УДК 519.171.1

Анализ неопределенности детерминистических моделей с помощью аппроксимации гауссовскими процессами

Р.Ш. Кальметьев[1], ***Ю.Н. Орлов***[1]

[1]Московский физико-технический институт (государственный университет)

Введение и постановка задачи

В последнее время задачам анализа неопределенности математических моделей, используемых в различных расчетах при анализе безопасности объектов ядерной энергетики, придают все большее значение. Повышенное внимание к подобным задачам обусловлено, как практической необходимостью в оценке качества получаемых с помощью этих моделей результатов, так и в связи с развитием математического аппарата статистических методов, методов построения аппроксимаций и ростом вычислительных мощностей.

В данной работе рассматриваются задачи анализа неопределенности детерминистических моделей.

Под детерминистической моделью подразумевается функция $f: \Omega_x \to \Omega_y, \Omega_x \subseteq \mathbb{R}^n, \Omega_y \subseteq \mathbb{R}^m, n \in \mathbb{N}, m \in \mathbb{N}$, где n - число входных параметров модели, а m - число выходных параметров модели. Вид функции предполагается неизвестным, то есть детерминистическая модель рассматривается как «черный ящик» и ее внутренние свойства не используются при решении задач неопределенности модели.

Постановку задачи неопределенности детерминистических моделей можно сформулировать следующим образом. Пусть имеется набор детерминистических моделей $\{f_1, f_2, \dots, f_N\}$, описывающих одно и то же физическое явление. Необходимо построить оценку близости данных детерминистических моделей, причем сами зависимости $\{f_1, f_2, \dots, f_N\}$ предполагаются неизвестными, а оценка должна быть построена на основе имеющихся конечных выборок для каждой модели вида $\{(x_i, y_i) | i = 1, \dots, M\}$, где M - число точек в выборке, x - вектор входных параметров, y - вектор выходных параметров.

Предполагается, что значение функций задано неточно: $y = f(x) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$, а множества точек, в которых известны значения, не совпадают для различных моделей.

В работе [1] в оценки близости для двух детерминистических моделей был предложен коэффициент стохастической аппроксимации SAR следующего вида:

$$SAR = \left(1 - \frac{\sqrt{a_{f_1 - f_2}}}{\sqrt{a_{f_1}} + \sqrt{a_{f_2}}}\right)^2, \quad (1)$$

где $a_{f_1 - f_2}, a_{f_1}, a_{f_2}$ - моменты второго порядка:

$$a_{f_1 - f_2} = \int (f_1(x) - f_2(x))^2 \mathbb{P}(dx), \quad (2)$$

$$a_f = \int (f_1(x))^2 \mathbb{P}(dx), \quad (3)$$

$$a_{f_{\text{approx}}} = \int (f_2(x))^2 \mathbb{P}(dx), \quad (4)$$

и $\mathbb{P}(dx))$ - некоторая вероятностная мера.

Вероятностная мера $\mathbb{P}(dx)$может выбираться различными способами, в частности в работе [2] предлагается использовать для этого ядерную оценку плотности (метод Парзеновского окна) с ядром Бартлетта-Епанечникова, восстановленную по доступным выборкам данных, также может

использоваться некоторое априорное распределение, задаваемое исходя из интересующей с практической точки зрения области пространства параметров моделей.
Значение коэффициента стохастической аппроксимации принадлежит отрезку $[0,1]$. Если значение SAR близко к 1, то это означает, что соответствующие модели f близки и неопределенность модели низка. Если $SAR \ll 1$, то это означает, что результаты эксперимента и соответствующая модель не близки, и неопределенность высока.
Моменты в формуле (1) оцениваются статистически с использованием имеющихся выборок данных.
В случае, когда наборы входных параметров, на которых известны значения различных моделей, не совпадают, возникает необходимость строить аппроксимации рассматриваемых моделей для получения оценки второго момента в числителе формулы (1).
В работе [2] для построения приближений (аппроксимирующего отображения) детерминистических моделей или данных экспериментов используется метод стохастической аппроксимации, являющийся частным случаем метода аппроксимации Шепарда или метода взвешенных обратных расстояний [3]. Выбор данного метода аппроксимации обусловлен рядом полезных свойств получаемой аппроксимирующей функции [2], таких как непрерывность, асимптотическое стремление к среднему значению, возможность расчета ограничения сверху для ошибки аппроксимации.
Необходимо заметить, что значение коэффициента стохастической аппроксимации существенным образом зависит от способа построения аппроксимирующих моделей.
Данная работа посвящена разработке байесовского подхода к построению аппроксимаций используемых моделей. Аппроксимация строится в предположении о том, что аппроксимируемые модели могут приближенно рассматриваться как гауссовские процессы. Данный подход позволяет в части случаев улучшить «разрешающую способность» коэффициента стохастической аппроксимации (имеется в виду возможность отличить с помощью расчета данного коэффициента более точную модель от менее точной), а также построить доверительные множества для коэффициента стохастической аппроксимации.
Для расчета коэффициента стохастической аппроксимации необходимо наличие метода оценки расстояния между моделями, понимаемыми как отображения из пространства входных параметров в пространство целевых параметров $f: \Omega_x \to \Omega_y, \Omega_x \subseteq \mathbb{R}^n, \Omega_y \subseteq \mathbb{R}^m, n \in \mathbb{N}, m \in \mathbb{N}$, где n - число входных параметров модели, а m - число выходных параметров модели, в норме L^2, при этом на пространство параметров может быть наложена мера $\mathbb{P}(dx)$, отличная от равномерной. В случае если наборы значений входных параметров, для которых известны значения рассматриваемых детерминистических моделей, совпадают, то эта оценка может быть построена как выборочное среднее по выборке доступных значений:

$$\hat{a}_{f_1-f_2} = \frac{1}{M}\sum_{i=1}^{M}(y_i^1 - y_i^2)^2 \qquad (5)$$

В случае пространства параметров с заданной мерой $\mathbb{P}(dx)$ слагаемые в формуле (5) необходимо брать с соответствующими весами.
Но в общем случае наборы значений входных параметров, для которых известны значения рассматриваемых детерминистических моделей, не совпадают. В этом случае значение расстояния между моделями $\rho\big(f_1(x), f_2(x)\big) = |f_1(x) - f_2(x)|$ неизвестно ни в одной точке, и построение выборочной оценки SAR невозможно без построения аппроксимаций рассматриваемых детерминистических моделей.

