

На правах рукописи

Черноусова Елена Олеговна

**БЕСПЕРЕБОРНЫЕ МЕТОДЫ КРОСС-ВАЛИДАЦИИ ДЛЯ
ОЦЕНИВАНИЯ ОБОБЩАЮЩЕЙ СПОСОБНОСТИ
РЕГРЕССИОННЫХ МОДЕЛЕЙ**

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2013

Работа выполнена на кафедре «Интеллектуальные системы» Федерального государственного автономного образовательного учреждения высшего профессионального образования «Московский физико-технический институт (государственный университет)».

Научный руководитель: доктор технических наук, профессор
Моттль Вадим Вячеславович

Официальные оппоненты: доктор физико-математических наук, профессор
Вьюгин Владимир Вячеславович, заведующий лабораторией №1 им. М.С. Пинскера «Теория передачи информации и управления», Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук.

кандидат физико-математических наук, доцент
Сулимова Валентина Вячеславовна, кафедра информационной безопасности, Государственное образовательное учреждение высшего профессионального образования "Тульский государственный университет".

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук.

Защита состоится « 26 » декабря 2013 г. в 14:00 часов на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном учреждении науки «Вычислительный центр им. А.А. Дородницына Российской академии наук», расположенном по адресу 119991, г. Москва, ул. Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан « 22 » ноября 2013 г.

Ученый секретарь
диссертационного совета

доктор физико-математических наук



Рязанов В.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Задача восстановления объективно существующей зависимости между наблюдаемыми свойствами объектов реального мира и их некоторой скрытой характеристикой, доступной для наблюдения лишь в пределах конечной обучающей совокупности, является центральной задачи интеллектуального анализа данных. В частности, если скрытая характеристика принимает значения из конечного неупорядоченного множества, то такую задачу принято называть задачей обучения распознаванию образов, а в случае числовой скрытой характеристики говорят о задаче восстановления регрессионной зависимости. В диссертационной работе рассматривается последний случай.

Вероятностная интерпретация задачи обучения основана на предположении, что с каждым объектом из некоторого множества объектов реального мира, привлекающего внимание наблюдателя, объективно связаны значения двух его характеристик, одна из которых доступна для непосредственного наблюдения, а другая скрыта. В теории обучения обычно предполагается, что природа, случайным образом выбирая один объект, генерирует, тем самым, случайную пару значений его наблюдаемой и скрытой характеристик, причем соответствующее совместное распределение вероятностей объективно существует, но наблюдателю неизвестно. Наблюдатель всякий раз видит значение лишь наблюдаемой характеристики, в то время как природа требует, чтобы он «угадывал» значение скрытой характеристики объекта, штрафую неправильное оценивание в соответствии с известной функцией потерь. Наблюдатель вынужден выработать решающее правило, связывающее с каждым наблюденным значением доступной характеристики объекта предполагаемое значение его скрытой характеристики. Как правило, наблюдатель формирует свое решающее правило на основе предположения о некотором параметрическом классе зависимостей, так что выбор наблюдателем конкретного решающего правила полностью задается выбором параметра. Естественной объективной оценкой «качества» решающего правила является математическое ожидание потерь, которое в теории обучения принято называть средним риском ошибки. Очевидно, что наблюдатель, выбирая вариант решающего правила, а именно значение параметра, не может вычислить средний риск ошибки, поскольку совместное распределение наблюдаемой и скрытой характеристик случайно появляющегося объекта ему неизвестно.

Единственную объективную информацию о свойствах природы, доступную наблюдателю, несет обучающая совокупность, под которой понимается конечное множество пар значений как наблюдаемой, так и скрытой характеристики объектов, случайно выбранных природой в соответствии с объективно существующим распределением вероятностей. Выбирая решающее правило, наблюдатель может лишь вычислить для всякого его варианта среднее арифметическое значение функции потерь, называемое эмпирическим риском ошибки. При выборе решающего правила общепринятым соображением, основанным на данных, является минимизация эмпирического риска в некотором классе решающих правил (*variance minimization* в англоязычной литературе).

Однако параметрический класс решающих правил, изначально принятый наблюдателем, может оказаться слишком широким для ограниченного объема обучающей совокупности, и средний риск ошибки результата обучения по критерию минимума эмпирического риска может оказаться неприемлемо большим. Такое явление принято называть переобучением. Другим общепринятым соображением, направленным

на уменьшение опасности переобучения, является использование априорной (регуляризирующей) информации об «ожидаемом» решающем правиле восстановления зависимости. Другими словами, наблюдатель пытается сузить параметрический класс зависимостей, накладывая на параметр априорные регуляризирующие требования, в свою очередь контролируемые структурным параметром. Как правило, априорная информация выражена в виде некоторого функционала на классе решающих правил, подлежащего минимизации, причем обычно такой функционал содержит параметр, контролирующий отклонение решающего правила от некоторого подмножества наиболее «простых» правил, и называемый структурным параметром «сложности» класса решающих правил. Это дополнительное соображение при построении метода обучения касательно выбора решающего правила называется в англоязычной литературе *bias*, поскольку управляет «смещением» выбираемого решающего правила от выбранного на основе минимизации эмпирического риска.

В современной теории обучения эти два соображения объединяются в единый критерий обучения, получая тем самым регуляризованный критерий минимизации эмпирического риска. Естественно, что результат обучения, т.е. решающее правило, получаемое в качестве решения задачи минимизации, зависит от структурного параметра, отвечающего за сложность зависимости между ненаблюдаемой и наблюдаемой компонентами объекта.

Очевидным показателем «качества» выбора структурных параметров и, следовательно, получаемого решающего правила, является средний риск ошибки оценивания скрытой характеристики нового случайного объекта, не входящего в обучающую совокупность. Однако, вычисление среднего риска принципиально невозможно, поскольку наблюдателю неизвестно совместное распределение вероятностей на множестве пар значений наблюдаемой и скрытой характеристик объектов в генеральной совокупности. В качестве общепринятого компромисса на практике обычно заменяют критерий минимума среднего риска ошибки при выборе структурного параметра на его суррогат, вычисленный путем кросс-валидации единственной обучающей совокупности, доступной наблюдателю. Метод кросс-валидации (*Cross-Validation*)¹ заключается в том, что обучающая совокупность многократно разбивается на две части, по одной из которых определяется решающее правило для каждого пробного значения структурного параметра, а по другой оценивается среднее значение ошибки.

Проблемная ситуация заключается в том, что методы кросс-валидации требуют многократного повторения обучения при разных разбиениях обучающей совокупности, что определяет их чрезвычайно высокую вычислительную сложность. В частности, наиболее популярными видами кросс-валидации являются блочная кросс-валидация, заключающаяся в разбиении обучающей совокупности на достаточно большое число частей и поочередном использовании каждой части в качестве контрольной при обучении по остальным частям (*K-fold Cross-Validation*), и скользящий контроль², в котором поочередно выделяется один объект в качестве контрольного, а обучение проводится по оставшимся объектам (*Leave-one-out Cross-Validation*). При этом число повторений обучения равно кратности разбиения обучающей сово-

¹ P.A. Devijver, J. Kittler. *Pattern Recognition: A Statistical Approach*, Prentice-Hall, London, GB, 1982.

² Бонгард М.М., Вайнцвайг М.Н. Об оценках ожидаемого качества признаков. *Проблемы кибернетики*, 1968, вып. 20, с. 151-157.

купности на блоки, а в методе скользящего контроля совпадает с числом объектов в обучающей совокупности.

Для разрешения этой проблемной ситуации в диссертации предлагается общий метод, основанный на некотором предположении наблюдателя о возможном параметрическом классе совместных распределений наблюдаемой и скрытой характеристик случайно появляющегося объекта, и назван в диссертации методом неявной кросс-валидации. Метод основан на мысленном эксперименте, заключающемся в получении двух независимых выборок, по первой из которых находится решающее правило как по обучающей совокупности, а на второй измеряется эмпирический риск ошибки восстановления скрытой характеристики объекта. В качестве критерия выбора значений структурных параметров предлагается использовать математическое ожидание эмпирического риска ошибки.

В диссертации доказано, что в случае квадратичной функции потерь, адекватной широкому классу задач восстановления регрессионных зависимостей, и квадратичного регуляризующего штрафа, налагаемого на вектор искомых коэффициентов регрессии, несмещенная оценка математического ожидания эмпирического риска ошибки выражается через элементы обучающей совокупности в виде простой формулы. Показано, что частным случаем такого критерия выбора структурных параметров при некоторых специальных предположениях о модели данных является известный информационный критерий Акаике³.

