

Регрессия на основе нестационарных гауссовских процессов с байесовской регуляризацией¹

Е.В. Бурнаев, М.Е. Панов, А.А. Зайцев

Институт проблем передачи информации, Российская академия наук, Москва, Россия

Поступила в редколлегию 26.06.2015

Аннотация—Одним из основных методов построения нелинейных регрессионных моделей является регрессия на основе гауссовских процессов. В данной работе предложен подход к байесовской регуляризации параметров ковариационной функции гауссовского процесса, который позволяет увеличить вычислительную устойчивость алгоритма построения регрессионной модели и её точность. Также предложен метод моделирования нестационарной ковариационной функции гауссовского процесса на основе линейного разложения по словарю параметрических функций. Использование такой ковариационной функции позволяет более эффективно моделировать пространственно неоднородные зависимости. Сочетание приведенных выше методов позволяет получить общий подход к регрессии на основе гауссовских процессов, который может быть легко реализован на практике и не требует настройки каких-либо параметров. Тестирование на большом наборе функций показало эффективность предложенного подхода.

Ключевые слова: гауссовские процессы, регрессия, байесовская регуляризация, априорное распределение, байесовская регрессия

1. ВВЕДЕНИЕ

Одной из основных задач, решаемых при построении суррогатных моделей (метамоделей, моделей на основе данных), является задача аппроксимации неизвестной зависимости по данным (построение функции регрессии) [1], которую можно описать следующим образом. Пусть $y = f(\mathbf{x})$ — некоторая неизвестная функция со входом $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$ и выходом $y \in \mathbb{R}$. Значения этой функции известны в точках обучающей выборки $D_{train} = (X, \mathbf{y}) = \{(\mathbf{x}_i, y_i = f(\mathbf{x}_i)), i = 1, \dots, N\}$. Необходимо по обучающей выборке D_{train} построить аппроксимацию $\hat{y} = \hat{f}(x) \approx f(x)$ для целевой функции $f(x)$. Качество аппроксимации оценивается по независимой тестовой выборке $D_{test} = (X_*, \mathbf{y}_*) = \{(\mathbf{x}_j^*, y_j^* = f(\mathbf{x}_j^*)), j = 1, \dots, N_*\}$, например, с помощью средней абсолютной ошибки

$$\varepsilon(\hat{f}|D_{test}) = \frac{1}{N_*} \sum_{j=1}^{N_*} |y_j^* - \hat{f}(\mathbf{x}_j^*)|, \quad (1)$$

95% и 99% квантилей абсолютной ошибки аппроксимации.

Широко распространенный подход к описанию класса моделей аппроксимации, в котором ищется решение поставленной задачи, состоит во введении априорного распределения на пространстве функций. В литературе обычно используется априорное гауссовское распределение, в результате чего задача сводится к непараметрической регрессии на основе гауссовских процессов [2–8]. Данный метод широко используется при решении прикладных задач, включая

¹ Исследование выполнено в ИППИ РАН исключительно за счет гранта Российского научного фонда (проект № 14-50-00150)

концептуальное проектирование [9], структурную оптимизацию [10], многокритериальную оптимизацию при проектировании [11], конструирование в аэрокосмической [12] и автомобильной отраслях [13], и т.д.

Общий подход к построению моделей с помощью регрессии на основе гауссовских процессов описан в литературе [14], однако мало внимания уделено его практической реализации. Повсеместно используемые программные пакеты, например, DACE [15], в некоторых случаях порождают вырожденные аппроксимации, интерпретация которых с точки зрения исходной задачи является затруднительной. Имеющиеся теоретические результаты [16] показывают, что выбор гамма-распределения в качестве априорного распределения на параметры ковариационной функции приводит к минимаксно оптимальной скорости концентрации апостериорного распределения, при этом такое априорное распределение играет роль своего рода регуляризации, которая потенциально должна препятствовать вырождению модели аппроксимации. Однако практическая реализация соответствующего алгоритма регуляризации требует аккуратного выбора гипераприорного распределения следующего уровня на параметры гамма-распределения и надежной процедуры оптимизации, которая позволит получить робастный локальный максимум правдоподобия. В данной работе предлагается ввести гипераприорное логнормальное распределение с фиксированными параметрами на параметры априорного гамма-распределения и оптимизировать совместное правдоподобие данных и модели как по параметрам ковариационной функции, так и по параметрам априорного гамма-распределения. Результаты экспериментов показывают, что это позволяет существенно повысить точность аппроксимации и избежать случаев её вырождения.

Другой важной особенностью общего подхода к регрессии на основе гауссовских процессов является преобладание методов, основанных на модели со стационарной ковариационной функцией [14]. Существующие подходы к моделированию нестационарной ковариационной функции как правило являются избыточно параметризованными [17], что приводит к большим вычислительным затратам при оценке параметров модели и итоговой нестабильности соответствующей оптимизационной процедуры. В данной работе мы предлагаем метод моделирования нестационарной ковариационной функции на основе линейного разложения по словарю параметрических функций. При заданном словаре предлагаемая процедура позволяет автоматически определить вклад стационарной и нестационарной частей ковариационной функции в общую изменчивость аппроксимационной модели.

Работа состоит из следующих частей:

- В разделе 2 описана регрессия на основе гауссовских процессов, а также описан предложенный нами метод построения нестационарной ковариационной функции,
- Раздел 3 содержит описание предложенного нами алгоритма оценки параметров регрессии на основе гауссовских процессов с помощью байесовского подхода,
- В разделе 4 описаны результаты проведенных вычислительных экспериментов.

2. АППРОКСИМАТОР НА ОСНОВЕ ГАУССОВСКОГО ПРОЦЕССА

Регрессия на основе гауссовских процессов позволяет строить нелинейные аппроксимационные модели и обеспечивает аналитическое выражение для оценки неопределенности аппроксимации в заданной точке $\mathbf{x}$.

Гауссовский процесс $f(\cdot)$ является стохастическим процессом, в котором с каждой точкой пространства $\mathbf{x}$ ассоциирована случайная нормальная величина. Более того, любой конечный набор значений гауссовского процесса в различных точках пространства имеет многомерное нормальное распределение. Таким образом, случайный процесс $f(\cdot)$ определяется средним $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ и ковариационной функцией $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. Для

нулевого среднего значения $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = 0$ и фиксированной ковариационной функции апостериорное среднее значений гауссовского процесса в точках тестовой выборки X_* имеет вид [14]:

$$\hat{f}(X_*) = K_* K^{-1} \mathbf{y}, \quad (2)$$

где

$$K_* = K(X_*, X) = [k(\mathbf{x}_j^*, \mathbf{x}_i), i = 1, \dots, N; j = 1, \dots, N_*],$$

$$K = K(X, X) = [k(\mathbf{x}_i, \mathbf{x}_j), i, j = 1, \dots, N].$$

Обычно при моделировании мы имеем доступ только к зашумленным значениям функции:

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad (3)$$

где шум $\varepsilon(\mathbf{x})$ моделируется гауссовским белым шумом с дисперсией $\tilde{\sigma}^2$. В таком случае наблюдения $y(\mathbf{x})$ будут гауссовским процессом с нулевым средним и ковариационной функцией $cov(y(\mathbf{x}), y(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}') + \tilde{\sigma}^2$. Тогда апостериорное среднее значений гауссовского процесса $f(\mathbf{x})$ в точках тестовой выборки X_* имеет вид:

$$\hat{f}(X_*) = K_*(K + \tilde{\sigma}^2 I_N)^{-1} \mathbf{y}, \quad (4)$$

где I_N — единичная матрица размера $N \times N$. Апостериорная ковариационная функция гауссовского процесса в точках тестовой выборки имеет вид:

$$\mathbb{V}[\hat{f}(X_*)] = K(X_*, X_*) + \tilde{\sigma}^2 I_{N_*} - K_*(K + \tilde{\sigma}^2 I_N)^{-1} K_*^T \quad (5)$$

где $K(X_*, X_*) = [k(\mathbf{x}_i^*, \mathbf{x}_j^*), i, j = 1, \dots, N_*]$.

