Многомерный анализ данных вариаций интенсивности мюонов в атмосфере

В. Л. Янчуковский, А. Ю. Белинская

Институт нефтегазовой геологии и геофизики им. А. А. Трофимука СО РАН
Новосибирск, 630090, Россия
E-mails: YanchukovskiyVL@ipgg.sbras.ru, BelinskayaAY@ipgg.sbras.ru

Данные непрерывных наблюдений мюонных телескопов космических лучей подлежат коррекции на вариации атмосферного происхождения: барометрический и температурный эффекты. Температурный эффект интенсивности мюонов, в отличие от барометрического, определяется многими параметрами, характеризующими состояние атмосферы от слоя генерации до уровня регистрации мюонов (температура и распределение масс). Вариации температуры различных слоёв атмосферы коррелированы, поэтому применение методов многофакторной регрессии при оценке температурного эффекта для мюонов некорректно. При исследовании температурного эффекта мюонов в атмосфере проанализированы возможности методов регрессии на главные компоненты и метода проекций на скрытые структуры (ПЛС). Рассмотрены способы выбора оптимального значения числа главных компонент. С применением ПЛС-алгоритма выполнена оценка связи вариаций интенсивности мюонов и изменений температуры атмосферы на 16 изобарах.

Ключевые слова: космические лучи, атмосфера, мюоны, температурный эффект, метод регрессии на главные компоненты, метод проекций на скрытые структуры

Одобрена к печати: 24.04.2023
DOI: 10.21046/2070-7401-2023-20-3-301-306

Вариации интенсивности мюонов, наблюдаемые в глубине атмосферы, обусловлены не только изменениями энергетического спектра первичного потока космических лучей за пределами атмосферы, но также изменениями параметров самой атмосферы. Атмосферная составляющая вариации в основном представляет собой барометрический и температурный эффекты. Барометрический эффект для мюонов учитывается просто, так как зависит от одного параметра — атмосферного давления на уровне наблюдения. Температурный эффект для мюонов определяется многими параметрами, характеризующими состояние атмосферы от слоя генерации до уровня регистрации мюонов (температура и распределение масс). Для регулярного учёта температурного эффекта предварительно необходимо оценить величину воздействия этих параметров на интенсивность мюонов в атмосфере. Поскольку вариации температуры различных слоёв атмосферы коррелированы, применение методов многофакторной регрессии (МФР) при оценке распределения плотности температурных коэффициентов некорректно. Поэтому при исследовании температурного эффекта мюонов были использованы методы регрессии на главные компоненты (РГК) (Gorban et al., 2007; Jolliffe, 2002). При построении системы линейных уравнений (СЛУ) в пространстве главных компонент (ГК) привлекался метод проекций на скрытые (латентные) структуры (ПЛС1 и ПЛС2) (Померанцев, 2014; Эсбенсен, 2005).

Суть методов РГК и ПЛС заключается в построении пространства N неявных параметров, имеющих между собой нулевые коэффициенты корреляции, таким образом, что доли информативности исходных данных распределяются на каждом из них от большего к меньшему, что в сумме даёт 100 % информации исходной выборки. Для построения этого пространства необходимо такое ортогональное преобразование в новую систему координат, для которого были бы верны следующие условия:

- выборочная дисперсия данных вдоль первой координаты максимальна (эту координату называют первой главной компонентой);
- выборочная дисперсия данных вдоль второй координаты максимальна при условии ортогональности первой координате (вторая главная компонента);

- выборочная дисперсия данных вдоль значений k-й координаты максимальна при условии ортогональности первым $k-1$ координатам.