Чрезвычайная актуальность автоматического сокращения размерности представления объектов непосредственно в ходе обучения приводит к необходимости применения более сложной регуляризующей функции от вектора искомых коэффициентов регрессии, нежели квадратичная, а именно, квадратично-модульной функции (в англоязычной литературе соответствующий критерий обучения получил название *Elastic Net*⁴). Решающее правило наблюдателя, получаемое в результате обучения, характеризуется двумя структурами параметрами, отвечающими за квадратичную и модульную регуляризацию, но для их выбора исходный метод неявной (беспереборной) кросс-валидации в чистом виде оказывается неприменимым в силу неквадратичности критерия обучения. Для того, чтобы избежать применения обычных переборных методов кросс-валидации, в диссертации используется тот факт, что с каждой парой значений числовых структурных параметров однозначно связано разбиение множества числовых признаков объектов на три непересекающихся подмножества, полученных с учетом знака и обнуления коэффициентов регрессии в точке минимума критерия. Именно такое разбиение, полученное согласно решению задачи обучения, предлагается использовать в качестве вторичного нечислового структурного параметра модели, подлежащего кросс-валидации. При фиксации такого структурного параметра критерий обучения, в исходном варианте не являющийся квадратичным, становится квадратичным по активным (ненулевым) коэффициентам регрессии, и к нему полностью применим разработанный ранее метод неявной (беспереборной) кросс-валидации.

³ Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, Vol. 19, pp. 716-723.

⁴ H. Zou, T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67:301–320, 2005.

Объект исследования: задачи восстановления числовых зависимостей между скрытой и наблюдаемой характеристиками объектов реального мира по эмпирическим данным.

Предмет исследования: выбор структурных параметров решающих правил, восстанавливающих числовую зависимость между скрытой и наблюдаемой характеристиками объектов, с помощью беспереборных методов кросс-валидации для оценивания адекватности решения на генеральной совокупности по единственно доступной исследователю выборке.

Цели и задачи диссертации:

1. Показать, что идея информационного критерия Акаике для выбора структурного параметра основана на принципе неявной кросс-валидации.
2. Разработать беспереборный критерий кросс-валидации для квадратичной задачи оценивания линейной регрессии, в котором классический критерий Акаике являлся бы частным случаем.
3. Разработать беспереборный критерий кросс-валидации для неквадратичной задачи оценивания линейной регрессии с квадратично-модульной регуляризацией.

Общая методика исследования: Исследование базируется на использовании классических понятий теории восстановления регрессионных зависимостей, теории вероятности, математической статистики, теории оптимизации.

Научная новизна. В работе предложены два варианта нового беспереборного метода кросс-валидации для оценивания обобщающей способности регрессионных моделей, отличающиеся областью применимости. Оба варианта являются альтернативами классическим способам оценивания обобщающей способности, основанным на принципе кросс-валидации.

Положения, выносимые на защиту.

1. Принцип неявной кросс-валидации для оценивания обобщающей способности линейно-квадратичных моделей числовых зависимостей.
2. Исследование природы классического информационного критерия Акаике как простейшего частного случая критерия неявной кросс-валидации.
3. Критерий неявной кросс-валидации для выбора степени волатильности модели нестационарной регрессии.
4. Критерий неявной кросс-валидации для выбора степени подавления нерелевантных регрессоров в линейно-квадратичной модели числовой регрессии.
5. Критерий неявной кросс-валидации для выбора уровня селективности формирования подмножества релевантных регрессоров в квадратично-модульной модели Elastic Net.

Достоверность полученных результатов подтверждается доказательствами сформулированных теорем и проверкой полученных результатов на модельных экспериментах и на реальных данных.

Практическая значимость результатов диссертации заключается в том, что предложенные беспереборные методы кросс-валидации для оценивания обобщающей способности регрессионных моделей являются (в силу беспереборности) вычислительно эффективными в сравнении с классическими (переборными) методами кросс-валидации, основанными на многократном повторении процедуры обучения и контроля качества решающего правила, полученного на этапе обучения, на различных разбиениях исходной выборки.

Связь с плановыми научными исследованиями. Работа выполнена при поддержке грантов Российского фонда фундаментальных исследований №№ 11-07-00409-а, 11-07-00634-а, 12-07-13142-офи-м и при поддержке Лаборатории структурных методов анализа данных в предсказательном моделировании.

Апробация работы. Основные положения и результаты диссертации докладывались на конференциях «Интеллектуализация обработки информации ИОИ - 2010» (Республика Кипр, г. Пафос, 2010 г.), «Интеллектуализация обработки информации ИОИ - 2012» (Черногория, г. Будва, 2012 г.), «Математические методы распознавания образов ММРО - 2009» (г. Суздаль, 2009 г.), «Математические методы распознавания образов ММРО - 2013» (г. Казань, 2013 г.).

Публикации. По тематике работы опубликовано 8 статей, в том числе 2 статьи в журналах, рекомендованных ВАК.

Структура и объем работы. Диссертация состоит из введения, 4 глав основного содержания, заключения и библиографии. Работа содержит 87 страниц основного текста.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследований по разработке методов эффективного оценивания обобщающей способности регрессионных моделей данных, указывается степень разработанности исследуемой проблемы, приведены цели и задачи проводимых исследований, указывается на научную новизну полученных результатов и их теоретическую и практическую значимость, приведены положения, выносимые на защиту, а также сведения о структуре диссертации, степени достоверности результатов исследований и их апробации.

В первой главе рассматривается общая постановка типичной задачи оценивания числовой зависимости между двумя характеристиками объектов реальной природы. А именно, пусть имеется некоторое множество объектов реального мира Ω , привлекающее внимание наблюдателя. Пусть с каждым объектом $\omega \in \Omega$ связаны значения двух его характеристик $x(\omega): \Omega \rightarrow \mathbb{X}$ и $y(\omega): \Omega \rightarrow \mathbb{Y}$, первая из которых доступна для непосредственного наблюдения, а наблюдение второй невозможно либо, по крайней мере, затруднено. Цель исследователя – анализируя конечную обучающую совокупность объектов, в пределах которой обе характеристики известны

$$(\mathbf{x}, \mathbf{y}) = \left\{ (x_j, y_j) = (x(\omega_j), y(\omega_j)), j = 1, \dots, N \right\}, \quad (1)$$

сформировать для нового объекта $\omega \in \Omega$, не представленного в обучении, оценку его скрытой характеристики по наблюдаемой $\hat{y}(x): \mathbb{X} \rightarrow \mathbb{Y}$. Такая задача называется задачей обучения по прецедентам.

Предположим, что природа случайным образом многократно и независимо выбирает один объект из множества Ω , генерируя, тем самым, пару значений

$$\omega \in \Omega: (x, y) = (x(\omega), y(\omega)), x \in \mathbb{X}, y \in \mathbb{Y}, \quad (2)$$

согласно неизвестной наблюдателю «истинной» совместной плотности распределения $f^*(x, y) \geq 0: \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{R}$, $\int \int_{\mathbb{X} \times \mathbb{Y}} f^*(x, y) dy dx = 1$.

Штаф за неверное оценивание скрытой характеристики объекта измеряется действительным числом $Loss \in \mathbb{R}$, являющимся известной функцией от фактического и оцененного значений ненаблюдаемой характеристики объекта $Loss(y, \hat{y}): \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$. Такую функцию принято называть функцией потерь. Предполагается,

что наблюдатель формирует свое решающее правило в пределах выбранного им некоторого параметрического класса зависимостей

$$\hat{y}(x, \mathbf{a}): \mathbb{X} \rightarrow \mathbb{Y}, \mathbf{a} \in \mathbb{R}^n, \quad (3)$$

так что выбор оценки ненаблюдаемой компоненты полностью определяется выбором параметра $\mathbf{a}$, а функция потерь $Loss(y, \hat{y}(x, \mathbf{a}))$ есть функция от трех переменных $q(y, x, \mathbf{a}) \equiv Loss(y, \hat{y}(x, \mathbf{a}))$.

Вообще говоря, естественно желание наблюдателя выбирать вектор параметров решающего правила из условия минимума средних потерь при оценивании скрытой характеристики (среднего риска ошибки решающего правила):

$$r_{av}(\mathbf{a}) = r_{av}[\hat{y}(\cdot, \mathbf{a})] = \int \int_{\mathbb{X} \times \mathbb{Y}} q(x, y, \mathbf{a}) f^*(x, y) dy dx, \quad (4)$$

$$r_{av}(\mathbf{a}) \rightarrow \min(\mathbf{a} \in \mathbb{R}^n). \quad (5)$$

Понятно, однако, что такой выбор принципиально невозможен, поскольку неизвестно «истинное» распределение $f^*(x, y)$.

