线性情况下全局隐变量后验分布更新计算量大的缺点. 第 2.3 节将介绍利用该方法更新维纳模型的模型参数.

2.2 收敛性分析

根据 SVBI 的思路, 在第 k 次迭代时刻, 首先固定 $\mathcal{V}^{k-1}$, 依次更新各隐变量的变分后验分布 q^{k-1}, 从而最大化下界函数 $\mathcal{L}(q^{k-1}, \mathcal{V}^{k-1})$. 在最大化过程中各个隐变量的后验分布互相独立, 因此对于下界函数 $\mathcal{L}(q^{k-1}, \mathcal{V}^{k-1})$ 的最大化等价于最小化 $KL(q^{k-1}, \mathcal{V}^{k-1})$. 而又已知 $KL(\cdot) \geq 0$, 当 $KL(\cdot)$ 最小化时, 有

$$q^k = \arg\max_{q^{k-1}} \mathcal{L}(q^{k-1}, \mathcal{V}^{k-1})$$

$$KL(q^k, \mathcal{V}^{k-1}) \leq KL(q^{k-1}, \mathcal{V}^{k-1})$$

在这一步骤中, 通过固定 $\mathcal{V}^{k-1}$ 来更新 $q^k(\mathcal{Z}_m)$, 图 2 为迭代更新示意图, 注意到每更新一个隐变量之后会使得下界函数 $\mathcal{L}(q)$ 增大, 即

$$\mathcal{L}(q^k(\mathcal{Z}_2), \mathcal{V}^{k-1}) > \mathcal{L}(q^k(\mathcal{Z}_1), \mathcal{V}^{k-1})$$

$$\mathcal{L}(q^k(\mathcal{Z}_m), \mathcal{V}^{k-1}) > \mathcal{L}(q^k(\mathcal{Z}_{m-1}), \mathcal{V}^{k-1})$$

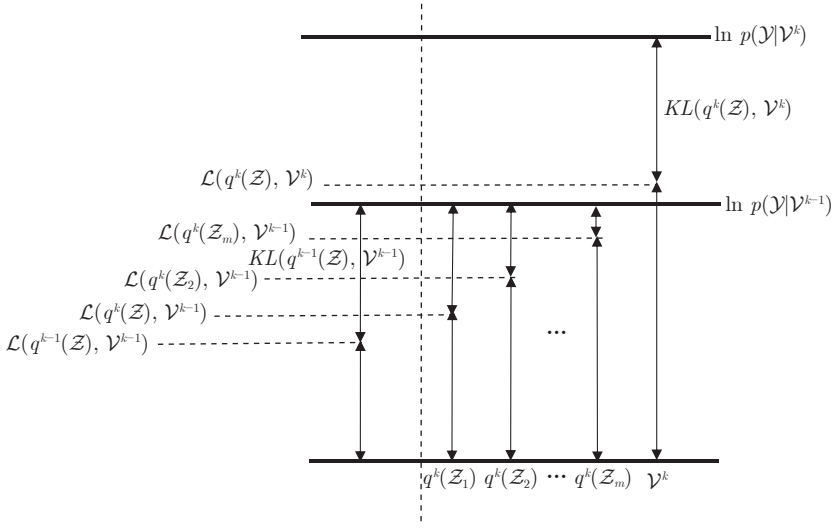


图 2 SVBI 中目标函数的更新示意图

Fig. 2 The update process of the objective function in SVBI

在下一步骤中, 固定前一步骤更新所得到的各个隐变量对应的变分后验分布 q^k, 更新结构参数 $\mathcal{V}^{k-1}$, 从而继续最大化下界函数 $\mathcal{L}(q^k, \mathcal{V}^{k-1})$. 此时如果 $\mathcal{V}^{k-1}$ 发生变换, 对数似然函数 $\ln p(\mathcal{Y}|\mathcal{V}^{k-1})$ 将会发生变化, 且有 $KL(q^k, \mathcal{V}^k) \geq KL(q^k, \mathcal{V}^{k-1})$, 这一过程可以描述为

$$\mathcal{V}^k = \arg\max_{\mathcal{V}^{k-1}} \mathcal{L}(q^k, \mathcal{V}^{k-1})$$

$$KL(q^k, \mathcal{V}^k) \geq KL(q^k, \mathcal{V}^{k-1})$$

图 2 同样描述了这一过程. 在 SVBI 框架中, 通过最大化似然函数下界 $\mathcal{L}(q^{k-1}, \mathcal{V}^{k-1})$, 实现似然函数的最大化. 在每一次迭代中, 下界函数 $\mathcal{L}(q^{k-1}, \mathcal{V}^{k-1})$ 通过两个步骤进行最大化. 在步骤 1 中, 通过更新 q^{k-1} 来最大化下界 $\mathcal{L}(q^{k-1}, \mathcal{V}^{k-1})$, 此时似然函数不会发生改变; 在步骤 2 中新的下界函数 $\mathcal{L}(q^k, \mathcal{V}^{k-1})$ 通过 $\mathcal{V}^{k-1}$ 的更新而最大化, 在这一步中似然函数增加, 可以证明 $\ln p(\mathcal{Y}|\mathcal{V}^k) \geq \ln p(\mathcal{Y}|\mathcal{V}^{k-1})$ (如图 2 所示). 每次迭代都基于以上两个步骤, 保证下界逐渐增加, 直至收敛于 $\ln p(\mathcal{Y})$.

2.3　基于 SVBI 的模型参数辨识

2.3.1　$q(\boldsymbol{\Theta})$ 和 $q(\boldsymbol{\Lambda})$ 更新

根据前文所述, 变分后验分布 $q(\boldsymbol{\Theta})$ 的形式为

$$q(\boldsymbol{\Theta}) \propto p(\mathcal{Y}, \mathcal{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \propto \prod_{n=1}^{N} p(x_n|\boldsymbol{\Theta}, \delta_w)p(\boldsymbol{\Theta}|\alpha)$$

其中,

$$p(x_n|\boldsymbol{\Theta}, \delta_w) = \frac{\sqrt{\delta_w}}{\sqrt{2\pi}} \exp\left\{-\frac{\delta_w}{2}(x_n - x_n^0)^2\right\} = \frac{\sqrt{\delta_w}}{\sqrt{2\pi}} \exp\left\{\left[\delta_w x_n \boldsymbol{U}_n^{\mathrm{T}}, -\frac{1}{2}\delta_w \varphi^{\mathrm{T}}(\boldsymbol{U}_n)\right] \times \begin{bmatrix}\boldsymbol{\Theta} \\ \varphi(\boldsymbol{\Theta})\end{bmatrix} - \frac{1}{2}\delta_w x_n^2\right\}$$

$$p(\boldsymbol{\Theta}|\alpha) = \frac{(\sqrt{\alpha})^L}{(\sqrt{2\pi})^L} \exp\left\{-\frac{\alpha}{2}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}\right\} = \frac{(\sqrt{\alpha})^L}{(\sqrt{2\pi})^L} \exp\left\{\left[0, -\frac{1}{2}\varphi^{\mathrm{T}}(\alpha \boldsymbol{I})\right] \begin{bmatrix}\boldsymbol{\Theta} \\ \varphi(\boldsymbol{\Theta})\end{bmatrix}\right\}$$

其中, $\varphi(\boldsymbol{U}_n) = 2dvec(\boldsymbol{U}_n\boldsymbol{U}_n^{\mathrm{T}}) - dvec\big(sdiag(\boldsymbol{U}_n)\big)$, $\varphi(\boldsymbol{\Theta}) = dvec(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}})$, $\varphi(\alpha \boldsymbol{I}) = dvec(\alpha \boldsymbol{I})$. 其中, $dvec(\cdot)$ 定义为矩阵向量化操作, $sdiag(\cdot)$ 定义为向量对角矩阵化操作, 具体形式见附录 A 和附录 B.