Байесовский подход, аппроксимация гауссовскими процессами

Широко применяемый сегодня метод восстановления регрессии гауссовскими процессами является методом построения вероятностной меры, заданной на пространстве функций, при этом используется предположение, что восстанавливаемая функция принадлежит классу гауссовских случайных процессов [4].

Краткое введение в тему использования гауссовских процессов для различных задач аппроксимации можно найти в книге [5].
Гауссовским процессом ($\mathcal{GP}$) называется случайный процесс, чьи конечномерные распределения гауссовские.
Регрессия неизвестной функции в классе гауссовких процессов задается следующим образом:

$$\hat{f}(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{6}$$

где

$$m(x) = \mathbb{E}f(x),$$
$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].$$

Для восстановления регрессионной зависимости делается параметрическое предположение о виде функций $m(x)$ и $k(x, x')$.
Так как регрессия строится в классе гауссовских процессов, то совместное распределение известных значений функции и значений функции в интересующих точках будет нормальным:

$$\begin{bmatrix} Y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \tag{7}$$

После выбора вида корреляционной функции $k(x, x')$ решение задачи регрессии можно записать в явном виде:

$$\begin{gathered} f_* | X, Y, X_* \sim \mathcal{N}(\overline{f_*}, cov(f_*)) \\ \overline{f_*} = K(X_*, X)[K(X,X) + \sigma_n^2 I]^{-1} Y \\ cov(f_*) = K(X_*, X_*) - K(X_*, X)[K(X,X) + \sigma_n^2 I]^{-1} K(X, X_*) \end{gathered} \tag{8}$$

Выбор корреляционной функции является ключевым шагом в алгоритме восстановления регрессии с помощью гауссовских полей и происходит следующим образом. Выбирается некоторое достаточно обширное параметрическое семейство корреляционных функций, например,

$$k(x_p, x_q) = \sigma_f^2 exp(-\frac{1}{2}(x_p - x_q)^T diag(l)^{-2}(x_p - x_q) \tag{9}$$

Для определения оптимальных значений гиперпараметров σ_f^2 и l существует несколько общепринятых методик: метод максимального правдоподобия, кросс-валидация, leave-one-out как частный случай кросс-валидации. При этом как правило численно решается соответствующая оптимизационная задача [4].
При решении задач анализа неопределенности нет необходимости максимально точно строить регрессионную модель рассматриваемых детерминистических моделей, непосредственной целью является построение регрессии для расстояния между моделями $\rho(f_1(x), f_2(x))$. Основная сложность состоит в том, что восстановление зависимости $\rho(f_1(x), f_2(x))$ нельзя свести к решению классической задачи регрессии, так как значения $\rho(f_1(x), f_2(x))$ неизвестны, а доступны лишь значения $f_1(x)$ и $f_2(x)$ на разных наборах точек.
В случае если $f_1(x)$ и $f_2(x)$ принадлежат классу гауссовских процессов, то $f_1(x) - f_2(x)$ тоже принадлежит этому классу. Тогда для $f_1(x) - f_2(x)$ можно явно выписать функцию

правдоподобия при известных значениях гиперпараметров. При этом из общих соображений логичным видится взять для аппроксимации $f_1(x)$ и $f_2(x)$ одни и те же значения некоторых гиперпараметров, так как эти детерминистические модели описывают одно и то же физическое явление.

Оптимизация гиперпараметров на основе данных нескольких моделей

Оптимизация гиперпараметров проводится на основе всех имеющихся выборок для различных детерминистических моделей. Данный подход позволяет с большей точностью оценивать оптимальные значения гиперпараметров. Также при данном подходе все аппроксимирующие модели строятся при помощи ковариационных матриц с одинаковыми свойствами, определяемыми общими значениями гиперпараметров.

Пусть даны две выборки значений $(x^1, y^1)\{(x_i, y_i)|i = 1, \dots, M_1\}$ и $(x^2, y^2) = \{(x_j, y_j)|j = 1, \dots, M_2\}$, где M_1 - число точек в первой выборке, M_2 - число точек во второй выборке, x - вектор входных параметров, y - вектор выходных параметров.

Предполагается следующая модель генерации данных:

$$\begin{aligned} y_i &= f_1(x_i) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_1^2) \\ y_j &= f_2(x_j) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_2^2) \end{aligned} \quad (10)$$

f_1 и f_2 - являются реализациями гауссовских процессов с одинаковыми ковариационными функциями:

$$k(x_p, x_q) = \sigma_f^2 exp(-\frac{1}{2}(x_p - x_q)^T diag(l)^{-2}(x_p - x_q) \quad (11)$$

Тогда функция правдоподобия записывается следующим образом:

$$\begin{aligned} &\log p(y^1, y^2|x^1, x^2) = \\ -\ &\frac{1}{2}(y^1)^T(K_1 + \sigma_1^2 I)y^1 - \log|K_1 + \sigma_1^2 I| - \frac{M_1}{2}\log 2\pi - \\ -\ &\frac{1}{2}(y^2)^T(K_2 + \sigma_2^2 I)y^2 - \log|K_2 + \sigma_2^2 I| - \frac{M_2}{2}\log 2\pi \end{aligned} \quad (12)$$

Оптимальные значения гиперпараметров $(\sigma_1, \sigma_2, \sigma, l)$ находятся численно с помощью поиска точки максимума этой функции (рисунок 1).