Взамен минимизации среднего риска (5) наблюдатель, в качестве естественного компромисса, может организовать обучение, то есть выбор параметра решающего правила оценивания скрытой характеристики объекта $\mathbf{a} \in \mathbb{R}^n$, на основе минимизации эмпирического риска ошибки, построенного по обучающей совокупности:

$$r_{emp}(\mathbf{a}) = \frac{1}{N} \sum_{j=1}^N q(x_j, y_j, \mathbf{a}) = \frac{1}{N} Q(\mathbf{x}, \mathbf{y}, \mathbf{a}), \quad (6)$$

$$r_{emp}(\mathbf{a}) \rightarrow \min(\mathbf{a} \in \mathbb{R}^n), \text{ т.е. } \sum_{j=1}^N q(x_j, y_j, \mathbf{a}) \rightarrow \min(\mathbf{a} \in \mathbb{R}^n). \quad (7)$$

Если предполагаемый параметрический класс решающих правил $\hat{y}(x, \mathbf{a}): \mathbb{X} \rightarrow \mathbb{Y}$, $\mathbf{a} \in \mathbb{R}^n$, достаточно широк, то наблюдатель сможет достаточно хорошо аппроксимировать числовую зависимость между характеристиками объектов и обучающей совокупности. Однако, в силу конечности множества прецедентов выбранная модель зависимостей может быть неадекватной на генеральной совокупности. Во избежание этой проблемы, наблюдатель пытается сузить параметрический класс зависимостей, накладывая на параметр $\mathbf{a}$ априорные регуляризующие требования, формулируемые в терминах минимизации некоторого функционала

$$V(\mathbf{a}, \lambda) \rightarrow \min(\mathbf{a}), \quad (8)$$

где λ – некоторый структурный скалярный или векторный параметр, контролирующий нежелательность отклонения параметра $\mathbf{a} \in \mathbb{R}^n$, т.е. решающего правила $\hat{y}(x, \mathbf{a})$, от некоторого подмножества наиболее «простых» правил $\mathbf{A} \in \mathbb{R}^n$.

Итоговое компромиссное решение для выбора вектора параметров $\mathbf{a}$ решающего правила осуществляется с помощью критерия минимизации регуляризованного эмпирического риска

$$\hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y}) = \arg \min_{\mathbf{a}} \{Q(\mathbf{x}, \mathbf{y}, \mathbf{a}) + V(\mathbf{a}, \lambda)\}. \quad (9)$$

Ясно, что результат обучения (9) зависит от параметра λ . Выбор адекватного значения структурного параметра λ , отвечающего за степень регуляризации параметрического класса зависимостей $\hat{y}(x, \mathbf{a}): \mathbb{X} \rightarrow \mathbb{Y}$ с помощью минимизации функционала $V(\mathbf{a}, \lambda) \rightarrow \min(\mathbf{a})$, является одной из ключевых задач машинного обучения.

Действительно, если рассмотреть «наименее регуляризующее» априорное требование $V(\mathbf{a}, \lambda') \approx \text{const}$, $\mathbf{a} \in \mathbb{R}^n$, то процедуры машинного обучения, как правило, стремятся подобрать параметрическую зависимость $\hat{y}(x, \mathbf{a})$, которая «слишком хорошо» связывает наблюдаемую и скрытую характеристики для объектов из обучающей совокупности, но оказывается неадекватной для произвольных объектов генеральной совокупности. В литературе такое явление получило термин «переобучения» (overfitting). Напротив, выбор «сильного регуляризующего» априорного требования $V(\mathbf{a}, \lambda'') \approx 0$, $\mathbf{a} \notin \mathbb{A} \in \mathbb{R}^n$, приводит к проблеме «недообучения» (underfitting), когда регрессионная модель слишком проста и не способна хорошо описать эмпирические данные.

В теории машинного обучения есть ряд классических методов выбора структурного параметра модели на основе единственной обучающей совокупности.

В теории В.Н. Вапника и А.Я. Червоненкиса⁵ сформулирован принцип минимизации структурного риска в задачах обучения распознаванию образов $y \in \{-1, 1\}$ по обучающей совокупности размера N

$$\lambda^* = \arg \min \left[(1/N)Q(\mathbf{x}, \mathbf{y}, \hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y})) + \Delta(\lambda, N) \right],$$

понимаемого как сумма минимального эмпирического риска, достижимого в выбранном классе решающих правил $Q(\mathbf{x}, \mathbf{y}, \hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y}))$, определяемом в наших терминах выбором структурного параметра λ , и верхней границы превышения среднего риска для этого класса на генеральной совокупности над эмпирическим риском $\Delta(\lambda, N)$ ⁶. Зависимость этой верхней границы от класса решающих правил получена лишь для простейшей функции потерь, адекватной только задачам распознавания образов, под структурным параметром понимается эффективная размерности множества решающих правил, получившая в литературе название VC-размерности (Vapnik-Chervonenkis Dimension). К тому же, верхняя граница $\Delta(\lambda, N)$ определена из общих неравенств теории вероятностей на основе созданной авторами теории равномерной близости эмпирических средних к математическим ожиданиям⁵, и является чрезвычайно завышенной. В исходном виде этот принцип неприменим к задачам восстановления числовых зависимостей.

Однако сама идея минимизации суммы эмпирического риска и некоторого штрафа за сложность класса моделей была сформулирована японским математиком Хиротугу Акаике³ (с. 10). Применительно к задаче восстановления зависимости $\hat{y}(x)$ предположим, что наблюдатель рассматривает вероятностную связь между скрытой и наблюдаемой переменными в виде параметрического семейства плотностей распределений $\phi(y | x, \mathbf{a})$. Пусть размерность вектора $\mathbf{a} \in \mathbb{R}^n$ слишком велика для объема N доступной обучающей совокупности (1), так что оценка максимального правдоподобия $\hat{\mathbf{a}}(\mathbf{x}, \mathbf{y}) = \arg \max \ln \Phi(\mathbf{y} | \mathbf{x}, \mathbf{a}) = \arg \max \sum_{j=1}^N \ln \phi(y_j | x_j, \mathbf{a})$ недостаточно надежна.

Простота ситуации, рассматриваемой Акаике, заключается в том, что компоненты вектора параметров $\mathbf{a} = (a_1, \dots, a_n)$ предполагаются априори упорядоченными по

⁵ Вапник В. Н., Червоненкис А. Я. *Теория распознавания образов*. М.: Наука, 1974.

⁶ Вапник В. Н. *Принцип структурной минимизации в задачах восстановления зависимостей по эмпирическим данным*. Диссертация д.ф.-м.н. Институт проблем управления АН СССР, 1984.

«важности», и структурный параметр λ понимается как целое число $0 \leq \lambda \leq n$, ограничивающее количество компонент, отличных от нуля

$$\mathbf{a} = (\mathbf{a}_\lambda, \mathbf{a}_{n-\lambda}) = (a_1, \dots, a_\lambda, a_{\lambda+1}=0, \dots, a_n=0). \quad (10)$$

При дополнительном предположении, что логарифмическая функция правдоподобия $\ln \Phi(\mathbf{y} | \mathbf{x}, \mathbf{a})$ допускает квадратичную аппроксимацию, Хиротугу Акаике доказал, что решение

$$\lambda^* = \arg \min \left\{ \underbrace{-\ln \Phi(\mathbf{y} | \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y}))}_{\text{максимум правдоподобия}} + \lambda \right\}, \quad (11)$$

эмпирический риск

основанное на штрафе, равном просто числу активных параметров λ , в некотором усредненном смысле максимизирует количество информации Кульбака о неизвестном распределении природы, доступном лишь в виде конечной выборки $(\mathbf{x}, \mathbf{y})$, содержащееся в оценке параметра по самой этой выборке.

Формулу (11) Акаике получил в предположении, что гессиан $\nabla_{\mathbf{a}\mathbf{a}}^2 \ln \Phi(\mathbf{y} | \mathbf{x}, \mathbf{a})$ является матрицей полного ранга в точке максимума правдоподобия, и, как следствие, оценка $\hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y})$ единственна. Чтобы учесть более общий случай, штраф λ должен быть заменен рангом этой матрицы:

$$\lambda^* = \arg \min \left\{ -\ln \Phi(\mathbf{y} | \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y})) + \text{Rank} \left[\nabla_{\mathbf{a}\mathbf{a}}^2 \ln \Phi(\mathbf{y} | \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{x}, \mathbf{y})) \right] \right\}. \quad (12)$$

Очевидно, что предположение об априорной упорядоченности компонент вектора коэффициентов регрессии неадекватно большинству практических задач.

