Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук (ИППИ РАН)

На правах рукописи

Панин Иван Игоревич

Методы оценки точности показателей чувствительности Соболя на основе метамоделей

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата компьютерных наук

Москва – 2023

Работа выполнена в *Институте проблем передачи информации РАН*

Научный руководитель: *доктор физико-математических наук,*
профессор,
заведующий лабораторией ИППИ РАН,
Вьюгин Владимир Вячеславович

Тема диссертации

Актуальность темы исследования. Анализ чувствительности — это важный инструмент исследования вычислительных моделей в инженерии и других областях. Он позволяет выяснить, как различные входные параметры модели влияют на ее выход и оценить такое влияние количественно; в частности, он позволяет разделить входные параметры на важные, относительно важные и незначимые [13].

Анализ чувствительности включает в себя широкий спектр методов, таких как метод Морриса, методы на основе линейной регрессии и дисперсионного анализа и другие (см. обзор Иосса [8]). Среди всех методов анализа чувствительности мы рассматриваем *показатели чувствительности Соболя*, которые оценивают, какая доля выходной дисперсии модели объясняется изменчивостью ее различных входных параметров и их групп [15]. Изменчивость входных параметров модели при этом задается с помощью наперед выбранного вероятностного распределения. Преимуществом данного метода является то, что он дает возможность проводить анализ сложных нелинейных и немонотонных моделей, а его результаты легко интерпретируемы (эффективность метода в инженерии показана в работах Судрэ [17]).

Методы оценки показателей Соболя также многообразны и обычно делятся на *Монте-Карло* и *метамодельные* подходы. Первые используют анализируемую модель напрямую для численного интегрирования с применением явных формул, таких как приведенные в работах Соболя [14], Оуэна и других. Вопросы точности и сходимости Монте-Карло методов для оценки показателей Соболя хорошо изучены, и можно заключить, что такие методы просты и надежны; однако они требуют большого количества запусков анализируемой модели. Метамодельные подходы, в свою очередь, позволяют сократить необходимое количество запусков модели. Следуя им, исходную вычислительно тяжелую модель заменяют вычислительно легким приближением, называемым *метамоделью* (или просто *аппроксимацией*), которое строится на основе обучающей выборки и лучше подходит для расчета показателей Соболя. Для этой цели часто используют метамодели на основе разложения полиномиального[1] хаоса, поскольку они позволяют получить оценку показателей Соболя в явном виде из коэффициентов разложения [16].

На сегодняшний день вопросы точности и сходимости[2] метамодельных оценок показателей Соболя достаточно плохо изучены, а существующие практические способы определения их погрешности не имеют строгого теоретического обоснования. В частности, в случае метамоделей на основе разложения полиномиального хаоса не известны границы риска оценок показателей Соболя, а единственный известный метод [7] определения погрешности показателей, оцененных

[1] Основу данной аппроксимации составляют полиномы, ортогональные относительно распределения входных параметров модели. Использование именно полиномов не является принципиальным, вместе с тем они удобны с вычислительной точки зрения и для многих типичных распределений уже построены.

[2] С ростом размера обучающей выборки, используемой для построения метамодели.

с помощью таких метамоделей, не обосновывается математически.

Другой круг вопросов при использовании метамодельного подхода связан с выбором плана эксперимента для построения метамодели. В настоящее время в анализе чувствительности на основе метамоделей обычно применяют методы планирования, которые стремятся в некотором смысле “равномерно” заполнить пространство входных параметров, такие как сэмплирование из распределения или на основе *латинского гиперкуба*, а также *квазислучайные последовательности*. Однако подобные методы никак не учитывают особенности метамодели и специально не адаптированы для расчета показателей Соболя. Можно было бы ожидать, что теория оптимального планирования эксперимента позволит выбирать подходящие планы для анализа чувствительности [12]. Однако ее классические методы для регрессионных моделей (A-, D-, I-оптимальные планы и другие) связаны с максимизацией качества самой аппроксимации и напрямую не учитывают точность оценок показателей Соболя на ее основе.

Таким образом, **актуальна** разработка как теоретических, так и практических методов контроля качества оценок показателей Соболя при метамодельном подходе на основе разложения полиномиального хаоса и эффективных методов планирования эксперимента, обеспечивающих высокую точность этих оценок.

Цель данной работы заключается в создании методов контроля качества и планирования эксперимента для оценки показателей чувствительности Соболя с помощью метамоделей на основе разложения полиномиального хаоса. Для достижения поставленной цели были определены следующие **задачи** исследования:

1. Исследовать зависимость качества произвольной аппроксимации анализируемой модели и точности оценок показателей Соболя, полученных на основе этой аппроксимации.
2. Разработать алгоритм оценки погрешности показателей Соболя, рассчитанных при помощи метамоделей на основе разложения полиномиального хаоса.
3. Выполнить теоретический анализ точности оценок показателей Соболя, полученных с помощью разложения полиномиального хаоса.
4. Разработать алгоритм построения плана эксперимента, эффективный для оценки показателей Соболя на основе разложения полиномиального хаоса.
5. Реализовать предложенные алгоритмы в виде комплекса программ.

Основные результаты

Положения, выносимые на защиту:

1. Установлено соотношение между ошибкой оценок показателей Соболя и ошибкой аппроксимации, с помощью которой эти оценки были получены. Показано, что соответствующая верхняя граница ошибки оценок показателей Соболя достижима.

2. Для метамоделей, использующих разложение полиномиального хаоса, разработан метод контроля погрешности метамодельных оценок показателей Соболя, основанный на доказанной верхней границе ошибки этих оценок.

3. Доказаны неасимптотические верхние границы риска метамодельных оценок показателей Соболя в условиях случайного плана эксперимента для аппроксимации на основе разложения полиномиального хаоса. Для этих границ в случае конкретных семейств многомерных ортогональных полиномов найдены скорости сходимости.

4. Найдено асимптотическое распределение оценок показателей Соболя и разработан метод последовательного планирования эксперимента, позволяющий улучшить среднюю точность этих оценок по сравнению со стандартными методами планирования.

5. С помощью разработанного программного комплекса был решен ряд инженерных задач. В частности, его применение для решения задачи анализа факторов, которые влияют на величину прогиба находящейся под действием внешней нагрузки стержневой конструкции (фермы), позволило на 10% увеличить среднюю точность оценок показателей Соболя по сравнению со стандартными методами.

Научная новизна. В работе впервые ставится и решается вопрос о соотношении между погрешностью оценок показателей Соболя и теоретической ошибкой произвольной аппроксимации, с помощью которой эти оценки были получены; что открывает новые возможности анализа точности метамодельных оценок показателей Соболя на основе произвольных типов аппроксимаций.

Исходя из полученного соотношения между погрешностью показателей и ошибкой аппроксимации, предложен прикладной метод контроля качества оценок показателей Соболя. Это соотношение также впервые позволило оценить верхние границы риска показателей Соболя на основе разложения полиномиального хаоса в условиях случайного плана эксперимента.

Кроме того, в работе предложена идея использовать теорию оптимального планирования эксперимента для регрессионных моделей для более эффективного расчета показателей Соболя. В частности, для случая полиномиального хаоса на основе критерия D-оптимальности разработан новый метод последовательного планирования эксперимента, эффективный для оценивания показателей чувствительности. Предложенный подход может быть распространен и на другие критерии оптимальности.

Личный вклад автора. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад диссертанта в опубликованные работы.

Работа [10], в которой проведен анализ точности оценок показателей Соболя, предложен метод контроля их качества и получены границы риска для них, выполнена без соавторов.

Подготовка к публикации серии работ [5, 4, 3], посвященных планированию эксперимента для анализа чувствительности, проводилась совместно с соавторами, причем вклад диссертанта был определяющим. В этой серии работ Е. В. Бурнаевым была предложена общая постановка задачи, общие подходы к ее решению и идеи некоторых экспериментов. Кроме того, совместно с Е. В. Бурнаевым было получено доказательство Теоремы 1 в работе [5]; остальные результаты в этой работе, включая предложенный алгоритм планирования эксперимента, принадлежат лично диссертанту. Программный код конечно-элементных вычислительных моделей для тестирования предложенного в работах [5, 4] метода планирования эксперимента был предоставлен Б. Судрэ. Разработка программного комплекса, реализующего предложенные методы, и все вычислительные эксперименты были выполнены диссертантом.

Общая методика исследования. Для решения поставленных задач в работе используются методы математической статистики, теории вероятностей, теории приближения функций, матричной алгебры и аппарат анализа Фурье.

Теоретическая и практическая значимость. С теоретической точки зрения результаты диссертации дают основу для анализа точности оценок показателей Соболя, полученных на основе различных типов аппроксимаций и с использованием различных планов эксперимента. С практической точки зрения результаты дополняют и улучшают существующие подходы для анализа чувствительности математических моделей в инженерии и других областях.