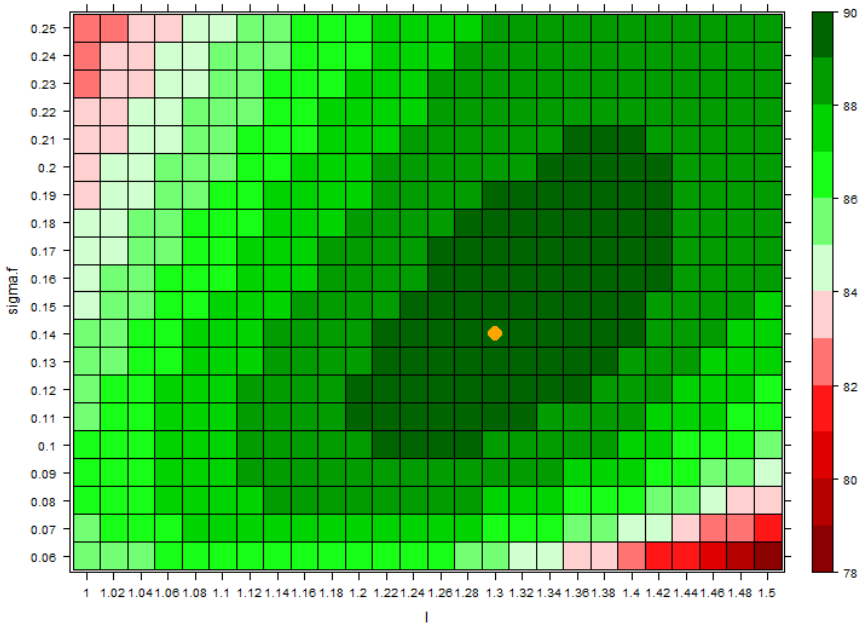

Рис. 1. Зависимость функции правдоподобия от значений гиперпараметров (σ, l)

Регрессию функции ошибки между двумя моделями в этом случае можно выписать явно:

$$\begin{aligned} \mu_{difference} &:= \mathbb{E}(f_1 - f_2)_*(x) \quad = \quad \mathbb{E}f_{1*} - \mathbb{E}f_{2*} \\ \Sigma_{difference} &:= cov((f_1 - f_2)_*) \quad = \quad cov(f_{1*}) + cov(f_{1*}) \end{aligned} \tag{13}$$

Алгоритм легко обобщается на большее число моделей, при этом просто в формуле (12) добавляются соответствующие слагаемые.
Оценка для второго момента в числителе формулы (1) равна:

$$\tilde{a}_{f_1 - f_2} = tr(\Sigma_{difference}) + \mu^T_{difference}\mu_{difference} \tag{14}$$

Дисперсия получаемой оценки:

$$D(\tilde{a}_{f_1 - f_2}) = 2tr(\Sigma_{difference})^2 + 4\mu^T_{difference}\Sigma_{difference}\mu_{difference} \tag{15}$$

Вывод формул (14) и (15) можно найти в книге [6].
Возможность получить оценку дисперсии величины $\tilde{a}_{f_1 - f_2}$ является преимуществом данного метода по сравнению с подходом, описанным в [1], т.к. при применении метода стохастической аппроксимации возможна лишь оценка точности аппроксимации сверху с использованием оценки для константы Липшица рассматриваемых моделей [2].

Численный эксперимент

Для сравнения «разрешающих способностей» коэффициентов стохастической аппроксимации, рассчитанных с помощью аппроксимации методом Шепарда и с помощью построения гауссовской модели, был проведен описываемый ниже численный эксперимент.
Пусть имеются три независимые выборки данных для следующих моделей:

$$\begin{aligned} y^{experiment} &= f(x) \\ y^{accuratemodel} &= f_1(x) = f(x) + \mathcal{GP}^{accuratemodel}(x_j) \\ y^{crudemodel} &= f_2(x) = f(x) + \mathcal{GP}^{crudemodel}(x_j), \end{aligned} \quad (17)$$

Выборки данных представляют из себя наборы пар (значения входных параметров, значение выходного параметра) для каждой модели. Множества точек в пространстве входных параметров, в которых известны значения моделей не совпадают для различных моделей.
Предполагается, что между входными параметрами и выходным параметром существует некоторая детерминистическая зависимость $f(x)$. При этом в эксперименте имеется возможность измерить значения $f(x)$. Точная и грубая расчетные детерминистические модели предполагают возможность вычисления функций $f_1(x)$ и $f_2(x)$, отличающихся от истинной $f(x)$ наличием смещений, представленных гауссовскими процессами $\mathcal{GP}^{accuratemodel}(x_j)$ и $\mathcal{GP}^{crudemodel}(x_j)$.
Ковариационная функция этих гауссовских процессов представляется в виде:

$$k(x_p, x_q) = \sigma_f^{model} exp(-\frac{1}{2}(x_p - x_q)^T diag(l)^{-2}(x_p - x_q) \quad (18)$$

$\mathcal{GP}^{accuratemodel}(x_j)$ и $\mathcal{GP}^{crudemodel}(x_j)$.отличаются различными значениями σ_f^{model}. Для грубой модели значение σ_f^{model} выше.
В численном эксперименте производится многократное моделирование независимых выборок из (17) и тестируется возможность отличить грубую модель от точной по сгенерированным выборкам с помощью различных методик, описанных выше, в зависимости от размерности рассматриваемой задачи (размерности векторов x) и количества точек в генерируемых выборках.
Гауссовская модель строится с помощью оптимизации гиперпараметров общей для всех моделей функции логарифмического правдоподобия.