Основная задача метода главных компонент (МГК) (Айвазян и др., 1989) состоит в замене исходного описания n образцов с помощью p переменных на новую форму, представленную в пространстве ГК. ГК-модель представляется как $\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E}$. В методе ГК исходная матрица $\mathbf{X}$ разбивается на сумму произведений матриц $\mathbf{TP}^{\mathrm{T}}$ и матрицу остатков $\mathbf{E}$. Здесь $\mathbf{T}$ — это матрица счетов; $\mathbf{P}^{\mathrm{T}}$ — соответствующая матрица нагрузок (транспонированная). Надо найти матрицы $\mathbf{T}$ и $\mathbf{P}$, чтобы использовать их произведение вместо матрицы $\mathbf{X}$, освобождая исходную матрицу от ошибок $\mathbf{E}$. Структурная часть матрицы $\mathbf{X}$ заключена в матрице $\mathbf{TP}^{\mathrm{T}}$, а шум (остатки) — в матрице $\mathbf{E}$. При этом подходе выбор числа A (количества ГК) соответствует установлению границы между структурной частью и шумом. Из этого следует, что для оптимального моделирования необходим правильный (обоснованный) выбор величины A. При переходе в пространство ГК не только меняется система координат, но и убирается шум, который описывается старшими ГК. Одновременно с переводом данных в новую систему координат решается две задачи: уменьшение размерности (использование только первых ГК, отражающих структуру данных) и понижение шума. В результате возникает важный вопрос: какое число ГК надо использовать в данной задаче? Программа The Unscrambler X (http://www.camo.com/rt/Products/Unscrambler/unscrambler.html) позволяет считать методом ПЛС2 с использованием четырёх алгоритмов. Применение трёх из них было рассмотрено нами ранее (Кузьменко, Янчуковский, 2015). Четвёртый из алгоритмов — KERNEL PLS — лучше всего подходит для большого количества образцов (тысячи образцов с несколькими переменными) (Dayal et al., 1997; de Jong, Ter Braak, 1994; Lindgren et al., 1993), поэтому он и станет использоваться при расчётах в настоящей работе.

Были привлечены данные наблюдений интенсивности мюонов на уровне моря (https://ikfia.ysn.ru/data/hecrlab/mt и https://ikfia.ysn.ru/data/hecrlab/ipm) и аэрологические данные (температура атмосферы на 16 изобарах) (http://crsa.izmiran.ru/phpmyadmin) за период с января 2016 г. по декабрь 2018 г.

Исходные данные нормировались. Находились средние значения интенсивности мюонов для каждого из каналов наблюдений за весь рассматриваемый период, а затем — изменения (вариации) текущих значений интенсивности (в %) для каждого из каналов относительно соответствующих средних значений (данные y). Для аэрологических данных также вычислялись средние значения температуры атмосферы для каждой из 16 изобар, а затем относительно этих средних находились изменения температуры (в градусах) на каждой изобаре (данные x). Предобработка в основном заключается в некотором «обезразмеривании» данных. Предварительная нормировка нужна также для обоснованного выбора метрики, в которой будет вычисляться наилучшая аппроксимация данных. Как правило, основная доля вариаций исходной выборки сосредотачивается на первой координате. Вариации, связанные со следующим параметром, сосредотачиваются на второй координате и т. д. Построение подобного пространства неявных параметров производится с условием максимизации связи откликов и входных данных. На основе такого базиса несложно получить модифицированное представление исходной регрессии. После чего с использованием рассчитанных значений информативности для каждой новой координаты определяется оптимальная размерность данных, отвечающая заданной точности. На *рис. 1* (см. с. 303) представлены результаты расчёта дисперсии для выбранных векторов ГК, собственных значений векторов ГК и меры информативности преобразованных данных при увеличении числа ГК.

Проблема мультиколлинеарности и уменьшения размерности матрицы входных данных может быть снята переходом в пространство ГК. Анализ расчётных значений мер информативности вкладов каждой компоненты позволяет выполнить оценку числа главных компонент A, вариации исходных данных на которых содержат основную информацию. Выбор величины A также можно определить при рассмотрении дисперсии остатков. По мере вычисления ГК и вычитания полученных значений из матрицы $\mathbf{X}$ остатки меняются. Эти остатки сравниваются с $\mathbf{E}_0$ — начальной точкой в уравнении $\mathbf{X} = \mathbf{TP}^{\mathrm{l}} + \mathbf{E}$, где $\mathbf{E}_0$ — это $\mathbf{X}$. Остатки

удобно выражать в относительных единицах, используя $\mathbf{E}_0$. Для $A = 0$ имеем $\mathbf{E} = 100\ \%$ от $\mathbf{E}_0$. На *рис. 2* показано детальное рассмотрение графика остатков, помогающее найти оптимальное значение A.

