DOI <https://doi.org/10.18551/rjoas.2017-03.13>

**АНАЛИЗ КОЛЛИНЕАРНОСТИ КАК ИНСТРУМЕНТ ПАРАМЕТРИЧЕСКОГО
СОПОСТАВЛЕНИЯ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЗАВИСИМОСТЕЙ**
COLLINEARITY ANALYSIS AS A TOOL FOR PARAMETRIC CO-ASSESSMENT
OF SOCIO-ECONOMIC RELATIONSHIPS

Прилуков А.Н., кандидат социологических наук

Prilukov A.N., Candidate of Sociological Sciences

Институт горного дела Дальневосточного отделения РАН, Хабаровск, Россия

Mining Institute of Russian Academy of Sciences, Khabarovsk, Russia

E-mail: a_prilukov@mail.ru

АННОТАЦИЯ

В статье описываются ключевые идеи, ход и результаты теоретико-прикладного исследования, направленного на создание высокоформализованного метода анализа коллинеарности, который предназначен для параметрического сопоставления функциональных и статистических зависимостей, относящихся к социально-экономической и другим областям деятельности. От родственных математико-статистических методов корреляционного и регрессионного анализа разработанный метод отличается рядом реализованных в нём принципиальных новшеств: (1) возможностью взаимной адаптации сопоставляемых зависимостей в многомерном признаковом пространстве, включающем ряд оптимизируемых параметров – соотношения масштабов зависимостей, их относительного взаимного смещения, способов интерполяции, экстраполяции, нормирования и цензурирования числовых значений зависимостей и их выборок; (2) применением варьируемой метрики Минковского, используемой для управления механизмом суммирования регрессионных ошибок или функциональных невязок; (3) применением эффективного комбинированного алгоритма поиска решения, заключающегося в нахождении глобального минимума целевой функции на основе сочетания методов Монте-Карло и скорейшего градиентного спуска. Кратко охарактеризованы результаты статистических испытаний метода и вариантов его программной реализации, отмечены предпочтительные направления дальнейшего совершенствования метода и использующихся в нём формализованных процедур.

ABSTRACT

In the article the ideas, progress, and results of a theoretic-and-applied research direction are described, aimed at highly formalized method creation for collinearity analysis, devoted to functional and statistical relationships co-assessment that belong to socio-economic and other human activities areas. With respect to related mathematical-statistical methods of correlation and regression analyses the new method differs by some principal novelties implemented there: (1) opportunity of dependencies being compared mutual adaptation in space of attributes that include several parameters being optimized – dependencies scales ratio, their mutual shift, methods of interpolation, extrapolation, normalization, and censoring with respect to dependencies numeric values and their samples; (2) variable Minkowsky metric, used for regression errors or functional residuals summation mechanism's management; (3) utilization of effective combined algorithm for solution attaining, which consists in finding the target function's global minimum, based on combination of Monte-Carlo and steepest descent methods. Results of the method and its program realization variants' statistical tests are briefly characterized, preferable directions of the method's and its inner formalized procedures' further improvement are pointed out.

КЛЮЧЕВЫЕ СЛОВА

Анализ коллинеарности, регрессия, взаимная корреляция, метод Монте-Карло.

KEY WORDS

Collinearity analysis, regression, cross-correlation, Monte-Carlo algorithm.

При анализе социально-экономической и других видов информации часто возникает необходимость сопоставления между собой функциональных либо статистических зависимостей – эмпирических или расчётных линий регрессии, полигонов распределения величин, динамических рядов и т. д. При проведении подобного рода анализов прибегают к способам, различающимся по уровню формализованности. Нередко выбор останавливается на частично формализованных эвристических способах.

При выполнении такого рода анализов широко применяемыми и в разной мере теоретически обоснованными подходами являются:

1) вычисление коэффициентов корреляции, ковариации и аналогичных им показателей, отражающих тесноту статистических связей между анализируемыми зависимостями;

2) сопоставление вместо исходных зависимостей их аналитических аппроксимаций, получаемых различными способами.

Первый из названных подходов обладает весьма ограниченными аналитическими возможностями. Рассчитываемые обыкновенные и ранговые коэффициенты корреляции дают возможность выразить численно лишь степень совпадения двух сравниваемых зависимостей по их форме и общей направленности. В то же время коэффициенты корреляции не дают представления ни о геометрической близости прямых или кривых, которые могут быть как угодно далеко разнесены одна от другой по осям абсцисс и ординат, ни об особенностях их формы.

К недостаткам корреляционного анализа относится то, что при этом оценивается лишь теснота линейной связи между случайными или детерминированными зависимостями, при наличии же жёстко детерминированной, но нелинейной связи рассчитываемые коэффициенты утрачивают свою результативность. Это, в частности, заставляет вводить при выполнении корреляционного анализа дополнительные операции для линеаризации исходных зависимостей, что затрудняет проведение анализа и не всегда способствует получению объективных результатов.

Существенный недостаток коэффициентов корреляции связан с процедурами их вычисления, которые строятся либо на основе знания аналитических выражений, описывающих сравниваемые функции $x(t)$ и $y(t)$, либо осуществляется приближённо по дискретным значениям этих функций x_i, y_j .