Для примера на рисунке 2 изображены графики зависимости доли правильных угадываний от числа точек для размерностей 9 и 10. Аналогичные численные эксперименты были проведены и для всех меньших размерностей. На рисунке 3 изображены поверхности доли правильных угадываний в зависимости от числа точек и размерности входных данных.

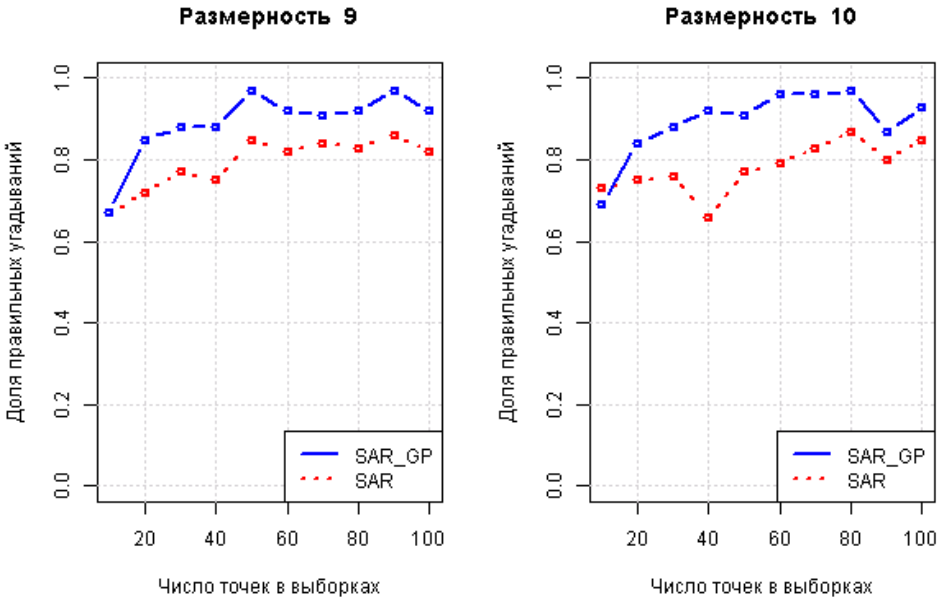

Рис. 2. Зависимость доли правильных угадываний от числа точек в выборках

Из результатов видно, что "разрешающая способность" коэффициента стохастической аппроксимации, рассчитанного с помощью аппроксимации гауссовскими процессами выше, чем для этого же коэффициента, рассчитанного с помощью аппроксимации Шепарда.

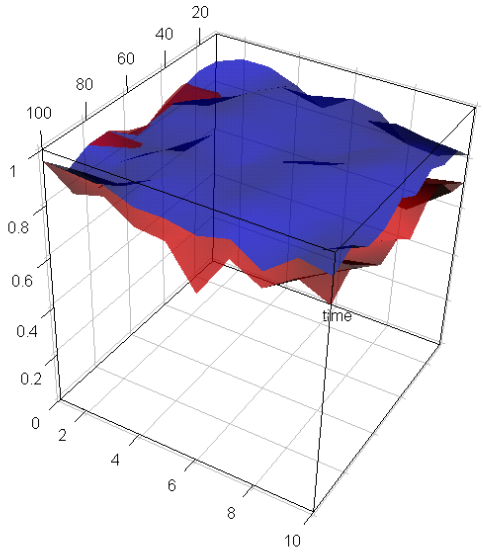

Рис. 3. Зависимость доли правильных угадываний от размерности моделей и от числа точек в генерируемых выборках, синий - с использованием аппроксимации Шепарда, красный - с использованием аппроксимации гауссовскими процессами

Кросс-верификация данных по ядерным реакциям

Международное сотрудничество в сфере накопления результатов по ядерным реакциям было инициировано МАГАТЭ в 60-х годах прошлого века.

Были созданы общепринятый формат хранения экспериментальных и расчетных данных EXFOR/ENDF и библиотеки, представляющие собой базы данных, содержащие всю накопленную информацию по ядерным реакциям.

Библиотеки по ядерным данным (данным по ядерным реакциям) содержат большое количество накопленных различными лабораториями количественных результатов ядерных физических исследований. В данные библиотеки включены как экспериментальные результаты, так и данные, полученные с помощью расчетов, проведенных согласно различным теоретическим моделям.

Данные из этих библиотек используются как для практических целей (моделирование реакций распада и синтеза в активной зоне ядерных реакторов, расчет эффективности радиоактивной защиты, оценка безопасности и т.д.), так и для построения новых теоретических моделей или уточнения существующих [7].

На сегодняшний день существует по несколько версий различных библиотек по ядерным данным. Сравнение данных из разных библиотек показывает, что в данные в них существенно различаются. Предлагаемый в данной работе метод позволяет количественно оценить различия в предоставляемых данных, выделить группы наиболее согласованных результатов, а также оценить статистическую значимость получаемых результатов по сравнению с заявленными методическими ошибками с помощью Монте-Карло моделирования.

Алгоритм анализа неопределенности библиотек по ядерным данным:

1. Выгрузка данных из доступных библиотек по конкретной ядерной реакции.
2. Аппроксимация всех выборок данных с помощью гауссовых процессов:
    - определение вида общей функции правдоподобия;
    - оптимизация гиперпараметров;
    - расчет средних значений и ковариационной матрицы в интересующих точках.