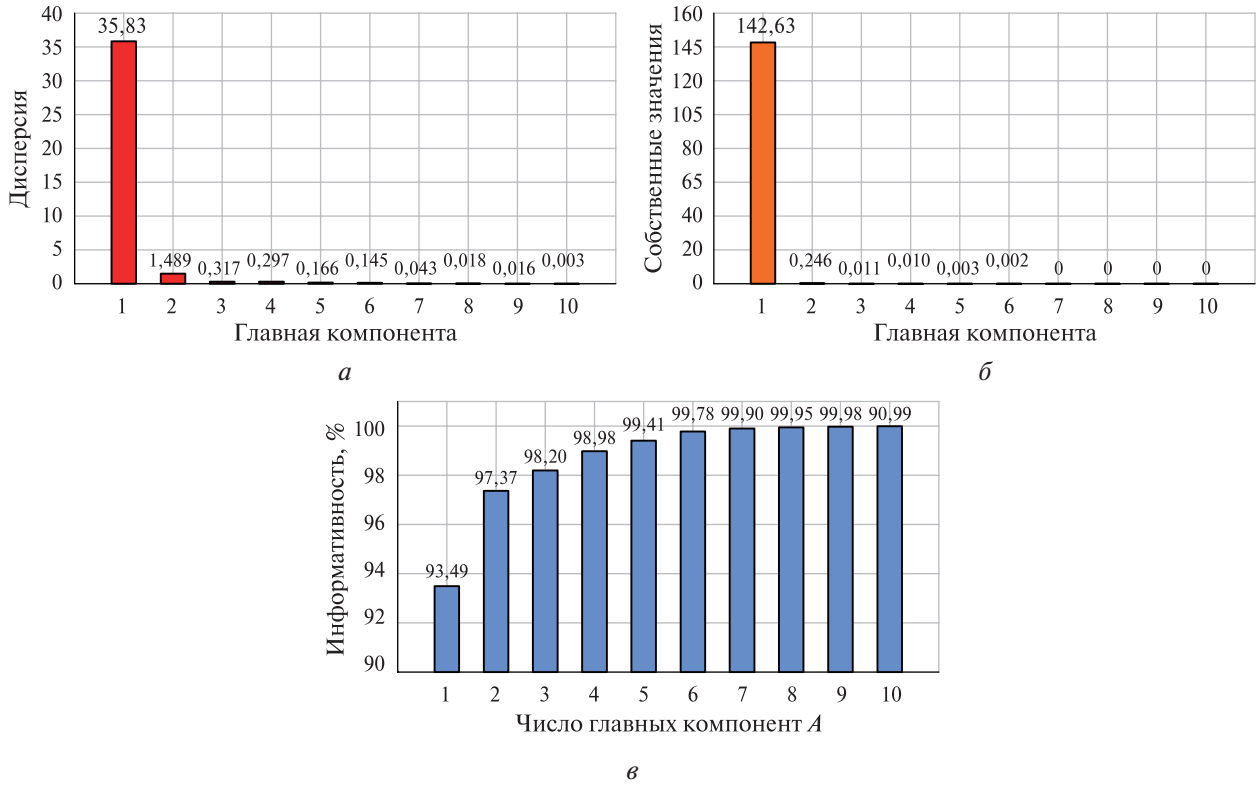

Рис. 1. Информативность преобразованных данных при увеличении числа ГК: *а* — дисперсия векторов ГК; *б* — собственные значения ГК; *в* — мера информативности преобразованных данных при увеличении числа ГК

Угол наклона (излома) кривой, описывающий дисперсию остатков, резко меняется при числе главных компонент A, равном 2. То есть для описания структурной части $\mathbf{TP}^{\mathrm{T}}$ достаточно двух ГК. При большем числе A, превышающем данное оптимальное значение, в основном будет вноситься шум старшими ГК. Поэтому выбор числа A превышающим оптимальное значение считается серьёзной ошибкой. При вовлечении данных y в декомпозицию $\mathbf{X}$ ПЛС-метод позволяет получить результаты прогноза на меньшем количестве ГК. ПЛС работает как два МГК-анализа, проводимых для $\mathbf{X}$ и для $\mathbf{Y}$: $\mathbf{X} = \sum_A \mathbf{TP}^{\mathrm{T}} + \mathbf{E}$ и $\mathbf{Y} = \sum_A \mathbf{UQ}^{\mathrm{T}} + \mathbf{F}$; $\mathbf{T}$ и $\mathbf{P}$ — счета и нагрузки, составляющие $\mathbf{X}$, а для $\mathbf{Y}$ они обозначены через $\mathbf{U}$ и $\mathbf{Q}$ соответственно. ПЛС-декомпозиция осуществляется не применением двух независимых операций МГК-анализа в двух разных пространствах, а с учётом тесной связи пространств $\mathbf{X}$ и $\mathbf{Y}$. Проекция строится согласованно — так, чтобы максимизировать корреляцию между соответствующими векторами $\mathbf{X}$-счетов $\mathbf{t}_a$ и $\mathbf{Y}$-счетов $\mathbf{u}_a$.