Публикации и апробация работы

Публикации повышенного уровня

1. Ivan Panin. «Risk of estimators for Sobol' sensitivity indices based on metamodels». В: *Electron. J. Statist.* 15.1 (2021), с. 235—281. ISSN: 1935-7524. DOI: 10.1214/20-EJS1793. URL: https://projecteuclid.org/euclid.ejs/1609902190

Публикации стандартного уровня

1. Evgeny Burnaev, Ivan Panin и Bruno Sudret. «Efficient design of experiments for sensitivity analysis based on polynomial chaos expansions». В: *Ann. Math. Artif. Intell.* 81.1-2 (2017), с. 187—207. ISSN: 1012-2443. DOI: 10.1007/s10472-017-9542-1. URL: https://doi.org/10.1007/s10472-017-9542-1

2. Evgeny Burnaev, Ivan Panin и Bruno Sudret. «Effective Design for Sobol Indices Estimation Based on Polynomial Chaos Expansions». В: *Conformal and Probabilistic Prediction with Applications.* Под ред. Alexander Gammerman и др. Cham: Springer International Publishing, 2016, с. 165—184. ISBN: 978-3-319-33395-3

3. Evgeny Burnaev и Ivan Panin. «Adaptive Design of Experiments for Sobol Indices Estimation Based on Quadratic Metamodel». В: *Statistical Learning and Data Sciences*. Под ред. Alexander Gammerman, Vladimir Vovk и Harris Papadopoulos. Cham: Springer International Publishing, 2015, с. 86—95. ISBN: 978-3-319-17091-6

Прочие публикации

1. Иван Панин и Павел Приходько. «Подходы к нахождению дисперсии оценок значимости признаков в задаче глобального анализа чувствительности». В: *Сборник трудов конференции "Информационные Технологии и Системы (ИТиС)"* (Петрозаводск, Россия). ИППИ РАН. 2012, с. 173—178. ISBN: 978-5-901158-19-7. URL: http://www.itas2012.iitp.ru/pdf/1569602539.pdf

Доклады на конференциях и семинарах

1. 5th International Symposium on Conformal and Probabilistic Prediction with Applications (2016, Мадрид, Испания);
2. 3rd International Symposium on Statistical Learning and Data Sciences (2015, Эгем, Великобритания);
3. 37-ая конференция-школа молодых ученых и специалистов "Информационные технологии и системы" (2013, Калининград, Россия);
4. 55-ая научная конференции МФТИ (2012, Долгопрудный, Россия);
5. 35-ая конференция молодых ученых и специалистов "Информационные технологии и системы" (2012, Петрозаводск, Россия);
6. Семинар по структурному обучению ИППИ РАН (2019, Москва, Россия).

Содержание работы

Во Введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследования, а также показана теоретическая и практическая значимость диссертационной работы.

В первой главе рассматривается постановка задачи глобального анализа чувствительности, дан пример подобной задачи из инженерной практики и описан подход к ее решению с использованием показателей чувствительности Соболя. Кроме того, рассмотрен способ расчета показателей Соболя с помощью метамоделей, основанных на разложении полиномиального хаоса.

В первой части вводятся базовые понятия и описан метод Соболя. Рассмотрим функцию $y = f(\mathbf{x})$, которая соответствует некоторой физической модели. Вектор входных переменных $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathscr{X} \subseteq \mathbb{R}^d$ — это *параметры модели (факторы)*, лежащие в *факторном пространстве*; а выходная переменная $y \in \mathbb{R}^1$ — *отклик модели.* Будем считать, что эта функция представляет собой "черный ящик", изучать который можно, подавая ему на вход различные наборы параметров и анализируя отклик. При этом предполагается, что расчет отклика модели может занимать продолжительное время.

Неформальная постановка задачи *глобального анализа чувствительности* заключается в том, чтобы при заданной каким-то образом изменчивости входа модели количественно оценить "значимость" различных входных параметров модели и их групп; и, таким образом, ранжировать их по степени влияния на отклик модели, выделив важные и несущественные параметры. При этом предполагается, что входные параметры могут меняться в некотором широком диапазоне, а не в узкой окрестности.

Один из способов формализации понятия "значимости" параметров связан с методом И. М. Соболя. Предположим, что на множестве $\mathscr{X}$ задана некоторая известная *a priori* вероятностная мера μ, которая является произведением мер: $\mu = \otimes_{i=1}^d \mu_i$, где μ_i — вероятностная мера на $\mathscr{X}_i \subseteq \mathbb{R}$, а множество $\mathscr{X}$ представимо в виде декартова произведения $\mathscr{X} = \mathscr{X}_1 \times \ldots \times \mathscr{X}_d$. Соответствующее вероятностное распределение описывает неопределенность и/или изменчивость входных параметров, которая моделируется случайным вектором $\mathbf{x} = (x_1, \ldots, x_d)^T$ с независимыми компонентами. Тогда выход модели $y = f(\mathbf{x})$ тоже становится случайной величиной.

Предположим, что функция f лежит в гильбертовом пространстве $L^2(\mathscr{X}, \mu)$ вещественнозначных функций на $\mathscr{X}$, квадратично-интегрируемых по мере μ. Тогда для нее существует следующее единственное разложение Соболя-Хёфдинга [18]

$$\begin{aligned} f(\mathbf{x}) &= f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{1 \le i < j \le d} f_{ij}(x_i, x_j) + \ldots + f_{1\ldots d}(x_1, \ldots, x_d) \\ &= \sum_{\mathcal{U} \subseteq \{1,\ldots,d\}} f_{\mathcal{U}}(\mathbf{x}_{\mathcal{U}}) \end{aligned} \tag{1}$$

с 2^d членами возрастающей размерности, которые удовлетворяют условию

$$\mathbb{E}_{\mu_i}[f_{\mathcal{U}}] = \int_{\mathscr{X}_i} f_{\mathcal{U}}(\mathbf{x}_{\mathcal{U}}) d\mu_i(x_i) = 0 \ \text{ для } \ \forall i \in \mathcal{U}, \tag{2}$$

где $\mathcal{U} \subseteq \{1, 2, \ldots, d\}$ — подмножество индексов входных переменных; $\mathbf{x}_{\mathcal{U}}$ — вектор с компонентами $(x_i, \ i \in \mathcal{U})^T$, а $f_\varnothing \triangleq f_0 = \mathbb{E}_\mu[f(\mathbf{x})]$. Слагаемые этого разложения ортогональны:

$$\mathbb{E}_\mu\big[f_\mathcal{U}(\mathbf{x}_\mathcal{U})f_\mathcal{V}(\mathbf{x}_\mathcal{V})\big] = 0, \text{ если } \mathcal{U} \neq \mathcal{V}, \text{ где } \mathcal{U}, \mathcal{V} \subseteq \{1, 2, \ldots, d\},$$

откуда можно получить следующее разложение дисперсии отклика модели:

$$\mathbb{V}_\mu[f(\mathbf{x})] = \sum_{\mathcal{U}\subseteq\{1,\ldots,d\}} \mathbb{V}_\mu[f_\mathcal{U}(\mathbf{x}_\mathcal{U})]. \tag{3}$$

Предполагая $\mathbb{V}_\mu[f] > 0$, введем *показатель чувствительности Соболя* (ПЧ), также известный как *индекс Соболя.*

Определение 1. *Показатель Соболя набора* $\mathbf{x}_\mathcal{U}$, $\mathcal{U} \subseteq \{1, \ldots, d\}$ *входных переменных функции определяется как*

$$S_\mathcal{U} = \frac{\mathbb{V}_\mu[f_\mathcal{U}(\mathbf{x}_\mathcal{U})]}{\mathbb{V}_\mu[f(\mathbf{x})]}. \tag{4}$$

Обозначим $\mu_\mathcal{U} \triangleq \otimes_{i\in\mathcal{U}}\mu_i$, $\mathbb{E}_\mathcal{U} \triangleq \mathbb{E}_{\mu_\mathcal{U}}$, $\mathbb{V}_\mathcal{U} \triangleq \mathbb{V}_{\mu_\mathcal{U}}$ и $\sim\!\mathcal{U} \triangleq \{1, \ldots, d\}\backslash\mathcal{U}$. Тогда для $\mathcal{U} = \{i\}$ показатель Соболя (4) можно представить как

$$S_i = \frac{\mathbb{V}_i\big[\mathbb{E}_{\sim i}[f(\mathbf{x})|x_i]\big]}{\mathbb{V}_\mu[f]}, \quad i = 1, \ldots, d, \tag{5}$$

Определим также величину, которая характеризует “полный” вклад группы переменных в изменчивость модели — *полный показатель чувствительности* (также известный как *показатель полного эффекта* и *total-индекс*).

Определение 2. *Полный показатель чувствительности набора* $\mathbf{x}_\mathcal{U}$, $\mathcal{U} \subseteq \{1, \ldots, d\}$ *входных переменных функции определяется как*

$$T_\mathcal{U} = \sum_{\mathcal{U}\cap\mathcal{V}\neq\varnothing} S_\mathcal{V} = \frac{\mathbb{E}_{\sim\mathcal{U}}\big[\mathbb{V}_\mathcal{U}[f(\mathbf{x})|\mathbf{x}_{\sim\mathcal{U}}]\big]}{\mathbb{V}_\mu[f]}. \tag{6}$$

Таким образом, введенные выше показатели чувствительности предлагают один из возможных путей, позволяющих формализовать и решить задачу глобального анализа чувствительности.

Во второй части рассмотрен способ расчета показателей Соболя и полных показателей чувствительности с помощью метамодельного подхода; в частности, с помощью аппроксимации на основе разложения полиномиального хаоса.

Непосредственный расчет показателей чувствительности приводит к вычислительно трудной задаче многомерного интегрирования. Чтобы упростить эту задачу, часто используют метамодельный подход, при котором исходная вычислительно тяжелая модель f заменяется вычислительно простой метамоделью $\widehat{f}$, лучше подходящей для расчета показателей Соболя.