Условиям рассматриваемой в статье задачи наиболее близко соответствует вычисление взаимной корреляции. Последняя обычно используется в качестве меры подобия двух переменных величин, из которых одна может смещаться относительно другой. В случае непрерывных функций $f(t)$ и $g(t)$ их взаимная корреляция описывается выражением:

$$(f * g)(\tau) \equiv \int_{-\infty}^{\infty} f^*(t)g(t+\tau)dt \quad (1),$$

где f^* – комплексное сопряжение f , τ – смещение g относительно f . Взаимная корреляция дискретных функций определяется выражением:

$$(f * g)[n] \equiv \sum_{m=-\infty}^{\infty} f^*[m]g[m+n] \quad (2),$$

здесь n – дискретное смещение [1].

Подбор смещения (лага) одной из сравниваемых зависимостей относительно другой применяется и при использовании других математико-статистических методов. Например, в модели APB (Adjusted Precipitation Benchmark) таким способом достигается приближающееся к оптимальному совмещение по времени графиков выпадения осадков и объёмов формирующихся на их основе водосборов на изучаемых территориях [2–3].

Недостатком методов, основывающихся на взаимном смещении зависимостей, является происходящая при этом утрата их концевых участков, тем более значительная, чем большей оказывается ожидаемая величина смещения. При использовании дискретных зависимостей к этому добавляется невозможность установления точной величины смещения, для которого требуются дробные доли шага дискретизации.

Реализация второго из названных методических подходов связана с аппроксимацией исходных зависимостей, то есть их приближённым представлением с помощью тех или иных математических функций – алгебраических многочленов от одной или нескольких переменных, тригонометрических многочленов, рядов Фурье и других [4]. Непосредственному сопоставлению затем подвергаются выбранные аналитические выражения – оцениваются, например, коэффициенты полиномиальных уравнений регрессии, коэффициенты разложений Фурье и т.д.

Этому направлению также присущи очевидные недостатки. Один из них состоит в том, что аппроксимация зависимостей сама по себе является непростой задачей, которая не всегда решается желаемым образом, зачастую требуя при этом приложения немалых усилий. Вдобавок, после получения аппроксимирующих выражений сопоставление последних, в зависимости от их конкретного вида, также может быть затруднено.

Поскольку аппроксимация связана с выбором конкретного вида аппроксимирующих функций, каждая из которых обладает присущими ей особенностями, перед выполнением аппроксимации требуется знание содержательных характеристик исходных зависимостей, что не всегда имеет место на практике. Положение часто не спасает даже точное знание форм анализируемых зависимостей, поскольку их геометрические очертания могут быть интерпретированы по-разному, приводя порой к неоправданным выводам. Для некоторых эмпирических зависимостей адекватных аппроксимирующих выражений подобрать не удаётся, что заставляет прибегать к поверхностным аналогиям, ставящим под сомнение результативность анализа.

Перечисленные трудности побуждают к поиску нетрадиционных способов сопоставления зависимостей. В одном из них, названном “диахронной корреляцией” [5], формулируются такие методические требования, как анализ развивающихся во времени объектов и выбор способа математического описания их “траектории движения”, соотношения между начальными моментами и скоростями изменения объектов, выявление степени нелинейности “траекторий” и др. Однако большинство этих содержательно важных методических вопросов рассматривается лишь в постановочном плане, без привлечения необходимого в таких случаях математического аппарата и выполнения соответствующих выкладок.

В другом конкретном методе, предложенном Ф.Харватом [6], степень совпадения зависимостей выражается с помощью так называемого индекса расстояния. При вычислении последнего два сравниваемых дискретных ряда x и y вначале нормируются так, чтобы в сумме каждый из пронормированных рядов давал 100:

$$\sum_i \hat{x}_i = \sum_i \hat{y}_i = 100. \quad (3)$$

Очевидно, этого можно добиться, если каждое из исходных значений x_i ($i=1, 2, \dots, m$) умножить на величину:

$$A = 100 / \sum_i x_i \quad (4)$$

и каждое из значений y_i на:

$$B = 100 / \sum_i y_i. \quad (5)$$

Затем попарно сопоставляются члены пронормированных рядов и вычисляется упомянутый выше индекс расстояния:

$$\rho(\hat{x}, \hat{y}) = 100 - \sum_i \min(\hat{x}_i, \hat{y}_i) = \sum_i \max(\hat{x}_i, \hat{y}_i) - 100. \quad (6)$$

Понятно, что если независимо от применённых единиц измерения исходные зависимости имеют одинаковые очертания, то в результате нормирования расчётные величины $\hat{x}_i$ и $\hat{y}_i$ окажутся попарно равными и ρ будет равен нулю. Если очертания у исходных зависимостей разные, ρ будет ненулевым, лежащим в диапазоне от 0 до 100. При $\rho=100$ наблюдается полное расхождение рядов, когда каждому нулевому члену ряда x соответствует ненулевой член ряда y и наоборот.

Подкупающими в предложенном Ф.Харватом методе являются наглядность и простота его реализации. В то же время очевидны и слабые стороны метода – его приложимость только к дискретным рядам, имеющим одинаковую длину, нечувствительность метода к характеру различий между сравниваемыми зависимостями.

