

# О способе построения динамически сопоставимых КОМПОЗИТНЫХ ИНДЕКСОВ

Борzych Д. А., Фурманов К. К., Чернышева И. К.

13 апреля 2017 г.

Предлагаемый вниманию доклад основан на результатах статьи [2, Борзых, Фурманов, Чернышева, 2016].

В статье предложен новый способ построения динамических композитных индексов, обладающих двумя важными свойствами:

- **свойством неотрицательности весов,**
- **свойством динамической сопоставимости.**

Апробирование методики проведено на примере построения **композитного индекса здоровья населения.**

Проведено сопоставление с существующими методами построения композитных индексов:

- с методом, предложенным в [1, Борзых, 2016],
- с **методом главных компонент (МГК),**
- с **модифицированным методом главных компонент (ММГК)** (см. [4, Айвазян, 2003]).

Обозначим через  $X_{i,t}^{(j)}$  показатель с номером  $j \in \{1, \dots, k\}$ , относящийся к объекту  $i \in \{1, \dots, n\}$  в момент времени  $t \in \{1, \dots, T\}$ .

Будем предполагать, что для каждого  $j \in \{1, \dots, k\}$  выполнены следующие условия:

- 1)  $\text{med}_{i,t} X_{i,t}^{(j)} > 0$ ,
- 2)  $\min_{i,t} X_{i,t}^{(j)} \geq 0$ ,
- 3)  $\min_{i,t} X_{i,t}^{(j)} < \max_{i,t} X_{i,t}^{(j)}$ ,

где через  $\text{med}_{i,t} X_{i,t}^{(j)}$  обозначена выборочная медиана массива данных

$$\{X_{i,t}^{(j)}\}_{i \in \{1, \dots, n\}, t \in \{1, \dots, T\}}.$$

Определим *композитный индекс*  $I_{i,t}$ , относящийся к объекту  $i$  в момент времени  $t$ , по формуле:

$$I_{i,t} = w^{(1)}Z_{i,t}^{(1)} + \dots + w^{(k)}Z_{i,t}^{(k)}, \quad (1)$$

где веса  $w^{(1)}, \dots, w^{(k)}$  вычисляются как отношения

$$w^{(j)} = \frac{v^{(j)}}{v^{(1)} + \dots + v^{(k)}}, \quad j = 1, \dots, k, \quad (2)$$

а величины  $v^{(1)}, \dots, v^{(k)}$  рассчитываются по формулам

$$v^{(j)} = \frac{\text{med}_{i,t} |X_{i,t}^{(j)} - \text{med}_{i,t} X_{i,t}^{(j)}|}{\text{med}_{i,t} X_{i,t}^{(j)}}, \quad j = 1, \dots, k. \quad (3)$$

При этом *нормализованные показатели*  $Z_{i,t}^{(1)}, \dots, Z_{i,t}^{(k)}$ , участвующие в формуле (1), определяются следующим образом:

$$Z_{i,t}^{(j)} := \begin{cases} \frac{X_{i,t}^{(j)} - \min_{i,t} X_{i,t}^{(j)}}{\max_{i,t} X_{i,t}^{(j)} - \min_{i,t} X_{i,t}^{(j)}}, & \text{если (A),} \\ \frac{\max_{i,t} X_{i,t}^{(j)} - X_{i,t}^{(j)}}{\max_{i,t} X_{i,t}^{(j)} - \min_{i,t} X_{i,t}^{(j)}}, & \text{если (B),} \end{cases} \quad (4)$$

(A): чем больше показатель  $X^{(j)}$ , тем лучше,

(B): чем меньше показатель  $X^{(j)}$ , тем лучше.

Если композитный индекс  $I$  является *многоярусным*, т. е. строится на основе подындексов  $I^{(1)}, \dots, I^{(k)}$ , то для его вычисления применяется та же процедура, с тем лишь отличием, что в качестве показателей  $X^{(1)}, \dots, X^{(k)}$  теперь используются подындексы.

Из формул (1) и (3) видно, что *веса*  $w^{(1)}, \dots, w^{(k)}$  *всегда неотрицательны.*

Заметим, что величина

$$\frac{\text{med}_{i,t} |X_{i,t}^{(j)} - \text{med}_{i,t} X_{i,t}^{(j)}|}{\text{med}_{i,t} X_{i,t}^{(j)}},$$

участвующая в формуле (3), представляет собой робастный (т. е. устойчивый к выбросам) аналог коэффициента вариации, который определяется как отношение стандартного отклонения случайной величины и её математического ожидания  $\sqrt{D(X)}/\mathbb{E}[X]$ .