В основном мы будем рассматривать метамодели на основе разложения полиномиального хаоса, которые часто встречаются в задачах анализа чувствительности. Введем их формально. Обозначим скалярное произведение и норму для $g, h \in L^2(\mathscr{X}, \mu)$ как $\langle g, h\rangle_\mu = \int_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x})h(\mathbf{x})d\mu(\mathbf{x})$ и $\|g\|_\mu^2 \triangleq \|g\|^2_{L^2(\mathscr{X},\mu)} = \int_{\mathbf{x}\in\mathscr{X}} g^2(\mathbf{x})d\mu(\mathbf{x})$; а *норму-супремум* как $\|g\|_{L^\infty} \triangleq \|g\|_{L^\infty(\mathscr{X})} \triangleq \sup_{\mathbf{x}\in\mathscr{X}} |g(\mathbf{x})|$. Кроме того, норма в евклидовых векторных пространствах обозначена как $\|\cdot\|$.

Предположим, что в $L^2(\mathscr{X}, \mu)$ существует набор $\{\Psi_{\boldsymbol{\alpha}}(\mathbf{x})\}$, параметризованный мультииндексом[3] $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, который состоит из μ-ортонормированных функций и имеет вид тензорного произведения d семейств μ_i-ортонормированных одномерных функций $\{\psi^{(i)}_{\alpha_i},\ \alpha_i \in \mathbb{N}\}$ с $\psi^{(i)}_0 \equiv 1$ и $\mathbb{E}_i\big[\psi^{(i)}_\alpha(x_i)\psi^{(i)}_\beta(x_i)\big] = \delta_{\alpha\beta}$ для $\alpha, \beta \in \mathbb{N}$, где δ — символ Кронекера. В результате имеем

$$\begin{gathered}\Psi_{\mathbf{0}}(\mathbf{x}) \equiv 1,\ \ \Psi_{\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{i=1}^d \psi^{(i)}_{\alpha_i}(x_i),\ \ \mathbf{x} \in \mathscr{X}, \\ \langle \Psi_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\beta}}\rangle_\mu = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}},\ \ \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^d.\end{gathered} \tag{7}$$

Определим метамодель в виде линейной комбинации N введенных выше функций $\{\Psi_{\boldsymbol{\alpha}}(\mathbf{x}),\ \boldsymbol{\alpha} \in \mathscr{L}_N\}$ для некоторого набора мультииндексов $\mathscr{L}_N \subset \mathbb{N}^d$:

$$\widehat{f}(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathscr{L}_N} \widehat{c}_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x}), \quad \mathbf{x} \in \mathscr{X},\ \ \widehat{c}_{\boldsymbol{\alpha}} \in \mathbb{R}. \tag{8}$$

Если функции $\{\Psi_{\boldsymbol{\alpha}}\}$ являются многомерными полиномами, то (8) соответствует метамодели на основе разложения полиномиального хаоса[4] (ПХ). Одним из важных преимуществ представленного типа аппроксимации является то, что он позволяет рассчитывать показатели Соболя аналитически по коэффициентам разложения [11]. Действительно, если $\mathbb{V}_\mu[\widehat{f}] > 0$, то показатель Соболя непустого набора входных переменных $\mathbf{x}_{\mathcal{U}}$, $\mathcal{U} \subseteq \{1, \ldots, d\}$ функции $\widehat{f}$ вида (8) выражается как

$$\widehat{S}_{\mathcal{U}}(\widehat{\mathbf{c}}) = \frac{\sum_{\boldsymbol{\alpha}\in\mathbb{L}_{\mathcal{U}}} \widehat{c}^2_{\boldsymbol{\alpha}}}{\sum_{\boldsymbol{\alpha}\in\mathscr{L}^+_N} \widehat{c}^2_{\boldsymbol{\alpha}}}, \tag{9}$$

где $\mathscr{L}^+_N \triangleq \mathscr{L}_N \backslash \mathbf{0}$, а $\mathbb{L}_{\mathcal{U}} \triangleq \mathbb{L}_{\mathcal{U}}[\mathscr{L}_N]$ — это подмножество $\mathscr{L}_N$, состоящее из мультииндексов, в которых отличны от нуля все индексы, соответствующие переменным $\mathbf{x}_{\mathcal{U}}$, и только они: $\mathbb{L}_{\mathcal{U}} = \{\boldsymbol{\alpha} \in \mathscr{L}^+_N\colon\ \alpha_i > 0$ для всех $i \in \mathcal{U};\ \alpha_i = 0$ для $i \notin \mathcal{U}\}$. Вектор коэффициентов $(\widehat{c}_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha} \in \mathscr{L}_N)^T$ обозначен как $\widehat{\mathbf{c}} \in \mathbb{R}^N$. Отметим, что $\mathbb{V}_\mu[\widehat{f}] = \sum_{\boldsymbol{\alpha}\in\mathscr{L}^+_N} \widehat{c}^2_{\boldsymbol{\alpha}}$. Полные показатели для $\widehat{f}$ рассчитываются аналогично.

В третьей части введена модель наблюдений и рассмотрены способы построения метамодели на основе разложения полиномиального хаоса для конечной

[3] Обозначим $\mathbb{N} \triangleq \{0, 1, 2, \ldots\}$, $\mathbb{N}_+ \triangleq \{0, 1, 2, \ldots\}$ и введем нулевой мультииндекс $\mathbf{0} \triangleq (0, \ldots, 0) \in \mathbb{N}^d$.

[4] Мы не ограничиваемся только полиномами. Вместе с тем, это важный, часто встречающийся частный случай, поэтому для простоты мы ссылаемся на аппроксимации типа (8) как на метамодели на основе разложения полиномиального хаоса.

обучающей выборки.

Модель наблюдений. Будем предполагать, что единственная информация о модели f поступает из наблюдений; точнее, что для некоторого *плана эксперимента* $\mathcal{D} = (\mathbf{x}_i \in \mathscr{X})_{i=1}^n \in \mathbb{R}^{n\times d}$ можно получить набор откликов модели и сформировать *обучающую выборку*:

$$\mathcal{S} = \big(\mathbf{x}_i,\ y_i = f(\mathbf{x}_i) + \eta_i\big)_{i=1}^n, \tag{10}$$

где η_i — это независимые и одинаково распределенные случайные ошибки измерения такие, что $\mathbb{E}\eta_i = 0$ и $\mathbb{V}\eta_i = \sigma^2 < \infty$; и независимые от $\mathbf{x}$. В матричной форме: $\mathcal{S} = \big(\mathcal{D} \in \mathscr{X}^n,\ Y = f(\mathcal{D}) + \boldsymbol{\eta} \in \mathbb{R}^n\big)$, где $f(\mathcal{D}) \triangleq \big(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\big)^T$ и $\boldsymbol{\eta} = \big(\eta_1, \ldots, \eta_n\big)^T$. Мы будем рассматривать как общий случай *шумных* наблюдений, так и специальный *бесшумный* случай, соответствующий $\sigma^2 = 0$.

Построение метамодели. Рассмотрим некоторую фиксированную последовательность вложенных наборов d-мерных мультииндексов:

$$\{\mathbf{0}\} = \mathscr{L}_1 \subset \ldots \subset \mathscr{L}_N \ldots \subset \mathscr{L}_\infty = \mathbb{N}^d, \tag{11}$$

где $|\mathscr{L}_N| = N$ для всех $N \in \mathbb{N}_+$. Каждому элементу этой последовательности соответствует своя метамодель вида (8). Будем называть каждый набор $\mathscr{L}_N$ *усечением* (*truncation set*), так как для соответствующей N-ой аппроксимации $\widehat{c}_{\boldsymbol{\alpha}} \triangleq 0$ при $\boldsymbol{\alpha} \notin \mathscr{L}_N$. Фиксируем некоторое усечение $\mathscr{L}_N$. Обозначим подпространство всех линейных комбинаций $\{\varPsi_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha} \in \mathscr{L}_N\}$ как $V_N \triangleq span\{\varPsi_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha} \in \mathscr{L}_N\}$. *Теоретическая ортогональная проекция* f на V_N относительно μ-нормы определяется как

$$f_N \triangleq \underset{\widehat{f} \in V_N}{\operatorname{argmin}} \|f - \widehat{f}\|_\mu. \tag{12}$$

Определим также $e_N(\mathbf{x}) \triangleq f(\mathbf{x}) - f_N(\mathbf{x})$. Функция f_N соответствует наилучшему возможному приближению модели f в пространстве V_N с точки зрения μ-нормы.

Рассмотрим теперь построение аппроксимации для конечной обучающей выборки. Пусть $\varPhi \triangleq \boldsymbol{\varPsi}(\mathcal{D}) = \big(\varPsi_{\boldsymbol{\alpha}}(\mathcal{D}),\ \ \boldsymbol{\alpha} \in \mathscr{L}_N\big) \in \mathbb{R}^{n\times N}$, а $\varPsi_{\boldsymbol{\alpha}}(\mathcal{D}) \triangleq \big(\varPsi_{\boldsymbol{\alpha}}(\mathbf{x}_1), \ldots, \varPsi_{\boldsymbol{\alpha}}(\mathbf{x}_n)\big)^T \in \mathbb{R}^n$. Среди всех методов оценки коэффициентов разложения полиномиального хаоса будем рассматривать следующие:

- Метод проекций на основе квази-регрессии [1]

$$\widehat{\mathbf{c}}^P = \frac{1}{n}\varPhi^T Y \in \mathbb{R}^N. \tag{13}$$

- Метод наименьших квадратов, МНК

$$\widehat{\mathbf{c}}^{LS} = (\varPhi^T\varPhi)^{-1}\varPhi^T Y, \text{ если } \det(\varPhi^T\varPhi) \neq 0. \tag{14}$$

Будем ссылаться на аппроксимации, построенные на основе этих двух методов, как на $\widehat{f}^P$ и $\widehat{f}^{LS}$. Показатели Соболя этих аппроксимаций обозначаются соответственно $\widehat{S}^P$ и $\widehat{S}^{LS}$.