ВЫБОР СПОСОБА И ПРОЦЕДУР РЕШЕНИЯ ЗАДАЧИ

Разработанный в ходе описываемого исследования метод *анализа коллинеарности* создан на стыке нескольких математических разделов: математического анализа, прикладной математической статистики, теории аппроксимации функций, построения алгоритмов и реализации численных методов на ЭВМ.

Термин “коллинеарность” (collinearity) известен давно и достаточно часто появляется в работах, основывающихся на использовании математико-статистических методов либо освещающих проблемы последних [2; 4; 7–10], хотя единства в понимании термина до сих пор не достигнуто. Иногда в данный термин вкладывается близкий к геометрическому смысл, как, например, в [9], где им обозначаются линии связи на генетических картах, характеризующих сельскохозяйственные культуры. Значительно чаще термин “коллинеарность” появляется в работах, связанных с использованием методов корреляционного и регрессионного анализа. При этом, как утверждается, например, в [10], на коллинеарность принято возлагать ответственность за всевозможные неприятности в ходе эмпирических исследований – неубедительные и слабые результаты, неожиданно сменяющиеся положительный и отрицательный знаки у расчётных коэффициентов, общий хаос в нелинейных оценках [10, p.1].

С другой стороны, называя числовые ряды или векторы коллинеарными, обычно имеют в виду их высокую взаимную коррелированность либо схожесть по форме. О. Михненко, например, утверждает, что факторные признаки, которыми описываются условия и результаты человеческой деятельности, чаще всего в большей или меньшей степени коллинеарны [8, с.52]. Аналогичное обстоятельство имеет в виду и автор [2, p.887].

Указанный позитивный смысл термина сохранён в названии, предложенном для обозначения описываемого метода. Однако здесь добавляется и новый смысловой оттенок, связанный с тем, что метод предназначен для исследования степени близости зависимостей, выявляемой по сходству очертаний линий либо гиперповерхностей, отображающих эти зависимости. Полностью коллинеарными будут в этом смысле те статистические или функциональные зависимости, условные

многомерные гиперпространственные изображения которых после взаимосогласования их ключевых параметров становятся идентичными друг другу.

Для нахождения величины, численно характеризующей коллинеарность, используется известное в математическом анализе понятие *расстояния в функциональном пространстве L_p* . В дальнейшем параметр, определяющий свойства пространства L_p и называемый в некоторых литературных источниках *метрикой Минковского*, будем обозначать символом r . При сопоставлении непрерывных функциональных зависимостей названное расстояние описывается выражением [7; 11–12]:

$$\rho_r(x, y) = \left[\int_v w(t) |x(t) - y(t)|^r dt \right]^{1/r} \quad (7)$$

Здесь $x(t)$, $y(t)$ – заданные сравниваемые зависимости; t – их аргумент; $w(t)$ – весовая функция; v – область определения функций x , y , w ; r – показатель (метрика) Минковского.

При дискретном задании функций x , y , w выражение (7) принимает вид:

$$\hat{\rho}_r(x, y) = \left(\sum_{i=1}^n w_i |x(t_i) - y(t_i)|^r \right)^{1/r} \quad (8)$$

где $x_i = x(t_i)$, $y_i = y(t_i)$, $w_i = w(t_i)$ – дискретные значения функций x , y , w ; $i = 1, 2, \dots, n$ – номер очередной дискретной точки; n – общее количество точек, для которых известны конкретные значения функций x , y , w .

Исходя из вида функционала (7) или его дискретного аналога (8), можно заключить, что с геометрической точки зрения коллинеарность обратно пропорциональна площади “зазора”, остающегося между линиями (гиперповерхностями) функций $x(t)$ и $y(t)$ при их взаимном наложении на одном графике или, в случае многомерных функций x , y , w , – в гиперпространстве. Введение взвешивания на $1/v$ (или на $1/n$) устраняет зависимость расчётной величины от протяжённости области v (или количества дискретных точек n), а использование весовой функции w (или дискретных весовых коэффициентов w_i) позволяет вводить в расчёты априорные оценки значимости тех или иных участков сравниваемых функций. Если, например, известно, что крайние участки функций x , y обладают меньшей значимостью или заданы с меньшей достоверностью по сравнению с их срединными участками, $w(t)$ задаётся таким образом, чтобы она принимала большие значения в середине области v и меньшие значения по её краям. Кроме того, весовая функция или коэффициенты могут использоваться для нормирования и цензурирования сравниваемых зависимостей, в частности, корректировки либо выбраковки их случайных “выбросов”. На функцию $w(t)$ в функционалах (7) и (8) накладываются определённые ограничения. Во-первых, она предполагается неотрицательной в пределах действительных значений области v :

$$w(t) \geq 0 \\ t \in v$$

Во-вторых, она должна быть нормирована таким образом, чтобы:

$$\int_v w(t) dt = 1 \quad \text{или} \quad \sum_{i=1}^n w_i = 1$$

При несоблюдении этих условий весовая функция будет вносить существенные изменения в расчётные значения ρ .