Из формулы (1) и сказанного выше следует, что при построении композитных индексов мы руководствуемся *принципом максимума вариации: чем выше изменчивость показателя, тем бóльшую значимость мы придаем этому показателю в индексе.*

Этот подход основан на следующем соображении: *если значение некоторого показателя для всех объектов одно и то же, то по этому показателю мы не сможем предпочесть ни один из объектов по отношению к другим.* Аналогичная логика применима и в случае, если рассматриваемый показатель не является константой, но меняется достаточно слабо.

Между нашим подходом и методом главных компонент имеется существенное идейное сходство: ***так же, как и метод главных компонент, наш подход основан на максимизации изменчивости некоторых переменных.*** Однако в техническом плане реализации предлагаемого нами способа и метода главных компонент совершенно разные. Отсюда и столь существенные различия в свойствах получаемых композитных индексов:

- ***композитный индекс, построенный при помощи нашего подхода, обладает свойством неотрицательности весов и свойством динамической сопоставимости,***
- ***композитный индекс, построенный при помощи метода главных компонент, не обладает ни свойством неотрицательности весов, ни свойством динамической сопоставимости.***

## Определение динамической сопоставимости.

Пусть индекс  $l_{i,t}$  относящийся к объекту  $i$  в момент времени  $t$  вычисляется на основе показателей

$$X^{(1)} = (X_{i,t}^{(1)})_{i=1,\dots,n}^{t=1,\dots,T}, \dots, X^{(k)} = (X_{i,t}^{(k)})_{i=1,\dots,n}^{t=1,\dots,T}.$$

Назовем индекс  $l$  **динамически сопоставимым**, если выполнено условие: для любого  $i \in \{1, \dots, n\}$  и любых  $s, t \in \{1, \dots, T\}$ :

$$\left( \forall j \in \{1, \dots, k\}: \delta^{(j)} \cdot X_{i,s}^{(j)} \leq \delta^{(j)} \cdot X_{i,t}^{(j)} \right) \implies l_{i,s} \leq l_{i,t},$$

где

$$\delta^{(j)} := \begin{cases} 1, & \text{если чем больше показатель } X^{(j)}, \text{ тем лучше,} \\ -1, & \text{если чем меньше показатель } X^{(j)}, \text{ тем лучше.} \end{cases}$$

Несложно видеть, что в силу монотонности функций

$$f(x) = \frac{x - \min_{i,t} X_{i,t}^{(j)}}{\max_{i,t} X_{i,t}^{(j)} - \min_{i,t} X_{i,t}^{(j)}} \quad \text{и} \quad g(x) = \frac{\max_{i,t} X_{i,t}^{(j)} - x}{\max_{i,t} X_{i,t}^{(j)} - \min_{i,t} X_{i,t}^{(j)}}$$

определяемый формулами (1)–(4) композитный индекс  $I$  обладает свойством динамической сопоставимости.

# Построение динамически сопоставимого композитного индекса здоровья населения — 1

Ниже приведены результаты апробирования описанной выше методики.

В качестве примера рассмотрена задача построения композитного индекса здоровья населения по региональной статистике РФ за 2005–2013 гг.

Источник данных — статистический бюллетень [3, Регионы России. Социально-экономические показатели. 2013].

Подробности см. [2, Борзых, Фурманов, Чернышева, 2016].

*Композитный индекс здоровья населения* рассчитывается на основе двух подындексов:

- *индекс заболеваемости,*
- *индекс продолжительности жизни.*

# Построение динамически сопоставимого композитного индекса здоровья населения — 2

*Индекс заболеваемости* рассчитывается на основе 15 показателей:

- некоторые инфекционные и паразитарные болезни,
- новые образования,
- болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм,
- болезни эндокринной системы, расстройства питания и нарушения обмена веществ,
- болезни нервной системы,
- болезни глаза и его придаточного аппарата,
- болезни уха и сосцевидного отростка,
- болезни системы кровообращения,
- болезни органов дыхания,
- болезни органов пищеварения,
- болезни кожи и подкожной клетчатки,

# Построение динамически сопоставимого композитного индекса здоровья населения — 3

- болезни костно-мышечной системы и соединительной ткани,
- болезни мочеполовой системы,
- врожденные аномалии (пороки развития), деформации и хромосомные нарушения,
- травмы, отравления и некоторые другие последствия воздействия внешних причин.

*Индекс продолжительности жизни* рассчитывается на основе 15 показателей:

- ожидаемая продолжительность жизни мужчин,
- ожидаемая продолжительность жизни женщин.