Дисперсия шума $\tilde{\sigma}^2$ в формуле (4) регуляризует ковариационную матрицу K и позволяет контролировать обобщающую способность аппроксиматора. Диагональные элементы матрицы $\mathbb{V}[\hat{f}(X_*)]$ из формулы (5) могут быть использованы как оценки ожидаемой ошибки аппроксимации в соответствующих точках X_* .

2.1. Оценка параметров гауссовского процесса

При работе с реальными данными ковариационная функция породившего их гауссовского процесса не известна. Предположим, что ковариационная функция гауссовского процесса принадлежит параметрическому семейству $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}' | \mathbf{a})$, где $\mathbf{a} \in \mathbb{R}^K$ — вектор параметров ковариационной функции. В качестве семейства $k(\mathbf{x}, \mathbf{x}' | \mathbf{a})$ зачастую используется класс стационарных ковариационных функций, т.е. функций значение которых зависит только от разности значений аргументов $k(\mathbf{x}, \mathbf{x}' | \mathbf{a}) = k(\mathbf{x} - \mathbf{x}' | \mathbf{a})$. Значение параметра $\mathbf{a}$ оценивается по обучающей выборке D_{train} с помощью метода максимума правдоподобия. Логарифм правдоподобия гауссовского процесса в точках обучающей выборки [14] имеет вид:

$$\log p(\mathbf{y} | X, \mathbf{a}, \tilde{\sigma}) = -\frac{1}{2} \left(\mathbf{y}^T (K + \tilde{\sigma}^2 I_N)^{-1} \mathbf{y} + \log |K + \tilde{\sigma}^2 I_N| + N \log 2\pi \right), \quad (6)$$

где $|K + \tilde{\sigma}^2 I|$ — детерминант матрицы $K + \tilde{\sigma}^2 I$. Логарифм правдоподобия зависит от параметров $\mathbf{a}$ ковариационной функции и дисперсии шума $\tilde{\sigma}^2$. Таким образом, оценка $\mathbf{a}$ и $\tilde{\sigma}^2$ получается с помощью максимизации логарифма правдоподобия:

$$\log p(\mathbf{y} | X, \mathbf{a}, \tilde{\sigma}) \rightarrow \max_{\mathbf{a}, \tilde{\sigma}}. \quad (7)$$

Выбор конкретного семейства ковариационных функций $k(\mathbf{x}, \mathbf{x}'|\mathbf{a})$ обычно определяется априорными представлениями о свойствах аппроксимируемой зависимости. В данной работе мы используем ковариационные функции из семейства

$$k_{p,q}(\mathbf{x} - \tilde{\mathbf{x}}|\mathbf{a}) = \sigma^2 \exp \left(- \left\{ \sum_{i=1}^n \theta_i^2 |x^i - \tilde{x}^i|^p \right\}^q \right), \quad \mathbf{a} = \{\boldsymbol{\theta} = (\theta_i)_{i=1}^n, \sigma\}, \quad p \in (0, 2], \quad q \in (0, 1].$$

При этом обычно рекомендуют использовать следующую параметризацию $\boldsymbol{\theta}$: $\theta_i = e^{\tau_i}, i = 1, \dots, n$, и оптимизировать правдоподобие по параметрам $\boldsymbol{\tau} = (\tau_i)_{i=1}^n$, см. [18, 19].

Заметим, что наряду с вышеописанным семейством ковариационных функций также можно рассматривать семейство ковариационных функций на основе расстояния Махаланобиса

$$k(\mathbf{x} - \tilde{\mathbf{x}}|\Lambda, \sigma) = \sigma^2 \exp \left((\mathbf{x} - \tilde{\mathbf{x}})^T \Lambda^2 (\mathbf{x} - \tilde{\mathbf{x}}) \right),$$

где Λ^2 — некоторая положительно определенная матрица. Данное семейство ковариационных функций позволяет аппроксимировать широкий класс зависимостей, однако квадратичная зависимость числа параметров от размерности входного пространства приводит к тому, что задача оптимизации правдоподобия становится сложной вычислительной задачей, построение эффективного алгоритма решения которой заслуживает отдельной работы.

Использование стационарной ковариационной функции ограничивает точность аппроксимации если моделируемая зависимость пространственна неоднородна. Далее будет предложен метод построения нестационарной ковариационной функции для гауссовского процесса.

2.2. Модель нестационарного гауссовского процесса

Рассмотрим нестационарный гауссовский процесс следующего вида

$$y = f(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad (8)$$

где $f(\mathbf{x}) = \sum_{i=1}^Q \alpha_i \psi_i(\mathbf{x}) + \tilde{f}(\mathbf{x})$, $\tilde{f}(\mathbf{x})$ — гауссовский процесс с нулевым средним и стационарной ковариационной функцией $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}'|\mathbf{a})$, $\{\alpha_i, i = 1, \dots, Q\}$ — независимые одинаково распределенные нормальные случайные величины с нулевым средним и дисперсией $\frac{\bar{\sigma}^2}{Q}$, $\{\psi_i(\mathbf{x}), i = 1, \dots, Q\}$ — некоторый набор функций. Для такого процесса ковариационная функция принимает вид:

$$\text{cov}(y(\mathbf{x}), y(\mathbf{x}')) = k(\mathbf{x} - \mathbf{x}'|\mathbf{a}) + \bar{\sigma}^2 \boldsymbol{\psi}(\mathbf{x})^T \boldsymbol{\psi}(\mathbf{x}') + \bar{\sigma}^2, \quad (9)$$

где $\boldsymbol{\psi}(\mathbf{x}) = \{\psi_i(\mathbf{x}), i = 1, \dots, Q\}$.

Введем обозначения $\Psi(X) = \{\boldsymbol{\psi}(\mathbf{x}_i), i = 1, \dots, N\}$, $\Psi(X_*) = \{\boldsymbol{\psi}(\mathbf{x}_j^*), j = 1, \dots, N_*\}$, $\bar{K}_* = K_* + \bar{\sigma}^2 \Psi(X_*)^T \Psi(X_*)$, $\bar{K} = K + \bar{\sigma}^2 \Psi(X)^T \Psi(X)$. Тогда апостериорное математическое ожидание процесса (8) в точках тестовой выборки имеет вид

$$\hat{f}(X_*) = \bar{K}_* (\bar{K} + \bar{\sigma}^2 I_N)^{-1} \mathbf{y}. \quad (10)$$

Соответственно, апостериорная ковариация предсказаний в точках тестовой выборки имеет вид

$$\mathbb{V}[\hat{f}(X_*)] = K(X_*, X_*) + \bar{\sigma}^2 I_{N_*} + \bar{\sigma}^2 \Psi(X_*)^T \Psi(X_*) - \bar{K}_* (\bar{K} + \bar{\sigma}^2 I_N)^{-1} \bar{K}_*^T. \quad (11)$$