Последнее из названных условий легко может быть реализовано в самих функционалах (7) и (8). Функционал (8), например, в этом случае представляется в виде:

$$\hat{\rho}_r(x, y) = \left(\frac{\sum_{i=1}^n w_i |x_i - y_i|^r}{\sum_{i=1}^n w_i} \right)^{1/r}. \quad (9)$$

Одним из наиболее важных параметров в выражениях (7) – (9) является метрика Минковского r . В обсуждении вопросов, связанных с её использованием, участвовали многие авторы. В теоретических дискуссиях применение указанного параметра связывается с так называемой теорией метрических пространств. В этой теории в частности утверждается, что выражения, подобные приведенным выше, сохраняют признаки мер в многомерных признаковых пространствах и функциональных пространствах L_p в тех случаях, когда r лежит в диапазоне от 1 до ∞ . При некоторых конкретных значениях r образуются широко известные разновидности метрик: канберровская ($r = 1$), евклидова ($r = 2$), метрика доминирования ($r \rightarrow \infty$). В регрессионном анализе, например, особой популярностью пользуется так называемый *метод наименьших квадратов (МНК)* или в англоязычной транскрипции *Ordinary Least Squares estimation (OLS)* [13–20], позволяющий в простых случаях обходиться алгебраическими вычислениями, не прибегая к помощи ЭВМ.

Выход r за пределы указанного диапазона обычно априори отвергается, так как в противном случае не гарантируется соблюдение одной из аксиом теории меры – аксиомы треугольника. Последняя формулируется следующим образом [21].

$$\rho_r(x, y) \leq \rho_r(x, z) + \rho_r(z, y). \quad (10)$$

Интерпретируется это выражение следующим образом. Расстояние между зависимостями x и y , измеренное в метрике r , всегда не больше, чем сумма расстояний между каждой из двух зависимостей x , y и произвольной третьей зависимостью z . Максимальный интерес здесь, очевидно, представляет случай, когда зависимость z занимает в признаковом пространстве промежуточное положение между зависимостями x и y , хотя самой аксиомой треугольника ограничений на пространственное положение функции не накладывается. С помощью математических выкладок можно показать, что аксиома (10) строго выполняется только тогда, когда $r \geq 1$. Это обстоятельство и лежит в основе ограничений, на которое ссылаются упомянутые авторы.

Нужно заметить, что не все авторы склонны придерживаться рассматриваемого ограничения. Роскам, например, утверждает, что метрика Минковского должна подчиняться менее жёсткому ограничению, а именно $r \geq 0$ [22]. Автору статьи на основе больших серий компьютерных математико-статистических экспериментов удалось продемонстрировать, что без ущерба для точности и правильности расчётов r можно придавать значения, существенно меньшие единицы. В этой же области, то есть “левее” единицы, оказалось и оптимальное, отыскиваемое с помощью различных критериев, значение метрики Минковского при решении некоторых классов статистических задач [12; 23].

Объяснений этому феномену может быть несколько. Одно из них заключается в том, что аксиома треугольника не имеет прямого отношения к рассматриваемой задаче попарного сопоставления зависимостей. Во всяком случае, несоблюдение этой аксиомы не несёт в себе очевидных отрицательных последствий для результатов решения типичных задач рассматриваемого класса. Другое возможное объяснение

может заключаться в нарушении требования ортогональности признакового пространства, на котором основывается условие взаимной статистической независимости дискретных значений рассматриваемых функций и тем самым оправдывается использование теории меры. Нарушение же постулата происходит потому, что уже самими исходными предпосылками решаемой задачи предполагается взаимозависимость дискретных величин, отображающих сопоставляемые зависимости.

Снятие априорного ограничения на величину r позволяет существенно расширить область, необходимую для поиска оптимума. Часто используемые значения метрики Минковского 1 или 2 не всегда соответствуют характеру исходных данных. Теоретически обосновано [24], что значение r равное 2 должно использоваться, когда суммируемые величины (в данном случае – разности между ординатами x и y) распределены по нормальному статистическому закону. В других конкретных случаях оптимальными могут оказаться иные значения r . Поэтому принятие соглашения о том, что метрика Минковского является переменной величиной, устанавливаемой непосредственно перед выполнением расчётов или на основе их промежуточных результатов, в значительной степени устраняет затруднения, связанные с неизвестностью закона распределения суммируемых величин.

Рассчитываемая в соответствии с выражениями (7) и (8) скалярная величина ρ позволяет произвести “одномоментное” сопоставление зависимостей. Как и коэффициент корреляции, она даёт представление о степени сходства зависимостей, но в этом случае ничего не говорит ни об их особенностях, ни о характере расхождения зависимостей.

Для получения представлений о взаимном расхождении процессов или явлений, описываемых зависимостями, необходимо иметь возможность смещать одну из них относительно другой вдоль оси абсцисс. Поскольку сопоставляемые зависимости могут кроме этого различаться динамикой фиксируемых ими изменений или строиться на различных аргументах, для их взаимной “подгонки” в общем случае требуется варьирование соотношения масштабов по оси абсцисс.

Эти две операции, выполняющиеся в процессе анализа коллинеарности, названы “смещением” и “масштабированием”. Их описания вводятся в выражения (7) – (9) путём включения в последние дополнительных переменных величин: смещения s и масштаба m . Помимо некоторого усложнения функционалов при этом происходят принципиальные изменения, связанные с превращением скалярной величины ρ в многомерную функцию, аргументами (параметрами) которой являются s и m .