于是得到:

$$p(\mathcal{Y}, \mathcal{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \propto \prod_{n=1}^{N} p(x_n|\boldsymbol{\Theta}, \delta_w)p(\boldsymbol{\Theta}|\alpha) \propto \exp\left\{\left[\delta_w \sum_{n=1}^{N} x_n \boldsymbol{U}_n^{\mathrm{T}}, -\frac{1}{2}\delta_w \sum_{n=1}^{N} \varphi^{\mathrm{T}}(\boldsymbol{U}_n) - \frac{1}{2}\varphi^{\mathrm{T}}(\alpha \boldsymbol{I})\right] \begin{bmatrix}\boldsymbol{\Theta} \\ \varphi(\boldsymbol{\Theta})\end{bmatrix}\right\}$$

因此, $q(\boldsymbol{\Theta})$ 属于高斯分布, 即

$$q(\boldsymbol{\Theta}) \propto \exp\left\{\beta_{\boldsymbol{\Theta}}^{\mathrm{T}} \begin{bmatrix}\boldsymbol{\Theta} \\ \varphi(\boldsymbol{\Theta})\end{bmatrix}\right\}$$

其中, $\beta_{\boldsymbol{\Theta}}$ 表示全局隐变量 $\boldsymbol{\Theta}$ 对应的全局变分参数.

将 $q(\boldsymbol{\Theta})$ 代入下界函数, 根据式 (21), 可以得到下界函数在第 k 次迭代关于全局变分参数 $\beta_{\boldsymbol{\Theta}}$ 的自然梯度估计值为

$$\hat{\nabla}\mathcal{L}(\beta_{\boldsymbol{\Theta}}^{(k-1)}) = \left[\frac{N}{Z}\sum_{I_z=1}^{Z}\left[\langle\delta_w\rangle\langle x_{I_z}\rangle \boldsymbol{U}_{I_z}^{\mathrm{T}}\right], \frac{N}{Z}\sum_{I_z=1}^{Z}\left[\frac{\langle\delta_w\rangle}{2}\varphi^{\mathrm{T}}(\boldsymbol{U}_{I_z})\right] + \frac{1}{2}\varphi^{\mathrm{T}}(\langle\alpha\rangle \boldsymbol{I})\right]^{\mathrm{T}} - \beta_{\boldsymbol{\Theta}}^{(k-1)} \tag{23}$$

其中, $\langle\cdot\rangle$ 表示期望运算. 此时, 全局隐变量 $\boldsymbol{\Theta}$ 对应的全局变分参数 $\beta_{\boldsymbol{\Theta}}$ 按照式 (22) 进行更新, 即

$$\beta_{\boldsymbol{\Theta}}^{(k)} = \beta_{\boldsymbol{\Theta}}^{(k-1)} + \rho_k \hat{\nabla}\mathcal{L}(\beta_{\boldsymbol{\Theta}}^{(k-1)}) \tag{24}$$

按照高斯分布性质, 得到全局隐变量 $\boldsymbol{\Theta}$ 在第 k 次迭代时的期望和方差, 即

$$\begin{cases}\mathrm{var}(\boldsymbol{\Theta}) = idvec^{-1}\big(2\beta_{\boldsymbol{\Theta}}^{(k)}(L+1:end)\big) \\ \langle\boldsymbol{\Theta}\rangle = \beta_{\boldsymbol{\Theta}}^{(k)}(1:L)\mathrm{var}(\boldsymbol{\Theta})\end{cases} \tag{25}$$

其中, $\beta_m^{(k)}(i)$ 表示 $\beta_m^{(k)}$ 的第 i 个元素, $idvec(\cdot)$ 定义为向量矩阵化操作, end 表示向量最后一个元素的索引, 向量矩阵化操作见附录 C.

同样地, 对于 $q(\boldsymbol{\Lambda})$, 可以得到下界函数在第 k 次迭代关于全局变分参数 $\beta_{\boldsymbol{\Theta}}$ 的自然梯度估计值为

$$\hat{\nabla}\mathcal{L}(\beta_{\boldsymbol{\Lambda}}^{(k-1)}) = \left[\frac{N}{Z}\sum_{I_z=1}^{Z}\left[\langle r_{I_z}\rangle\langle\delta_e\rangle y_{I_z}\langle \boldsymbol{F}^{\mathrm{T}}(x_{I_z})\rangle\right], \frac{N}{Z}\sum_{I_z=1}^{Z}\left[\frac{\langle r_{I_z}\rangle\langle\delta_w\rangle}{2}\langle\varphi^{\mathrm{T}}\big(\boldsymbol{F}(x_{I_z})\big)\rangle\right] + \frac{1}{2}\varphi^{\mathrm{T}}(\langle\alpha\rangle \boldsymbol{I})\right]^{\mathrm{T}} - \beta_{\boldsymbol{\Lambda}}^{(k-1)} \tag{26}$$

其中,

$$\varphi(\boldsymbol{F}(x_{I_z})) = 2dvec((\boldsymbol{F}(x_{I_z}))(\boldsymbol{F}^{\mathrm{T}}(x_{I_z}))) - dvec(sdiag((\boldsymbol{F}(x_{I_z}))))$$

此时, 全局隐变量 $\boldsymbol{\Lambda}$ 对应的全局变分参数 $\beta_{\boldsymbol{\Lambda}}$ 按照式 (22) 进行更新, 即

$$\beta_{\boldsymbol{\Lambda}}^{(k)} = \beta_{\boldsymbol{\Lambda}}^{(k-1)} + \rho_k \hat{\nabla}\, \mathcal{L}(\beta_{\boldsymbol{\Lambda}}^{(k-1)}) \tag{27}$$

按照高斯分布性质, 得到全局隐变量 $\boldsymbol{\Lambda}$ 在第 k 次迭代时的期望和方差, 即

$$\begin{cases} \mathrm{var}(\boldsymbol{\Lambda}) = idvec^{-1}\big(2\beta_{\boldsymbol{\Lambda}}^{(k)}(M+1:end)\big) \\ \langle \boldsymbol{\Lambda} \rangle = \beta_{\boldsymbol{\Lambda}}^{(k)}(1:M)\mathrm{var}(\boldsymbol{\Lambda}) \end{cases} \tag{28}$$

2.3.2 $q(\delta_w)$ 和 $q(\delta_e)$ 更新

变分后验分布 $q(\delta_w)$ 的形式为

$$\begin{aligned} q(\delta_w) \propto\ & p(\mathcal{Y}, \mathcal{X}|\delta_w)p(\delta_w) \propto \\ & \prod_{n=1}^{N} p(x_n|\boldsymbol{\Theta}, \delta_w)p(\delta_w|a_0, b_0) \propto \\ & \exp\left\{ \left[a_0 + \frac{N}{2} - 1, \right.\right. \\ & -\frac{1}{2}\sum_{n=1}^{N}(x_n^2 - 2x_n\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_n + \\ & \left.\left. \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_n\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_n) - b_0 \right] \begin{bmatrix} \ln\delta_w \\ \delta_w \end{bmatrix} \right\} \end{aligned}$$

因此, $q(\delta_w)$ 属于伽马分布, 即

$$q(\delta_w) \propto \exp\left\{ \beta_{\delta_w}^{\mathrm{T}} \begin{bmatrix} \ln\delta_w \\ \delta_w \end{bmatrix} \right\}$$

其中, β_{δ_w} 表示全局隐变量 δ_w 对应的全局变分参数.