В качестве универсального метода борьбы с «проклятием единственной выборки» в ходе оценки обобщающей способности класса моделей широкую популярность получили метод кросс-валидации¹ и его наиболее часто используемый вариант, известный как скользящий контроль² (с. 4), свободные от каких бы то ни было предположений о неизвестном распределении, породившем анализируемую выборку. Эти методы, в свою очередь, можно рассматривать как частные случаи методов, известных под общим названием *Resampling*⁷. В основе этих методов лежит идея многократного дробления исходной выборки на две части, одна из которых используется для обучения, т.е. для оценивания параметров модели искомой зависимости, а другая – для оценивания качества обучения. Неизбежной платой за универсальность метода кросс-валидации является его высокая вычислительная сложность.

Первая глава завершается постановкой задачи диссертационного исследования, сформулированной как разработка принципа выбора структурного параметра модели числовой зависимости по единственной обучающей совокупности, доступной наблюдателю, который сочетал бы, с одной стороны, универсальность общей идеи кросс-валидации и, с другой стороны, сохранял бы беспереборный характер, являющийся вычислительным достоинством информационного критерия Акаике.

Во второй главе изложена основная математическая идея диссертации, названная принципом неявной кросс-валидации.

Представим неизвестную наблюдателю совместную плотность распределения случайной выборки $(\mathbf{x}, \mathbf{y})$ (обучающей совокупности) (1)

⁷ B. Efron. *The jackknife, the bootstrap, and other resampling plans*. Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38, 1982.

$$F^*(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^N f^*(x_j, y_j)$$

как произведение маргинальной плотности распределения наблюдаемой характеристики и условной плотности распределения ненаблюдаемой характеристики:

$$f^*(x, y) = \varphi^*(y|x)g^*(x), \quad F^*(\mathbf{x}, \mathbf{y}) = \underbrace{\prod_{j=1}^N \varphi^*(y_j|x_j)}_{\Phi^*(\mathbf{y}|\mathbf{x})} \underbrace{\prod_{j=1}^N g^*(x_j)}_{G^*(\mathbf{x})} = \Phi^*(\mathbf{y}|\mathbf{x})G^*(\mathbf{x}). \quad (13)$$

Дальнейшие построения основаны на предположении о структуре неизвестного условного распределения $\Phi^*(\mathbf{y}|\mathbf{x})$, названном в диссертации первой эвристикой наблюдателя.

Первая эвристика наблюдателя о незлонамеренности природы. Наблюдатель предполагает, что это распределение является неизвестной смесью $\Psi^*(\mathbf{a})$ известных распределений $\Phi(\mathbf{y}|\mathbf{x}, \mathbf{a})$

$$\Phi^*(\mathbf{y}|\mathbf{x}) = \int_{\mathbb{R}^n} \Phi(\mathbf{y}|\mathbf{x}, \mathbf{a}) \Psi^*(\mathbf{a}) d\mathbf{a}, \quad (14)$$

причем природа чаще генерирует пары с низким значением штрафа

$$\Phi(\mathbf{y}|\mathbf{x}, \mathbf{a}) \propto \exp\{-dQ(\mathbf{y}, \mathbf{x}, \mathbf{a})\}. \quad (15)$$

Иными словами, первая эвристика наблюдателя заключается в предположении, что он правильно «угадал» функцию потерь (6) и, следовательно, класс решающих правил (9).

Дополнительный коэффициент d в (15) выражает предполагаемую степень обоснованности такого предположения – чем больше d , тем сильнее уверенность наблюдателя. Заметим, что при этом наблюдатель по-прежнему не делает никаких предположений о характере распределения скрытого параметра решающего правила $\Psi^*(\mathbf{a})$ в (14).

Предлагаемый в диссертации принцип неявной кросс-валидации основан на следующем мысленном эксперименте наблюдателя. Пусть природа разыграла конкретное значение параметра согласно $\Psi^*(\mathbf{a})$, а также выборку наблюдаемых характеристик объектов согласно $G^*(\mathbf{x})$ (13). Затем, дважды применив условное распределение $\Phi(\mathbf{y}|\mathbf{x}, \mathbf{a})$ (15), получила две реализации $\mathbf{y} = (y_1, \dots, y_N) \in \mathbb{R}^N$, $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_N) \in \mathbb{R}^N$, образующие две воображаемые совокупности которые наблюдатель мысленно рассматривает как контрольную $(\mathbf{x}, \mathbf{y})$ и обучающую $(\mathbf{x}, \tilde{\mathbf{y}})$. Если бы наблюдатель знал реализации $\mathbf{y}$ и $\tilde{\mathbf{y}}$, то мог бы для всякого значения λ вычислить оценку по обучающей совокупности $\hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x})$ и подставить в функцию потерь для контрольной совокупности, вычислив потери на мысленном контроле $Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x}))$. Идея скрытой кросс-валидации заключается в минимизации математического ожидания потерь:

$$\int \int \int Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x})) \left\{ \int \Phi(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{a}) \Phi(\mathbf{y}|\mathbf{x}, \mathbf{a}) \Psi^*(\mathbf{a}) d\mathbf{a} \right\} G^*(\mathbf{x}) d\tilde{\mathbf{y}} d\mathbf{y} d\mathbf{x} \rightarrow \min(\lambda). \quad (16)$$

Согласно первой эвристике наблюдателя $(1/\sigma^2)Q(\mathbf{x}, \mathbf{y}, \mathbf{a}) = -\ln \Phi(\mathbf{y}|\mathbf{x}, \mathbf{a}) + \text{const}$ (15), поэтому идея скрытой кросс-валидации $\int [\ln \Phi(\mathbf{y}|\mathbf{x}, \hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x}))] \Phi(\mathbf{y}|\mathbf{x}, \mathbf{a}) d\mathbf{y} \rightarrow \max$ (16) есть максимизация информации Кульбака о распределении $\Phi(\mathbf{y}|\mathbf{x}, \mathbf{a})$, содержащейся в его оценке по другой выборке $\Phi(\mathbf{y}|\mathbf{x}, \hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x}))$. В силу этого обстоятельства крите-

рии неявной кросс-валидации уместно называть информационными и рассматривать их как обобщение классической идеи Хиротугу Акаике.

В реальности у наблюдателя имеется единственная обучающая выборка $(\mathbf{x}, \mathbf{y})$, и он может подставить в функцию потерь лишь оценку, вычисленную по той же выборке $Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x}))$. Насколько испортится критерий, подлежащий максимизации, по сравнению с идеей (16)? Каким должен быть штраф за использование оценки параметра, вычисленного по той же выборке?

Теорема 1. Критерий максимизации по значению структурного параметра λ (16) допускает эквивалентную запись:

$$\lambda^* = \arg \min_{\lambda} \left\{ \iint \left[\int Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x})) \Phi(\mathbf{y} | \mathbf{x}, \mathbf{a}) d\mathbf{y} + \int \left(Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x})) - Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x})) \right) \underbrace{\Phi(\tilde{\mathbf{y}} | \mathbf{x}, \mathbf{a}) \Phi(\mathbf{y} | \mathbf{x}, \mathbf{a})}_{\propto \exp\{-(1/\sigma^2)[Q(\mathbf{y}, \mathbf{x}, \mathbf{a}) + Q(\mathbf{y}, \mathbf{x}, \mathbf{a})]\}} d\tilde{\mathbf{y}} d\mathbf{y} \right] \Psi^*(\mathbf{a}) G^*(\mathbf{x}) d\mathbf{a} d\mathbf{x} \right\}. \quad (17)$$

функционал от функции потерь $Q(\cdot, \mathbf{x}, \mathbf{a})$ и критерия обучения $\hat{\mathbf{a}}_\lambda(\cdot, \mathbf{x})$

Основная идея принципа неявной кросс-валидации базируется на том факте, что, как показано далее во второй главе, для многих типичных функций потерь $Q(\mathbf{y}, \mathbf{x}, \mathbf{a})$ и критериев обучения $\hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x})$, адекватных широкому классу практических задач, функционал во втором слагаемом в (17), играющий роль штрафа за использование оценки параметра, вычисленного по той же выборке, не зависит от $\mathbf{a}$:

$$\iint \left(Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x})) - Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x})) \right) \Phi(\tilde{\mathbf{y}} | \mathbf{x}, \mathbf{a}) \Phi(\mathbf{y} | \mathbf{x}, \mathbf{a}) d\tilde{\mathbf{y}} d\mathbf{y} = \sigma^2 \Delta(\lambda, \mathbf{x}). \quad (18)$$

Для этого класса задач обучения критерий неявной кросс-валидации (17), установленный Теоремой 1 принимает следующий простой вид:

$$\lambda^* = \arg \min_{\lambda} \left\{ \iint \left[dQ(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x})) + \Delta(\lambda, \mathbf{x}) \right] F^*(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} \right\}. \quad (19)$$

Однако и в таком виде критерий по-прежнему неприменим, так как распределение генеральной совокупности $F^*(\mathbf{x}, \mathbf{y})$ в (19) неизвестно.