Логарифм правдоподобия гауссовского процесса (8) совпадает с выражением (6), в котором матрица K заменена на матрицу $\bar{K}$:

$$\log p(\mathbf{y}|X, \mathbf{a}, \tilde{\sigma}, \bar{\sigma}) = -\frac{1}{2} \mathbf{y}^T (\bar{K} + \bar{\sigma}^2 I_N)^{-1} \mathbf{y} - \frac{1}{2} \log |\bar{K} + \bar{\sigma}^2 I_N| - \frac{N}{2} \log 2\pi. \quad (12)$$

Если набор базисных функции задан, то для идентификации аппроксимационной модели логарифм правдоподобия (12) необходимо максимизировать по параметрам $\mathbf{a}$ ковариационной функции $k(\mathbf{x} - \mathbf{x}'|\mathbf{a})$, а также по параметрам $\tilde{\sigma}$ и $\bar{\sigma}$, т.е.

$$\log p(\mathbf{y}|X, \mathbf{a}, \tilde{\sigma}, \bar{\sigma}) \rightarrow \max_{\mathbf{a}, \tilde{\sigma}, \bar{\sigma}}. \quad (13)$$

Заметим, что при оптимизации по параметрам $\tilde{\sigma}$ и $\bar{\sigma}$ автоматически будет определен вклад стационарной и нестационарной частей ковариационной функции в общую изменчивость аппроксимационной модели.

Построение набора функций $\{\psi_i(\mathbf{x}), i = 1, \dots, Q\}$ по обучающей выборке представляет собой отдельную задачу. В цикле работ [20–23] был предложен метод построения аппроксимации неизвестной многомерной зависимости с помощью линейного разложения по словарию параметрических функций различного типа:

1. Сигмоидные базисные функции $\psi_j(\mathbf{x}) = h(\sum_{i=1}^n \beta_{j,i} x^i)$, где

$$\mathbf{x} = (x^1, \dots, x^n), \quad h(z) = g(|z|^{g(\gamma_j)+1} \text{sign}(z)), \quad g(v) = \frac{e^v - 1}{e^v + 1},$$

γ_j и $\beta_{j,i}$ — настраиваемые параметры,

2. Радиальные базисные функции $\psi_j(\mathbf{x}) = \exp(-\|\mathbf{x} - \mathbf{d}_j\|^2/t_j^2)$, $\mathbf{d}_j$ и t_j — настраиваемые параметры,

3. Линейные базисные функции $\psi_j(\mathbf{x}) = x^j$, $j = 1, \dots, n$.

Параметры функций из словаря настраиваются с помощью методов градиентной оптимизации, бустинга и беггинга [20–23]. Получаемый таким образом словарь и предлагается использовать для моделирования нестационарного гауссовского процесса.

2.3. Моделирование многомерного выхода

В предыдущих разделах рассматривалось построение аппроксиматора на основе модели гауссовского процесса для случая одномерного выхода $y \in \mathbb{R}^1$. В данном разделе мы обобщим эти результаты на случай многомерного выхода $y = (y^1, \dots, y^m) \in \mathbb{R}^m$, $m > 1$. Простейший подход состоит в построении такого аппроксиматора для каждой компоненты выхода y^l , $l = 1, \dots, m$ в отдельности. Такой подход, однако, никак не учитывает взаимную корреляцию компонент выхода. Моделирование не только корреляционных свойств каждой компоненты выхода, но и кросскорреляций между ними может существенно повысить качество аппроксимации, и возможно, например, за счет использования описанного в геостатистической литературе алгоритма “кокригинг” [24].

Мы же в данной работе предлагаем промежуточный подход, который предполагает, что гауссовские процессы

$$y^l(\mathbf{x}) = \sum_{i=1}^Q \alpha_{i,l} \psi_i(\mathbf{x}) + \tilde{f}_l(\mathbf{x}) + \varepsilon_l(x),$$

моделирующие разные компоненты выхода, независимы, но имеют при этом “общую” ковариационную функцию. Будем предполагать, что для l -ой компоненты выхода ковариационная функция соответствующего гауссовского процесса $\tilde{f}_l(x)$ имеет вид

$$k_l(\mathbf{x} - \tilde{\mathbf{x}}|\mathbf{a}_l) = \sigma_l^2 \exp\left(-\left\{\sum_{i=1}^n \theta_i^2 |x^i - \tilde{x}^i|^p\right\}^q\right), \quad \mathbf{a}_l = \{\boldsymbol{\theta} = (\theta_i)_{i=1}^n, \sigma_l\}, \quad p \in (0, 2], \quad q \in (0, 1],$$

где разные компоненты выхода имеют разные масштабирующие коэффициенты σ_l^2 , а параметры $\boldsymbol{\theta} = (\theta_i)_{i=1}^n$ одинаковы для всех компонент выхода. Предположим также, что коэффициенты $\{\alpha_{i,l}, i = 1, \dots, Q\}_{l=1}^m$ — независимые одинаково распределенные нормальные случайные величины с нулевым средним и дисперсией $\mathbb{E}\alpha_{i,l}^2 = \frac{\bar{\sigma}_l^2}{Q}$, а процессы $\varepsilon_l(\mathbf{x})$, $l = 1, \dots, m$ — независимы между собой и от $\{\alpha_{i,l}, i = 1, \dots, Q\}_{l=1}^m$, $\{\hat{f}_l(x)\}_{l=1}^m$, и моделируются с помощью процессов белого гауссовского шума с дисперсией $\bar{\sigma}^2$.

Правдоподобие такой модели равно сумме правдоподобий вида (12) по всем каналам выхода. Оптимальные значения параметров находятся максимизацией правдоподобия:

$$\log p(Y|X, \bar{\sigma}, \{\mathbf{a}_l, \bar{\sigma}_l\}_{l=1}^m) = \sum_{l=1}^m \log p(\mathbf{y}_l|X, \bar{\sigma}, \mathbf{a}_l, \bar{\sigma}_l) \rightarrow \max_{\{\bar{\sigma}_l, \mathbf{a}_l\}_{l=1}^m, \bar{\sigma}}. \quad (14)$$

Данный подход позволяет вычислительно эффективно строить модель зависимости и при этом учитывать взаимосвязь между различными компонентами выхода. Недостаток такого подхода состоит в том, что хотя апостериорные средние значения соответствующих гауссовских процессов и будут различными для разных компонент выхода

$$\hat{f}_l(X_*) = \bar{K}_{l,*}(\bar{K}_l + \bar{\sigma}^2 I_N)^{-1} \mathbf{y}_l, l = 1, \dots, m,$$

но при этом все компоненты выхода будут моделироваться с помощью независимых гауссовских процессов, ковариационные матрицы которых отличаются только значениями масштабирующих множителей $\{\sigma_l\}_{l=1}^m$. Соответственно, например, для случая интерполяции апостериорные ковариации (11) разных компонент выхода будут отличаться только масштабом.