将 $q(\delta_w)$ 代入下界函数, 根据式 (21), 可以得到下界函数在第 k 次迭代关于全局变分参数 β_{δ_w} 的自然梯度估计值为:

$$\begin{aligned} \hat{\nabla}\, \mathcal{L}(\beta_{\delta_w}^{(k-1)}) = & \left[a_0 + \frac{N}{2} - 1, \right. \\ & \frac{N}{2Z}\sum_{I_z=1}^{Z}\left[\langle \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_{I_z}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_{I_z} \rangle + \langle x_{I_z}^2 \rangle - \right. \\ & \left.\left. 2\langle x_{I_z} \rangle \langle \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_{I_z} \rangle \right] + b_0 \right]^{\mathrm{T}} - \beta_{\delta_w}^{(k-1)} \end{aligned} \tag{29}$$

此时, 全局隐变量 δ_w 对应的全局变分参数 β_{δ_w} 按照式 (22) 进行更新, 即

$$\beta_{\delta_w}^{(k)} = \beta_{\delta_w}^{(k-1)} + \rho_k \hat{\nabla}\, \mathcal{L}(\beta_{\delta_w}^{(k-1)}) \tag{30}$$

按照伽马分布性质, 得到全局隐变量 δ_w 在第 k 次迭代时的期望和方差, 即

$$\langle \delta_w \rangle = \frac{\beta_{\delta_w}^{(k)}(1) + 1}{\beta_{\delta_w}^{(k)}(2)},\ \mathrm{var}(\delta_w) = \frac{\langle \delta_w \rangle}{\beta_{\delta_w}^{(k)}(2)} \tag{31}$$

同样地, 对于 $q(\delta_e)$, 可以得到下界函数在第 k 次迭代关于全局变分参数 β_{δ_e} 的自然梯度估计值为

$$\begin{aligned} \hat{\nabla}\, \mathcal{L}(\beta_{\delta_e}^{(k-1)}) = & \left[a_0 + \frac{N}{2} - 1, \right. \\ & \frac{N}{2Z}\sum_{I_z=1}^{Z}\langle r_{I_z} \rangle [y_{I_z}^2 - 2y_{I_z}\langle \boldsymbol{\Lambda} \rangle^{\mathrm{T}} \langle \boldsymbol{F}(x_{I_z}) \rangle + \\ & \left. \langle \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{F}(x_{I_z})\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{F}(x_{I_z}) \rangle] + b_0 \right]^{\mathrm{T}} - \beta_{\delta_e}^{(k-1)} \end{aligned} \tag{32}$$

此时, 全局隐变量 δ_e 对应的全局变分参数 β_{δ_e} 按照式 (22) 进行更新, 即

$$\beta_{\delta_e}^{(k)} = \beta_{\delta_e}^{(k-1)} + \rho_k \hat{\nabla}\, \mathcal{L}(\beta_{\delta_e}^{(k-1)}) \tag{33}$$

按照伽马分布性质, 得到全局隐变量 δ_e 在第 k 次迭代时的期望和方差, 即

$$\langle \delta_e \rangle = \frac{\beta_{\delta_e}^{(k)}(1) + 1}{\beta_{\delta_e}^{(k)}(2)},\ \mathrm{var}(\delta_e) = \frac{\langle \delta_e \rangle}{\beta_{\delta_e}^{(k)}(2)} \tag{34}$$

2.3.3 $q(\alpha)$ 更新

变分后验分布 $q(\alpha)$ 的形式为

$$\begin{aligned} q(\alpha) \propto\ & p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}|\alpha)p(\alpha|a_0, b_0) \propto \\ & \exp\left\{ \left[\frac{1}{2}(L + M + 2) + a_0 - 1, \right.\right. \\ & \left.\left. \frac{1}{2}[\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Lambda}^{\mathrm{T}}]\boldsymbol{I}\begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{\Lambda} \end{bmatrix} + b_0\boldsymbol{I} \right] \begin{bmatrix} \ln\alpha \\ \alpha \end{bmatrix} \right\} \end{aligned}$$

因此, $q(\alpha)$ 属于伽马分布, 即

$$q(\alpha) \propto \exp\left\{ \beta_{\alpha}^{\mathrm{T}} \begin{bmatrix} \ln\alpha \\ \alpha \end{bmatrix} \right\}$$

其中, β_{α} 表示全局隐变量 α 对应的全局变分参数.

将 $q(\alpha)$ 代入下界函数, 根据式 (21), 可以得到下界函数在第 k 次迭代关于全局变分参数 β_{α} 的自然梯度估计值为

$$\begin{aligned} \hat{\nabla}\, \mathcal{L}(\beta_{\alpha}^{(k-1)}) = & \left[\frac{1}{2}(L + M + 2) + a_0 - 1, \right. \\ & \left. \frac{1}{2}[\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Lambda}^{\mathrm{T}}]\boldsymbol{I}\begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{\Lambda} \end{bmatrix} + b_0\boldsymbol{I} \right]^{\mathrm{T}} - \beta_{\alpha}^{(k-1)} \end{aligned} \tag{35}$$

此时, 全局隐变量 α 对应的全局变分参数 β_{α} 按照式 (22) 进行更新, 即

$$\beta_{\alpha}^{(k)} = \beta_{\alpha}^{(k-1)} + \rho_k \hat{\nabla}\, \mathcal{L}(\beta_{\alpha}^{(k-1)}) \tag{36}$$

按照伽马分布性质, 得到全局隐变量 α 在第 k 次迭代时的期望和方差, 即

$$\langle \alpha \rangle = \frac{\beta_{\alpha}^{(k)}(1) + 1}{\beta_{\alpha}^{(k)}(2)},\ \mathrm{var}(\alpha) = \frac{\langle \alpha \rangle}{\beta_{\alpha}^{(k)}(2)} \tag{37}$$

不难发现在对全局隐变量的变分后验分布进行

更新时会使用到局部隐变量的信息, 在此我们给出局部隐变量 $\mathcal{X}=\{x_{1:N}, r_{1:N}\}$ 变分后验分布的更新方式.

2.3.4　$q(x_n)$ 更新

对于 $x_{1:N}$ 中每个 x_n, $n\in(1, N)$ 互相独立, 有 $q^k(x_{1:N})=\prod_{n=1}^{N} q_n^k(x_n)$, 不失一般性, 考虑第 n 个数据点, 记 $x_{-n}=\{x_1, \cdots, x_{n-1}, x_{n+1}, \cdots, x_N\}$. 考虑与 x_n 有关的项为 $p(y_n|x_n, r_n, \boldsymbol{\Lambda}, \delta_e, v)$ 和 $p(x_n|\boldsymbol{\Theta}, \delta_w)$, 其对数概率分布为

$$\ln p(y_n|x_n, r_n, \boldsymbol{\Lambda}, \delta_e, v)=\ln\frac{\sqrt{r_n\delta_e}}{\sqrt{2\pi}}-\frac{r_n\delta_e}{2}\Big(y_n-\sum_{i=0}^{M}\lambda_i f_i(x_n)\Big)^2 \tag{38}$$

和

$$\ln p(x_n|\boldsymbol{\Theta}, \delta_w)=\ln\frac{\sqrt{\delta_w}}{\sqrt{2\pi}}-\frac{\delta_w}{2}(x_n-x_n^0)^2 \tag{39}$$

将式 (38) 和式 (39) 代入式 (13), 有

$$q^k(x_n)=\frac{\exp\left[B(x_n)\right]}{\int\exp\left[B(x_n)\right]\mathrm{d}x_n} \tag{40}$$