Введение дополнительных операций влечёт за собой и ещё ряд существенных преобразований в функционалах. Дело в том, что варьирование параметров s и m сопровождается смещением друг относительно друга первоначально имевших одинаковые абсциссы точек x_i и y_i . Кроме того, за счёт смещения и масштабирования первоначально совпадавшие области задания функций x и y также перестают совпадать: “начало” одной функции может оказаться “левее” начала другой; такая же картина наблюдается и на крайних правых участках функций.

К наиболее ощутимым последствиям эти осложнения приводят при дискретном варианте задания функций x и y , описываемом функционалом (9). Здесь помимо способа продолжения “укороченных” слева или справа участков функций x и y требуется ещё поиск способа заполнения промежутков между заданными дискретными точками функций, так чтобы для каждой переместившейся точки x_i и y_i была найдена соответствующая, имеющая ту же абсциссу расчётная точка $\hat{y}_i$ или $\hat{x}_i$.

Задачи поиска продолжения функций и заполнения промежутков между их дискретными точками осуществляются процедурами экстраполяции и интерполяции. Они не имеют однозначных решений и, более того, допускают множество способов решений, различающихся между собой как по формальным признакам, так и по достигаемым результатам. По отношению к анализу коллинеарности это означает, что в число его аргументов должны быть включены также параметры, которыми будут

задаваться способы экстраполирования и интерполирования функций в ходе решения задач. Преобразуя соответствующим образом функционал (9), можем записать

$$\hat{\rho}_K(x, y; \Theta) = \left(\frac{\sum_{i=1}^n p_i \hat{q}_i |x_i - \hat{y}_i|^r + \sum_{j=1}^n \hat{p}_j q_j |\hat{x}_j - y_j|^r}{\sum_{i=1}^n p_i \hat{q}_i + \sum_{j=1}^n \hat{p}_j q_j} \right)^{1/r}. \quad (11)$$

Здесь x_k, y_k , как и прежде, – заданные дискретные значения функций $x(t)$ и $y(t)$; $\hat{x}_k, \hat{y}_k$ – расчётные значения функций $x(t), y(t)$, определяемые методами интерполяции и экстраполяции; p_k, q_k – дискретные значения весовых коэффициентов, задаваемых соответственно для функций $x(t)$ и $y(t)$; $\hat{p}_k, \hat{q}_k$ – значения весовых коэффициентов, получающиеся при интерполяции и экстраполяции; $\Theta = \{s, m, J, E\}$ – вектор параметров, где s – параметр смещения, m – параметр масштаба, J – параметр, определяющий способ интерполяции, E – параметр, определяющий способ экстраполяции функций $x(t)$ и $y(t)$.

Несмотря на видимое усложнение, в (11) сохранена в основном структура функционала (9). Основные изменения произошли в связи с введением операций смещения и масштабирования. Так, единый набор весовых коэффициентов $\{w\}$ в данном случае заменён на два отдельных набора $\{p\}$ и $\{q\}$, что вызывается смещением точек x_i и y_i друг относительно друга. Взаимным смещением точек вызвано и разложение единых сумм в числителе и знаменателе (9) на две части в (11). Первая сумма в числителе (11), например, получается при просмотре всех заданных дискретных точек зависимости $x(t)$ и поиске соответствующих им расчётных точек зависимости $y(t)$; вторая сумма, наоборот, отыскивается в результате просмотра всех исходных точек зависимости $y(t)$ и поиска соответствующих им расчётных точек зависимости $x(t)$. Аналогично отыскиваются суммы, составляющие знаменатель (11).

Как известно из работ, посвящённых регрессионному анализу, вычисления, связанные с использованием функционалов рассматриваемого типа, бывают эффективными, если случайные отклонения, наблюдающиеся в суммируемых величинах, распределены независимо от значений исходных функций. Если такая зависимость существует, она обычно устраняется введением нормирования суммируемых величин. Последнее может осуществляться различными способами [25]. В созданной автором программной реализации метода анализа коллинеарности нормирование осуществляется делением каждой из суммируемых величин на модуль той из двух величин x, y (исходной или расчётной), которая принимает в данной точке максимальное значение. В соответствии с этим функционал (11) преобразуется:

$$\rho_r(x, y; \Theta) = \left\{ \frac{\sum_{i=1}^n p_i \hat{q}_i \left[\frac{|x_i - \hat{y}_i|}{\max(|x_i|, |\hat{y}_i|)} \right]^r + \sum_{j=1}^n \hat{p}_j q_j \left[\frac{|\hat{x}_j - y_j|}{\max(|\hat{x}_j|, |y_j|)} \right]^r}{\sum_{i=1}^n p_i \hat{q}_i + \sum_{j=1}^n \hat{p}_j q_j} \right\}^{1/r} \quad (12)$$

(обозначения те же, что и выше).

В конечном счете, решение задач анализа коллинеарности сводится к тому, чтобы при выбранных параметрах r, J, E путём варьирования переменных величин s (смещения) и m (масштабного коэффициента) добиться минимального рассогласования по форме между сопоставляемыми зависимостями. Иными словами, формально решается задача

$$\rho(x, y; \Theta) \Rightarrow \min \quad (13)$$

или, в развёрнутом виде,

$$\rho(x, y; s, m \mid J = J_0, E = E_0, r = r_0) \Rightarrow \min. \quad (14)$$

• – проекция глобального минимума функционала (14) на координатную плоскость (s, m)

Рисунок 1 – Вариация формы парной функции коллинеарности, обусловленная особенностями сравниваемых зависимостей

Рисунок 2 – Результат последовательного выполнения операций смещения и масштабирования: а) исходное положение зависимостей ($c=0$; $m=1$); б) относительное смещение ($c=-0,72$); в) масштабирование ($m=1,07$).