3. Расчет коэффициента SAR.
4. При необходимости оценка точности рассчитанного коэффициента SAR с помощью Монте-Карло моделирования различных реализаций аппроксимации выборок данных, представляющих собой реализации гауссовских процессов.
5. При необходимости оценка значимости рассчитанного коэффициента SAR с помощью Монте-Карло моделирования различных реализаций аппроксимации выборок данных, представляющих собой реализации гауссовских процессов.

Для иллюстрации работы алгоритма выбрана реакция рассеяния нейтрона на ядре урана 235 с образованием еще одного нейтрона: U-235(N,2N)U-234.

Значения полученных коэффициентов стохастической аппроксимации сведены в таблицу на рисунке 4. При этом использована следующая цветовая шкала: зеленый соответствует хорошо согласующимся данным, красный - случаям сильных расхождений в данных, и желтый - промежуточным случаям.

	CENDL-3.1	ENDFB-V.2	ENDFB-VI.8	ENDFB-VII.0	ENDFB-VII.1	JEFF-3.2	JENDL-3.3	JENDL-4.0	ROSFOND-2010	NSE	AWRE
CENDL-3.1	0.93	0.54	0.92	0.93	0.93	0.68	0.92	0.92	0.93	0.73	0.49
ENDFB-V.2	0.54	0.90	0.56	0.54	0.54	0.46	0.50	0.53	0.54	0.48	0.36
ENDFB-VI.8	0.92	0.56	0.94	0.93	0.93	0.69	0.90	0.92	0.93	0.70	0.49
ENDFB-VII.0	0.93	0.54	0.93	0.94	0.94	0.70	0.90	0.93	0.94	0.69	0.49
ENDFB-VII.1	0.93	0.54	0.93	0.94	0.94	0.70	0.90	0.93	0.94	0.69	0.49
JEFF-3.2	0.68	0.46	0.69	0.70	0.70	0.96	0.68	0.68	0.70	0.59	0.50
JENDL-3.3	0.92	0.50	0.90	0.90	0.90	0.68	0.95	0.92	0.90	0.65	0.41
JENDL-4.0	0.92	0.53	0.92	0.93	0.93	0.68	0.92	0.93	0.93	0.66	0.40
ROSFOND-201	0.93	0.54	0.93	0.94	0.94	0.70	0.90	0.93	0.94	0.69	0.49
NSE	0.73	0.48	0.70	0.69	0.69	0.59	0.65	0.66	0.69	0.94	0.72
AWRE	0.49	0.36	0.49	0.49	0.49	0.50	0.41	0.40	0.49	0.72	0.87

Рис. 4. Парные значения вычисленных коэффициентов SAR

С помощью алгоритма кластеризации можно получить иерархическое дерево моделей и выделить группы близких друг к другу моделей. Пример дерева, построенный по полученным расчетным результатам коэффициентов стохастической аппроксимации представлен на рисунке 5.

Из полученных данных можно сделать следующие выводы: большинство расчетных моделей лучше согласуются друг с другом нежели с данными экспериментов (AWRE и NSE - экспериментальные данные), расчетные данные модели ENDFB-V2 значительно отличаются от данных в следующих версиях этой библиотеки и от данных из других библиотек, также отдельно выделяется модель JEFF-3.2, которая имеет низкие коэффициенты стохастической аппроксимации как с экспериментальными данными, так и с расчетными данными других моделей.

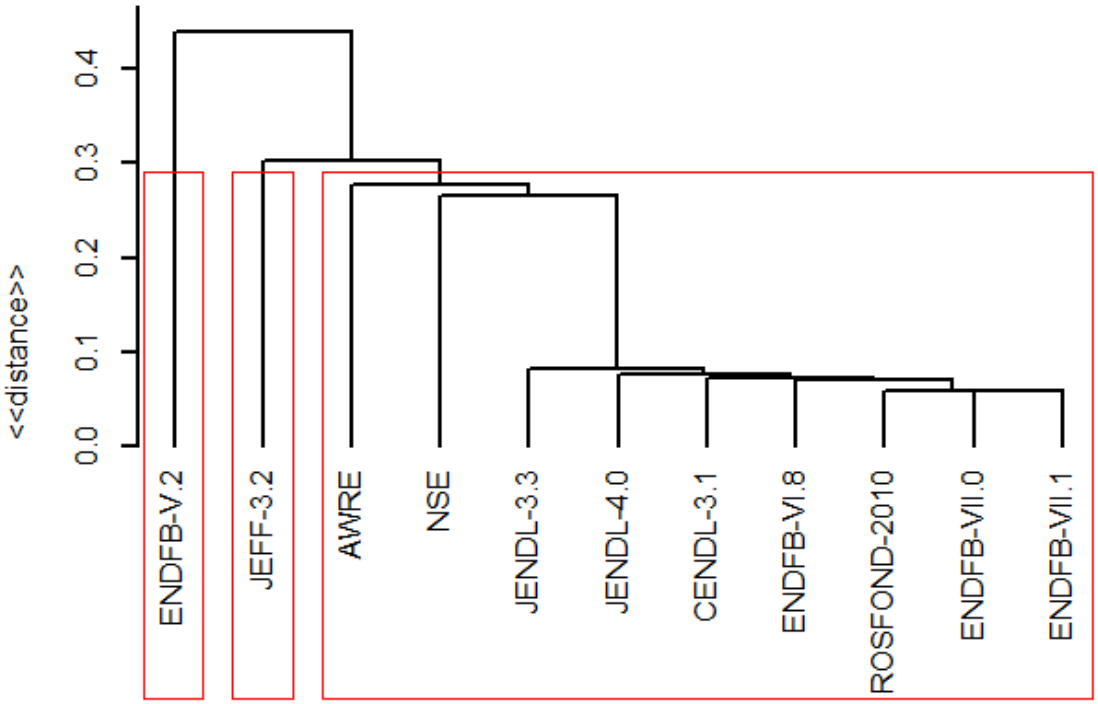

Рис. 5. Кластеризация расчетных моделей и данных экспериментов

Данным алгоритм оценки близости моделей может быть применен и другим ядерным реакциям, имеющимся в базах, что позволит глубже понять степень схожести или различия в данным, предоставляемых данными библиотеками.