其中

$$\begin{aligned}B(x_n)=\mathrm{E}_{q(x_{-n})q(\mathcal{Z})}\Big[&\ln p(y_n|x_n, r_n, \boldsymbol{\Lambda}, \delta_e, v)+\\&\ln p(x_n|\boldsymbol{\Theta}, \delta_w)\Big]=-\frac{\langle r_n\rangle\langle\delta_e\rangle}{2}\\&\Big(\boldsymbol{F}^{\mathrm{T}}(x_n)\langle\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}\rangle\boldsymbol{F}(x_n)-2y_n\langle\boldsymbol{\Lambda}\rangle^{\mathrm{T}}\boldsymbol{F}(x_n)\Big)-\\&\frac{\langle\delta_w\rangle}{2}(x_n^2-2x_n\langle\boldsymbol{\Theta}\rangle^{\mathrm{T}}\boldsymbol{U}_n)+const\end{aligned}$$

其中, $\boldsymbol{F}(x_n)=[f_0(x_n), \cdots, f_M(x_n)]^{\mathrm{T}}$, $\boldsymbol{U}_n=[u_n, u_{n-1}, \cdots, u_{n-L}]^{\mathrm{T}}$, $\langle\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}\rangle=\langle\boldsymbol{\Lambda}\rangle\langle\boldsymbol{\Lambda}\rangle^{\mathrm{T}}+\mathrm{var}(\boldsymbol{\Lambda})$, 为全局隐变量 $\boldsymbol{\Lambda}$ 的协方差.

由于参数的非线性, 无法直接得到 $q^k(x_n)$ 的解析形式, 故使用重要性采样[17] 的方法来近似 $q^k(x_n)$. 假设已知 x_n 的分布, 由 $\widetilde{p}_n(x_n)$ 表示, 则对于式 (40) 的分母项, 可将其近似为

$$\int\exp\left[B(x_n)\right]\mathrm{d}x_n\approx\frac{1}{C}\sum_{c=1}^{C}\frac{\exp\left[B(x_{n,c})\right]}{\widetilde{p}_n(x_{n,c})}$$

选择一个合适的分布 $\widetilde{p}_n(x_n)$ 对于计算上式积分十分重要, 直接影响计算的精度, 为了更好地近似 $q^k(x_n)$, 本文选择 $\widetilde{p}_n(x_n)=\mathcal{N}(\tilde{\mu}_n, \tilde{\sigma}_n)$, 其中

$$\tilde{\mu}_n=\arg\max_{x_n}B(x_n),\ \tilde{\sigma}_n=\langle\delta_w\rangle^{-1}$$

记近似概率密度函数为 $q^k(x_{n,c})$, 有

$$q^k(x_{n,c})=\frac{\exp\left[B(x_{n,c})\right]}{\widetilde{p}_n(x_{n,c})}\left(\sum_{c=1}^{C}\frac{\exp\left[B(x_{n,c})\right]}{\widetilde{p}_n(x_{n,c})}\right)^{-1} \tag{41}$$

故 $q^k(x_n)$ 可以表示为

$$q^k(x_n)\approx\sum_{c=1}^{C}\delta(x_n-x_{n,c})q^k(x_{n,c}) \tag{42}$$

其中, $\delta(\cdot)$ 表示 δ-函数. 至此, 可以得到局部隐变量 x_n 在第 k 次迭代时的期望和方差, 即

$$\begin{cases}\langle x_n\rangle=\sum\limits_{c=1}^{C}x_{n,c}q^k(x_{n,c})\\ \mathrm{var}(x_n)=\sum\limits_{c=1}^{C}(x_{n,c}-\langle x_n\rangle)q^k(x_{n,c})\end{cases} \tag{43}$$

2.3.5　$q(r_n)$ 更新

不失一般性, 考虑第 n 个数据点, 与 r_n 有关的项为 $p(y_n|x_n, r_n, \boldsymbol{\Lambda}, \delta_e, v)$ 和 $p(r_n|v)$. 其中, $p(r_n|v)$ 的对数概率分布为

$$\ln p(r_n|v)=\frac{v}{2}\ln\frac{v}{2}-\ln\Gamma\left(\frac{v}{2}\right)+\left(\frac{v}{2}-1\right)\ln r_n-\frac{v}{2}r_n \tag{44}$$

其中, $\Gamma(\cdot)$ 表示伽马函数.

根据式 (13), 可以得到关于 $q^k(r_n)$ 的变分解的形式:

$$q^k(r_n)\propto\exp\left\{\left(\frac{v^{k-1}+1}{2}-1\right)\ln r_n-\frac{\langle\delta_e\rangle A_n+v^{k-1}}{2}r_n\right\} \tag{45}$$

其中, A_n 表达式为

$$\begin{aligned}A_n=\mathrm{E}_{q(x_n)q(\boldsymbol{\Lambda})}\Big[(y_n-\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{F}(x_n))^2\Big]=&\\y_n^2-2y_n\langle\boldsymbol{\Lambda}\rangle^{\mathrm{T}}\langle\boldsymbol{F}(x_n)\rangle+\sum_{i=0}^{M}\langle\lambda_i^2\rangle\langle f_i^2(x_n)\rangle+&\\2\sum_{i=0}^{M-1}\sum_{j=0}^{M}\langle\lambda_i\rangle\langle\lambda_j\rangle\langle f_i(x_n)\rangle\langle f_j(x_n)\rangle&\end{aligned}$$

其中, $f_i(x_n)$ 和 $f_i^2(x_n)$ 的期望可以表示为

$$\langle f_i(x_n)\rangle=\sum_{c=1}^{C}f_i(x_{n,c})q^k(x_{n,c})$$

$$\langle f_i^2(x_n)\rangle=\sum_{c=1}^{C}f_i^2(x_{n,c})q^k(x_{n,c})$$

易知 $q^k(r_n)$ 服从伽马分布, 即

$$q^k(r_n)=\mathcal{G}\left(\frac{v^{k-1}+1}{2},\frac{\langle\delta_e\rangle A_n+v^{k-1}}{2}\right) \tag{46}$$

根据伽马分布的性质, 可以得到局部隐变量 r_n 在第 k 次迭代时的期望和方差, 即

$$\begin{cases}\langle r_n\rangle=\dfrac{v^{k-1}+1}{\langle\delta_e\rangle A_n+v^{k-1}}\\ \mathrm{var}(r_n)=\dfrac{v^{k-1}+1}{(\langle\delta_e\rangle A_n+v^{k-1})^2}\end{cases} \tag{47}$$

2.3.6 $v^{(k)}$ 更新

对数似然函数关于 v^k 有关的项可以表示为

$$\ln p(y_{1:N},x_{1:N},r_{1:N},\boldsymbol{\Theta},\boldsymbol{\Lambda},\delta_w,\delta_e,\alpha|v)=\frac{v^k}{2}\ln\frac{v^k}{2}-\ln\Gamma\left(\frac{v^k}{2}\right)+\left(\frac{v^k}{2}-1\right)\ln r_n-\frac{v^k}{2}r_n \tag{48}$$

下界函数 $\mathcal{L}(v^k)$ 关于 v^k 的变分解的形式可以表示为

$$\begin{aligned}\mathcal{L}(v^k)\propto&\left(\frac{v^k}{2}-1\right)\sum_{t=1}^{N}\left[\Psi\left(\frac{v^{k-1}+1}{2}\right)-\right.\\&\left.\ln\left(\frac{\langle\tau_2\rangle A_n+v^{k-1}}{2}\right)\right]-\frac{v^k}{2}\sum_{t=1}^{N}\langle r_n\rangle+\\&N\frac{v^k}{2}\ln\frac{v^k}{2}-N\ln\Gamma(\frac{v^k}{2})\end{aligned} \tag{49}$$

其中, $\Psi(\cdot)$ 表示 $\ln\Gamma(\cdot)$ 的微分, 可以通过 MATLAB 中的 fminbnb 函数求解上述单变量优化问题.