Вторая эвристика наблюдателя заключается в замене математического ожидания его несмещенной оценкой по единственной доступной наблюдателю выборке $\mathbf{y} \in \mathbb{R}^N$:

$$\hat{\lambda}(\mathbf{y}, \mathbf{x}) = \arg \min_{\lambda} \left\{ d \underbrace{Q(\mathbf{y}, \mathbf{x}, \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{x}))}_{\text{эмпирический риск}} + \Delta(\lambda, \mathbf{x}) \right\} \quad (20)$$

структурный риск

Далее во второй главе рассмотрен важнейший частный случай **линейно-квадратичной модели данных**, адекватной многим практическим задачам восстановления числовых зависимостей, и доказано, что этот частный случай удовлетворяет основному предположению о независимости штрафа за использование оценки параметра решающего правила, вычисленного по той же выборке, от самого значения параметра (18). Класс линейно-квадратичных моделей определяется следующими предположениями.

Во-первых, предполагается, что наблюдатель выбрал линейный класс решающих правил оценивания скрытой характеристики объектов (3)

$$\hat{\mathbf{y}}(\mathbf{x}) = \mathbf{a}^T \mathbf{x} : \mathbb{R}^n \rightarrow \mathbb{R}, \quad \mathbf{a} = (a_1 \cdots a_n)^T, \quad \mathbf{x} = (x_1 \cdots x_n)^T \in \mathbb{R}^n, \quad y \in \mathbb{R}. \quad (21)$$

Это предположение позволяет предварительно центрировать и нормировать обучающую совокупность размера N (1)

$$(\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_j, y_j), j=1, \dots, N\}, \quad \mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N) - \text{матрица } (n \times N), \quad \mathbf{y} \in \mathbb{R}^N, \\ \frac{1}{N} \sum_{j=1}^N x_{ij} = 0, \quad \frac{1}{N} \sum_{j=1}^N x_{ij}^2 = 1 \quad \text{для всех } i = 1, \dots, n. \quad (22)$$

Во-вторых, предполагается, что наблюдатель выбрал квадратичную функцию потерь (6)

$$Q(\mathbf{y}, \mathbf{X}, \mathbf{a}) = \sum_{j=1}^N (y_j - \mathbf{a}^T \mathbf{x}_j)^2 = (\mathbf{y} - \mathbf{X}^T \mathbf{a})^T (\mathbf{y} - \mathbf{X}^T \mathbf{a}). \quad (23)$$

Наконец, в-третьих, предполагается квадратичная регуляризация (8)

$$V(\mathbf{a}, \lambda) = \mathbf{a}^T \mathbf{B}_\lambda \mathbf{a}, \quad (24)$$

полностью определяемая неотрицательно определенной квадратной матрицей $\mathbf{B}_\lambda$ ($n \times n$), произвольным образом зависящей от выбора структурного параметра λ , скалярного или векторного.

При таких предположениях задача обучения по предъявленной обучающей совокупности (9) становится квадратичной

$$\hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{X}) = \arg \min_{\mathbf{a} \in \mathbb{R}^n} \left\{ \mathbf{a}^T \mathbf{B}_\lambda \mathbf{a} + (\mathbf{y} - \mathbf{X}^T \mathbf{a})^T (\mathbf{y} - \mathbf{X}^T \mathbf{a}) \right\} \quad (25)$$

и численно сводится, вообще говоря, к решению системы n линейных уравнений.

Теорема 2. В методе неявной кросс-валидации для линейно-квадратичной модели (21)-(24) штраф за подстановку в функцию потерь оценки параметра, вычисленной по той же выборке (18), не зависит от исходного значения параметра $\mathbf{a} \in \mathbb{R}^n$ и определяется выражением

$$\Delta(\lambda, \mathbf{X}) = \text{Tr} \left[\mathbf{X} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \mathbf{B}_\lambda)^{-1} \right], \quad \mathbf{X} \mathbf{X}^T (n \times n). \quad (26)$$

В модели с квадратичной функцией потерь первая эвристика наблюдателя включает в себя предположение об условно-нормальном распределении наблюдаемых характеристик объектов (15), поэтому уместно использовать обозначение $d = 1/2\sigma^2$, прямо указывающее на дисперсию этого распределения

$$\Phi(\mathbf{y} | \mathbf{x}, \mathbf{a}, \sigma) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} Q(\mathbf{y}, \mathbf{x}, \mathbf{a}) \right\}. \quad (27)$$

Таким образом, критерий неявной кросс-валидации (20) для общей линейно-квадратичной модели принимает вид:

$$\hat{\lambda}(\mathbf{y}, \mathbf{X}) = \arg \min_{\lambda} \left\{ \underbrace{\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}^T \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{X}))^T (\mathbf{y} - \mathbf{X}^T \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{X}))}_{\text{эмпирический риск}} + \underbrace{\frac{\Delta(\lambda, \mathbf{X})}{\sigma^2}}_{\text{штраф за сложность}} \right\}, \quad (28)$$

$$\Delta(\lambda, \mathbf{X}) = \text{Tr} \left[\mathbf{X} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \mathbf{B}_\lambda)^{-1} \right]. \quad \text{структурный риск}$$

В частности, предположение об априорной упорядоченности коэффициентов регрессии (10) при отсутствии априорных предположений об их значениях, лежащее в основе классического информационного критерия Акаике, выражается специальным видом матрицы квадратичной регуляризации (24):

$$\mathbf{B}_\lambda = \begin{matrix} & 1 & \dots & \lambda \\ \begin{matrix} 1 \\ \vdots \\ \lambda \end{matrix} & \begin{pmatrix} 1/\rho & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1/\rho & 0 & \dots & 0 \\ 0 & \dots & 0 & \rho & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & \rho \end{pmatrix} \end{matrix}, \rho \rightarrow \infty. \quad (29)$$

Введем также обозначение $\mathbf{x}_{\lambda j} = (x_{1j} \dots x_{\lambda j})^T \in \mathbb{R}^\lambda$ для первых λ компонент векторов наблюдаемых характеристик объектов обучающей совокупности в дополнение к (22) и $\mathbf{X}_\lambda = (\mathbf{x}_{\lambda 1} \dots \mathbf{x}_{\lambda N})$ для матрицы $(\lambda \times N)$, составленной из них как из столбцов, соответственно, $\mathbf{X}_\lambda \mathbf{X}_\lambda^T$ – матрица $(\lambda \times \lambda)$.

Теорема 3. Величина штрафа (26) неявной кросс-валидации с квадратичной регуляризацией (29) определяется выражением $\Delta(\lambda, \mathbf{X}) = \text{Rank}(\mathbf{X}_\lambda \mathbf{X}_\lambda^T)$. Если $\mathbf{X}_\lambda \mathbf{X}_\lambda^T$ есть матрица полного ранга, то $\Delta(\lambda, \mathbf{X}) = \lambda$.

Эти утверждения эквивалентно классическому критерию Акаике (11)-(12).

Во второй главе рассмотрен также способ оценивания дисперсии наблюдений σ^2 , входящей как свободный параметр в (27) и (28). Для его оценивания вместе с параметром решающего правила $\mathbf{a}$ предлагается дополнить мысленный эксперимент наблюдателя предположением, что природа применила условное распределение $\Phi(\mathbf{y} | \mathbf{x}, \mathbf{a}, \sigma)$ (27) не дважды, а трижды, и получила три независимые реализации: $\mathbf{y} = (y_1, \dots, y_N) \in \mathbb{R}^N$, $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_N) \in \mathbb{R}^N$, $\tilde{\tilde{\mathbf{y}}} = (\tilde{\tilde{y}}_1, \dots, \tilde{\tilde{y}}_N) \in \mathbb{R}^N$, образовав, таким образом три независимые обучающие совокупности $(\mathbf{X}, \mathbf{y})$, $(\mathbf{X}, \tilde{\mathbf{y}})$ и $(\mathbf{X}, \tilde{\tilde{\mathbf{y}}})$, рассматриваемые соответственно как контрольная $(\mathbf{X}, \mathbf{y})$, обучающая для $\hat{\mathbf{a}}_\lambda(\tilde{\mathbf{y}}, \mathbf{x})$ и обучающая для $\hat{\sigma}_\lambda^2(\mathbf{X}, \tilde{\tilde{\mathbf{y}}}) = (1/N) (\tilde{\tilde{\mathbf{y}}} - \mathbf{X}^T \hat{\mathbf{a}}_\lambda(\mathbf{X}, \tilde{\tilde{\mathbf{y}}}))^T (\tilde{\tilde{\mathbf{y}}} - \mathbf{X}^T \hat{\mathbf{a}}_\lambda(\mathbf{X}, \tilde{\tilde{\mathbf{y}}}))$. Однако в реальности у наблюдателя имеется единственная выборка $(\mathbf{X}, \mathbf{y})$.