Критерий согласия для нестационарных временных рядов

Рассмотрим следующую задачу. Пусть имеется два временных ряда $x_1(t)$ и $x_2(t)$ (под временным рядом будем понимать реализацию некоторого случайного процесса с дискретным временем). Предполагается, что значения рядов являются наборами реализаций независимых, но не одинаково распределенных случайных величин. Задача заключается в определении являются ли данные временные ряды реализациями одного нестационарного случайного процесса или нет.

Центральным понятием в рассматриваемом подходе будет распределение случайного процесса в конкретный момент времени, меняющееся с течением времени, и восстанавливаемое из выборки предыдущих значений.

Восстановление функции распределения в конкретный момент времени по известным предыдущим значениям временного ряда является самостоятельной задачей и не будет подробно рассмотрено, методы решения данной задачи можно найти в [8].

В данной главе под функцией распределения, восстановленной по предыдущим значениям, будем понимать просто выборочную функцию распределения, построенную по некоторому множеству предыдущих значений:

$$\hat{F}_n(x,t) = \frac{1}{n}\sum_{i=t-n}^{t-1} \mathbf{1}_{\{x(i)\le x\}} = \frac{1}{n}\sum_{i=1}^{n} H(x - x(i)) \qquad (19)$$

где $\mathbf{1}_A$ - индикатор события A, $H(x)$ - функция Хевисайда.

Для общего случая нестационарного временного ряда функция распределения в конкретный момент времени вообще говоря не может быть восстановлена из выборки исторических значений с какой-либо точностью, так как на это распределение не наложено никаких ограничений, и оно может не иметь ничего общего с распределениями в предыдущие моменты времени.
Поэтому целесообразно ввести некоторое свойство квазистационарности на рассматриваемые процессы. Идеологически это свойство заключается в том, что функция распределения равномерно непрерывно меняется со временем.
В монографии [8] предложено следующее определение:
Выборочную плотность функции распределения (гистограмму) временного ряда $f_T(x,t)$, построенную в скользящем окне длины T, будем называть $\theta-\varepsilon$-стационарной на временном промежутке θ, если

$$\forall\tau: 1 \leqslant \tau \leqslant \theta, \forall t \int_0^1 |f_T(x,t+\tau) - f_T(x,t)|dx \leqslant \varepsilon \qquad (20)$$

Для решения вышеописанной задачи по определению того, являются ли два временных ряда реализациями одного и того же случайного процесса, можно предложить следующий алгоритм:

- с помощью минимизации усредненного значения интеграла из (20) определить оптимальные объемы выборок для восстановления плотностей рассматриваемых процессов;
- определить расстояния между восстановленными плотностями в каждой точке t;
- далее необходимо некоторым образом осуществить переход от оценки расстояний между плотностями распределения в каждый момент времени к некоторому интегральному значению для всего рассматриваемого временного промежутка, простейшим способ является просто усреднение всех расстояний;
- определенную таким образом статистику можно использовать для построения статистического критерия, критерии значимости могут быть построены методом бутстрэпа.

Стоит отметить, что определение $\theta-\varepsilon$ -стационарности не накладывает каких-либо ограничений на рассматриваемые ряды, т.к. любой временной ряд будет $\theta-\varepsilon$ -стационарным при надлежащем выборе T. Этот эффект возникает из-за того, что выборочные плотности распределений в моменты времени, отстающие друг от друга на τ, строятся по выборкам, имеющим $(T-\tau)$ общих точек. Соответственно доля общих точек $\frac{T-\tau}{T}$ растет при увеличении T, и расстояния между выборочными плотностями можно сделать сколь угодно малым. Влияние этого эффекта становится критическим при прогнозировании функции распределения на один шаг вперед.
Введем несколько другое определение квазистационарности, которое позволит избежать подобных проблем при прогнозировании распределения на небольшое число шагов.
Пусть в каждой точке временного ряда имеется восстановленная по предыдущим точкам функция распределения $\hat{F}_n(x,t)$. Тогда можно построить критерий, проверяющий простую гипотезу, заключающуюся в том, что значения ряда во всех точках принадлежат соответствующим восстановленным распределениям в этих точках.
В случае если ряд является стационарным, в узком смысле, и оценка функции распределения в каждой точке является одинаковой $(\hat{F}_n(x,t) = \hat{F}_n(x))$, то задача сводится к проверке критерия согласия между выборочным распределением построенным на окне определенной длины и значениями ряда. В случае же нестационарного ряда оценка функции распределения будет меняться от точки к точке, из-за этого применить какой-либо критерий согласия в явном виде становится невозможным.
Сначала введем критерий согласия для нестационарного ряда $x(t)$ с некоторым распределением $F(x,t)$. Напомним, что мы изначально предполагаем, что значения ряда являются независимыми величинами, а, следовательно, можно ограничится рассмотрением распределений, представимых в виде произведения маржинальных

$$F(x(t_1), x(t_2)) = F(x, t_1) * F(x, t_2). \quad (21)$$

Если бы рассматриваемый процесс был стационарным, то

$$F(x, t_1) = F(x, t_2) = F(x) \quad (21)$$

и в качестве критерия согласия ряда и распределения $F(x,t)$ может выступать критерий согласия Колмогорова для эмпирической функции распределения и $F(x)$.
В случае нестационарного ряда