2.4 SVBI 算法更新步骤

步骤 1. 设定初始迭代时刻 $k=1$, 初始化各变量 $\{x_{1:N},r_{1:N},\boldsymbol{\Theta},\boldsymbol{\Lambda},\delta_w,\delta_e,\alpha\}$ 的分布以及全局隐变量 $\{\boldsymbol{\Theta},\boldsymbol{\Lambda},\delta_w,\delta_e,\alpha\}$ 对应的自然参数; 分别设定超参数 a_0 和 b_0 以及结构参数 v 的初始值.

步骤 2. 根据式 (19) 适当地设定步长 ρ_k.

步骤 3. 从原始数据点均匀分布地采样 Z 个数据点 I_z.

步骤 4. 根据式 (43) 和式 (47) 分别计算第 I_z 个数据点对应的局部隐变量 $q(x_{I_z})$ 和 $q(r_{I_z})$.

步骤 5. 根据式 (24) 和式 (27) 分别计算全局隐变量 $q(\boldsymbol{\Theta})$ 和 $q(\boldsymbol{\Lambda})$ 对应的全局变分参数.

步骤 6. 根据式 (30) 和式 (33) 分别计算全局隐变量 $q(\delta_w)$ 和 $q(\delta_e)$ 对应的全局变分参数.

步骤 7. 根据式 (36) 计算全局隐变量 $q(\alpha)$ 对应的全局变分参数.

步骤 8. 根据式 (49) 求解优化问题更新 v^k.

步骤 9. 当下界函数 $\mathcal{L}(q)$ 收敛时停止迭代; 否则重复执行步骤 2.

3 实验验证

3.1 仿真实验

使用一个数值仿真的实例来说明本文所提出 SVBI 方法的有效性, 考虑以下维纳模型:

$$\begin{cases}x_n^0=\dfrac{1}{1+0.5q^{-1}}u_n\\ x_n=x_n^0+w_n\\ y_n=x_n+x_n^2+e_n\end{cases} \tag{50}$$

其中, $w_n\sim\mathrm{N}(0,0.3^2)$, $e_n\sim\mathrm{N}(0,0.3^2)$, 则 $\boldsymbol{\Lambda}$ 的真实值为 $\boldsymbol{\Lambda}=[0,1,1]$, 将传递函数根据式 (3) 和式 (4) 改写为 FIR 模型, 取 $L=10$, 有 $x_n^0=\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{U}_n$, 可以得到 $\boldsymbol{\Theta}$ 的真实值为

$$\boldsymbol{\Theta}=[1,-0.5,0.25,-0.125,0.062\,5,-0.031\,25,\cdots]^{\mathrm{T}} \tag{51}$$

在仿真之前, 将系统引入 5% 的异常值, 其都来自于 $[-20,-15]\cup[15,20]$ 之间的均匀分布, 设定遗忘速率 $\gamma=0.3$, 延迟因子 $\tau=5$. 在实验过程中, 从 $[-2,2]$ 的均匀分布中采样 300 个数据点作为系统的激励信号. 为获得唯一解, 对数据进行归一化操作, 固定线性环节的第一个参数 $\theta_0=1$. 在全局参数更新时, 分别在每次迭代时随机采样 1 个点、5%、10%、20% 以及全部的局部隐变量来估计下界函数关于全局隐变量后验分布的分布参数的自然梯度值, 每次迭代时刻使算法循环 500 次, 从而实现对全局隐变量的更新, 将辨识到的参数集列入表 1.

根据表 1 可知本文所提出的 SVBI 对所考虑的维纳模型辨识的有效性, 当每次迭代时刻子采样数据点的个数增加时, 对模型参数的辨识更加准确, 但相应地会降低算法的速度优势.

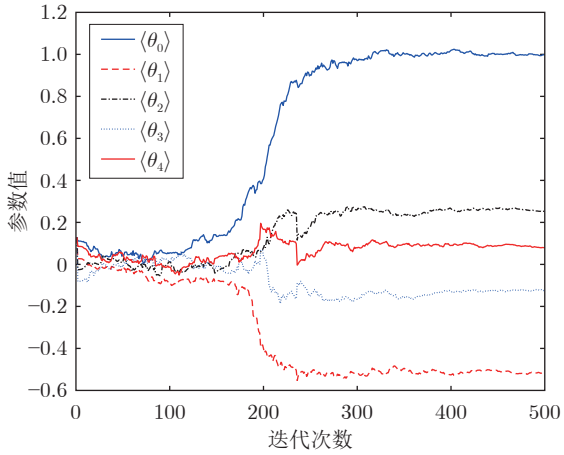
为进一步说明所提出 SVBI 方法的有效性, 同时兼顾模型的准确性和算法的效率, 在每次循环中随机均匀分布地选择 5% 的局部隐变量进行更新, 图 3 为线性环节参数 $\boldsymbol{\Theta}$ 的前 5 个参数以及非线性环节参数 $\boldsymbol{\Lambda}$ 的收敛情况, 随着迭代次数的增加, 可以看出各参数逐渐收敛到真值. 图 4 为下界函数 $\mathcal{L}(q)$ 随着迭代次数增加的收敛情况. 图 5 为当系统存在 5% 的异常值情况下使用本文所提出的 SVBI 方法得到的预测输出, 同时绘出了 PEM 方法和 MLE 方法的输出结果, 通过对比表明 SVBI 方法对于参数辨识的有效性.

表 2 列出了系统存在不同程度异常值时使用本文所提出方法对参数的辨识情况. 将本文提出的 SVBI 与 VBEM、PEM、MLE 方法进行比较, 假设分别有 5%、10% 的测量值被异常值所影响, 使用 50 次蒙特卡洛实验来验证辨识方法的有效性, 将 4

表 1　不同子采样数据点对应的参数辨识情况

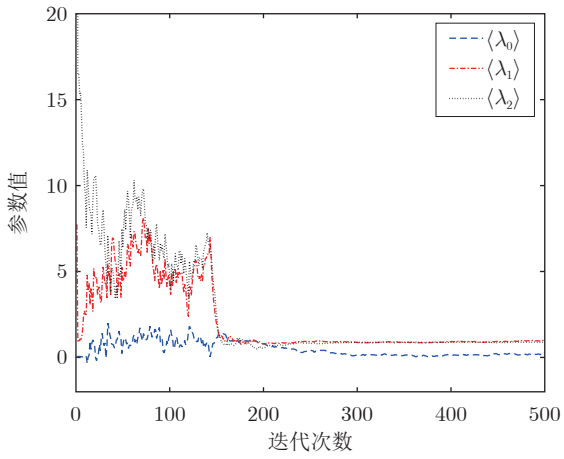
Table 1　Identification of parameters corresponding to different sub-sampling data points