2.4. Интерполирующие свойства в случае нестационарной ковариационной функции

Модель зависимости на основе гауссовских процессов обладает важным свойством: в моделях (3) и (8) с шумом при $\bar{\sigma}^2 = 0$ мы получаем интерполирующее поведение апостериорного среднего, то есть в точках выборки X значение аппроксимации будет совпадать с наблюдаемым значением зависимости [14]. Такое поведение важно для некоторых инженерных задач. При $\bar{\sigma} > 0$ интерполирующее свойство вообще говоря не выполняется, однако для модели (8) можно показать, что при достаточно малых значениях $\bar{\sigma}^2 \approx 0$ всегда найдутся такие значения параметров $\{\bar{\sigma}, (\theta_i)_{i=1}^n, \sigma, \bar{\sigma}\}$, что поведение аппроксиматора будет интерполирующим с любой наперед заданной точностью. Формально это утверждение описывается следующей теоремой [25]:

Теорема 1. Пусть задано $\epsilon > 0$, тогда, если

$$\begin{aligned} & - \|\Psi(X)^T \Psi(X)\| < C, \quad C \in \mathbb{R}_+, \\ & - \bar{\sigma}^2 < \frac{1}{8} \min \left(\min_{i=1, N} \lambda_i, \epsilon \right), \\ & - \sigma^2 > \frac{1}{2}, \\ & - \bar{\sigma}^2 < \frac{1}{8} \cdot \frac{\min_{i=1, N} \lambda_i}{C+1}, \end{aligned}$$

где λ_i , $i = 1, \dots, N$ — собственные числа матрицы K , а норма понимается как стандартная евклидова, то

$$\|\mathbf{y} - \hat{f}(X)\| < \epsilon \|\mathbf{y}\|.$$

Таким образом, для предложенной нами модели нестационарного гауссовского процесса можно получать также и интерполирующие реализации, что бывает необходимо в прикладных задачах.

3. ИСПОЛЬЗОВАНИЕ БАЙЕСОВСКОГО ПОДХОДА ДЛЯ ОЦЕНКИ ПАРАМЕТРОВ МОДЕЛИ РЕГРЕССИИ НА ОСНОВЕ ГАУССОВСКОГО ПРОЦЕССА

3.1. Недостатки стандартного подхода к идентификации модели

В разделе 2 описан базовый алгоритм построения модели на основе гауссовских процессов. На большом классе задач этот алгоритм обеспечивает хорошее качество аппроксимации моделируемой зависимости, однако в некоторых случаях может приводить к вырожденной модели. Для примера рассмотрим случай использования ковариационной функции со взвешенным евклидовым расстоянием:

$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \sigma^2 \exp \left(- \sum_{i=1}^n \theta_i^2 (x^i - \tilde{x}^i)^2 \right).$$

При оптимизации правдоподобия по параметрам ковариационной функции $\mathbf{a} = \{\boldsymbol{\theta} = (\theta_i)_{i=1}^n, \sigma\}$ и дисперсии шума $\tilde{\sigma}$ могут наблюдаться следующие эффекты:

- Оптимум может лежать в области, где $\|\boldsymbol{\theta}\| \sim 0$, при этом матрица K в этой области плохо обусловлена,
- При $\|\boldsymbol{\theta}\| \gg 1$ матрица $K \rightarrow I_N$, что приводит к вырожденной аппроксимации,
- Правдоподобие $\log p(\mathbf{y}|\mathbf{X}, \mathbf{a}, \tilde{\sigma})$ может слабо зависеть от значений параметров $\mathbf{a}$ и $\tilde{\sigma}$.

В случае, когда максимум правдоподобия лежит в области малых значений параметров $\|\boldsymbol{\theta}\| \sim 0$, матрица ковариаций K становится плохо обусловленной. Из формулы (4) следует, что аппроксимация $\hat{f}(\mathbf{x})$ может быть представлена в виде $\hat{f}(\mathbf{x}) = \sum_{i=1}^N \alpha_i k(\mathbf{x}, \mathbf{x}_i)$, где $(\alpha_i)_{i=1}^N = (K + \tilde{\sigma}^2 I_N)^{-1} \mathbf{y}$. Если значение $\tilde{\sigma}$ недостаточно большое, то из-за плохой обусловленности матрицы K норма вектора $(\alpha_i)_{i=1}^N$ принимает большие значения и в результате в аппроксимации $\hat{f}(\mathbf{x})$ становится существенно заметным численный шум. В случае, когда максимум правдоподобия лежит в области больших значений параметров $\|\boldsymbol{\theta}\| \gg 1$, также происходит вырождение аппроксимационной модели: при $\tilde{\sigma} \approx 0$ для $(\mathbf{x}_i, y_i) \in D_{train}$ выполнено, что $\hat{f}(\mathbf{x}_i) \approx y_i$, при этом $\hat{f}(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^N y_i$ для $\mathbf{x} \notin D_{train}$.

3.2. Байесовский вывод

Для преодоления описанных выше проблем предлагается ввести байесовскую регуляризацию на параметры ковариационной функции. Ниже рассматриваются два варианта введения такой регуляризации: на основе априорного нормального распределения и на основе априорного гамма-распределения.

Предположим, что параметры $\boldsymbol{\theta}$ распределены с некоторой априорной плотностью $\bar{p}(\boldsymbol{\theta}|\gamma)$, где γ — гиперпараметры априорного распределения параметров ковариационной функции. Тогда логарифм совместной плотности распределения данных и параметров модели имеет вид:

$$\log P(\mathbf{y}, \boldsymbol{\theta}|\mathbf{X}, \sigma, \tilde{\sigma}, \gamma) = \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma, \tilde{\sigma}) + \log \bar{p}(\boldsymbol{\theta}|\gamma). \quad (15)$$

Слагаемое $\log \bar{p}(\boldsymbol{\theta}|\gamma)$ выполняет роль дополнительной регуляризации на значения параметров ковариационной функции $\boldsymbol{\theta}$, причем выбор конкретного распределения будет определять характер этой регуляризации. Выражение (15) можно максимизировать по параметрам $\sigma, \tilde{\sigma}, \gamma, \boldsymbol{\theta}$ с целью нахождения их оптимальных значений. Такой подход соответствует нахождению МАР-оценки (Maximum A Posteriori Probability) параметров модели $\sigma, \tilde{\sigma}, \boldsymbol{\theta}$ и параметров априорного распределения γ . В данной работе в качестве априорного распределения $\bar{p}(\boldsymbol{\theta}|\gamma)$ используются нормальное и гамма-распределения.

В случае нормального распределения рассмотрим следующее преобразование вектора θ :

$$\theta = \theta_{\parallel} + \theta_{\perp},$$

где $\theta_{\parallel} = |\theta_{\parallel}| \cdot \mathbf{e}$, $\mathbf{e} = (\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$, а вектор $\theta_{\perp}$ ортогонален вектору $\mathbf{e}$. Предположим, что величина $|\theta_{\parallel}|$ распределена нормально со средним $\mu_{\parallel} > 0$ и дисперсией $d_{\parallel}$, а вектор $\theta_{\perp}$ не зависит от $|\theta_{\parallel}|$ и имеет нормальное распределение с нулевым средним и диагональной ковариационной матрицей $d_{\perp} I_{\perp}$, где $I_{\perp}$ — единичная матрица размера $n - 1$, $d_{\perp}$ — дисперсия компонент вектора $\theta_{\perp}$. Такое предположение основано на нашем желании, чтобы все компоненты вектора θ были не слишком велики и не слишком малы по величине, а также не слишком сильно отличались по величине друг от друга. Таким образом в данном случае $\gamma = \{\mu_{\parallel}, d_{\parallel}, d_{\perp}\}$, а $\log \bar{p}(\theta|\gamma) = \log p_{N(\mu_{\parallel}, d_{\parallel})}(\theta) + \log p_{N(\vec{0}, d_{\perp} I_{\perp})}(\theta)$, где $p_{N(\mu, \Sigma)}(\cdot)$ — функция плотности нормального распределения со средним μ и ковариационной матрицей Σ .