Во второй главе выполнен теоретический анализ погрешности оценок показателей Соболя на основе произвольных метамоделей и установлена связь этой погрешности с точностью метамоделей. Исходя из этого анализа, предложен новый метод контроля качества для таких оценок. Также показано, что наши оценки сверху для погрешности показателей достижимы.

Если не указано иное, все дальнейшие результаты верны как для показателей Соболя, так и для полных показателей чувствительности, причем всех порядков. Чтобы избежать дублирования, мы используем в формулировках теорем обозначение $S_{\mathcal{U}}$ для показателей обоих типов.

Для начала убедимся, что близость функции и ее произвольного[5] приближения $f \approx \widehat{f}$ приводят к близости их показателей Соболя $S_{\mathcal{U}}$ и $\widehat{S}_{\mathcal{U}}$. Заметим, что обратное в общем случае неверно. Чтобы охарактеризовать близость функций, используется следующая относительная ошибка:

$$\mathscr{E} \triangleq \frac{\|f - \widehat{f}\|_\mu}{\mathbb{V}_\mu^{1/2}[f]}. \tag{15}$$

Теорема 1. *Для любых функций $f, \widehat{f} \in L^2(\mathscr{X}, \mu)$ таких, что $\mathbb{V}_\mu[f] > 0$, $\mathbb{V}_\mu[\widehat{f}] > 0$, и их показателей Соболя и полных показателей для $\mathcal{U} \subseteq \{1, \ldots, d\}$ справедливы соотношения*

$$\left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \ \le \ \left\{\sqrt{S_{\mathcal{U}}(1 - \widehat{S}_{\mathcal{U}})} + \sqrt{\widehat{S}_{\mathcal{U}}(1 - S_{\mathcal{U}})}\right\} \cdot \mathscr{E}, \tag{16}$$

$$\max_{\mathcal{U}} \left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \ \le \ \mathscr{E}. \tag{17}$$

Следствие 1. *В условиях Теоремы 1 для всех $\mathcal{U} \subseteq \{1, \ldots, d\}$ справедливо*

$$\left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \le \min\left(1,\ \mathscr{E} + 2\sqrt{S_{\mathcal{U}}},\ \mathscr{E} + 2\sqrt{1 - S_{\mathcal{U}}}\right) \cdot \mathscr{E}. \tag{18}$$

В частности, если для некоторого $\mathcal{U} \subseteq \{1, \ldots, d\}$ показатель Соболя или полный показатель $S_{\mathcal{U}} \in \{0, 1\}$, то

$$\left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \le \mathscr{E}^2. \tag{19}$$

Следующее следствие дает оценку для суммы ошибок показателей Соболя для всех 2^d различных подгрупп входных переменных (неверно для полных показателей).

Следствие 2. *Для любых функций $f, \widehat{f} \in L^2(\mathscr{X}, \mu)$ таких, что $\mathbb{V}_\mu[f] > 0$,*

[5] В частности, аппроксимация может быть не связана с разложением полиномиального хаоса.

$\mathbb{V}_\mu[\widehat{f}] > 0$, *и их показателей Соболя для* $\mathcal{U} \subseteq \{1,\dots,d\}$ *справедливо*

$$\sum_{\mathcal{U}} \left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \;\le\; 2\cdot\mathscr{E}, \tag{20}$$

$$\sum_{\mathcal{U}} \left(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right)^2 \;\le\; 2\cdot\mathscr{E}^2. \tag{21}$$

Следствие 1 позволяет предложить новый метод контроля качества оценок показателей Соболя на основе метамоделей (см. Алгоритм 1). Метод использует оценку

$$\left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \le \min\left(1,\ \mathscr{E}_2 + 2\sqrt{\widehat{S}_{\mathcal{U}}},\ \mathscr{E}_2 + 2\sqrt{1-\widehat{S}_{\mathcal{U}}}\right)\cdot\mathscr{E}_2, \tag{22}$$

где $\mathscr{E}_2 \triangleq \|f-\widehat{f}\|_\mu \cdot \min\left\{\mathbb{V}_\mu^{-1/2}[f], \mathbb{V}_\mu^{-1/2}[\widehat{f}]\right\}$, которая следует из симметрии Теоремы 1 относительно f и $\widehat{f}$. Величины в правой части (22), которые не могут быть вычислены аналитически, заменяются выборочными оценками. Погрешность аппроксимации оценивается с помощью отложенного контроля.

Для простоты предполагается, что используется метамодель типа (8), а в откликах нет дополнительного случайного шума ($\sigma^2 = 0$). Однако метод легко обобщается и на произвольные метамодели, и на шумный случай. Асимптотическая вычислительная сложность[6] Алгоритма 1 составляет $\mathcal{O}(n_t)$, n_t — размер тестовой выборки.

Покажем, что верхние оценки ошибок в Теореме 1 достижимы. Для этого с небольшими модификациями будут использованы две "модельные" функции вида

$$\begin{aligned} f(\mathbf{x}) &= c_1\cdot\Psi_{(1,0,0,\dots)}(\mathbf{x}) + c_2\cdot\Psi_{(0,1,0,\dots)}(\mathbf{x}),\\ \widehat{f}(\mathbf{x}) &= \widehat{c}_1\cdot\Psi_{(1,0,0,\dots)}(\mathbf{x}) + \widehat{c}_2\cdot\Psi_{(0,1,0,\dots)}(\mathbf{x}). \end{aligned} \tag{23}$$

Теорема 2. *Для любого подмножества* $\mathcal{U} \subseteq \{1,\dots,d\}$ *и любых значений* $S_{\mathcal{U}}, \widehat{S}_{\mathcal{U}} \in (0,1)$ *существуют* $f, \widehat{f} \in L^2(\mathscr{X},\mu)$, *имеющие показатели Соболя (полные показатели) набора переменных* $\mathbf{x}_{\mathcal{U}}$ *равные* $S_{\mathcal{U}}$ *и* $\widehat{S}_{\mathcal{U}}$ *соответственно, такие, что*

$$\left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| \;=\; \left\{\sqrt{S_{\mathcal{U}}(1-\widehat{S}_{\mathcal{U}})} + \sqrt{\widehat{S}_{\mathcal{U}}(1-S_{\mathcal{U}})}\right\}\cdot\mathscr{E}. \tag{24}$$

Теорема 3. *Для любого* $\varepsilon \in [0,1]$ *существуют* $f, \widehat{f} \in L^2(\mathscr{X},\mu)$ *такие, что для их показателей Соболя (и для полных показателей) выполнено*

$$\max_{\mathcal{U}\subseteq\{1,\dots,d\}} \left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right| = \varepsilon\cdot\mathscr{E}. \tag{25}$$

[6] Число регрессоров N, размерность d и время расчета отклика $f(\mathbf{x}_i)$ предполагаются константными.

Алгоритм 1. Оценка ошибок показателей Соболя / полных показателей.

Параметры: размер тестовой выборки n_t; построенная аппроксимация $\widehat{f} = \sum_{\boldsymbol{\alpha}\in\mathscr{L}_N} \widehat{c}_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}$.

1. Получить выборку $\mathcal{D}_t = (\mathbf{x}_i \in \mathscr{X})_{i=1}^{n_t}$ из распределения μ и отклики $f(\mathbf{x}_i)$ и $\widehat{f}(\mathbf{x}_i)$ для каждого $\mathbf{x}_i \in \mathcal{D}_t$.
2. $\mathscr{M} \leftarrow \frac{1}{n_t}\sum_{i=1}^{n_t}\left[f(\mathbf{x}_i) - \widehat{f}(\mathbf{x}_i)\right]^2$, где все $\mathbf{x}_i \in \mathcal{D}_t$.
3. $\mathscr{V}_1 \leftarrow \frac{1}{n_t}\sum_{i=1}^{n_t}\left[f(\mathbf{x}_i) - \frac{1}{n_t}\sum_{j=1}^{n_t} f(\mathbf{x}_j)\right]^2$, где все $\mathbf{x}_i$, $\mathbf{x}_j \in \mathcal{D}_t$
4. $\mathscr{V}_2 \leftarrow \sum_{\boldsymbol{\alpha}\in\mathscr{L}_N^+} \widehat{c}_{\boldsymbol{\alpha}}^2$.
5. $\widehat{\mathscr{E}}_2 \leftarrow \sqrt{\mathscr{M} / \max(\mathscr{V}_1, \mathscr{V}_2)}$.
6. Для каждого $\mathcal{U} \subseteq \{1, \ldots, d\}$:
   - 6.1. Вычислить $\widehat{S}_{\mathcal{U}}$ из коэффициентов разложения $\widehat{c}_{\boldsymbol{\alpha}}$.
   - 6.2. $\mathscr{Q}_{\mathcal{U}} \leftarrow \min\left(1,\ \widehat{\mathscr{E}}_2 + 2\sqrt{\widehat{S}_{\mathcal{U}}},\ \widehat{\mathscr{E}}_2 + 2\sqrt{1 - \widehat{S}_{\mathcal{U}}}\right) \cdot \widehat{\mathscr{E}}_2$.

Выход: оценки сверху для $\left|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}\right|$ для всех $\mathcal{U} \subseteq \{1, \ldots, d\}$, $\{\mathscr{Q}_{\mathcal{U}}\}$.

В целом, оценки ошибок в Теореме 1 могут быть завышены, но они в принципе достижимы. Таким образом, проблема точности оценок показателей Соболя сводится к оценке качества аппроксимации.

В третьей главе получены верхние неасимптотические границы риска оценок показателей Соболя для случайного плана эксперимента. Исследованы асимптотические свойства этих границ и рассмотрены факторы, которые обусловливают быструю сходимость.