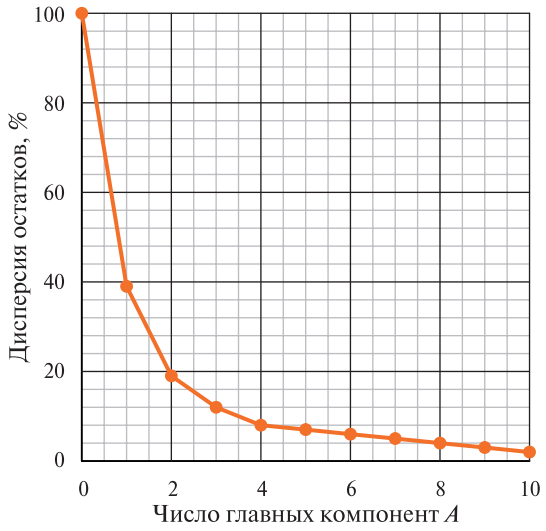

Рис. 2. Общая остаточная дисперсия

С целью тестирования метода ПЛС при оценке температурного эффекта интенсивности мюонов в атмосфере использовались результаты непрерывных наблюдений с помощью мюонного телескопа «Вертикаль» на уровне моря с направления вертикаль и данные аэрологического зондирования. При этом закладывалось разное число ГК. Результаты показаны на *рис. 3*.

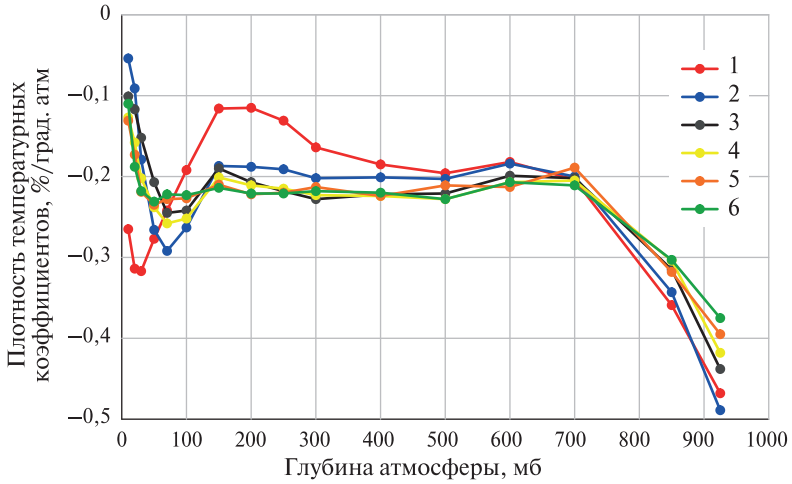

Рис. 3. Распределение плотности температурных коэффициентов для телескопа «Вертикаль», найденные при использовании различного числа ГК: 1 — для одной ГК; 2 — для двух ГК; 3 — для трёх ГК; 4 — для пяти ГК; 5 — для 16 ГК; 6 — результат, полученный методом МФР

При максимальном числе ГК, равном общему числу переменных (температура на 16 изобарах), результат (кривая 5) полностью совпадает с результатом, полученным методом многофакторной регрессии (кривая 6). Одна главная компонента содержит 93 % информации, которая при двух ГК увеличивается всего на 4 %, а при трёх ГК — на 0,8 % (см. *рис. 1*). При дальнейшем увеличении числа ГК информативность увеличивается незначительно, в то время как возрастает вероятность вклада шумов. При большом числе ГК метод ПЛС2 теряет свои преимущества перед методом МФР, поэтому полученный результат оказывается практически одинаков (см. *рис. 3*). Это также указывает, что число ГК в данной задаче выбрано верно — равным 2.

Показано, что проблема мультиколлинеарности и уменьшения размерности матрицы входных данных может быть снята переходом в пространство ГК. Отмечается важность выбора числа ГК при моделировании. Выбор недостаточного количества ГК в модели будет свидетельствовать, что информация, заложенная в данных, использована не полностью. Эта ошибка может быть учтена при последующем анализе. Однако при использовании большого числа ГК в модель будет включён шум, вклад которого приводит к ошибочной интерпретации, и анализ будет частично ошибочным. Анализ расчётных значений мер информативности вкладов каждой компоненты и рассмотрение дисперсии остатков позволяют выполнить оценку оптимального числа ГК A, вариации исходных данных на которых содержат основную информацию. Применение ПЛС-алгоритма обеспечивает моделирование **X-** и **Y-**пространств взаимозависимо, что позволяет оценить связь вариаций интенсивности мюонов и изменений температуры атмосферы на 16 изобарах.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (проект FWZZ-2022-0019). Результаты получены с использованием оборудования УНУ-85 «Российская национальная сеть станций космических лучей» (http://www.ckp-rf.ru/usu/433536).