Теорема 4. Критерий неявной кросс-валидации для неизвестной дисперсии наблюдений:

$$\hat{\lambda}(\mathbf{y}, \mathbf{X}) = \arg \min_{\lambda} \left\{ \frac{N}{2} \ln \hat{\sigma}_\lambda^2(\mathbf{X}, \mathbf{y}) + \frac{1}{\hat{\sigma}_\lambda^2(\mathbf{X}, \mathbf{y})} \Delta(\lambda, \mathbf{X}) \right\}, \quad (30)$$

$$\hat{\sigma}_\lambda^2(\mathbf{X}, \mathbf{y}) = \frac{1}{N} (\mathbf{y} - \mathbf{X}^T \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{X}))^T (\mathbf{y} - \mathbf{X}^T \hat{\mathbf{a}}_\lambda(\mathbf{y}, \mathbf{X})), \quad \Delta(\lambda, \mathbf{X}) = \text{Tr} \left[\mathbf{X} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \mathbf{B}_\lambda)^{-1} \right].$$

В третьей главе рассмотрены особенности применения метода неявной кросс-валидации для трех частных видов квадратичной модели линейной регрессии.

Нестационарная регрессия с неизвестной степенью нестационарности отличается тем, что вместо обучающей совокупности независимых данных анализу подлежит пара процессов на оси дискретного времени

$$\{(\mathbf{x}_t, y_t), t = 1, \dots, N\}, \quad \mathbf{x}_t \in \mathbb{R}^n, \quad y_t \in \mathbb{R}.$$

Требуется найти линейную зависимость скалярной компоненты процесса y_t от векторной компоненты, предполагая, что искомый вектор коэффициентов регрессии сам изменяется во времени:

$$\hat{y}_t(\mathbf{x}_t, \mathbf{a}_t) = \mathbf{a}_t^T \mathbf{x}_t, \quad \mathbf{a}_t \in \mathbb{R}^n, \quad t = 1, \dots, N.$$

Заметим, что совокупность искоемых коэффициентов регрессии многократно превышает число наблюдений, поэтому бессмысленно пытаться минимизировать квадратичную функцию потерь

$$\sum_{t=1}^N q(y_t, \mathbf{x}_t, \mathbf{a}_t) = \sum_{t=1}^N (y_t - \mathbf{a}_t^T \mathbf{x}_t)^2 \rightarrow \min_{(\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^{nN}}.$$

В качестве квадратичного условия регуляризации выступает предположение о медленном изменении модели регрессии во времени

$$\frac{1}{\lambda} \sum_{t=2}^N (\mathbf{a}_t - \mathbf{V}_t \mathbf{a}_{t-1})^T (\mathbf{a}_t - \mathbf{V}_t \mathbf{a}_{t-1}) \rightarrow \min_{(\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^{nN}},$$

где последовательность заданных квадратных матриц $\mathbf{V}_2, \dots, \mathbf{V}_N$ выражает понимание «медленного изменения» в смысле конкретного процесса реального мира, подлежащего изучению, но это условие также не имеет смысла как отдельная задача оптимизации.

Регуляризованный критерий оценивания модели нестационарной регрессии естественно построить как баланс между этими взаимно противоречивыми условиями

$$\sum_{t=1}^N (y_t - \mathbf{a}_t^T \mathbf{x}_t)^2 + \frac{1}{\lambda} \sum_{t=2}^N (\mathbf{a}_t - \mathbf{V}_t \mathbf{a}_{t-1})^T (\mathbf{a}_t - \mathbf{V}_t \mathbf{a}_{t-1}) \rightarrow \min_{(\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^{nN}}, \quad (31)$$

где структурный параметр λ определяет степень изменчивости во времени последовательности оценок векторных коэффициентов регрессии – чем больше λ , тем более волатильна получаемая последовательность.

Несмотря на большое число переменных, задача (31) легко численно решается за время, пропорциональное длине временного ряда N , с помощью фильтра-интерполятора Калмана-Бьюси, реализующего принцип квадратичного динамического программирования⁸.

Задача (31) является частным случаем общей линейно-квадратичной задачи (25) с матрицей регуляризации специального вида

$$\mathbf{B}_\lambda = \frac{1}{\lambda} \begin{pmatrix} \mathbf{V}_2^T \mathbf{V}_2 & -\mathbf{V}_2 & \mathbf{0} & \dots & \mathbf{0} \\ -\mathbf{V}_2^T & \mathbf{V}_3^T \mathbf{V}_3 + \mathbf{I} & -\mathbf{V}_3 & \ddots & \vdots \\ \mathbf{0} & -\mathbf{V}_3^T & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \mathbf{V}_N^T \mathbf{V}_N + \mathbf{I} & -\mathbf{V}_N \\ \mathbf{0} & \dots & \mathbf{0} & -\mathbf{V}_N^T & \mathbf{I} \end{pmatrix} (nN \times nN). \quad (32)$$

Теорема 5. С учетом обозначения (32) критерии неявной кросс-валидации, соответственно, для задачи оценивания нестационарной регрессии с заданной и оцениваемой дисперсией наблюдений имеют вид, аналогичный (28) и (30).

Однако зависимость штрафного члена от структурного параметра волатильности оценок требует специального исследования.

Теорема 6. Эффективная размерность $\lim_{\lambda \rightarrow 0} \Delta(\lambda, \mathbf{X}) = \begin{cases} n, & \lambda \rightarrow 0, \\ N, & \lambda \rightarrow \infty. \end{cases}$

Штрафной член $\Delta(\lambda, \mathbf{X})$ играет здесь роль *эффективной размерности* задачи. При малых значениях λ модель становится стационарной, и вся последовательность оценок определяется первым вектором коэффициентов регрессии $\mathbf{a}_1 \in \mathbb{R}^n$, поэтому

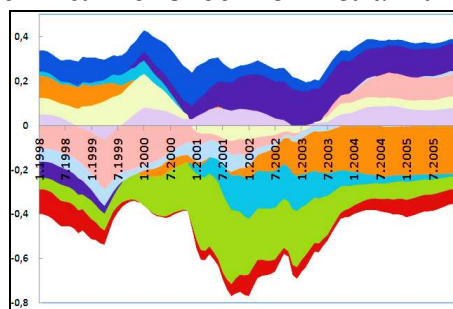
⁸ M. Markov, O. Krasotkina, V. Mottl, I. Muchnik. Time-varying regression model with unknown time-volatility for nonstationary signal analysis. *Proceedings of the 8th IASTED International Conference on Signal and Image Processing*. Honolulu, Hawaii, USA, August 14-16, 2006, paper 534-196.

эффективная размерность совпадает с числом регрессоров n . При очень больших значениях λ нет никакой априорной информации об nN значениях $(\mathbf{a}_1 \cdots \mathbf{a}_N)$, но по N наблюдениям все равно можно оценить не более N параметров, поэтому эффективная размерность не превышает длины временного ряда.

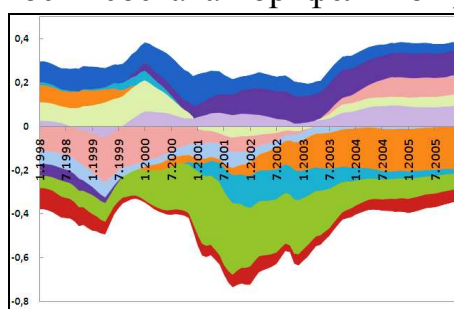
С вычислительной точки зрения преимущества неявной кросс-валидации перед прямым скользящим контролем очевидны – для каждого пробного значения параметра λ временной ряд обрабатывается один раз вместо N раз. Однако сравнить получаемые результаты можно лишь экспериментально.

Модельные эксперименты дают возможность придать абсолютный смысл понятию «подходящего» значения параметра нестационарности в модели нестационарной регрессии. Качество конкретной процедуры анализа временного ряда естественно оценивать по относительному отклонению восстановленной последовательности коэффициентов регрессии от истинной последовательности, в точности известной в модельных данных. Серия модельных экспериментов показала практически одинаковый разброс оценок последовательностей коэффициентов регрессии, полученных при подборе волатильности модели по методу неявной кросс-валидации и по скользящему контролю.

