Метод главных компонент и его применение

Работу выполнил
студент 4 курса 7 группы
Стец Вадим

Содержание метода главных компонент

Метод главных компонент в настоящее время представляет эффективный аппарат комплексного анализа геоданных, и его программное обеспечение является составной частью многих компьютерных технологий по обработке геофизической информации.

Математической моделью метода служит, как и для корреляционно-регрессионного анализа, система случайных величин $x_1, x_2, \ldots, x_N$. При этом каждая случайная величина обычно содержит

n – наблюдений, т.е. исходный массив геоданных представлен матрицей X размерностью $n * N$:

$$X = \begin{vmatrix} x_{11} x_{21} \ldots . x_{N1} \\ x_{12} x_{22} \ldots . x_{N2} \\ \ldots\ldots\ldots\ldots \\ x_{1n} x_{2n} \ldots . x_{Nn} \end{vmatrix}$$

В качестве числа N может быть число профилей съемки, и тогда с помощью главных компонент решается задача оценки регионального тренда (региональной составляющей). Если число N представлено совокупностью различных методов и атрибутов, то путем метода главных компонент решается задача комплексного анализа по разделению исследуемой территории на классы.

Значение n определяет число точек наблюдений по отдельным профилям съемки. Суть метода главных компонент состоит в переходе от системы случайных величин $x_1, \ldots, x_N$ к новой системе случайных величин $y_1, \ldots, y_N$, ориентируясь на поведение дисперсий yi . При этом главная компонента определяется как линейная комбинация исходных случайных величин xi ,

$$y_j = \sum_{i=1}^{N} a_{ij} x_i ; \; j = 1, \ldots, N \qquad (1.1)$$

причем первая главная компонента $y1$: $y_1 = \sum_{i=1}^{N} a_{i1} x_i$, обладает максимальной дисперсией среди всех возможных линейных комбинаций вида (1.1).

Величины a *ij* являются коэффициентами перехода от одной системы случайных величин *xi* к другой системе случайных величин *yi* . Дисперсии линейных комбинаций *yi* располагаются в убывающем порядке, т.е.: $\sigma^2(y_1) > \sigma^2(y_2) > \ldots > \sigma^2(y_N)$

Переход от системы величин *xi* к системе *yj* сопровождается нормировкой коэффициентов a*ij* в виде:

$$\sum_{i-1}^{N} a_{ij}^2 = 1$$

Математически метод главных компонент сводится либо к вычислению ковариационной матрицы *B* системы случайных величин *xi* , если все *xi* измерены в одних и тех же единицах, либо к вычислению корреляционной матрицы *R*, если случайные величины *xi* измерены в разных физических единицах.

Первый случай соответствует измерениям одного и того же поля по *N*-профилям съемки, **второй случай** соответствует измерениям *N*-разных полей и (или) их атрибутов. Далее для матрицы *B* или матрицы *R* находятся их собственные значения λ_1 и соответствующие этим собственным значениям собственные векторы, которыми являются коэффициенты перехода a *ij* .

Обычно ограничиваются вычислением первых двух-трех главных компонент, поскольку в этих компонентах сосредоточена основная энергия исходных данных.

Физическое истолкование главных компонент является весьма неоднозначным. Однако первая главная компонента практически всегда имеет однозначное истолкование, поскольку ее дисперсия отражает основную энергию поля при обработке данных по площади или энергию нескольких полей и (или) их атрибутов при комплексном анализе данных.

Метод главных компонент при оценке региональной составляющей поля

Выделение регионального тренда является распространенной процедурой обработки практически для всех методов геофизики. Однако при решении этой задачи приходится задавать те или иные параметры. Так, при осреднении поля в скользящем окне надо задать размеры окна, при пересчете поля на высоту надо задать высоту пересчета, при оценке региональной составляющей путем регрессии надо задать степень полинома регрессии и т.д.

Метод главных компонент не требует задания той или иной априорной информации. Единственное предположение, при котором происходит применение метода главных компонент, состоит в том, что региональная составляющая обладает наибольшей дисперсией по сравнению с локальными составляющими, что обычно на практике выполняется.

Алгоритм оценки региональной составляющей на основе метода главных компонент сводится к реализации следующих процедур:

1.Вычисление средних значений поля по каждому профилю $\overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$, где n – число точек наблюдений, i=1,.....N, N – число профилей, и ковариаций данных различных пар профилей :

$$b_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{ki} - \overline{x}_j), \quad i,j=1,\ldots.N.$$

Поскольку система $x1,\ldots.x\,N$ представлена измерениями одного и того же поля, нет необходимости рассчитывать коэффициенты корреляции, которые используются при комплексном анализе данных, измеренных в разных физических единицах.