В первой части рассматривается риск оценок $\widehat{S}^P$ и $\widehat{S}^{LS}$, связанных с двумя способами расчета коэффициентов разложения полиномиального хаоса. Все границы риска в этой главе получены при условии случайного плана эксперимента:

Условие 1 (случайного плана). *Пусть план эксперимента $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ состоит из независимых и одинаково распределенных случайных величин с распределением μ.*

Предположим, что существует некоторая детерминированная процедура обучения, которая строит приближение на основе обучающей выборки (10):

$$\mathcal{L}\colon \left(\mathcal{D},\ Y = f(\mathcal{D}) + \boldsymbol{\eta}\right) \to \widehat{f} \tag{26}$$

такая, что фиксированная реализация обучающей выборки всегда приводит к одной и той же аппроксимации $\widehat{f} \in L^2(\mathscr{X}, \mu)$.

Следующая теорема устанавливает связь между риском оцениваемых показателей Соболя и квадратичным риском аппроксимации $\mathbb{E}\|f - \widehat{f}\|_\mu^2$, где матожидание берется по плану эксперимента и случайным ошибкам измерения. В дальнейшем будем считать, что $\mathbb{V}_\mu[f] > 0$. Дополнительно определим показатели Соболя $\widehat{S}_\mathcal{U} = 2^{-d}$ в случае $\mathbb{V}_\mu[\widehat{f}] = 0$.

Теорема 4. *Пусть $\widehat{f}$ — произвольное приближение f, построенное согласно процедуре $\mathcal{L}$, удовлетворяющей (26). Предположим, что при Условии 1 случайного плана существует $\mathbb{E}\|f - \widehat{f}\|_\mu^2 < \infty$. Тогда для соответствующих показателей Соболя и полных показателей функций f и $\widehat{f}$ для $\mathcal{U} \subseteq \{1, \ldots, d\}$ справедливо*

$$\max_{\mathcal{U}} \mathbb{E}\big(S_\mathcal{U} - \widehat{S}_\mathcal{U}\big)^2 \;\leq\; \mathcal{R}^2, \tag{27}$$

$$\mathbb{E}\big|S_\mathcal{U} - \widehat{S}_\mathcal{U}\big| \;\leq\; \mathcal{R}\left(\mathcal{R} + 2\sqrt{S_\mathcal{U}}\right), \tag{28}$$

$$\text{где } \mathcal{R}^2 \;\triangleq\; \frac{\mathbb{E}\|f - \widehat{f}\|_\mu^2}{\mathbb{V}_\mu[f]}. \tag{29}$$

Следствие 3. *В условиях Теоремы 4 для соответствующих показателей Соболя[7] функций f и $\widehat{f}$ для $\mathcal{U} \subseteq \{1, \ldots, d\}$ выполнено*

$$\mathbb{E}\Big[\sum_{\mathcal{U}} \big(S_\mathcal{U} - \widehat{S}_\mathcal{U}\big)^2\Big] \leq 2 \cdot \mathcal{R}^2. \tag{30}$$

Метод проекций. Рассмотрим теперь не аппроксимацию общего вида, а метамодель на основе разложения полиномиального хаоса.

Условие 2 (ограниченности)**.** *Дополнительно потребуем, чтобы f была ограничена на $\mathscr{X}$:*

$$\big|f(\mathbf{x})\big| \leq L \quad \text{для } \mathbf{x} \in \mathscr{X}. \tag{31}$$

Теорема 5. *При Условии 1 случайного плана и Условии 2 ограниченности для соответствующих показателей Соболя и полных показателей функций f и $\widehat{f}^P$ для $\mathcal{U} \subseteq \{1, \ldots, d\}$ справедливо*

$$\max_{\mathcal{U}} \mathbb{E}\big(S_\mathcal{U} - \widehat{S}_\mathcal{U}^P\big)^2 \;\leq\; \mathcal{R}_p^2, \tag{32}$$

$$\mathbb{E}\big|S_\mathcal{U} - \widehat{S}_\mathcal{U}^P\big| \;\leq\; \mathcal{R}_p\left(\mathcal{R}_p + 2\sqrt{S_\mathcal{U}}\right), \tag{33}$$

$$\text{где } \mathcal{R}_p^2 \;\triangleq\; \frac{1}{\mathbb{V}_\mu[f]} \cdot \|e_N\|_\mu^2 + \frac{L^2 + \sigma^2}{\mathbb{V}_\mu[f]} \cdot \frac{N}{n}.$$

[7] Неверно для полных показателей чувствительности.

Следствие 4. *В условиях Теоремы 5 предположим дополнительно, что* $\lim_{N\to\infty} \|e_N\|_\mu = 0$. *Пусть* $N = N(n)$,

$$\frac{N}{n} \to 0 \ \ \text{и} \ \ N \to \infty \ \ \text{при} \ \ n \to \infty,$$

тогда

$$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}^P_{\mathcal{U}}\big)^2 \underset{n\to\infty}{\longrightarrow} 0.$$

Метод наименьших квадратов. Введем числовую характеристику, которую часто используют для контроля “устойчивости” МНК-оценок:

Определение 3. *Для ортонормированного набора функций* $\{\Psi_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha} \in \mathscr{L}_N\}$, *который удовлетворяет (7), и для некоторой фиксированной последовательности усечений (11) определим*

$$K_N \triangleq \sup_{\mathbf{x}\in\mathscr{X}} \Big[\sum_{\boldsymbol{\alpha}\in\mathscr{L}_N} \Psi^2_{\boldsymbol{\alpha}}(\mathbf{x})\Big]. \tag{34}$$

Обозначим *спектральную норму* матрицы $A \in \mathbb{R}^{m\times p}$ как

$$|||A||| = \max_{\mathbf{z}\in\mathbb{R}^p:\ \|\mathbf{z}\|\neq 0} \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|}. \tag{35}$$

Нам потребуется следующий результат, который получил Коэн [6].

Лемма 1 (Коэн, 2013)**.** *При Условии 1 случайного плана и* $\delta \in (0,1)$ *справедливо*

$$P\Big\{|||\Phi^T\Phi/n - I_N||| > \delta\Big\} \le 2N \cdot \exp\left[-\frac{c_\delta \cdot n}{K_N}\right], \tag{36}$$

где $c_\delta \triangleq (1+\delta)\ln(1+\delta) - \delta > 0$.

Лемма 1 приводит к условию на размер обучающей выборки n и число регрессоров N, которое с высокой вероятностью позволяет исключить возможность плохой обусловленности нормализованной информационной матрицы $\Phi^T\Phi/n$.

Условие 3 (устойчивости)**.** *Пусть для некоторого фиксированного* $r > 0$ *соотношение* N *и* n *удовлетворяет условию*

$$K_N \le \varkappa_r \cdot \frac{n}{\ln n}, \ \ \text{где} \ \ \varkappa_r = \frac{3\cdot\ln(3/2) - 1}{2+2r}. \tag{37}$$

При Условии 3 устойчивости, опираясь на Лемму 1, имеем

$$P\Big\{|||\Phi^T\Phi/n - I_N||| > 1/2\Big\} \le 2n^{-r}. \tag{38}$$

Теорема 6. *При Условии 1 случайного плана и Условии 3 устойчивости для соответствующих показателей Соболя и полных показателей функций f и $\widehat{f}^{LS}$ для $\mathcal{U} \subseteq \{1, \ldots, d\}$ справедливо*

$$\max_{\mathcal{U}} \mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big)^2 \ \leq\ \mathcal{R}_{LS}^2 + 2n^{-r}, \tag{39}$$

$$\mathbb{E}\big|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big| \ \leq\ \mathcal{R}_{LS}\left(\mathcal{R}_{LS} + 2\sqrt{S_{\mathcal{U}}}\right) + 2n^{-r}, \tag{40}$$

$$\text{где } \mathcal{R}_{LS}^2 \ \triangleq\ \frac{1.2}{\mathbb{V}_\mu[f]} \cdot \|e_N\|_\mu^2 + \frac{4\sigma^2}{\mathbb{V}_\mu[f]} \cdot \frac{N}{n}.$$

Следствие 5. *В предположениях Теоремы 6 для случая бесшумных наблюдений, т.е. при $\sigma^2 = 0$, справедливо*

$$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big)^2 \ \leq\ \frac{1.2}{\mathbb{V}_\mu[f]}\|e_N\|_\mu^2 + 2n^{-r}, \tag{41}$$

$$\mathbb{E}\big|S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big| \ \leq\ \frac{1.2}{\mathbb{V}_\mu[f]}\|e_N\|_\mu^2 + \frac{2.2 \cdot S_{\mathcal{U}}^{1/2}}{\mathbb{V}_\mu^{1/2}[f]}\|e_N\|_\mu + 2n^{-r}. \tag{42}$$

Следствие 6. *В условиях Теоремы 6 за исключением Условия 3 предположим дополнительно, что $\lim_{N\to\infty}\|e_N\|_\mu = 0$. Пусть $N = N(n)$,*

$$\frac{K_N \cdot \ln N}{n} \to 0 \ \text{ и } \ N \to \infty \ \text{ при } \ n \to \infty,$$

тогда

$$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big)^2 \underset{n\to\infty}{\longrightarrow} 0.$$

Во второй части выполнен асимптотический анализ полученных верхних границ риска с ростом размера выборки для аппроксимаций на основе конкретных семейств многомерных полиномов. Будем предполагать, что анализируемая функция f является *p-гладкой.*

В качестве основы для аппроксимации рассмотрим три семейства полиномов: Лежандра, тригонометрические и Чебышева, первые два из которых ортогональны относительно непрерывного равномерного распределения, а последнее — относительно распределения арксинуса. В соответствии с (7) неконстантные элементы этих семейств дополнительно нормализованы так, чтобы иметь единичную дисперсию относительно соответствующих распределений.