Теперь рассмотрим случай гамма-распределения. Предположим, что все компоненты вектора θ независимы и распределены с одним и тем же гамма-распределением с параметрами α и β . В таком случае

$$\log \bar{p}(\theta|\gamma) = \log \bar{p}(\theta|\alpha, \beta) = \sum_{i=1}^n [\alpha \log \beta + (\alpha - 1) \log \theta_i - \beta \theta_i - \log \Gamma(\alpha)], \quad (16)$$

где $\alpha > 1$, $\beta > 0$, а $\Gamma(\alpha)$ — гамма-функция. Такой выбор априорного распределения параметров также штрафует слишком большие и слишком маленькие значения компонент вектора θ_i . В разделе 4.1 приведен ряд экспериментальных результатов, которые показывают что оба предложенных метода байесовской регуляризации приводят к заметному улучшению надежности работы алгоритма и качеству финальной модели. Отметим, что аналогичным образом можно регуляризовать и параметр σ .

3.3. Многоуровневый байесовский вывод

Для того, чтобы процедуры, основанные на байесовском подходе, могли использоваться для решения прикладных задач, необходимо обеспечить разумный выбор параметров априорного распределения [14]. Зачастую у исследователя нет возможности вручную выбирать параметры априорного распределения для каждой задачи — если, например, число задач велико, или предполагается, что разработанный алгоритм будет использоваться как черный ящик, то есть, пользователь не будет обладать необходимыми знаниями для выбора параметров априорного распределения в байесовском подходе.

В данном разделе предложен многоуровневый байесовский вывод [26] для модели регрессии на основе гауссовских процессов, позволяющий избежать описанной выше проблемы. Идея подхода состоит во введении дополнительного гипераприорного распределения на параметрах априорного распределения параметров исходной гауссовской модели. Процедура выбора модели в таком случае состоит в выборе параметров модели, параметров априорного распределения и параметров гипераприорного распределения исходя из принципа максимума апостериорной вероятности. При этом выбор параметров гипераприорного распределения меньше влияет на функционал правдоподобия, а вводимая этим распределением регуляризация позволяет избежать получения вырожденных значений параметров априорного распределения первого уровня при оптимизации (15) и, как следствие, получения вырожденных значений параметров ковариационной функции.

Таким образом, предлагается ввести

- априорное гамма-распределение $\Gamma(\alpha, \beta)$ на параметры ковариационной функции θ ,

- априорное гамма-распределение $\Gamma(\alpha_\sigma, \beta_\sigma)$ на параметр ковариационной функции σ ,
- гипераприорное лог-нормальное распределение с заранее фиксированными параметрами μ, σ_{prior}^2 на каждый из параметров $\alpha, \beta, \alpha_\sigma, \beta_\sigma$ байесовской регуляризации.

При этом совместная плотность распределения будет пропорциональна:

$$p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2, \alpha, \beta, \alpha_\sigma, \beta_\sigma, \mu, \sigma_{prior}^2) \propto p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2) \cdot \Gamma(\boldsymbol{\theta}|\alpha, \beta) \cdot \Gamma(\sigma^2|\alpha_\sigma, \beta_\sigma) \times \\ \times \log \mathcal{N}(\alpha|\mu, \sigma_{prior}^2) \cdot \log \mathcal{N}(\beta|\mu, \sigma_{prior}^2) \cdot \log \mathcal{N}(\alpha_\sigma|\mu, \sigma_{prior}^2) \cdot \log \mathcal{N}(\beta_\sigma|\mu, \sigma_{prior}^2),$$

где $\Gamma(\cdot)$ — плотность гамма-распределения, а $\log \mathcal{N}(\cdot)$ — плотность лог-нормального распределения. Подставляя явные выражения для распределений получаем явное выражение для максимизации:

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2, \alpha, \beta, \alpha_\sigma, \beta_\sigma, \mu, \sigma_{prior}^2) \propto \\ \propto -\frac{1}{2} \left[N \log 2\pi + \log |K_{\boldsymbol{\theta}, \sigma}| + \mathbf{y}^T K_{\boldsymbol{\theta}, \sigma}^{-1} \mathbf{y} \right] - \\ - \log \Gamma(\alpha) + \alpha \log \beta + (\alpha - 1) \sum_{i=1}^n \log \theta_i - \beta \sum_{i=1}^n \theta_i - \\ - \log \Gamma(\alpha_\sigma) + \alpha_\sigma \log \beta_\sigma + (\alpha_\sigma - 1) \log \sigma^2 - \beta_\sigma \sigma^2 - \\ - \frac{1}{2} \left[2 \log \alpha + \frac{1}{\sigma_{prior}^2} (\log \alpha - \mu)^2 + 2 \log \beta + \frac{1}{\sigma_{prior}^2} (\log \beta - \mu)^2 \right] - \\ - \frac{1}{2} \left[2 \log \alpha_\sigma + \frac{1}{\sigma_{prior}^2} (\log \alpha_\sigma - \mu)^2 + 2 \log \beta_\sigma + \frac{1}{\sigma_{prior}^2} (\log \beta_\sigma - \mu)^2 \right].$$

Здесь $K_{\boldsymbol{\theta}, \sigma}$ — ковариационная матрица выборки, а $\Gamma(\alpha)$ — значение гамма-функции при значении аргумента α .

4. ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

4.1. Байесовская регуляризация

В данном разделе приведено два примера использования байесовской регуляризации для оценки параметров ковариационной функции регрессии на основе гауссовских процессов: аппроксимация функции Растригина и сравнение байесовской оценки параметров с оценкой параметров на основе метода максимума правдоподобия на большом наборе различных тестовых функций.

В некоторых случаях максимум правдоподобия соответствует значениям параметров, которые не имеют хорошей физической интерпретации. Рассмотрим, например, аппроксимацию функции Растригина [27] с помощью регрессии на основе гауссовских процессов по 20 случайным точкам, при этом параметры ковариационной функции оцениваются с помощью метода максимума правдоподобия (функция изображена на рисунке 1 слева). Полученная аппроксимация изображена на рисунке 1 по центру — регрессионная модель состоит из узких пиков вокруг точек обучающей выборки и не может быть использована на практике. Если же при построении аппроксимации использовать байесовскую регуляризацию с априорным гауссовским или гамма-распределением, то получаются одинаковые аппроксимации, вид которых изображен на рисунке 1 слева. Видно, что аппроксимация с использованием байесовской регуляризации точнее передает вид функции Растригина, хотя и довольно сильно отличается от нее в силу малого объема обучающей выборки.

В ходе экспериментов сравнивалась работа трех алгоритмов построения аппроксимации, основанных на гауссовских процессах, а именно:

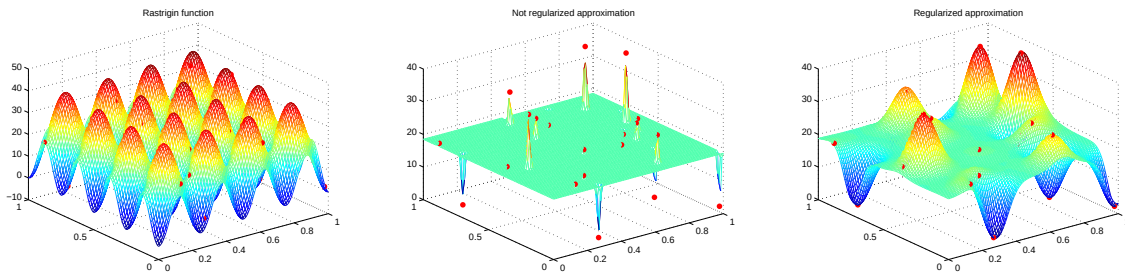


Рис. 1. Функция Растригина и ее аппроксимация. Слева изображена целевая функция, по центру — ее аппроксимация с оценкой параметров по методу максимума правдоподобия, справа — ее аппроксимация с байесовской оценкой параметров.