1) Вычисление средних значений поля по каждому профилю $\overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$, где n – число точек

наблюдений, i=1,…….N, N – число профилей, и ковариаций данных различных пар профилей :

$$b_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{ki} - \overline{x}_j), \quad i,j=1,\ldots.N.$$

Поскольку система $x1$,….$x\ N$ представлена измерениями одного и того же поля, нет необходимости рассчитывать коэффициенты корреляции, которые используются при комплексном анализе данных, измеренных в разных физических единицах.

2) Составление ковариационной матрицы исходных данных по их коэффициентам ковариаций bij :

$$B = \begin{vmatrix} b_{11} & b_{12} & \dots & b_{1N} \\ b_{21} & b_{22} & \dots & b_{2N} \\ \dots & \dots & \dots & \dots \\ b_{N1} & b_{N2} & \dots & b_{NN} \end{vmatrix}$$

Матрица *B* симметрична относительно главной диагонали, т.е. $b_{ij} = b_{ji}$, а по диагонали расположены дисперсии значений поля каждого профиля.

3) Нахождение максимального собственного значения λmax из уравнения:

$$\left|B-\lambda_{\max}I\right|=0 \quad \text{или} \quad \left|B-\lambda_{\max}I\right|=\begin{vmatrix} b_{11}-\lambda_{\max} & b_{12} & \cdots & b_{1N} \\ b_{21} & b_{22}-\lambda_{\max} & \cdots & b_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ b_{N1} & b_{N2} & \cdots & b_{NN}-\lambda_{\max} \end{vmatrix}=0$$

т.е. после раскрытия определителя из этого уравнения достаточно найти его корень с максимальным значением λmax.

- 4) Вычисление значений собственного вектора матрицы (B-λmax*I), соответствующего максимальному собственному значению λmax из системы линейных уравнений

$$(B-\lambda_{\max}I)\vec{a}_1=0 \text{ или } \begin{bmatrix} b_{11}-\lambda_{\max} & b_{12} & \dots & b_{1N} \\ b_{21} & b_{22}-\lambda_{\max} & \dots & b_{2N} \\ \dots & \dots & \dots & \dots \\ b_{N1} & b_{N2} & \dots & b_{NN}-\lambda_{\max} \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ . \\ a_{N1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix}$$

Значения собственного вектора $\vec{a}_1$ определяются с учетом нормировки $\sum a_{i1}^2 = 1$

Физический смысл этой нормировки состоит в том, чтобы преобразованные данные, т.е. значения региональной составляющей, не отличались бы по масштабу от исходных значений поля, а физический смысл значений a_{i1} заключается в определении весовых коэффициентов для каждого профиля.

5) Нахождение значений первой главной компоненты $y_1 = \sum_{i=1}^{N} a_{i1} x_i$, то есть

$$y_{1k} = (a_{11}, a_{21}, \ldots, a_{N1}) \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ x_{N1} & x_{N2} & \ldots & x_{Nn} \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{21} \\ . \\ . \\ y_{1n} \end{bmatrix}$$

Физический смысл значений первой главной компоненты $y_{1k}(k_i = 1, \ldots, n)$ состоит в том, что они определяют весовые коэффициенты для каждого пикета исходных данных, аналогично тому, как значения $a_{i1}(i = 1, \ldots, N)$ определяют весовые коэффициенты для каждого профиля съемки.

6) Оценка региональной составляющей исходного поля характеризующейся y_{1i} наибольшей дисперсией. Эта оценка равна произведению вектора-столбца на вектор-строку a_{i1} с добавлением к каждому элементу образующейся матрицы среднего значения поля по профилю $\overline{x}_i$, то есть

$$x_{ki}^{рег} = \begin{bmatrix} y_{11} \\ y_{21} \\ . \\ . \\ y_{1n} \end{bmatrix} (a_{11}, a_{21}, \ldots, a_{N1}) + \overline{x}_1 = \begin{bmatrix} y_{11}a_{11}+\overline{x}_1 & y_{11}a_{21}+\overline{x}_2 & \ldots y_{11}a_{N1}+\overline{x}_N \\ y_{21}a_{11}+\overline{x}_1 & y_{12}a_{21}+\overline{x}_2 & \ldots y_{12}a_{N1}+\overline{x}_N \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ y_{N1}a_{11}+\overline{x}_1 & y_{1n}a_{21}+\overline{x}_2 & \ldots y_{1n}a_{N1}+\overline{x}_N \end{bmatrix}$$

Поскольку значения $x_{ki}^{рег}$ представляют оценку региональной составляющей, то разность $x_{ki}^{ост} = x_{ki} - x_{ki}^{рег}$ оценивает поле локальных составляющих.