Замечание 1. *Для случая тригонометрических полиномов дополнительно потребуем, чтобы функцию f можно было продолжить вне $\mathscr{X} = [0,1]^d$ так, чтобы она была 1-периодической по каждому входному аргументу.*

Для асимптотического анализа будет использована *схема усечения (truncation scheme)* на основе максимальной степени одномерных полиномиаль-

Таблица 1: Асимптотические верхние границы квадратичного риска оценок показателей Соболя и полных показателей в зависимости от размера выборки n, размерности d и гладкости p.

<table>
<tr><td colspan="2">Полиномы</td><td>Лежандра</td><td>Чебышева</td><td>Тригонометрические</td></tr>
<tr><td colspan="2">Распределение</td><td>$U([-1,1]^d)$</td><td>$Arc\,([-1,1]^d)$</td><td>$U([0,1]^d)$</td></tr>
<tr><td rowspan="2">$\sigma^2 = 0$</td><td>$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big)^2 \lesssim$</td><td>$\left(\frac{n}{\ln n}\right)^{-p/d}$</td><td>$\left(\frac{n}{\ln n}\right)^{-2p/d}$</td><td>$\left(\frac{n}{\ln n}\right)^{-2p/d}$</td></tr>
<tr><td>$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{P}\big)^2 \lesssim$</td><td colspan="3">$n^{-\frac{2p}{2p+d}}$</td></tr>
<tr><td rowspan="2">$\sigma^2 > 0$</td><td>$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{LS}\big)^2 \lesssim$</td><td>$n^{-\frac{2p}{2p+d}}, \quad p/d > 1/2$
$\left(\frac{n}{\ln n}\right)^{-p/d}, \; p/d \le 1/2$</td><td>$n^{-\frac{2p}{2p+d}}$</td><td>$n^{-\frac{2p}{2p+d}}$</td></tr>
<tr><td>$\mathbb{E}\big(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}}^{P}\big)^2 \lesssim$</td><td colspan="3">$n^{-\frac{2p}{2p+d}}$</td></tr>
</table>

ных множителей. Для некоторого $\alpha_{max} \in \mathbb{N}_+$ зададим

$$\mathscr{L}_N = \{\boldsymbol{\alpha} \in \mathbb{N}^d \colon \max_{i=1,\ldots,d} \{\alpha_i\} \le \alpha_{max}\}, \tag{43}$$

где $N = |\mathscr{L}_N| = (\alpha_{max} + 1)^d$.

Таблица 1, основанная на результатах Теорем 5 и 6, суммирует асимптотические[8] границы риска оценок показателей Соболя и полных показателей для двух методов расчета коэффициентов разложения полиномиального хаоса и трех видов полиномов. При получении этих границ предполагалось, что число регрессоров N выбирается асимптотически оптимальным образом, то есть так, чтобы минимизировать результирующий риск.

Можно заключить, что отсутствие дополнительного шума при измерении отклика анализируемой функции, а также ее высокая гладкость и низкая размерность являются ключевыми факторами, обеспечивающими возможность быстрой сходимости оценок показателей Соболя.

В четвертой главе рассмотрен еще один – асимптотический – подход для анализа качества оценок показателей Соболя и предложен метод последовательного планирования эксперимента на его основе, обеспечивающий высокую точность этих оценок.

В этой главе мы рассматриваем только показатели Соболя вида $S_i \triangleq S_{\{i\}}$,

[8] Если заданы две последовательности положительных чисел $\{a_n\}$ и $\{b_n\}$, то $a_n \lesssim b_n$ означает, что последовательность $\{a_n/b_n\}$ ограничена.

где $i = 1, \ldots, d$, называемые *показателями чувствительности первого порядка*. Кроме этого, мы исходим из упрощенной[9] модели данных, в которой план эксперимента фиксирован (не случаен), а анализируемая функция имеет вид разложения полиномиального хаоса (8) с конечным число членов $f(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathscr{L}_N} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathbf{x})$, где $\mathbb{V}_\mu[f] = \sum_{\boldsymbol{\alpha} \in \mathscr{L}_N^+} c_{\boldsymbol{\alpha}}^2 > 0$. Отклик модели формируется как

$$y = \mathbf{c}^T \boldsymbol{\Psi}(\mathbf{x}) + \eta, \tag{44}$$

где $\eta \sim N(0, \sigma^2)$, $\sigma^2 > 0$ — независимый *гауссовский* шум, а $\boldsymbol{\Psi}(\mathbf{x}) \triangleq \big(\Psi_{\boldsymbol{\alpha}}(\mathbf{x}),\ \boldsymbol{\alpha} \in \mathscr{L}_N\big)^T \in \mathbb{R}^N$, а метамодель имеет вид $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{c}}^T \boldsymbol{\Psi}(\mathbf{x})$. При этом коэффициенты разложения оцениваются по обучающей выборке только с помощью МНК.

Определим *информационную матрицу* $A_n \in \mathbb{R}^{N \times N}$ как

$$A_n = \sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{x}_i) \boldsymbol{\Psi}^T(\mathbf{x}_i). \tag{45}$$

Зададим вектор-функцию показателей чувствительности первого порядка с компонентами вида (9) как $\mathbf{S}(\mathbf{c}) \triangleq \big(S_1(\mathbf{c}), \ldots, S_d(\mathbf{c})\big)^T$ и определим для нее матрицу Якоби (размера d на N)

$$B \triangleq B(\mathbf{c}) = \frac{\partial \mathbf{S}(\mathbf{c})}{\partial \mathbf{c}}. \tag{46}$$

Теорема 7.
Пусть выполнены условия:

1. *В факторном пространстве существует некоторая бесконечная детерминированная последовательность точек $\{\mathbf{x}_i \in \mathscr{X}\}_{i=1}^{\infty}$ такая, что*

$$\frac{1}{n} A_n \underset{n \to \infty}{\longrightarrow} H, \tag{47}$$

   где $H \in \mathbb{R}^{N \times N}$ — некоторая симметричная, положительно определенная матрица.

2. *Точки из этой последовательности $\{\mathbf{x}_i \in \mathscr{X}\}_{i=1}^{\infty}$ и соответствующие отклики (44) итеративно добавляются в обучающую выборку.*

3. *Для истинных коэффициентов $\mathbf{c}$ выполнено $\det\big(B(\mathbf{c}) H^{-1} B^T(\mathbf{c})\big) \neq 0$.*

Тогда имеет место следующая сходимость по распределению:

$$\sqrt{n}\big(\mathbf{S}(\mathbf{c}) - \mathbf{S}(\widehat{\mathbf{c}}_n)\big) \underset{n \to \infty}{\overset{d}{\longrightarrow}} N\big(0,\ \sigma^2 B H^{-1} B^T\big). \tag{48}$$

[9]Таким образом, отличие от представленной ранее модели данных состоит в том, что здесь $e_N(\mathbf{x}) \equiv 0$, шум гауссовский, а Условие 1 случайного плана не накладывается. Кроме того, структура метамодели фиксирована.

Алгоритм 2. План эксперимента для оценки показателей Соболя.

Параметры: начальное число точек в плане эксперимента m и их конечное число $n > m$; набор точек-кандидатов Ξ.

Инициализация: начальная обучающая выборка $(\mathcal{D}, Y)$ из m примеров такая, что план $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^m \subset \Xi$ приводит к невырожденной информационной матрице $A_m = \boldsymbol{\Psi}^T(\mathcal{D}) \cdot \boldsymbol{\Psi}(\mathcal{D})$; оценка $\widehat{\mathbf{c}}_m = A_m^{-1} T_m$, где $T_m = \boldsymbol{\Psi}^T(\mathcal{D})\, Y$.

Итерации: пока план $\mathcal{D}$ содержит меньше n примеров:

1. $\mathbf{x} \leftarrow \arg\min_{\mathbf{x}\in\Xi} \det\left[B(\widehat{\mathbf{c}}) \cdot \left\{A + \boldsymbol{\Psi}(\mathbf{x}) \cdot \boldsymbol{\Psi}^T(\mathbf{x})\right\}^{-1} \cdot B^T(\widehat{\mathbf{c}})\right]$. // см. (49)
2. $\mathcal{D} \leftarrow \mathcal{D} \cup \{\mathbf{x}\}, \quad Y \leftarrow Y \cup \{f(\mathbf{x})\}$.
3. $A \leftarrow A + \boldsymbol{\Psi}(\mathbf{x}) \cdot \boldsymbol{\Psi}^T(\mathbf{x}), \quad T \leftarrow T + f(\mathbf{x}) \cdot \boldsymbol{\Psi}(\mathbf{x})$.
4. $\widehat{\mathbf{c}} \leftarrow A^{-1} T$. // обновляем оценку согласно (14)

Выход: план эксперимента, $\mathcal{D}$.