- MLE GP — регрессия на основе гауссовских процессов с методом максимума правдоподобия для оценки параметров;
- Gamma GP — байесовский метод оценки параметров с априорным гамма-распределением;
- Normal GP — байесовский метод оценки параметров с априорным нормальным распределением.

Сравнение алгоритмов проводилось с помощью кривых Долана-Мора [28], которые позволяют наглядно представить результаты тестирования нескольких алгоритмов на наборе задач следующим образом. Пусть у нас есть набор алгоритмов $a_i, i = 1, \dots, A$ и набор задач $t_j, j = 1, \dots, T$. Тогда для каждого алгоритма мы можем измерить качество его работы (например, среднюю абсолютную ошибку на тестовой выборке) на каждой задаче и получить набор чисел q_{ij} . Пусть лучший алгоритм для рассматриваемой задачи имеет качество $q_j, j = 1, \dots, T$, такое что $\forall i = 1, \dots, A: q_{ij} \geq q_j$. Определим $r_{ij} = \frac{q_{ij}}{q_j} \geq 1$. Чем ближе r_{ij} к единице — тем лучше сработал алгоритм a_i на задаче t_j . Будем менять параметр τ от 1 до ∞ . Для каждого τ определим точку $c_i(\tau)$ на кривой Долана-Мора для алгоритма a_i как долю задач, на которых r_{ij} меньше, чем τ . С увеличением τ величина $c_i(\tau)$ стремится к единице, и чем выше лежит кривая $c_i(\tau)$ — тем лучше работает алгоритм по сравнению с другими алгоритмами.

Для демонстрации экспериментальных результатов использовался большой набор тестовых функций, которые применяются для тестирования задач оптимизации [29–31]. Тестирование проводилось на 30 различных функциях, для каждой из которых генерировалось по 2 случайные обучающие выборки с размерами 10, 20, 40, 80 и 160 точек. 95% квантиль абсолютной ошибки на контрольных выборках размером 10000 точек служила мерой качества алгоритмов построения аппроксимации. Результаты представлены в виде кривых Долана-Мора на рисунке 2. Также в качестве меры качества работы алгоритма использовалась площадь под кривой Долана-Мора, значения которой для трех рассматриваемых алгоритмов приведены в подписи к рисунку 2. Заметим, что оба алгоритма с байесовской регуляризацией показали результаты лучше, чем алгоритм без регуляризации.

4.2. Нестационарная ковариационная функция

В данном разделе приведено два примера использования нестационарной ковариационной функции: аппроксимация пространственно неоднородной одномерной функции и сравнение трех различных реализаций регрессии на основе гауссовских процессов на большом наборе различных функций.

Рассмотрим одномерный пример, на котором видно, что использование нестационарной ковариационной функции позволяет заметно увеличить качество аппроксимации. Будем аппроксимировать функцию Хевисайда на отрезке $[0, 1]$ по 20 случайным точкам из этого отрезка.

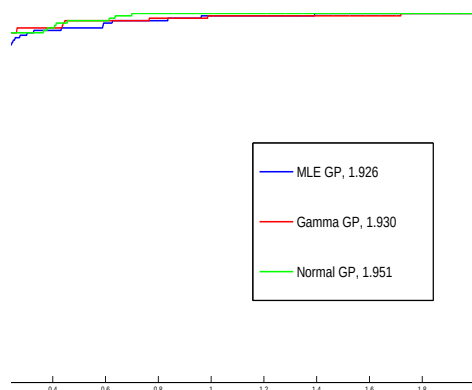


Рис. 2. Кривые Долан-Мора для тестовых функций. Площади под кривыми указаны в легенде

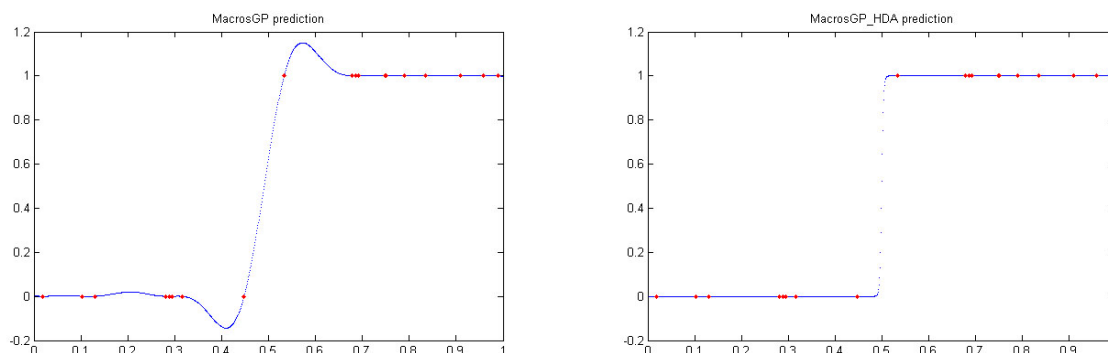


Рис. 3. Аппроксимация функции Хевисайда. Аппроксимация, которая получается при использовании стационарной ковариационной функции, изображена на рисунке слева, при использовании нестационарной ковариационной функции — на рисунке справа.

Графики аппроксимации при использовании стационарной и нестационарной ковариационных функций показаны на рисунке 3. Введение нестационарной ковариационной функции существенно улучшает качество аппроксимации.

В ходе экспериментов сравнивалась работа трех алгоритмов, основанных на гауссовских процессах, а именно:

- DACE — реализация регрессии на основе гауссовских процессов в пакете DACE [15],
- GP — алгоритм на основе гауссовских процессов без использования нестационарной ковариационной функции, разработанный авторами статьи,
- Nonstationary GP — алгоритм на основе гауссовских процессов с использованием нестационарной ковариационной функции, разработанный авторами статьи.

Для демонстрации экспериментальных результатов был использован большой набор тестовых функций, которые применяются для тестирования задач оптимизации [29–31]. Всего тестирование проводилось на 23 различных функциях, для каждой из которых генерировалось 2 случайные обучающие выборки с размером 160 точек каждая. Среднеквадратичная ошибка (1) на тестовых выборках размера 10000 точек служила мерой качества аппроксимации. Результаты представлены в виде кривых Долан-Мора [28] на рисунке 4. Чем выше кривая находится на графике, тем выше качество работы соответствующего алгоритма. Предложенная