Эффективность метода главных компонент иллюстрируется на (рис.3.1) , на котором приведены исходное поле силы тяжести (а), оценка региональной составляющей (б) и локальная составляющая (в). Пунктиром на рисунке показана область рудного объекта.

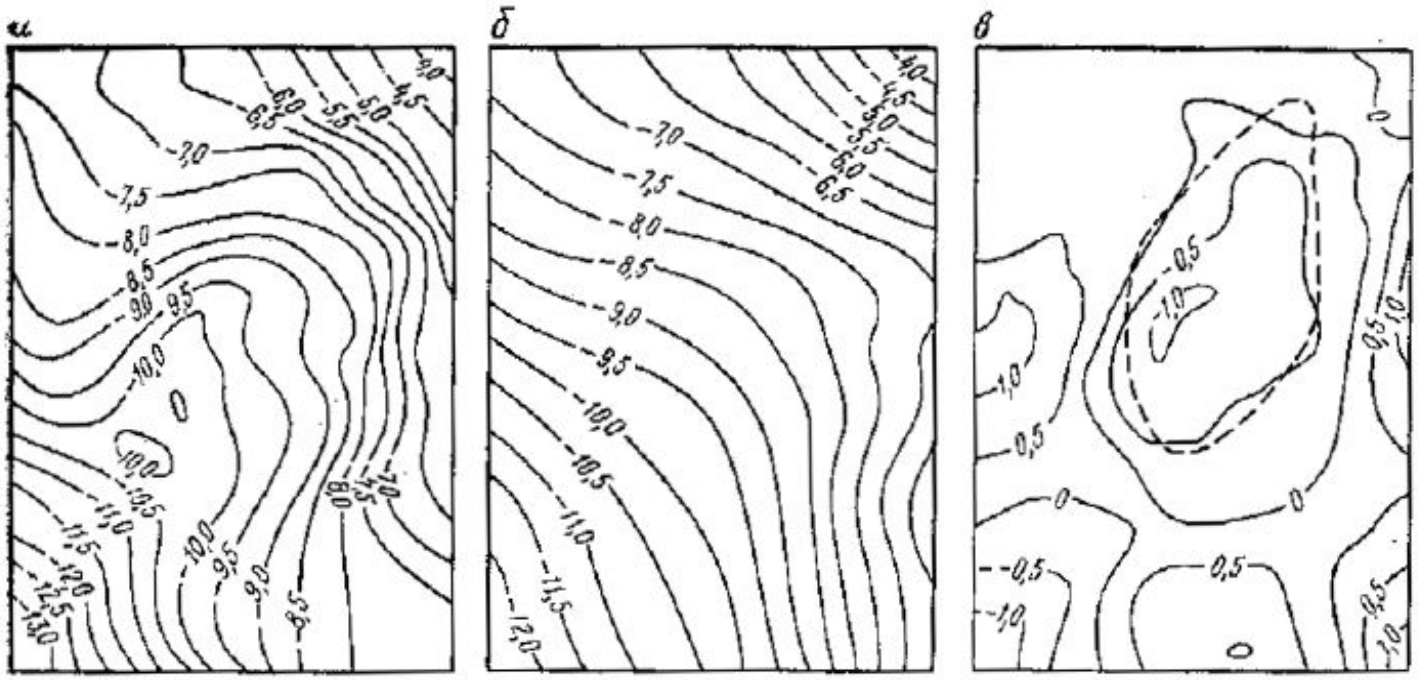

Рис.3.1.Исходное поле силы тяжести (а), оценка региональной составляющей (б) и локальной составляющей (в) с использованием метода главных компонент.

Метод главных компонент эффективен при обработке данных на достаточно ограниченных площадях, поскольку не учитывается изменение корреляционных свойств, т.е. структуры корреляционных матриц, по площади.

Следует отметить эффективность применения метода главных компонент при решении задач интерполяции. При этом задача интерполяции физического поля, представленного в виде функции двух переменных *x* и *y*, сводится к интерполяции функций, зависящих от одного аргумента.

Однако чаще всего метод главных компонент используется при решении задач комплексного анализа данных.

Ну всё, чё