Литература

1. *Айвазян С. А.*, *Бухштабер В. М.*, *Енюков И. С.*, *Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.
2. *Кузьменко В. С.*, *Янчуковский В. Л.* Определение плотности температурных коэффициентов для мюонов в атмосфере // Солнечно-земная физика. 2015. Т. 1. № 2. С. 91–96. DOI: 10.12737/10403.
3. *Померанцев А. Л.* Хемометрика в Excel: учеб. пособие. Томск: Изд-во ТПУ, 2014. 435 с.
4. *Эсбенсен К.* Анализ многомерных данных. Избранные главы / пер. с англ. С. В. Кучерявского; под ред. О. Е. Родионовой. Черноголовка: Изд-во ИПХФ РАН, 2005. 160 с.
5. *Dayal B. S.*, *McGregor J. F.* Improved PLS Algorithms // J. Chemometrics. 1997. V. 11. P. 73–65.
6. *De Jong S.*, *Ter Braak C.* Comments on the PLS kernel algorithm // J. Chemometrics. 1994. V. 8. P. 169–174.
7. *Gorban A. N.*, *Kegl B.*, *Wunsch D.*, *Zinovyev A. Y.* Principal Manifolds for Data Visualization and Dimension Reduction: Lecture Notes in Computational Science and Engineering. Berlin; Heidelberg; N. Y.: Springer, 2007. 340 p.
8. *Jolliffe I. T.* Principal Component Analysis. Series in Statistics. N. Y.: Springer, 2002. 487 p.
9. *Lindgren F.*, *Geladi P.*, *Wold S.* The kernel algorithm for PLS // J. Chemometrics. 1993. V. 7. P. 45–59.

Multivariate data analysis of variations of Earth's atmosphere muons

V. L. Yanchukovsky, A. Yu. Belinskaya

Trofimuk Institute of Petroleum Geology and Geophysics SB RAS
Novosibirsk 630090, Russia
E-mails: YanchukovskiyVL@ipgg.sbras.ru, BelinskayaAY@ipgg.sbras.ru

The data of continuous observations of muon telescopes of cosmic rays are subject to correction for variations of atmospheric origin: barometric and temperature effects. The temperature effect of muon intensity, unlike the barometric one, is determined by many parameters characterizing the state of the atmosphere from the generation layer to the muon registration level (temperature and mass distribution). Temperature variations of different layers of the atmosphere are correlated, so the use of multivariate regression methods in assessing the temperature effect for muons is not correct. The possibilities of regression methods on the main components (RGC) and the method of projections on hidden structures (PLC) in the study of the temperature effect of muons in the atmosphere are analyzed. The ways of choosing the optimal value of the number of principal components are considered. Using the PLC algorithm, the relationship between muon intensity variations and atmospheric temperature changes on 16 isobars was estimated.

Keywords: cosmic rays, atmosphere, muons, temperature effect, regression method on principal components, projection method on hidden structures

Accepted: 24.04.2023
DOI: 10.21046/2070-7401-2023-20-3-301-306

References

1. Aivazyan S. A., Bukhshtaber V. M., Enyukov I. S., Meshalkin L. D., *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti* (Applied statistics. Classification and dimensionality reduction), Moscow: Finansy i statistika, 1989, 607 p. (in Russian).
2. Kuzmenko V. S., Yanchukovsky V. L., Determination of density of temperature coefficients for the Earth's atmosphere muons, *Solar-Terrestrial Physics*, 2015, Vol. 1, Issue 2, pp. 91–96 (in Russian), DOI: 10.12737/10403.
3. Pomerantsev A. L., *Khemometrika v Excel: uchebnoe posobie* (Chemometrics in Excel: a tutorial), Tomsk: TSU Publ. House, 2014, 435 p. (in Russian).

4. Esbensen K. H., *Multivariate data analysis in practice*, Oslo: CAMO Process AS, 2002, 589 p.
5. Dayal B. S., McGregor J. F., Improved PLS Algorithms, *J. Chemometrics*, 1997, Vol. 11, pp. 73–65.
6. De Jong S., Ter Braak C., Comments on the PLS kernel algorithm, *J. Chemometrics*, 1994, Vol. 8, pp. 169–174.
7. Gorban A. N., Kegl B., Wunsch D., Zinovyev A. Y., *Principal Manifolds for Data Visualization and Dimension Reduction: Lecture Notes in Computational Science and Engineering*, Berlin; Heidelberg; New York: Springer, 2007, 340 p.
8. Jolliffe I. T., *Principal Component Analysis. Series in Statistics*, New York: Springer, 2002, 487 p.
9. Lindgren F., Geladi P., Wold S., The kernel algorithm for PLS, *J. Chemometrics*, 1993, Vol. 7, pp. 45–59.