	$\langle\theta_0\rangle$	$\langle\theta_1\rangle$	$\langle\theta_2\rangle$	$\langle\theta_3\rangle$	$\langle\theta_4\rangle$	$\langle\lambda_0\rangle$	$\langle\lambda_1\rangle$	$\langle\lambda_2\rangle$	时间 (s)
真实值	1	−0.5000	0.2500	−0.1250	0.0625	0	1	1	—
采样 1 个点	1±0	−0.5463±0.3604	0.2507±0.2471	−0.2446±0.2655	0.0358±0.2882	0.5434±0.4180	0.6625±0.2907	0.3803±0.2185	0.6005
采样 5%	1±0	−0.5060±0.0330	0.2693±0.0497	−0.1252±0.0323	0.0633±0.0323	0.0908±0.2707	0.9871±0.1480	0.9103±0.1246	3.1829
采样 10%	1±0	−0.5055±0.0248	0.2571±0.0257	−0.1341±0.0255	0.0594±0.0256	0.0631±0.0504	0.9684±0.0498	0.9499±0.0459	7.7402
采样 20%	1±0	−0.5077±0.0204	0.2544±0.0202	−0.1287±0.0289	0.0659±0.0291	0.0575±0.0540	0.9813±0.0518	0.9574±0.0451	11.4620
采样全部	1±0	−0.5078±0.0278	0.2541±0.0283	−0.1299±0.0271	0.0685±0.0246	0.0777±0.0726	0.9439±0.1183	0.9252±0.1326	9.0772



(a) 线性环节参数

(a) Parameters of linear part



(b) 非线性环节参数

(b) Parameters of nonlinear part

图 3　辨识参数的收敛状况

Fig. 3　Convergence of identified parameters

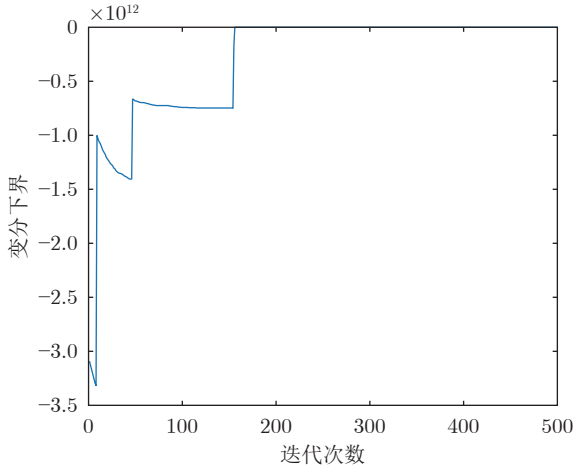


图 4　下界函数的收敛过程

Fig. 4　Convergence process of the lower bound function

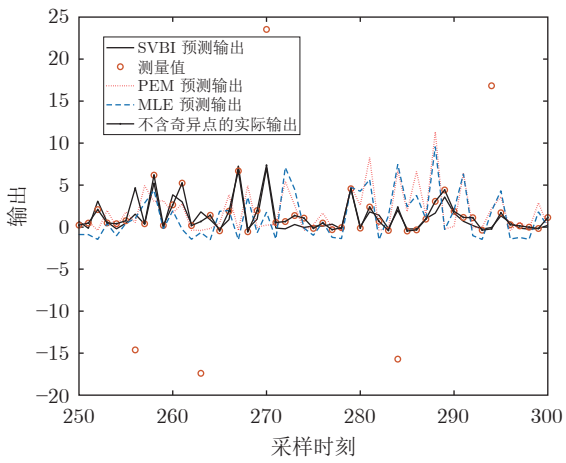


图 5　预测输出与实际输出比较

Fig. 5　Comparison of predicted output with actual output

种方法得到的系统非线性部分的参数列入表 3, 同时记录不同异常值存在时各方法的平均 CPU 时间.

对于 PEM 方法, 由于采用简单的优化求解方法就可以得到模型参数, 计算量很小, 但该方法未考虑测量噪声对辨识的影响, 在测量数据存在异常值时的输出预测效果相比于其他两类方法较差 (如图 5 所示), 同时在多次实验中该方法得到的参数的方差较大, 辨识的鲁棒性不好. SVBI、MLE、VBEM 方法均通过极大化测量值的似然函数辨识模型参数. 从比较结果来看, 在测量值无异常情况下, 三者给出的辨识精度相当. 而在存在测量异常时, MLE 算法因为直接计算似然函数, 从表 3 可以看出, 若异常测量占比较大时, MLE 已经无法给出正确的辨识结果. VBEM 和 SVBI 均能处理测量中的异常值, 辨识精度也相当, 但是 SVBI 的计算时间显著

表 2 不同异常值存在时的参数辨识情况

Table 2 Parameter identification when different outliers exist

	$\langle\theta_0\rangle$	$\langle\theta_1\rangle$	$\langle\theta_2\rangle$	$\langle\theta_3\rangle$	$\langle\theta_4\rangle$	$\langle\theta_5\rangle$	时间 (s)
真实值	1	−0.5000	0.2500	−0.1250	0.0625	−0.03125	—
无异常值	1±0	−0.4989±0.0292	0.2495±0.0293	−0.1254±0.0223	0.0611±0.0257	−0.0338±0.0262	2.9369
2% 异常值	1±0	−0.5097±0.0389	0.2672±0.0497	−0.1305±0.0426	0.0652±0.0452	−0.0291±0.0494	2.9480
5% 异常值	1±0	−0.5060±0.0330	0.2693±0.0497	−0.1252±0.0323	0.0633±0.0323	−0.0314±0.0523	3.1829
10% 异常值	1±0	−0.5349±0.0325	0.2627±0.0323	−0.1314±0.0330	0.0685±0.0389	−0.0377±0.0355	2.9057

表 3 不同辨识方法的性能比较

Table 3 Performance comparison of different recognition methods

		b_0	a_1	$\langle\lambda_0\rangle(\lambda_0)$	$\langle\lambda_1\rangle(\lambda_1)$	$\langle\lambda_2\rangle(\lambda_2)$	均方误差	时间 (s)
真实值		1	0.5	0	1	1	—	—
无异常值	SVBI	—	—	0.0648±0.0620	0.9633±0.0509	0.9766±0.0626	0.9136	**2.936 9**
	VBEM	—	—	0.0503±0.0346	0.9411±0.0393	0.9655±0.0459	0.8978	9.7046
	MLE	1±0	0.5102±0.0136	0.1054±0.0405	1.0154±0.0464	0.9490±0.0411	0.9130	9.0350
	PEM	1±0	0.4948±0.0172	0.0828±0.0524	0.9905±0.0373	1.0072±0.0449	0.9132	0.6474
5% 异常值	SVBI	—	—	0.0575±0.0540	0.9813±0.0520	0.9573±0.0450	5.4540	**2.935 2**
	VBEM	—	—	0.0503±0.0411	0.9770±0.0532	0.9748±0.0518	3.8695	9.7709
	MLE	1±0	0.4150±0.0711	−0.9407±0.1253	1.0019±0.1839	1.3715±0.1895	3.9574	9.6693
	PEM	1±0	0.4999±0.0549	0.1072±0.1871	0.9646±0.1926	0.9878±0.1558	3.8374	0.6580
10% 异常值	SVBI	—	—	0.1439±0.1065	0.9163±0.0924	0.8416±0.0924	7.5364	**2.9057**
	VBEM	—	—	0.0556±0.0468	0.9711±0.0538	0.9568±0.0553	5.5110	9.9245
	MLE	—	—	—	—	—	—	—
	PEM	1±0	0.4723±0.2004	0.1458±0.5211	0.9746±0.3091	1.0030±0.3253	5.4992	0.6620

低于 VBEM 方法, 该结果验证了本文提出的设想, 表明 SVBI 在不降低辨识精度下, 显著降低计算量, 为在线辨识或强实时情况下的系统辨识提供支撑, 解决贝叶斯学习计算量高的问题.