Здесь основными переменными величинами являются смещение s и коэффициент масштабирования m , которые изменяются хотя и дискретно, но, небольшими, постепенно уменьшающимися в процессе вычислительных итераций шагами. Параметры J, E, r также являются переменными величинами, наделёнными свойством приоритетности или “глобальности”. В отличие от двух предыдущих этим параметрам либо придаются априорно выбранные фиксированные значения, либо они автоматически изменяются в ходе вычислений, принимая ограниченное количество дискретных значений.

Вид некоторых двумерных функций $\rho_i(s, m)$, соответствующих различным парам сравниваемых зависимостей x, y при фиксированных значениях параметров J, E, r , представлен на Рис. 1. На Рис. 2 схематично показан процесс отыскания значений s и m , сводящих рассчитываемый коэффициент расхождения (неколлинеарности) между сопоставляемыми зависимостями $\rho_i(s, m)$, к минимуму.

Показанные на рисунке 2 фазы процесса поиска оптимального взаимного расположения зависимостей в значительной мере условны, поскольку фактически этот поиск ведётся посредством построения многомерной функции ρ , зависящей одновременно от всех её параметров, с завершающим расчётные процедуры нахождением глобального минимума этой многомерной функции.

Функционалом (12), как сказано выше, описывается многомерная функция, размерность которой, в соответствии с количеством использующихся параметров, равна пяти. Такая размерность оказывается слишком большой, чтобы функционалом в его полном виде можно было пользоваться для массовых текущих расчётов. Сокращения размерности функционала (и объёма вычислений) можно достичь, придавая некоторым из переменных параметров фиксированные значения. Целесообразнее всего это сделать по отношению к параметрам J, E , задающим способы интерполяции и экстраполяции сопоставляемых зависимостей, и метрике Минковского r , приближающееся к оптимальному конкретное значение которой может быть найдено на основе пробных расчётов или по результатам их аналогичных вариантов, выполненных ранее.

ОБСУЖДЕНИЕ СТАТИСТИЧЕСКОГО ЭКСПЕРИМЕНТА

Для определения оптимальных значений параметров, использующихся при решении задач методом анализа коллинеарности, на одном из этапов исследования была выполнена большая серия статистических расчётов, в качестве основного эмпирического материала которых использовались промежуточные результаты обширного межрегионального социологического исследования, охватившего разные группы населения. Эмпирическую базу анализа составили более 100 зависимостей, анализировавшихся по отдельности и в тематически сгруппированном виде. Статистическое сжатие результатов выполнялось как известными приёмами (построением диаграмм рассеяния, кумулят, нахождением медиан и квантилей), так и с помощью так называемой множественной функции коллинеарности, получаемой путём взвешенного векторного суммирования соответствующих парных функций.

Значения пяти параметров функций коллинеарности, фигурирующих в функционалах (12) – (14), задавались следующим образом.

1. Способ *интерполяции* (параметр J) во всех случаях был представлен одним вариантом – использовалась линейно-кусочная интерполяция. То есть, как это изображено на Рис. 2, одна из сравниваемых зависимостей просто растягивалась и смещалась по оси абсцисс для достижения её минимального расхождения с парной зависимостью, при этом ординаты узлов обеих линейно-кусочных функций оставались неизменными. Выбор такого способа связан с простотой его реализации и экономичностью в вычислительном отношении по сравнению с другими известными способами, например, аппроксимацией полиномами или сплайн-функциями.

В ряде опубликованных ранее работ показано, что несмотря на свою простоту кусочно-линейная интерполяция обладает довольно хорошими аппроксимирующими свойствами. Основные опасения, связанные с данным выбором, касались вида получающейся многомерной функции коллинеарности (Рис. 1). В частности, можно было ожидать появления на гиперповерхности функции коллинеарности значительных “шероховатостей”, вызываемых негладкостью аппроксимирующих линий.

Как показали многочисленные выполненные расчёты, дополнительная шероховатость функции коллинеарности, возникающая в связи с применяемым способом интерполяции, имеет не настолько выраженный характер, чтобы являться препятствием для решения основной процедурной задачи – нахождения координат c_0 и m_0 глобального минимума функционала (14).

2. Для *экстраполяции* зависимостей использовались три различных способа, выбор которых задаётся с помощью параметра E (рис. 3):

Рисунок 3 – Три способа экстраполяции зависимостей: а) продление конечного отрезка ($E=1$); б) экстраполяция из конечной точки параллельно линейному тренду ($E=2$); в) продление линейного тренда ($E=3$).

а) при $E=1$ происходит продление конечного (левого или правого) отрезка кусочно-линейной аппроксимации вплоть до необходимой длины; если при этом продлеваемый отрезок встречается с осью абсцисс, начиная с этой точки последующие значения функции считаются равными нулю;

б) при $E=2$ линия экстраполяции проводится через левую и правую конечные точки заданного участка функции; по своему направлению эта линия параллельна линейному тренду функции;

в) при $E=3$ экстраполируется общий линейный тренд функции.