Обработка реальных данных дала тот же результат. Мы применили динамическую регрессионную модель скрытого долевого состава инвестиционного портфеля⁹ к опубликованным временным рядам периодических доходностей инвестиционной компании Laudus Rosenberg Value Long/Short Fund $y_1, \dots, y_N$ в течении $N = 60$ месяцев в интервале с января 2001 года по декабрь 2005 года вместе с известными по биржевым сводкам временными рядами индексов доходностей 10 основных секторов экономики $\mathbf{x}_1, \dots, \mathbf{x}_N$, $\mathbf{x}_t \in \mathbb{R}^{10}$. Долевой состав портфеля относительно этих индексов $\mathbf{a}_1, \dots, \mathbf{a}_N$, $\mathbf{a}_t \in \mathbb{R}^{10}$, оценивался как последовательность коэффициентов нестационарной регрессии. В такой модели параметр нестационарности λ имеет смысл неизвестной волатильности состава портфеля во времени.



неявная кросс-валидация



скользящий контроль

Временные ряды восстановленного долевого состава портфеля с коэффициентом волатильности, выбранным двумя методами, показывают, что критерии дают весьма близкие результаты.

Вернемся к рассмотрению обычных задач оценивания регрессионных моделей, связанных с анализом обучающих совокупностей, рассматриваемых как результат многократных независимых экспериментов (21)–(22). Для практики типична ситуация, когда число пробных регрессоров n , доступных наблюдателю, превосходит число наблюдений N , и актуальна задача сокращения числа активных регрессоров и, соответственно, коэффициентов регрессии. Мы рассмотрим применение метода неявной кросс-валидации для выбора наиболее подходящего уровня сложности модели для двух спо-

⁹ M. Markov, I. Muchnik, V. Mottl, O. Krasotkina. Dynamic analysis of hedge funds. *Proceedings of the 8th IASTED International Conference on Financial Engineering and Applications*. MIT, Cambridge, Massachusetts, USA, October 9-11, 2006, paper 546-028.

собов отбора регрессоров, один из которых подавляет «лишние» регрессоры «мягким» путем, снижая степень их участия в модели, в то время как второй способ полностью удаляет неинформативные регрессоры из модели.

«Мягкое» подавление нерелевантных регрессоров. Соответствующий способ предложил аспирант Тульского государственного университета Нгуен Чонг Тинь¹⁰. В качестве основы он использовал диагональную квадратичную регуляризацию

$$\hat{\mathbf{a}}_{\lambda}(\mathbf{y}, \mathbf{X}) = \arg \min_{\mathbf{a} \in \mathbb{R}^n} \left\{ \mathbf{a}^T \mathbf{B}_{\lambda} \mathbf{a} + (\mathbf{y} - \mathbf{X}^T \mathbf{a})^T (\mathbf{y} - \mathbf{X}^T \mathbf{a}) \right\} \quad (33)$$

с векторным структурным параметром $\lambda = (\lambda_1, \dots, \lambda_n)$, отличающуюся от обычной ридж-регрессии только тем, что коэффициенты, штрафующие отклонение коэффициентов регрессии от нуля, приняты разными для разных регрессоров $\mathbf{a}^T \mathbf{B}_{\lambda} \mathbf{a} = \sum_{i=1}^n (1/\lambda_i) a_i^2$. Если $\lambda_i \rightarrow 0$, то $\hat{a}_i(\mathbf{y}, \mathbf{X}) \rightarrow 0$, и i -й регрессор существенно подавляется. Для автоматического индивидуального выбора коэффициента $1/\lambda_i$ для каждого регрессора, в диссертации¹⁰ предложено использовать модифицированный критерий обучения

$$(\hat{\mathbf{a}}_{\lambda}(\mathbf{y}, \mathbf{X}), \hat{\lambda}(\mathbf{y}, \mathbf{X})) = \arg \min_{\mathbf{a} \in \mathbb{R}^n, \lambda \geq 0} \left\{ \sum_{i=1}^n \frac{1}{\lambda_i} a_i^2 + \sum_{i=1}^n \left[\frac{1}{\mu} \frac{1}{\lambda_i} + \left(\frac{1}{\mu} + 1 \right) \ln \lambda_i \right] + (\mathbf{y} - \mathbf{X}^T \mathbf{a})^T (\mathbf{y} - \mathbf{X}^T \mathbf{a}) \right\}, \quad (34)$$

содержащий дополнительное второе слагаемое в регуляризирующем члене, и предложен алгоритм численного решения расширенной задачи оптимизации. Коэффициент $\mu > 0$ в дополнительном слагаемом играет роль вспомогательного структурного параметра селективности отбора регрессоров – чем больше μ , тем большее число штрафных коэффициентов $1/\lambda_i$ становятся очень большими, заставляя соответствующие коэффициенты регрессии приближаться к нулю в точке минимума критерия $\hat{a}_i \rightarrow 0$, подавляя соответствующие регрессоры и уменьшая тем самым эффективную размерность модели.

Для последовательности пробных значений селективности $\mu^{(1)} < \dots < \mu^{(m)}$ алгоритм обучения Нгуена дает последовательность моделей уменьшающейся сложности $\lambda^{(1)}, \dots, \lambda^{(m)}$. В своей диссертации¹⁰ он использовал обычный алгоритм скользящего контроля для выбора оптимального значения селективности, требующий для каждого пробного значения $\mu^{(k)}$ решать задачу обучения N раз. В настоящей диссертации мы используем для этого беспереборный метод неявной кросс-валидации.

Для каждого из полученных пробных значений векторного параметра регуляризации $\lambda^{(1)}, \dots, \lambda^{(m)}$ модель (33) является линейно-квадратичной (25) с диагональной матрицей $\mathbf{B}_{\lambda^{(k)}} = \text{Diag}(1/\lambda_1^{(k)} \dots 1/\lambda_n^{(k)})$. Дисперсию наблюдения σ^2 в (27) вряд ли можно считать известной, поэтому следует использовать критерий неявной кросс-валидации (30).

Теорема 7. Эффективная размерность $\lim \Delta(\lambda, \mathbf{X}) = \begin{cases} 0, & \lambda \rightarrow 0, \\ \min[n, \text{Rank}(\mathbf{X}\mathbf{X}^T)], & \lambda \rightarrow \infty. \end{cases}$

Мы исследовали такой способ выбора селективности отбора регрессоров, используя тот же модельный эксперимент, который описан в диссертации¹⁰.

¹⁰ Нгуен Чонг Тинь. Методы селективного комбинирования признаков информации в задаче оценивания регрессионной зависимости. Дисс. к.ф.-м.н. Тульский гос. унив., 2013.

Полное подавление нерелевантных регрессоров. Способ полного подавления нерелевантных регрессоров предложен в работе¹¹ под названием критерия обучения *Elastic Net*. Он опирается на обычную квадратичную модель ридж-регрессии

$$\hat{\mathbf{a}}_{\beta}(\mathbf{y}, \mathbf{X}) = \arg \min_{\mathbf{a} \in \mathbb{R}^n} \left\{ \beta \sum_{i \in I} a_i^2 + (\mathbf{y} - \mathbf{X}^T \mathbf{a})^T (\mathbf{y} - \mathbf{X}^T \mathbf{a}) \right\}, \text{ где } I = \{1, \dots, n\} - \left\{ \begin{array}{l} \text{множество} \\ \text{регрессоров,} \end{array} \right. \quad (35)$$

но отличается от нее наличием дополнительного регуляризующего слагаемого, штрафующего отклонение от нуля абсолютных значений коэффициентов регрессии:

$$\hat{\mathbf{a}}_{\beta\mu}(\mathbf{y}, \mathbf{X}) = \arg \min_{\mathbf{a} \in \mathbb{R}^n} \left\{ \beta \sum_{i \in I} a_i^2 + \mu \sum_{i \in I} |a_i| + (\mathbf{y} - \mathbf{X}^T \mathbf{a})^T (\mathbf{y} - \mathbf{X}^T \mathbf{a}) \right\}. \quad (36)$$

Обычно ридж-коэффициент $\beta > 0$ принимают достаточно малым, только чтобы преодолеть возможную вырожденность матрицы (26), если исходное число регрессоров превышает объем обучающей совокупности $n > N$ (22). В качестве рабочего структурного параметра выступает коэффициент селективности $\mu \geq 0$. Чем больше μ , тем большее число коэффициентов регрессии оказываются в точности равными нулю $\hat{a}_{i,\mu} = 0$ в точке минимума выпуклого критерия обучения (36).