$$F(x, t_1) \neq F(x, t_2). \quad (21)$$

Для построения критерия согласия необходимо перейти от значений самого ряда к квантильным значениям в каждый момент времени по распределению $F(x,t)$. Тогда в случае, если рассматриваемый временной ряд действительно является реализацией случайного процесса с распределением $F(x,t)$, то значения нового квантильного ряда должны быть иметь равномерное распределение на отрезке $[0,1]$. И соответственно в качестве критерия согласия можно взять тот же критерий Колмогорова, но примененный к выборке квантильных значений и равномерному распределению.
На практике нам как правило неизвестно $F(x,t)$, но тем не менее на основе значений ряда возможно аналогичным образом определить критерий согласия между значениями ряда и распределением $\hat{F}_n(x,t)$, восстанавливаемым из предыдущих значений.

Если восстановленное распределение $\hat{F}_n(x,t)$ близко к истинному, то значение квантиля, рассчитанного по этому восстановленному распределению, от значения ряда в данной точке будет иметь распределение близкое к равномерному на отрезке $[0,1]$.
Тогда становится возможным ввести следующее определение:
Временной ряд будем называть квазистационарным, если для значений квантилей восстановленных распределений $\hat{F}(x(t),t)$, выполняется критерий согласия с равномерным распределением на отрезке $[0,1]$ с уровнем значимости α.
Вернемся к задаче построения критерия для определения, являются ли различные временные ряды реализациями одного и того же нестационарного случайного процесса.
Введем понятие согласованности, аналогичное понятию квазистационарности, но в котором оценки функций распределений в различные моменты времени строятся по значениям одного временного ряда, а критерий согласия с этими восстановленными распределениями применяется для значений другого временного ряда.
Временной ряд $x_1(t)$ будем называть согласованным с временным рядом $x_2(t)$, если для значений квантилей $\hat{F}(x_1(t),t)$ распределений, восстановленных по точкам $x_2(t)$, выполняется критерий согласия с равномерным распределением на отрезке $[0,1]$ с уровнем значимости α.
Временные ряды $x_1(t)$ и $x_2(t)$ будем называть взаимно согласованными с уровнем значимости α, если $x_1(t)$ является согласованным с $x_2(t)$, и $x_2(t)$ является согласованным с $x_1(t)$ с уровнями значимости не больше α.
На практике большинство рассматриваемых временных рядов являются сильно нестационарными, и уровни значимости классических критериев согласия распределения значений квантилей восстановленных распределений $\hat{F}(x(t),t)$ (обозначим его $\mathcal{F}_q$) с равномерным распределением на отрезке $[0,1]$ (обозначим его $U[0,1]$)получаются очень низкими. Что делает их применение для решения вышеописанной задачи практически невозможным.
Далее описан метод построения критических значений для критерия согласия построенного на статистике разности распределений $\mathcal{F}_q$ и $U[0,1]$ в некоторой норме $(L_1, L_2, L_\infty, \ldots)$.

От значений ряда $x_1(t)$ перейдем к значениям квантилей по восстановленным распределениям $F_q(t)$. Как было описано выше распределение значений ряда, преобразованного таким образом, в случае квазистационарного ряда должны быть близки к $U[0,1]$. В каждом конкретном случае возможно посчитать конкретное значение расстояния ρ между $\mathcal{F}_q$ и $U[0,1]$, но для построения статистического критерия для сравнения различных временных рядов необходимо иметь оценку распределения для данного расстояния, получаемого в случае верности гипотезы о том, что ряда являются различными реализациями одного случайного процесса.

Это распределения может быть получено с помощью метода бутстрэпа [9]: из выборки $F_q(t)$ генерируется множество новых выборок того же размера с помощью выбора с возвращением. Для каждой генерируемой выборки рассчитывается расстояние ρ_{sample} между $\mathcal{F}_{q,sample}$ и $U[0,1]$. Для полученного множества расстояний строится выборочное распределение. Это распределение может быть использовано для построения критериев значимости для статистики ρ (построенный на основе этих критериев значимости статистический критерий обозначим как Cr).

Таким образом итоговый алгоритм решения рассматриваемой задачи можно описать следующим образом:

1. определить для ряда $x_1(t)$ оптимальную длину скользящей выборки для построения выборочного распределения с помощью минимизации расстояния между $F_q(t)$ и $U[0,1]$ в некоторой норме;
2. построить распределение для этого расстояния ρ методом бутстрэпа;
3. для значения расстояния между распределениями квантильных значений ряда $x_2(t)$ по выборочному распределению $x_1(t)$ и $U[0,1]$ рассчитать значение критерия Cr;
4. полученное значение критерия Cr определяет уровень значимости α_1 для согласованности $x_1(t)$ с $x_2(t)$;
5. аналогичным образом рассчитывается уровень значимости α_2 для согласованности $x_2(t)$ с $x_1(t)$;
6. гипотеза о том, рассмотренные временные ряды являются реализациями одного и того же нестационарного случайного процесса, может быть принята на уровне значимости $max(\alpha_1, \alpha_1)$.

Рассмотрим следующую модель генерации нестационарного временного ряда:

$$x(t) \sim \mathcal{N}(vt, 1), \tag{22}$$

где v - величина некоторой постоянной скорости смещения математического ожидания $x(t)$.

В качестве сильно нестационарного временного ряда будем рассматривать случайное блуждание:

$$x(t) = x(t-1) + \varepsilon_t, \varepsilon_t \sim \mathcal{N}(0,1) \tag{22}$$

На рисунке 6 представлен характерный профиль расстояний между $F_q(t)$ и $U[0,1]$ в норме L_1 для различных значений m и для случайного блуждания.