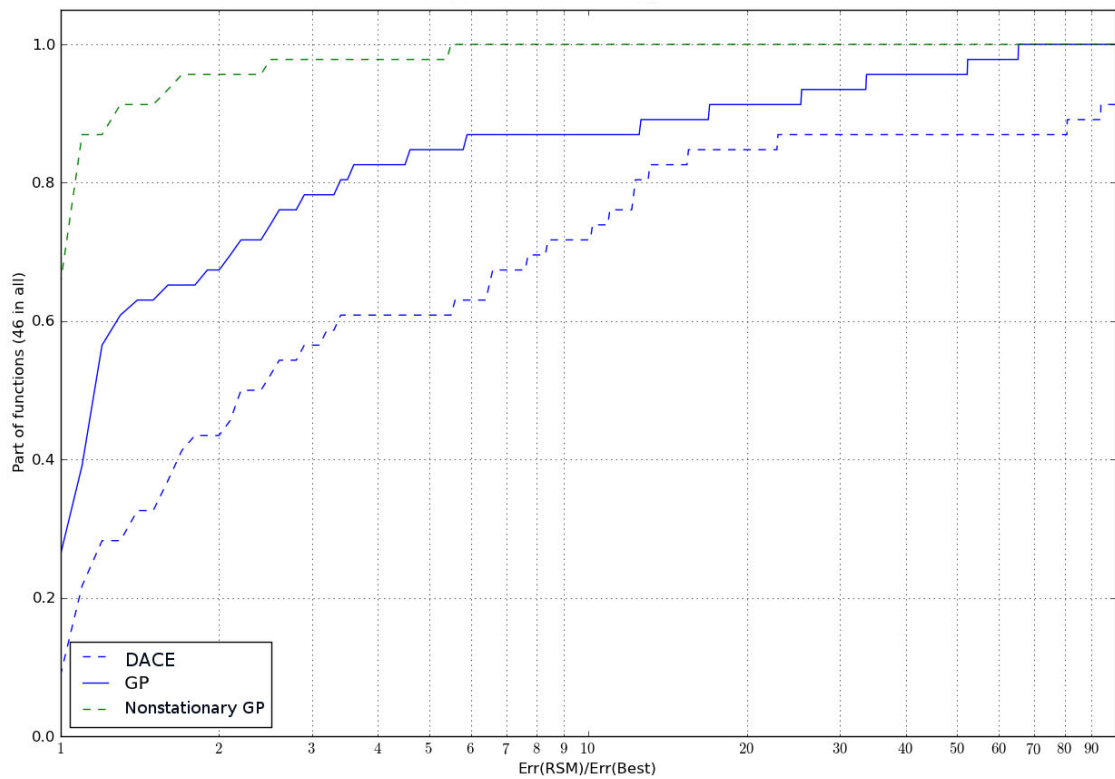


Рис. 4. Кривые Долан-Мора для тестовых функций

реализация аппроксиматора на основе гауссовских процессов заметно превосходит DACE, а добавление нестационарной ковариационной функции позволяет существенно улучшить качество результатов.

4.3. Многоуровневая байесовская регуляризация

Как следует из результатов тестирования (раздел 4.1), предложенный в разделе 3.2 алгоритм байесовской регуляризации позволяет повысить в среднем качество аппроксимации и повысить ее робастность. Однако, при оптимизации (15) по параметрам $\gamma = (\alpha, \beta)$ априорного распределения могут возникать ситуации, в которых, например, $\beta \approx 0$. Как следствие, регуляризация уже не накладывает штрафа на большие значения параметров $\theta = (\theta_i)_{i=1}^n$, и возможно вырождение итоговой аппроксимации в константу с “пиками” в точках обучающей выборки, см. пример на рисунке 1. Многоуровневый байесовский вывод из раздела 3.3, результаты тестирования которого представлены в данном разделе, как раз и позволяет избежать этой проблемы.

Тестирование проводилось на большом наборе функций разных входных размерностей. Параметр μ логнормального распределения был выбран равным 1.5. Сравнивались две меры качества аппроксимации: среднеквадратичная ошибка и мера изменчивости аппроксимации, позволяющая детектировать её вырождение. Мера изменчивости аппроксимации оценивалась

по значениям $\hat{\mathbf{y}} = \hat{f}(X_*)$ и $\mathbf{y}_*$ согласно формуле

$$\text{variability} = \frac{\max(\hat{\mathbf{y}}) - \min(\hat{\mathbf{y}})}{\max(\mathbf{y}_*) - \min(\mathbf{y}_*)},$$

где $\max(\mathbf{y})$ — максимальное значение вектора $\mathbf{y}$. Для каждого размера выборки подсчитывалось количество задач, для которых изменчивость меньше 10^{-4} (в этом случае полагаем, что аппроксимация выродилась в константу).

В экспериментах параметр σ_{prior}^2 брался равным 1.5, 1, 0.5, 0.3, 0.2. Кривая Долана-Мора, построенная по значению среднеквадратичной ошибки, приведена на рисунке 5. Видно, что при уменьшении значения σ_{prior}^2 точность аппроксимации уменьшается.

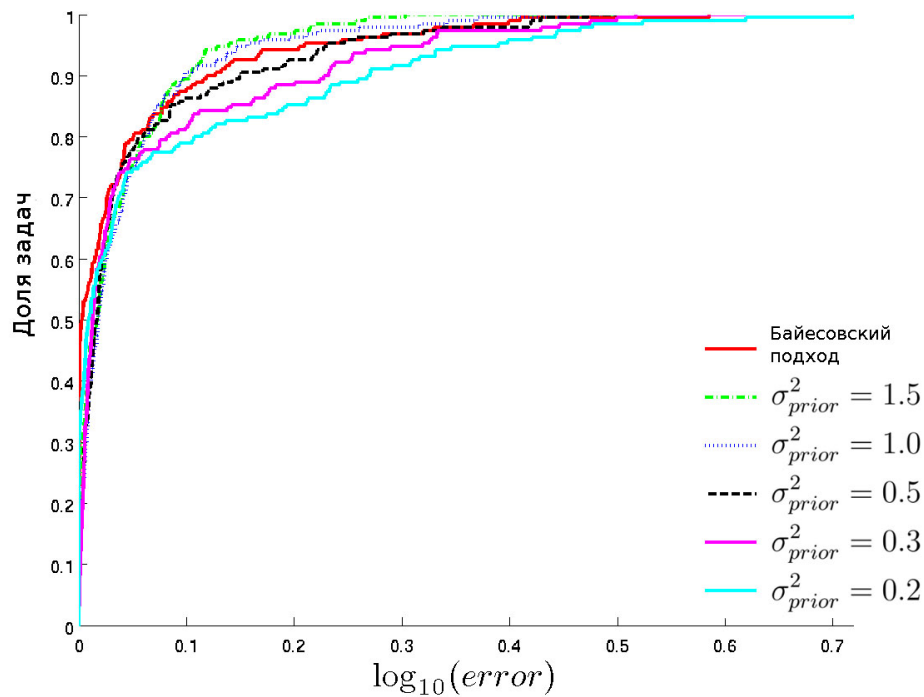


Рис. 5. Кривая Долана-Мора для байесовской регуляризации и многоуровневой байесовской регуляризации при различных значениях σ_{prior}^2

В таблице 1 приведены результаты статистики по количеству случаев построения моделей с ошибками аппроксимации, близкими к ошибкам аппроксимации константой. Уже для значения $\sigma_{prior}^2 = 0.5$ вырожденных регрессионных моделей нет, в то же время среднеквадратичная ошибка модели остается примерно такой же, как и при использовании обычного байесовского подхода.

Таким образом, предложенный подход позволяет избежать вырожденных моделей регрессии на основе гауссовских процессов, не ухудшая качество модели в среднем.

5. ВЫВОДЫ

В работе предложено два метода улучшения точности регрессионной модели на основе гауссовских процессов: использование байесовской регуляризации и использование новой нестационарной ковариационной функции, построенной по данным.

Размер выборки	Байесовская регуляризация	Многоуровневая байесовская регуляризация при разных σ_{prior}^2				
		1.5	1	0.5	0.3	0.2
10	16	0	0	0	0	0
20	2	0	0	0	0	0
40	31	0	0	0	0	0
80	37	14	0	0	0	0
160	24	15	1	0	0	0
320	3	1	0	0	0	0

Таблица 1. Выбор параметра σ_{prior}^2 : число задач, на которых изменчивость аппроксимации меньше 10^{-4} , то есть произошло вырождение итоговой аппроксимации. Всего рассматривалось 190 тестовых задач для каждого размера обучающей выборки N , который варьировался в пределах от 10 и до 320 точек.

Для байесовской регуляризации в качестве априорных распределений параметров использовались нормальное и гамма-распределения. Оба априорных распределения позволяют избежать вырожденных аппроксимаций, увеличивают обобщающую способность и надежность алгоритма.