В диссертации аспиранта МФТИ Николая Разина¹² доказано, что для всякого значения селективности μ полностью определено разбиение (*Partitioning*) $P_{\mu} = \hat{I}_{\mu}^{-} \cup \hat{I}_{\mu}^{+} \cup \hat{I}_{\mu}^0$ исходного полного множества регрессоров $I = \{1, \dots, n\}$ на три подмножества $\hat{I}_{\mu}^{-} = \{i \in I : \hat{a}_{i,\mu} < 0\}$, $\hat{I}_{\mu}^{+} = \{i \in I : \hat{a}_{i,\mu} > 0\}$ и $\hat{I}_{\mu}^0 = \{i \in I : \hat{a}_{i,\mu} = 0\}$, так что только объединение первых двух подмножеств $\hat{I}_{\mu} = \hat{I}_{\mu}^{-} \cup \hat{I}_{\mu}^{+} \subset I$ есть множество активных регрессоров в точке минимума критерия обучения (36), в то время как остальные регрессоры $\hat{I}_{\mu}^0 \subset I$ полностью удалены из модели.

Алгоритм Разина дает последовательность пробных значений селективности $\mu^{(1)} < \dots < \mu^{(m)}$ и соответствующую последовательность моделей в основном уменьшающейся сложности, хотя полную вложенность подмножеств активных регрессоров $I \supset \hat{I}^{(1)} \supset \dots \supset \hat{I}^{(m)} = \emptyset$ гарантировать, вообще говоря, нельзя. Остается вопрос о выборе «оптимального» значения селективности. Для ответа на него принято использовать алгоритм скользящего контроля, требующий повторения обучения по критерию Elastic Net (36) N раз для каждого пробного значения $\mu^{(k)}$.

Идея данной диссертации заключается в интерпретации подмножества активных (релевантных) регрессоров $\hat{I}^{(k)} \subset I$ как самостоятельного структурного параметра модели вместо исходного значения селективности $\mu^{(k)}$. Для каждого такого разбиения исходный критерий ридж-регрессии (35) принимает простой вид:

$$\hat{\mathbf{a}}^{(k)} = \arg \min_{\mathbf{a} \in \mathbb{R}^{n_k}} \left\{ \beta \sum_{i \in \hat{I}^{(k)}} a_i^2 + \sum_{j=1}^N \left(y_j - \sum_{i \in \hat{I}^{(k)}} a_i x_{ij} \right)^2 \right\}, \text{ где } n_k = |\hat{I}^{(k)}| - \text{число активных регрессоров.}$$

¹¹ H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 2005, Vol. 67, pp. 301-320.

¹² Разин Н.А. Выпуклые критерии и параллелизуемые алгоритмы селективного комбинирования разнородных представлений объектов в задачах восстановления зависимостей по эмпирическим данным. Дисс. к.ф.-м.н. ВЦ РАН, 2013.

Теорема 8. Критерий неявной кросс-валидации:

$$\hat{k} = \arg \min_k \left\{ \frac{N}{2} \ln(\hat{\sigma}^{(k)})^2 + \frac{\Delta^{(k)}}{(\hat{\sigma}^{(k)})^2} \right\}, \quad \text{где} \quad \Delta^{(k)} = \text{Tr} \left\{ \tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k^T \left(\tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k^T + \beta \mathbf{I}_{n_k} \right)^{-1} \right\},$$

$$\tilde{\mathbf{X}}_k = (\tilde{\mathbf{x}}_k^1 \dots \tilde{\mathbf{x}}_k^n)^T, \quad \tilde{\mathbf{x}}_k^j = (x_{ij}, i \in I_k) \in \mathbb{R}^{n_k}, \quad (\hat{\sigma}^{(k)})^2 = \frac{1}{N} \sum_{j=1}^N \left(y_j - \sum_{i \in \hat{I}_k} \hat{a}_i^{(k)} x_{ij} \right)^2$$

Такой способ выбора селективности отбора регрессоров исследован в том же модельном эксперименте, который описан в диссертации.

В заключении перечислены следующие основные результаты:

1. Разработан метод неявной кросс-валидации для оценивания обобщающей способности линейно-квадратичных моделей числовых зависимостей.
2. Исследована природа классического информационного критерия Акаике как простейшего частного случая критерия неявной кросс-валидации.
3. Разработан критерий неявной кросс-валидации для выбора степени волатильности модели нестационарной регрессии.
4. Разработан критерий неявной кросс-валидации для выбора степени подавления нерелевантных регрессоров в линейно-квадратичной модели числовой регрессии.
5. Разработан критерий неявной кросс-валидации для выбора уровня селективности формирования подмножества релевантных регрессоров в квадратично-модульной модели Elastic Net.
6. Экспериментально проверены полученные методы беспереборной кросс-валидации.

Список публикаций

- [1] Обобщение АИС для выбора значений непрерывных параметров / Е. О. Ежова (Черноусова), В. В. Моттль, О. В. Красоткина // *Таврический вестник информатики и математики*. – 2009. – No 1. – С. 61-71.
- [2] Обобщение информационного критерия Акаике для выбора значений непрерывного параметра в моделях данных / В. В. Моттль, О. В. Красоткина, Е. О. Ежова (Черноусова) // *Труды 51 научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»*. — 2008. — Т. 3. — С. 58-60.
- [3] Estimation of time-varying linear regression with unknown time-volatility via continuous generalization of the Akaike Information Criterion. / E. Ezhova (Chernousova), V. Mottl, O. Krasotkina // *Proceedings of World Academy of Science, Engineering and Technology*. – 2009. – March. - No.51. - P. 144-150.
- [4] Непрерывная коррекция информационного критерия Акаике для регуляризованного оценивания сверхбольшого числа параметров регрессионных моделей данных с неизвестной дисперсией наблюдения / Е. О. Ежова (Черноусова), О. В. Красоткина, В. В. Моттль // *Интеллектуализация обработки информации*:

8-я международная конференция. Сборник докладов. – 2010. - М.:МАКС Пресс. - С. 51-54.

- [5] Непрерывное обобщение информационного критерия Акаике для оценивания нестационарной регрессионной модели временного ряда с неизвестной степенью изменчивости коэффициентов / В. В. Моттль, О. В. Красоткина, Е. О. Ежова (Черноусова) // *Математические методы распознавания образов: 14-я Всероссийская конференция. Сборник докладов.* – 2009. - М.:МАКС Пресс. - С. 52-55.
- [6] Линейные модели числовых зависимостей на множествах объектов, представленных парными сравнениями / Е. О. Черноусова, О. В. Красоткина, М. Е. Панов, Е. В. Гребенников, В. В. Моттль // *Интеллектуализация обработки информации: 9-я международная конференция. Сборник докладов.* – 2012. - М.:Торус Пресс. - С. 58-62.
- [7] **Linear regression via Elastic Net: Non-enumerative leave-one-out verification of feature selection** / E.Chernousova, N. Razin, O. Krasotkina, V. Mottl, D Windridge // *Clusters, orders, trees: methods and applications. P. Pardalos, B. Goldengorin, F. Aleskerov (Eds).* Springer. – 2013. - V. 5 . – P. 35-46.
- [8] **Беспереборная кросс-валидация отбора признаков в линейной регрессионной модели** / О. В. Красоткина, В. В. Моттль, Н. А. Разин, Е. О. Черноусова // *Известия ТулГУ. Технические науки.* – 2013. – Вып. 7. – No. 2. – С. 88-98.

Жирным шрифтом выделены статьи в журналах, рекомендованы ВАК. Личный вклад соискателя в работах с соавторами заключается в следующем:

[1, 2] Предложен и доказан способ выбора структурных параметров моделей, принимающих непрерывные значения, как обобщение классического способа выбора структурного параметра, принимающего целочисленные значения, на основе информационного критерия Акаике (AIC).

[3, 5] Проведение экспериментов на реальных данных для нестационарной регрессионной модели временного ряда с неизвестной степенью изменчивости коэффициентов для проверки обобщенного информационного критерия Акаике. Эксперимент проведен на примере анализа деятельности американского паевого инвестиционного фонда Лаудис Розенберг.

[4] Предложена и доказана непрерывная коррекция информационного критерия Акаике для регуляризованного оценивания сверхбольшого числа параметров регрессионных моделей данных с неизвестной дисперсией наблюдения.

[6] Исследованы особенности неявной кросс-валидации в линейных моделях числовых зависимостей на множествах объектов, представленных парными сравнениями.

[7, 8] Предложена и доказана реализация беспереборного вычисления метода скользящего контроля в задаче оценивания линейной регрессионной зависимости с квадратично-модульной регуляризацией.

Черноусова Елена Олеговна

Беспереборные методы кросс-валидации для оценивания обобщающей
способности регрессионных моделей

АВТОРЕФЕРАТ

Подписано в печать 21.11.2013.

Формат $60 \times 84 \frac{1}{16}$. Усл. печ. л. 1,0. Тираж 100 экз. Заказ № 375 .

Федеральное государственное автономное образовательное учреждение высшего
профессионального образования «Московский физико-технический институт
(государственный университет)»

Отдел оперативной полиграфии «Физтех-полиграф»

141700, Московская обл., г. Долгопрудный, Институтский пер., 9