3.2 Wiener-Benchmark 问题辨识

本节利用提出的方法辨识 Wiener-Benchmark 模型. 该模型是一个典型的维纳过程, 由 Schoukens 等[14] 提出, 是验证维纳系统辨识的标准模型, 如图 1 所示, 该模型中的静态非线性环节由一个二极管构成, 线性环节由一个切比雪夫滤波器构成, 具体模型描述可参考文献 [14]. 本文用以下模型来描述此过程:

$$\begin{cases} x_n^0 = (\theta_0 + \theta_1 z^{-1} + \cdots + \theta_L z^{-L}) u_n \\ x_n = x_n^0 + w_n \\ f(x_n) = c_0 + c_1 x_n + c_2 x_n^2 \\ y_n = f(x_n) + e_n \end{cases} \tag{52}$$

其中, $w_n \sim \mathrm{N}(0, Q)$, $e_n \sim \mathrm{N}(0, R)$, 为了辨识此模型并体现出本文所提出方法的有效性, 设定 $L = 35$, 故有 $[\theta_0, \cdots, \theta_{34}, c_0, c_1, c_2, Q, R]$ 共 40 个参数需要辨识. 在文献 [14] 中, 一共包含了 188 000 组数据点, 其作者建议使用前 100 000 个数据点来辨识模型, 剩下的数据用来测试模型的准确性. 在此, 本文取 6 000 ~ 7 999 时刻以及 6 000 ~ 15 999 时刻两组数据点来辨识模型参数, 使用 150 000 ~ 151 999 时刻共 2 000 个数据点来验证模型的有效性.

表 4 列出了 SVBI 方法辨识到的模型的部分参数值, 图 6 为系统输出预测值与实际值的比较, 结果表明 SVBI 方法得到的辨识参数可以准确地预测输出值, 证明该方法在此 Benchmark 问题上的有效性. 表 5 将提出的方法与 VBEM 方法进行比较, 为了比较的公平性, 两种方法辨识的模型结构设置相同, 且辨识的数据点和模型验证的测试点均相同. 明显地, 本文所提出的 SVBI方法的辨识精度略高于 VBEM, 且所消耗时间显著低于 VBEM 方法, 为大规模数据下的非线性系统的贝叶斯学习提供了新的思路.

4 结束语

本文考虑了存在过程噪声、测量噪声以及参数不确定情况下的维纳模型的辨识问题, 提出 SVBI 方法, 根据随机优化的思想, 将模型参数分为全局隐变量和局部隐变量, 通过自然梯度下降方法计算全局隐变量对应的全局变分参数, 实现对模型信息的更新. 针对于文献 [20] 中所提出的 VBEM 方法的局限性, 本文所提出 SVBI 方法只需要部分局部

表 4　式 (52) 部分参数辨识结果

Table 4　The identification results of the part parameters of the process (52)

参数	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	c_0	c_1	c_2	Q	R
结果值	−0.0390	0.0648	−0.0547	0.0856	−0.0462	0.2613	0.0501	0.2041	0.3396	0.4154	−0.0188	0.1035	−0.0030	0.0034	0.0014

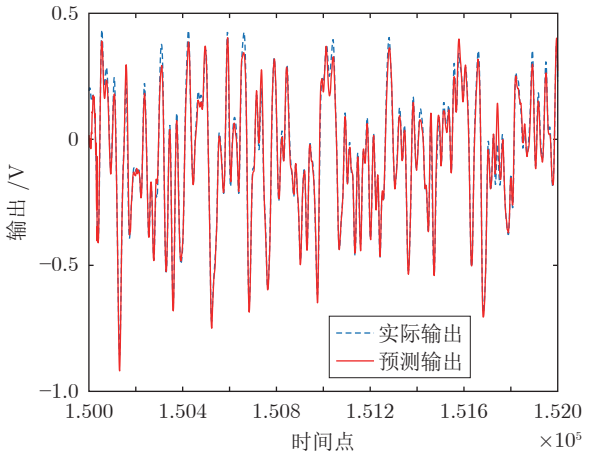


图 6　系统预测输出与实际输出

Fig. 6　Predicted output and actual output of the system

表 5　不同方法的性能比较

Table 5　Performance comparison of different methods

采样点数	方法	均方误差 (V)	参数个数	时间 (s)
2 000	SVBI	**0.056 95**	25	**256.12**
	VBEM	0.062 83	25	1 211.27
	SVBI	**0.034 07**	40	**264.27**
	VBEM	0.034 25	40	1 214.55
10 000	SVBI	**0.061 79**	25	**1 299.99**
	VBEM	0.093 34	25	6 347.28
	SVBI	**0.033 85**	40	**1 332.31**
	VBEM	0.034 04	40	6 442.98

隐变量的信息即可实现对全局隐变量后验分布的更新, 从而实现似然函数的极大化, 可以显著降低变分推理的计算量, 对于大规模数据系统的辨识具有很大的意义. 本文首先通过一个仿真实例, 并与 VBEM、MLE、PEM 三种方法进行了比较, 详细分析了该方法在辨识准确率和计算时间成本上的优势; 又通过一个非线性电路辨识的 Benchmark 问题验证了该方法在实际过程上应用的有效性, 结果表明本文所提出的方法在提高辨识准确度和提高计算效率方面均具有较大优势.

附录 A　矩阵向量化操作定义

考虑对称矩阵 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$, 将矩阵 $\boldsymbol{A}$ 的主对角线及上三角元素, 按行的顺序排列成一个列向量, 其向量化结果 $dvec(\boldsymbol{A})$ 是一个 $n(n+1)/2\times 1$ 维向量, 定义为

$$\boldsymbol{a}=dvec(\boldsymbol{A})=$$
$$[a_{11},\,a_{12},\,\cdots,\,a_{1n},\,a_{22},\,a_{23},\,\cdots,\,a_{2n},\,\cdots,\,a_{nn}]^{\mathrm{T}}\tag{A1}$$

其中, a_{ij} 表示矩阵 $\boldsymbol{A}$ 的第 $(i,\,j)$ 个元素, $dvec(\cdot)$ 为矩阵向量化操作.

附录 B　向量对角矩阵化操作定义

考虑 m 维列向量 $\boldsymbol{b}=[b_1,\,b_2,\,\cdots,\,b_m]^{\mathrm{T}}$, 则存在对角矩阵 $\boldsymbol{B}$ 的主对角线元素为 $\boldsymbol{b}$ 中各元素的平方, 定义为

$$\boldsymbol{B}=sdiag(\boldsymbol{b})=\begin{bmatrix}b_1^2 & 0 & \cdots & 0\\ 0 & b_2^2 & \cdots & \vdots\\ \vdots & 0 & \ddots & \vdots\\ 0 & \cdots & \cdots & b_m^2\end{bmatrix}\tag{B1}$$

其中, $sdiag(\cdot)$ 为向量对角矩阵化操作.