Теорема 7 приводит к идее Алгоритма 2 последовательного планирования эксперимента для оценки показателей Соболя. Алгоритм использует критерий D-оптимальности, при этом на каждой итерации алгоритма для вычисления детерминанта ковариационной матрицы $\sigma^2 \cdot B(\mathbf{c}) H^{-1} B^T(\mathbf{c})$ используются текущие оценки коэффициентов разложения и текущая нормализованная информационная матрица. Вместо того, чтобы считать детерминант для каждой точки-кандидата на шаге 1. Алгоритма 2, мы используем равносильную оптимизационную задачу с вычислительно эффективным решением:

$$\frac{\boldsymbol{\Psi}^T(\mathbf{x}) \cdot \left\{A^{-1} B^T (B A^{-1} B^T)^{-1} B A^{-1}\right\} \cdot \boldsymbol{\Psi}(\mathbf{x})}{1 + \boldsymbol{\Psi}^T(\mathbf{x}) \cdot A^{-1} \cdot \boldsymbol{\Psi}(\mathbf{x})} \to \max_{\mathbf{x}\in\Xi}. \tag{49}$$

Асимптотическая вычислительная сложность[10] Алгоритма 2 составляет $\mathcal{O}(n)$, где n — конечный размер плана после добавления всех новых точек.

В целом, наш подход заключается в том, чтобы получить (асимптотическое) нормальное распределение оценок показателей Соболя с ковариационной матрицей, зависящей от плана, и применить для планирования эксперимента один из критериев оптимальности. Алгоритм 2 иллюстрирует эту идею на примере критерия D-оптимальности, однако можно воспользоваться и другими критериями. Например, Пронзато [11] применил предложенный подход, используя A- и

[10] Число регрессоров N, размерность d, количество точек-кандидатов $|\Xi|$ и время расчета отклика $f(\mathbf{x}_i)$ предполагаются константными.

MV-критерий. Выбор конкретного критерия оптимальности плана остается за исследователем анализируемой модели и должен быть сделан исходя из специфики задачи. В частности, критерий D-оптимальности можно рекомендовать в случаях, когда более важна “средняя” ошибка оценок показателей Соболя, а не ее максимальное значение по всем группам входных параметров.

В пятой главе описан разработанный автором программный комплекс и даны результаты вычислительных экспериментов для предложенных алгоритмов.

В первой части приведено описание разработанного комплекса программ на языке Python, в который вошли созданные в работе алгоритмы контроля качества и планирования эксперимента. Помимо самих алгоритмов, комплекс также включает тестовое окружение для них (тестовые анализируемые функции и альтернативные методы из литературы), которое позволяет сравнить предложенные подходы с аналогами.

Во второй части выполнено тестирование метода контроля качества оценок показателей Соболя (Алгоритма 1). Для сравнения мы используем выборочные границы ошибок на основе *бутстрэп* метода [7]. Примеры результатов для 2-мерной g-функции Соболя показаны на Рисунке 1a. При оценке качества аппроксимации 15% выборки используется для отложенного контроля.

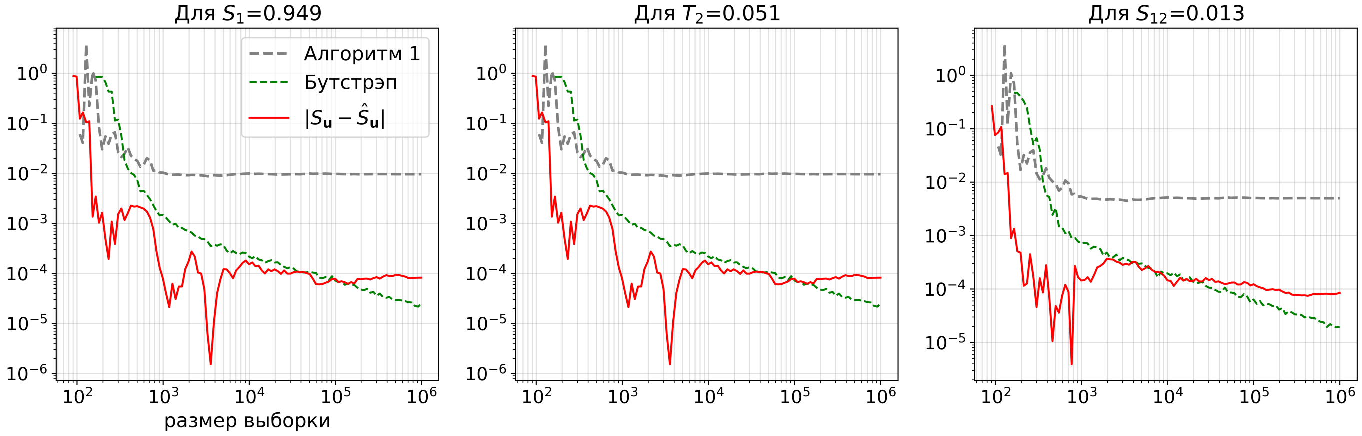
В третьей части дан экспериментальный анализ границ риска, полученных в Теоремах 5 и 6, и показано, что метамодельный подход действительно позволяет достичь высокой скорости сходимости оценок показателей Соболя к их истинным значениям.

В качестве иллюстрации на Рисунке 1б показана эмпирическая оценка риска $\max_{\mathcal{U}} \left\{\mathbb{E}(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}})^2,\ \mathbb{E}(T_{\mathcal{U}} - \widehat{T}_{\mathcal{U}})^2\right\}$ и компоненты его границ из упомянутых теорем $\|e_N\|^2_{\mu}/\mathbb{V}_{\mu}[f]$ и n^{-r} для g-функции Соболя в бесшумном случае.

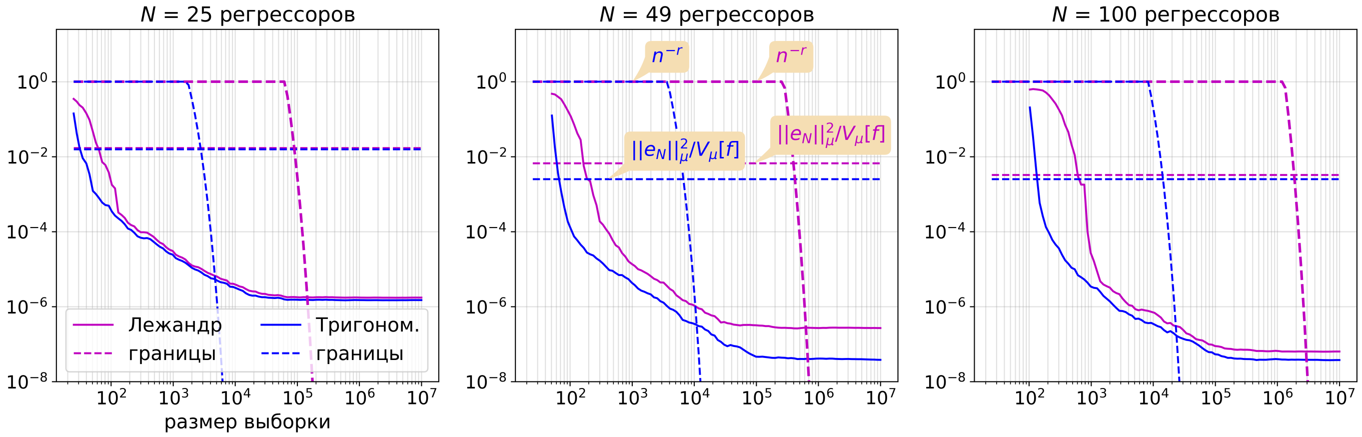
В четвертой части приведены результаты применения предложенного метода последовательного планирования эксперимента (Алгоритма 2) для решения серии искусственных и реальных инженерных задач (с использованием конечно-элементных моделей), размерность факторного пространства в которых варьируется от 2 до 53. Эксперимент предполагает, что в некоторый начальный случайный план итеративно добавляются новые точки. Предложенный метод сравнивается со следующими техниками планирования эксперимента:

- *Случайный план*, который предполагает последовательное добавление случайных точек из набора кандидатов Ξ.
- *Последовательный D-оптимальный план [12]*, основанный на итеративной максимизации детерминанта информационной матрицы: $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in\Xi} \det\left[A_n + \boldsymbol{\Psi}(\mathbf{x}) \cdot \boldsymbol{\Psi}^T(\mathbf{x})\right]$.
- *ЛГК* — сэмплирование на основе *латинского гиперкуба*. Техника является не итеративной, то есть на каждой итерации обновляются все точки плана.

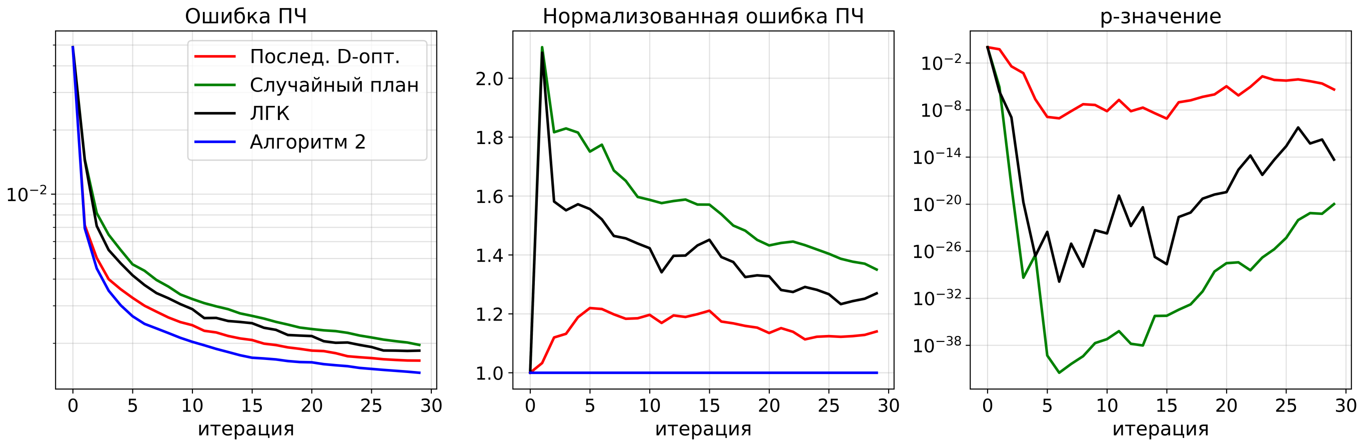
Эффективность предложенного подхода иллюстрирует Рисунок 1в, где показаны результаты перечисленных техник планирования для модели прогиба стержневой конструкции (фермы) под действием внешних сил из работ Судрэ [2] и Ли [9], у которой $d = 10$ входных параметров имеют непрерывное равномерное распределение. Метрикой качества плана эксперимента была ошибка показателей Соболя $\sqrt{\sum_{i=1}^{d}(S_i - \widehat{S}_i)^2}$, усредненная по нескольким запускам техники планирования эксперимента с различными начальными планами. Для удобства на рисунке также приведены значения ошибок для всех техник относительно Алгоритма 2. Кроме того, мы используем t-критерий Уэлча, чтобы убедиться, что разница усредненных ошибок показателей статистически значима. Видно, что оценки показателей Соболя, полученные на основе Алгоритма 2, в среднем являются более точными на всех итерациях; в частности, после добавления 29 новых точек средняя точность этих оценок на 10% превосходит результат, который дает наиболее точный из остальных сравниваемых методов — последовательный D-оптимальный план.