Нестационарная ковариационная функция гауссовского процесса в работе строится на основе словаря параметрических базисных функций, настроенного по данным. В отличие от других реализаций аппроксиматоров на основе гауссовских процессов, таких как DACE [15], предложенный метод полностью автоматизирован и не требует от применяющего его инженера каких-либо знаний в области машинного обучения. Экспериментальные результаты показывают высокую качество работы предложенного алгоритма по сравнению с обычными методами, особенно для пространственно неоднородных функций.

СПИСОК ЛИТЕРАТУРЫ

1. A.I.J. Forrester, A. Söbester, and A.J. Keane. Engineering design via surrogate modelling: a practical guide. Progress in Astronautics and Aeronautics. J. Wiley, 2008.
2. A.A. Zaytsev, E.V. Burnaev, and V.G. Spokoiny. Properties of the bayesian parameter estimation of a regression based on gaussian processes. Journal of Mathematical Sciences, 203(6):789–798, 2014.
3. E.V. Burnaev, A.A. Zaytsev, and V.G. Spokoiny. The Bernstein-von Mises theorem for regression based on Gaussian processes. Russian Mathematical Surveys, 68(5):954–956, 2013.
4. E. Burnaev and M. Panov. Adaptive design of experiments based on gaussian processes. In Lecture Notes in Artificial Intelligence. Proceedings of SLDS 2015. A. Gammerman et al. (Eds.), volume 9047, pages 116–126, London, UK, April 20–23 2015. Springer.
5. M. Belyaev, E. Burnaev, and Y. Kapushev. Gaussian process regression for structured data sets. In Lecture Notes in Artificial Intelligence. Proceedings of SLDS 2015. A. Gammerman et al. (Eds.), volume 9047, pages 106–115, London, UK, April 20–23 2015. Springer.
6. A. Giunta and L.T.A. Watson. Comparison of approximation modeling technique: Polynomial versus interpolating models. In 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, volume 1, pages 392–404, 1998.
7. T.W. Simpson, A.J. Booker, S. Ghosh, A. Giunta, P.N. Koch, and R.J. Yang. Approximation methods in multidisciplinary analysis and optimization: A panel discussion. Structural and Multidisciplinary Optimization, 27(5):302–313, 2004.
8. S. M. Batill, J. E. Renaud, and X. Gu. Modeling and simulation uncertainty in multidisciplinary design optimization. In 8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, pages 1–11, 2000.
9. J.E. Pacheco, C.H. Amon, and S. Finger. Bayesian surrogates applied to conceptual stages of the engineering design process. Journal of Mechanical Design, 125(4):664–672, 2003.

10. A.J. Booker, J.E. Dennis, P.D. Frank, D.B. Serafini, V. Torczon, and M. Trosset. A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization*, 17(1):1–13, 1999.
11. P.N. Koch, B.A. Wujek, O. Golovidov, and T.W. Simpson. Facilitating probabilistic multidisciplinary design optimization using kriging approximation models. In *9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, 2002.
12. T.W. Simpson, T.M. Maurey, J.J. Korte, and F. Mistree. Kriging metamodels for global approximation in simulation-based multidisciplinary design optimization. *AIAA journal*, 39(12):2223–2241, 2001.
13. T.W. Simpson, T.M. Maurey, J.J. Korte, and F. Mistree. Metamodeling development for vehicle frontal impact simulation. In *American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.-Design Automation Conf.*, DETC2001/DAC-21012, 2001.
14. C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.
15. S.N. Lophaven, H.B. Nielsen, and J. Sondergaard. Aspects of the matlab toolbox dace. Technical report, Technical University of Denmark, 2002.
16. A. Bhattacharya, D. Pati, and D. Dunson. Anisotropic function estimation using multi-bandwidth gaussian processes. *Ann. Statist.*, 42(1):352–381, 02 2014.
17. Ying Xiong, Wei Chen, Daniel Apley, and Xuru Ding. A non-stationary covariance-based kriging method for metamodeling in engineering design. *International Journal for Numerical Methods in Engineering*, 71(6):733–756, 2007.
18. B. Nagy, J.L. Loeppky, and W.J. Welch. Correlation parameterization in random function models to improve normal approximation of the likelihood or posterior. Technical Report 229, Department of Statistics, The University of British Columbia, 2007.
19. B. MacDonald, P. Ranjan, and H. Chipman. Gpfit: An r package for fitting a gaussian process model to deterministic simulator outputs. *Journal of Statistical Software*, 64(12), 2015.
20. E.V. Burnaev, M.G. Belyaev, and Prihodko P.V. About hybrid algorithm for tuning of parameters in approximation based on linear expansions in parametric functions. In *Proceeding of the 8th International Conference Intelligent Information Processing*, pages 200–203, Paphos, Cyprus, October 17–24 2011.
21. M. Belyaev and E. Burnaev. Approximation of a multidimensional dependency based on a linear expansion in a dictionary of parametric functions. *Informatics and Applications*, 7:114–125, 2013.
22. E. Burnaev and P. Prihodko. On a method for constructing ensembles of regression models. *Automation and Remote Control*, 74:1630–1644, 2013.
23. S. Grihon, E. Burnaev, M. Belyaev, and P. Prihodko. Surrogate Modeling of Stability Constraints for Optimization of Composite Structures, pages 359–391. *Springer Series in Computational Science & Engineering*. Springer, New York, 2013.
24. N.A.C. Cressie. *Statistics for spatial data*. Wiley New York, 1993.
25. E. Burnaev, E. Zaytsev, M. Panov, P. Prihodko, and Yu. Yanovich. Modeling of nonstationary covariance function of gaussian process using decomposition in dictionary of nonlinear functions. In *Information Technologies and Systems–2011*, pages 355–362, Moscow, Russia, October 2–7 2011.
26. C.M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.
27. A. Torn and A. Zilinskas. *Global Optimization*, volume 350 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 1989.
28. E.D. Dolan and J.J. More. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.
29. A. Saltelli and I.M. Sobol. About the use of rank transformation in sensitivity analysis of model output. *Reliab. Eng. Syst. Safety*, 50(3):225–239, 1995.
30. J. Ronkkonen and J. Lampinen. An extended mutation concept for the local selection based differential evolution algorithm. In *Genetic and Evolutionary Computation Conference*, 2007.

31. T. Ishigami and T. Homma. An importance qualification technique in uncertainty analysis for computer models. In Proceedings of the Isuma'90, First International Symposium on Uncertainty Modelling and Analysis, 1990.

REGRESSION BASED ON NONSTATIONARY GAUSSIAN PROCESS WITH BAYESIAN REGULARIZATION

Burnaev E.V., Panov M.E., Zaytsev A.A.

Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

We consider the regression problem, i.e. prediction of a real valued function. A Gaussian process prior is imposed on the function, and is combined with the training data to obtain predictions for new points. We introduce a Bayesian regularization on parameters of a covariance function of the process, which increases quality of approximation and robustness of the estimation. Also an approach to modeling nonstationary covariance function of a Gaussian process on basis of linear expansion in parametric functional dictionary is proposed. Introducing such a covariance function allows to model functions, which have non-homogeneous behaviour. Combining above features with careful optimization of covariance function parameters results in unified approach, which can be easily implemented and applied. The resulting algorithm is an out of the box solution to regression problems, with no need to tune parameters manually. The effectiveness of the method is demonstrated on various datasets.

KEYWORDS: Gaussian processes, regression, Bayesian regularization, prior distribution, Bayesian regression