附录 C　向量矩阵化操作定义

考虑 $n(n+1)/2$ 维列向量 $\boldsymbol{a}$,

$$\boldsymbol{a}=[a_{11},\,2a_{12},\,\cdots,\,2a_{1n},\,a_{22},$$
$$2a_{23},\,\cdots,\,2a_{2n},\,\cdots,\,a_{nn}]^{\mathrm{T}}\tag{C1}$$

将其按照行排列的顺序还原为对称矩阵 $\boldsymbol{A}$. 其矩阵化的结果 $idvec(\boldsymbol{a})$ 是一个 n 维的对称矩阵, 定义为

$$\boldsymbol{A}=idvec(\boldsymbol{a})=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{12} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{1n} & a_{2n} & \cdots & a_{nn}\end{bmatrix}\tag{C2}$$

References

1　Wang Le-Yi, Zhao Wen-Xiao. System identification: New paradigms, challenges, and opportunities. *Acta Automatica Sinica*, 2013, **39**(7): 933−942
(王乐一, 赵文虓. 系统辨识: 新的模式、挑战及机遇. 自动化学报, 2013, **39**(7): 933−942)

2　Liu Xin. Identification of linear time-delay systems with unknown delay distributions in its value range. *Acta Automatica Sinica*, 2023, **49**(10): 2136−2144
(刘鑫. 时滞取值概率未知下的线性时滞系统辨识方法. 自动化学报, 2023, **49**(10): 2136−2144)

3　Stoica P. On the convergence of an iterative algorithm used for Hammerstein system identification. *IEEE Transactions on Automatic Control*, 1981, **26**(4): 967−969

4　Zhang Ya-Jun, Chai Tian-You, Yang Jie. Alternating identification algorithm and its application to a class of nonlinear discrete-time dynamical systems. *Acta Automatica Sinica*, 2017, **43**(1): 101−113
(张亚军, 柴天佑, 杨杰. 一类非线性离散时间动态系统的交替辨识算法及应用. 自动化学报, 2017, **43**(1): 101−113)

5　Huang Yu-Long, Zhang Yong-Gang, Li Ning, Zhao Lin. An identification method for nonlinear systems with colored measurement noise. *Acta Automatica Sinica*, 2015, **41**(11): 1877−1892
(黄玉龙, 张勇刚, 李宁, 赵琳. 一种带有色量测噪声的非线性系统辨识方法. 自动化学报, 2015, **41**(11): 1877−1892)

6　Ljung L. Perspectives on system identification. *Annual Reviews in Control*, 2008, **34**(1): 1−12

7　Schön T B, Wills A, Ninness B. System identification of nonlinear state-space models. *Automatica*, 2011, **47**(1): 39−49

8　Billings S A. *Nonlinear System Identification: NARMAX Meth-*

ods in the Time, Frequency, and Spatio-Temporal Domains. Chichester: John Wiley & Sons, 2013.

9 Carini A, Orcioni S, Terenzi A, Cecchi S. Nonlinear system identification using Wiener basis functions and multiple-variance perfect sequences. *Signal Processing*, 2019, **160**: 137−149

10 Schoukens M, Tiels K. Identification of block-oriented nonlinear systems starting from linear approximations: A survey. *Automatica*, 2017, **85**: 272−292

11 Bershad N J, Celka P, McLaughlin S. Analysis of stochastic gradient identification of Wiener-Hammerstein systems for nonlinearities with Hermite polynomial expansions. *IEEE Transactions on Signal Processing*, 2001, **49**(5): 1060−1072

12 Valarmathi K, Devaraj D, Radhakrishnan T K. Intelligent techniques for system identification and controller tuning in pH process. *Brazilian Journal of Chemical Engineering*, 2009, **26**(1): 99−111

13 Zhu Y. Distillation column identification for control using Wiener model. In: Proceedings of the American Control Conference. San Diego, USA: IEEE, 1999. 3462−3466

14 Schoukens J, Suykens J, Ljung L. Wiener-Hammerstein benchmark. In: Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009). Malo, France: 2009.

15 Hagenblad A, Ljung L, Wills A. Maximum likelihood identification of Wiener models. *Automatica*, 2008, **44**(11): 2697−2705

16 Xu W Y, Bai E W, Cho M. System identification in the presence of outliers and random noises: A compressed sensing approach. *Automatica*, 2014, **50**(11): 2905−2911

17 Bottegal G, Castro-Garcia R, Suykens J A K. A two-experiment approach to Wiener system identification. *Automatica*, 2018, **93**: 282−289

18 Westwick D T, Schoukens J. Initial estimates of the linear subsystems of Wiener-Hammerstein models. *Automatica*, 2012, **48**(11): 2931−2936

19 Giordano G, Gros S, Sjöberg J. An improved method for Wiener-Hammerstein system identification based on the fractional approach. *Automatica*, 2018, **94**: 349−360

20 Liu Q, Lin W Y, Jiang S L, Chai Y, Sun L. Robust estimation of Wiener models in the presence of outliers using the VB approach. *IEEE Transactions on Industrial Electronics*, 2021, **68**(11): 11390−11399

21 Xie L, Yang H Z, Huang B. FIR model identification of multirate processes with random delays using EM algorithm. *AIChE Journal*, 2013, **59**(11): 4124−4132

22 Agamennoni G, Nieto J I, Nebot E M. Approximate inference in state-space models with heavy-tailed noise. *IEEE Transactions on Signal Processing*, 2012, **60**(10): 5024−5037

23 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

24 Amari S I. Natural gradient works efficiently in learning. *Neural Computation*, 1998, **10**(2): 251−276

25 Amari S I. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 1982, **10**(2): 357−385

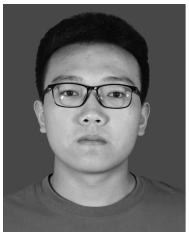
26 Bottou L. On-line learning and stochastic approximations. *On-line Learning in Neural Networks*. New York: Cambridge University Press, 1999.

27 Robbins H, Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951, **22**(3): 400−407

28 Hoffman M D, Blei D M, Wang C, Paisley J. Stochastic variational inference. *The Journal of Machine Learning Research*, 2013, **14**(1): 1303−1347

刘 切 重庆大学自动化学院副教授. 2016 年获得北京化工大学控制科学与工程专业博士学位. 主要研究方向为人工智能及其在复杂过程的控制和优化中的应用. 本文通信作者.
E-mail: qieliu@cqu.edu.cn
(**LIU Qie** Associate professor at the School of Automation, Chongqing University. He received his Ph.D. degree from Beijing University of Chemical Technology in 2016. His research interest covers artificial intelligence and its applications on the control and optimization of complex processes. Corresponding author of this paper.)

李俊豪 重庆大学自动化学院硕士研究生. 2020 年获得西安理工大学自动化与信息工程学院学士学位. 主要研究方向为系统辨识与人工智能.
E-mail: 202013021042@cqu.edu.cn
(**LI Jun-Hao** Master student at the School of Automation, Chongqing University. He received his bachelor degree from the School of Automation and Information Engineering, Xi′an University of Technology in 2020. His research interest covers system identification and artificial intelligence.)



王 浩 重庆大学自动化学院硕士研究生. 2021 年获得安徽师范大学物理与电子信息学院学士学位. 主要研究方向为模型预测与人工智能.
E-mail: 202113021007@cqu.edu.cn
(**WANG Hao** Master student at the School of Automation, Chongqing University. He received his bachelor degree from the School of Physics and Electronic Information, Anhui Normal University in 2021. His research interest covers model prediction and artificial intelligence.)

曾建学 重庆大学自动化学院硕士研究生. 2018 年获得华北科技学院电子信息工程学院学士学位. 主要研究方向为容器技术与人工智能.
E-mail: lc9zjx@126.com
(**ZENG Jian-Xue** Master student at the School of Automation, Chongqing University. He received his bachelor degree from the School of Electronic Information Engineering, North China Institute of Science and Technology in 2018. His research interest covers container and artificial intelligence.)

柴 毅 重庆大学自动化学院教授. 2001 年获得重庆大学博士学位. 主要研究方向为信息融合, 故障诊断, 智能控制系统.
E-mail: chaiyi@cqu.edu.cn
(**CHAI Yi** Professor at the School of Automation, Chongqing University. He received his Ph.D. degree from Chongqing University in 2001. His research interest covers information fusion, fault diagnosis, and intelligent control system.)