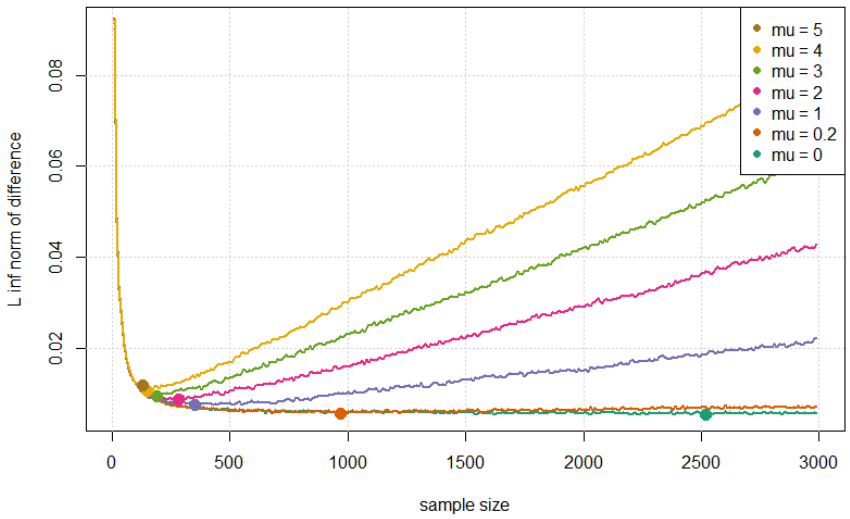

Рис. 6. Профиль расстояний между $F_q(t)$ и $U[0,1]$ в норме L_1 для различных значений m и для случайного блуждания

На рисунке 7 представлены результаты кластеризации кластеризации временных рядов с различным значением v.

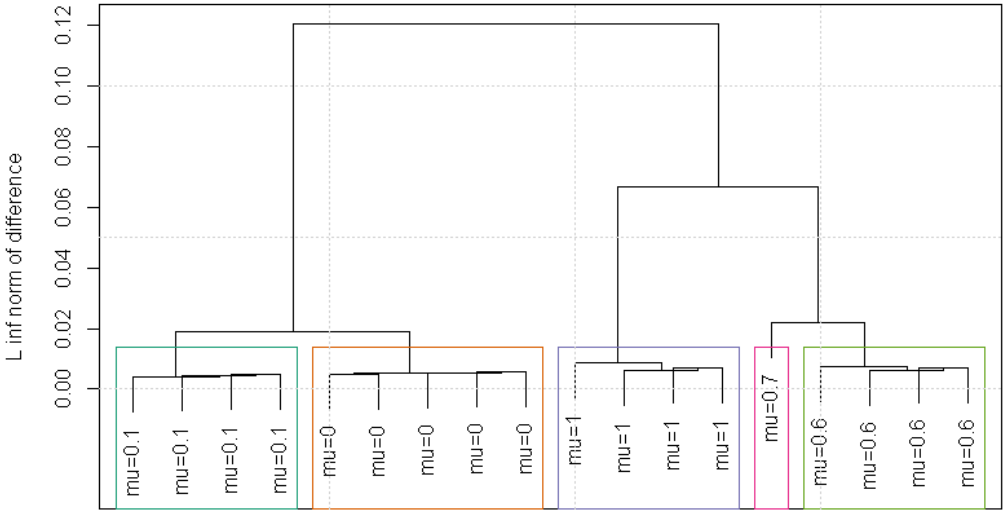

Рис. 7. Результаты кластеризации кластеризации временных рядов с различным значением v

Заключение

В данной работе предложен новый метод для анализа неопределенности детерминистических моделей на основе построения аппроксимации гауссовскими процессами. Данный метод является модификацией алгоритма, описанного в [1]. Новизна предложенного метода заключается в модели аппроксимации нескольких детерминистических моделей гауссовскими процессами с общей функцией логарифмического правдоподобия.

Преимуществами метода являются:

- возможность построения доверительных интервалов, а не точечных оценок для коэффициентов стохастической аппроксимации;

- также возможно построение оценки для распределения коэффициентов стохастической аппроксимации с помощью Монте-Карло моделирования;
- возможность неявного учета всех выборок данных для различных моделей при построении аппроксимации одной конкретной модели в свойствах корреляционной функции аппроксимирующих гауссовских процессов.

Нестационарный критерий согласия временных рядов является новым и продолжает идеи, описанные в [8].

Литература

1. R. Islamov V. Ustinov. Uncertainty analysis and stochastic approximation. Int. Mtg. Best-Estimate Methods in Nuclear Installation Safety Analysis (BE2000), Washington, DC. 2000.
2. А.А. Волков. Разработка математических моделей и методик стохастического моделирования для вероятностного анализа безопасности и надежности объектов энергетики, 2004.
3. Shepard Donald. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 ACM National Conference.
4. Rasmussen C. E., Williams C. Gaussian Processes for Machine Learning. the MIT Press, 2006.
5. MacKay D. J. C. Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge, UK, 2003.
6. A.M. Mathai Serge B. Provost. Quadratic Forms in Random Variables. Theory and Applications. Statistics, a Series of Textbooks and Monographs, 1992.
7. Trkov A. From Basic Nuclear Data to Applications. Nuclear Reaction Data and Nuclear Reactors, 2001.
8. Ю.Н. Орлов К.П. Осминин. Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. Математическое моделирование, 9, 2008.
9. Efron B.; Tibshirani R. An Introduction to the Bootstrap. Boca Raton, FL: Chapman and Hall/CRC, 1993.