Два из трёх перечисленных способов являются в некотором смысле экстремальными. В одном из них ($E=1$) безусловный приоритет отдан конечному участку линии, аппроксимирующей анализируемую функциональную зависимость. В другом ($E=3$), наоборот, значимость этого конечного участка практически игнорируется и приоритет в определении направления экстраполяции отдаётся общему линейному тренду зависимости. Третий способ ($E=2$) является компромиссом между двумя предыдущими и, как правило, именно он даёт наиболее приемлемые результаты.

3. Метрика Минковского r варьировалась в ходе статистических экспериментов дискретно, при этом диапазоны варьирования выбирались в каждом случае исходя из обнаруживаемых при предварительных расчётах приближённых значений, обеспечивающих наиболее устойчивые решения. Общий диапазон варьирования r лежал в пределах от 0,15 до 10.

4. Параметры смещения s и масштабирования m как это предусмотрено в программе для ЭВМ, изменялись автоматически в процессе решения задачи, обеспечивая отыскание глобального минимума функционала (14). В программном задании указывались диапазоны варьирования этих параметров и количества шагов для разбиения диапазонов. Последнее необходимо в связи с тем, что фактически в памяти ЭВМ при решении задачи анализа коллинеарности строится дискретный образ функционала (12), по которому с помощью особого алгоритма отыскивается глобальный минимум функции и соответствующие ему «координаты» – значения параметров s_0 и m_0 . На построенных ЭВМ 3D-проекциях (Рис. 1) эти координаты обозначены как S и M .

В связи с тем, что все параметры функционала (12) могут оказаться перекрёстно взаимосвязанными, задача отыскания их оптимальных значений должна решаться как многомерная, что связано с необходимостью одновременного варьирования всех изменяемых параметров. С учётом множества сопоставляемых зависимостей (которых, как сказано выше, было более ста) это потребовало выполнения более 15 тысяч расчётов.

При анализе и обобщении результатов использовались по отдельности и в сочетаниях различные приёмы. Например, по отношению к двумерным функциям коллинеарности, изображавшимся на выводимых из ЭВМ рисунках, применялся визуальный анализ. Различные графические методы применялись при получении обобщённых статистических характеристик; в этих же целях использовались и аналитические приёмы, в том числе выполнявшееся программным путём формирование множественных функций коллинеарности.

В целом проведённый статистический эксперимент продемонстрировал безусловную результативность метода анализа коллинеарности и одновременно с этим позволил усовершенствовать ряд применяемых при его реализации формализованных процедур.

ЗАКЛЮЧЕНИЕ

В статье описываются результаты исследования, основная постановочная и конструктивная часть которого выполнена более трёх десятилетий назад. Для объекта этого исследования было выбрано редко встречавшееся в отечественной научной публицистике название «анализ коллинеарности». Существенно более распространённый англоязычный термин *Collinearity Analysis* использовался в то

время, также как и сейчас [10], для обозначения совокупности процедур по выявлению и устранению взаимной линейной коррелированности (collinearity) между предикторами (компонентами регрессионных уравнений), мешающей осуществлению регрессионного анализа и снижающей его результативность.

За истекшие десятилетия интенсивность использования термина *анализ коллинеарности* мало изменилась, а термина *Collinearity Analysis* даже заметно возросла. При этом претензии к обоим терминам остались практически прежними. Во всяком случае, не вполне удачный выбор названия разрабатываемого математико-статистического метода в сложившихся обстоятельствах представляется очевидным, что побуждает к выбору в не слишком далёком будущем более адекватного и, главное, уникального названия.

Что касается существа описываемого в статье метода, то по своей предметной направленности и процедурному оснащению он содержит в себе элементы регрессионного, корреляционного и ковариационного анализов. Как отмечается в статье, по своей постановке описываемый метод наиболее близко соответствует вычислению взаимной корреляции (Cross-correlation). Вместе с тем предлагаемый метод содержит в себе ряд новаций, выгодно отличающих его от перечисленных математико-статистических методов. В их число входят.

1. Использование аппарата метрических пространств и являющейся их частью метрики Минковского. С помощью этих компонентов обеспечивается:

- варьирование в непрерывном широком диапазоне одного из ключевых параметров, определяющих способ суммирования регрессионных остатков (функциональных невязок) с возможностью получения в частных случаях методов наименьших модулей, наименьших квадратов, доминирования и других;
- возможность создания универсальных вычислительных процедур, совмещающих использование в качестве аргументов как дискретных переменных величин, так и непрерывных функциональных зависимостей;
- упрощение процедур интерполяции и экстраполяции, используемых при выполнении операций независимого масштабирования и взаимного сдвига анализируемых зависимостей.

2. Использование комбинированного вычислительного алгоритма, построенного на совмещении методов Монте-Карло и скорейшего градиентного спуска и обеспечивающего достаточно надёжное и относительно быстрое отыскание глобального минимума целевой функции, приводящее к решению задачи анализа коллинеарности независимо от конкретного значения, принимаемого варьируемой метрикой Минковского.

3. Расширение состава задаваемых априорно, вычисляемых и оптимизируемых параметров, что позволяет значительно диверсифицировать состав решаемых задач, переводя вычисления на новый качественный уровень.