(а) Ошибки оценок трех показателей Соболя и их выборочные границы сверху для Алгоритма 1 и бутстрэп метода. Метамодель на основе ПХ Лежандра, $N = 91$ регрессор, МНК.



(б) Квадратичный риск оценок показателей Соболя $\max_{\mathcal{U}} \left\{\mathbb{E}(S_{\mathcal{U}} - \widehat{S}_{\mathcal{U}})^2,\ \mathbb{E}(T_{\mathcal{U}} - \widehat{T}_{\mathcal{U}})^2\right\}$ и компоненты его границ из Теоремы 6. Метамодели на основе ПХ Лежандра и тригонометрических полиномов с различным числом регрессоров. МНК. Бесшумный случай, $\sigma^2 = 0$.



(в) Средняя ошибка оценок показателей Соболя и она же нормализованная; и p-значения t-критерия Уэлча для различных техник планирования. ПХ Лежандра, $N = 176$ регрессоров, МНК. Начальный план $m = 176$ точек.

Рис. 1: Некоторые результаты вычислительных экспериментов.

Выводы

1. Установлено соотношение между ошибкой оценок показателей Соболя и ошибкой аппроксимации, на основе которой эти оценки были получены. Указанное соотношение справедливо для показателей Соболя и полных показателей всех порядков. В частности показано, что максимальная по всем группам переменных абсолютная ошибка оценок показателей Соболя ограничена относительной ошибкой соответствующей аппроксимации, причем эта граница достижима.

2. Благодаря полученной теоретической границе ошибки, разработан метод контроля погрешности метамодельных оценок показателей Соболя.

3. В условиях случайного плана эксперимента для аппроксимации на основе разложения полиномиального хаоса получены неасимптотические верхние границы риска метамодельных оценок показателей Соболя. Кроме того, для этих границ найдены оценки скорости сходимости в случае анализируемых функций различной гладкости и аппроксимаций, использующих полиномы Лежандра, Чебышева и тригонометрические полиномы.

4. Найдено асимптотическое распределение оценок показателей Соболя, что позволило разработать метод последовательного планирования эксперимента для оценки показателей чувствительности с помощью аппроксимации на основе разложения полиномиального хаоса.

5. Разработан программный комплекс для решения задач, связанных с моделированием в инженерном проектировании, в который вошли предложенные методы контроля качества оценок показателей Соболя и планирования эксперимента.

6. Эффективность разработанного программного комплекса была продемонстрирована при решении ряда инженерных задач; в том числе задачи анализа факторов, которые влияют на величину прогиба стержневой конструкции (фермы), находящейся под действием внешних сил.

Список литературы

[1] Jian An и Art Owen. «Quasi-regression». В: т. 17. 4. Complexity of multivariate problems (Kowloon, 1999). 2001, с. 588—607. DOI: 10.1006/jcom.2001.0588. URL: https://doi.org/10.1006/jcom.2001.0588.

[2] Géraud Blatman и Bruno Sudret. «An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis». В: *Probabilistic Engineering Mechanics* 25 (апр. 2010), с. 183—197. DOI: 10.1016/j.probengmech.2009.10.003.

[3] Evgeny Burnaev и Ivan Panin. «Adaptive Design of Experiments for Sobol Indices Estimation Based on Quadratic Metamodel». В: *Statistical Learning and Data Sciences*. Под ред. Alexander Gammerman, Vladimir Vovk и Harris Papadopoulos. Cham: Springer International Publishing, 2015, с. 86—95. ISBN: 978-3-319-17091-6.

[4] Evgeny Burnaev, Ivan Panin и Bruno Sudret. «Effective Design for Sobol Indices Estimation Based on Polynomial Chaos Expansions». В: *Conformal and Probabilistic Prediction with Applications*. Под ред. Alexander Gammerman и др. Cham: Springer International Publishing, 2016, с. 165—184. ISBN: 978-3-319-33395-3.

[5] Evgeny Burnaev, Ivan Panin и Bruno Sudret. «Efficient design of experiments for sensitivity analysis based on polynomial chaos expansions». В: *Ann. Math. Artif. Intell.* 81.1-2 (2017), с. 187—207. ISSN: 1012-2443. DOI: 10.1007/s10472-017-9542-1. URL: https://doi.org/10.1007/s10472-017-9542-1.

[6] Albert Cohen, Mark A. Davenport и Dany Leviatan. «On the stability and accuracy of least squares approximations». В: *Found. Comput. Math.* 13.5 (2013), с. 819—834. ISSN: 1615-3375. DOI: 10.1007/s10208-013-9142-3. URL: https://doi.org/10.1007/s10208-013-9142-3.

[7] S. Dubreuil и др. «Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion». В: *Reliability Engineering & System Safety* 121 (2014), с. 263—275. ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2013.09.011. URL: http://www.sciencedirect.com/science/article/pii/S0951832013002688.

[8] Bertrand Iooss и Paul Lemaître. «A Review on Global Sensitivity Analysis Methods». В: *Operations Research/ Computer Science Interfaces Series* 59 (апр. 2014). DOI: 10.1007/978-1-4899-7547-8_5.

[9] Sang Hoon Lee и Byung Man Kwak. «Response surface augmented moment method for efficient reliability analysis». В: *Structural Safety* 28.3 (2006), с. 261—272. ISSN: 0167-4730. DOI: https://doi.org/10.1016/j.strusafe.2005.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0167473005000421.

[10] Ivan Panin. «Risk of estimators for Sobol' sensitivity indices based on metamodels». В: *Electron. J. Statist.* 15.1 (2021), с. 235—281. ISSN: 1935-7524. DOI: 10.1214/20-EJS1793. URL: https://projecteuclid.org/euclid.ejs/1609902190.

[11] Luc Pronzato. «Sensitivity analysis via Karhunen-Loève expansion of a random field model: Estimation of Sobol' indices and experimental design». В: *Reliability Engineering & System Safety* (янв. 2018). DOI: 10.1016/j.ress.2018.01.010.

[12] Luc Pronzato и Andrej Pázman. *Design of experiments in nonlinear models.* Т. 212. Lecture Notes in Statistics. Asymptotic normality, optimality criteria and small-sample properties. Springer, New York, 2013, с. xvi+399. ISBN: 978-1-4614-6362-7; 978-1-4614-6363-4. DOI: 10.1007/978-1-4614-6363-4. URL: https://doi.org/10.1007/978-1-4614-6363-4.

[13] Andrea Saltelli и др. *Global sensitivity analysis. The primer.* John Wiley & Sons, Ltd., Chichester, 2008, с. xii+292. ISBN: 978-0-470-05997-5.

[14] I. M. Sobol'. «Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates». В: т. 55. 1-3. The Second IMACS Seminar on Monte Carlo Methods (Varna, 1999). 2001, с. 271—280. DOI: 10.1016/S0378-4754(00)00270-6. URL: https://doi.org/10.1016/S0378-4754(00)00270-6.

[15] I. M. Sobol'. «Sensitivity estimates for nonlinear mathematical models». Англ. В: *Math. Modeling Comput. Experiment* 1.4 (1993), 407—414 (1995). ISSN: 1061-7590.

[16] Bruno Sudret. «Polynomial chaos expansions and stochastic finite element methods». В: *Risk and Reliability in Geotechnical Engineering.* Под ред. Jianye Ching Kok-Kwang Phoon. CRC Press, 2015, с. 265—300. URL: https://hal.archives-ouvertes.fr/hal-01449883.

[17] Bruno Sudret. «Uncertainty propagation and sensitivity analysis in mechanical models – Contributions to structural reliability and stochastic spectral methods». В: (янв. 2007).

[18] A.W. van der Vaart. *Asymptotic Statistics.* Asymptotic Statistics. Cambridge University Press, 2000. ISBN: 9780521784504. URL: https://books.google.ru/books?id=UEuQEM5RjWgC.

[19] Иван Панин и Павел Приходько. «Подходы к нахождению дисперсии оценок значимости признаков в задаче глобального анализа чувствительности». В: *Сборник трудов конференции "Информационные Технологии и Системы (ИТиС)"* (Петрозаводск, Россия). ИППИ РАН. 2012, с. 173—178. ISBN: 978-5-901158-19-7. URL: http://www.itas2012.iitp.ru/pdf/1569602539.pdf.