В отличие, например, от кросскорреляционных вычислений, приводящих к нахождению оптимального взаимного сдвига анализируемых зависимостей, в описываемом методе дополнительно отыскивается оптимальное соотношение масштабов, обеспечивающих более точную взаимную подгонку зависимостей. Тем самым создаются условия для сопоставления зависимостей, различающихся, например, не только моментами начала и окончания протекающих во времени процессов, но и их динамикой.

Дополнительные условия для диверсификации вычислений обеспечиваются возможностью использования варьируемой метрики Минковского, весовых функций либо коэффициентов, параметров, определяющих выбор механизмов интерполяции и экстраполяции анализируемых зависимостей.

В качестве одного из перспективных направлений дальнейшей работы по совершенствованию предлагаемого метода может быть названа замена в нём принятой в настоящее время кусочно-линейной аппроксимации анализируемых зависимостей их более “гладкой” аппроксимацией, основанной, например, на использовании сплайнов или аналогичных им функций.

БИБЛИОГРАФИЯ

1. Cross-correlation. – URL: <https://en.wikipedia.org/wiki/Cross-correlation>
2. Moriasi D.N., et al. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations // Transactions of the ASABE 2007. Vol. 50(3). P. 885–900.
3. Schaeffli B., Gupta H.V. Do Nash values have value? – URL: http://www.hydro-econ.citg.tudelft.nl/fileadmin/Faculteit/CiTG/Over_de_faculteit/Afdelingen/Afdeling_watermanagement/Secties/waterhuishouding/Leerstoelen/Hydrologie/People/Schaeffli,_B./doc/Schaeffli_Gupta_2007.pdf
4. Математика и кибернетика в экономике. – М.: Экономика, 1975. С.36.
5. Гальтунг Й. Диахронная корреляция // Математика в социологии. – М.: Мир, 1977. С.9-49.
6. Харват Ф. Сближение контекстов и контексты сближения. – Прага, 1983. 90 с.
7. Волков В.А. Численные методы. – М.: Наука, 1987. С. 81.
8. Михненко О. К вопросу применения регрессионных моделей // Вестник статистики. 1984. №9. С.50-57.
9. Liu F., et al. Collinearity analysis of allotetraploid *Gossypium tomentosum* and *Gossypium darwinii* // Genetics and Molecular Research. 2016. Vol.15(3). – URL: <http://www.funpecrp.com.br/gmr/year2016/vol15-3/pdf/gmr8391.pdf>
10. Adkins L.C., Waters M.S., Hill R.C. Collinearity Diagnostics in gretl. – URL: http://www.learneconometrics.com/pdf/Collin/collin_gretl.pdf
11. Корн Г., Корн Т. Справочник по математике (для научных работников и инженеров). – М.: Наука, 1973. С.458.
12. Прилуков А.Н. Идеи толерантного регрессионного анализа и их реализация // Успехи современной науки и образования. 2017. № 1. Т.1. С.134-136.
13. Goyal S., Goyal G.K. Soft computing single hidden layer models for shelf life prediction of burfi // Russian Journal of Agricultural and Socio-Economic Sciences. 2012. No.5. P. 28-32.
14. Acquah H.D. A threshold cointegration analysis of asymmetric adjustments in the ghanian maize markets // Russian Journal of Agricultural and Socio-Economic Sciences. 2012. No.8. P. 21-25.
15. Oyakhilomen O., Omadachi U.O., Zibah R.G. Cocoa production – agricultural credit guarantee scheme fund nexus in Nigeria: a cointegration approach // Russian Journal of Agricultural and Socio-Economic Sciences. 2012. No.9. P. 28-32.
16. Jannati N.N., Sultana N., Rayhan M.I. Are the real GDP series in Asian countries nonstationary or nonlinear stationary? // Russian Journal of Agricultural and Socio-Economic Sciences. 2013. No.6. P. 8-14.
17. Wongnaa C.A., et al. Economics of tomato marketing in Ashanti region, Ghana // Russian Journal of Agricultural and Socio-Economic Sciences. 2014. No.2(26). P. 3-13.
18. Akpan S.B., Udoh E.J., Patrick I.V. Assessment of economic policy variables that modeled agricultural intensification in Nigeria // Russian Journal of Agricultural and Socio-Economic Sciences. 2015. No.5(41). P. 9-29.
19. Shinta A. The influence of technical inefficiency level that involve farmer's behaviour on risk towards profit in rice production of Indonesia // Russian Journal of Agricultural and Socio-Economic Sciences. 2016. No.10(58). P. 3-12.
20. Acquah S., Kendie S., Agyenim J.B. Determinants of rural farmers' decision to adapt to climate change in Ghana // Russian Journal of Agricultural and Socio-Economic Sciences. 2017. №2. P. 195-204.
21. Шрейдер Ю.А. Что такое расстояние? – М.: Физматгиз, 1963. С.18.
22. Roskam E. Metric analysis of ordinal data in psychology. – Voorschoten, 1968. P.47.
23. Прилуков А.Н. Толерантный регрессионный анализ и пример его программной реализации. Деп. в ВИНТИ. №8527-В88. 40 с.
24. Мудров В.И. и др. Методы обработки измерений. – М.: Радио и связь, 1983. 304 с.
25. Анализ нечисловой информации в социологических исследованиях. – М.: Наука, 1985. С.176–180.