# Построение динамически сопоставимого композитного индекса здоровья населения — 4

## Веса нормализованных показателей в индексе уровня заболеваемости населения:

- (6%) некоторые инфекционные и паразитарные болезни,
- (5%) новые образования,
- (11%) болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм,
- (7%) болезни эндокринной системы, расстройства питания и нарушения обмена веществ,
- (8%) болезни нервной системы,
- (6%) болезни глаза и его придаточного аппарата,
- (5%) болезни уха и сосцевидного отростка,
- (6%) болезни системы кровообращения,
- (4%) болезни органов дыхания,
- (8%) болезни органов пищеварения,
- (5%) болезни кожи и подкожной клетчатки,



# Построение динамически сопоставимого композитного индекса здоровья населения — 5

- (6%) болезни костно-мышечной системы и соединительной ткани,
- (6%) болезни мочеполовой системы,
- (12%) врожденные аномалии (пороки развития), деформации и хромосомные нарушения,
- (5%) травмы, отравления и некоторые другие последствия воздействия внешних причин.

**Весы нормализованных показателей в индексе продолжительности жизни населения:**

- (62%) ожидаемая продолжительность жизни мужчин,
- (38%) ожидаемая продолжительность жизни женщин.

**Весы подындексов в композитном индексе здоровья населения:**

- (33%) индекс заболеваемости населения,
- (67%) индекс продолжительности жизни.

# Построение динамически сопоставимого композитного индекса здоровья населения — 6

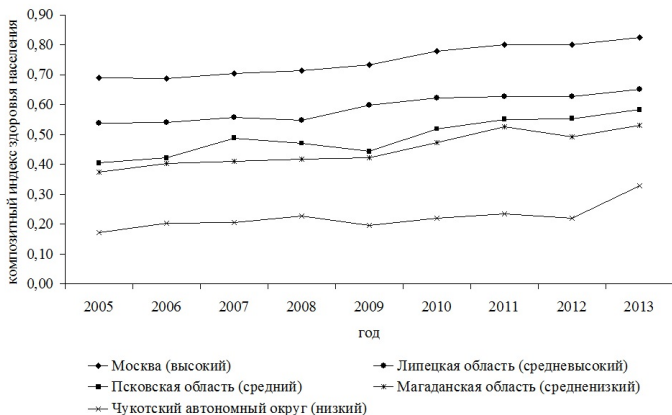


Рис.: Регионы с высоким, средневысоким, средним, средненизким и низким уровнем здоровья населения.

Ниже проводится сопоставление «нового» метода расчета композитных индексов из работы [2, Борзых, Фурманов, Чернышева, 2016] и «старого» метода из статьи [1, Борзых, 2016].

Будет показано, что «старый» способ, в отличие от «нового» не обладает свойством динамической сопоставимости.

## Сопоставление с методом из [1] — 2

В работе [1] композитный индекс  $I_{i,t}$ , относящийся к объекту  $i$  в момент времени  $t$ , определялся по формуле:

$$I_{i,t} = w_t^{(1)} Z_{i,t}^{(1)} + \dots + w_t^{(k)} Z_{i,t}^{(k)}, \quad (5)$$

где веса  $w_t^{(1)}, \dots, w_t^{(k)}$  находились как отношения

$$w_t^{(j)} = \frac{v_t^{(j)}}{v_t^{(1)} + \dots + v_t^{(k)}}, \quad j = 1, \dots, k, \quad (6)$$

а величины  $v_t^{(1)}, \dots, v_t^{(k)}$  рассчитываются по формулам

$$v_t^{(j)} = \frac{\operatorname{med}_i |X_{i,t}^{(j)} - \operatorname{med}_i X_{i,t}^{(j)}|}{\operatorname{med}_i X_{i,t}^{(j)}}, \quad j = 1, \dots, k. \quad (7)$$

При этом нормализованные показатели  $Z_{i,t}^{(1)}, \dots, Z_{i,t}^{(k)}$ , участвующие в формуле (5), определялись как

$$Z_{i,t}^{(j)} := \begin{cases} \frac{X_{i,t}^{(j)} - \min_i X_{i,t}^{(j)}}{\max_i X_{i,t}^{(j)} - \min_i X_{i,t}^{(j)}}, & \text{если (A),} \\ \frac{\max_i X_{i,t}^{(j)} - X_{i,t}^{(j)}}{\max_i X_{i,t}^{(j)} - \min_i X_{i,t}^{(j)}}, & \text{если (B),} \end{cases} \quad (8)$$

(A): чем больше показатель  $X^{(j)}$ , тем лучше,

(B): чем меньше показатель  $X^{(j)}$ , тем лучше.

## Сопоставление с методом из [1] — 4

Сравнивая формулы (4) и (8), видим, что оптимизация в формуле (4) производится как по  $i$ , так и по  $t$ , в то время как в формуле (8) — только по  $i$  (при фиксированном значении  $t$ ). Такая модификация позволяет сделать сопоставимыми значения нормализованных показателей в разные моменты времени.

Теперь обратимся к формулам (3) и (7). Как видно, в формуле (3) медиана находится по всему массиву данных  $\{X_{i,t}^j\}_{i=1,\dots,n}^{t=1,\dots,T}$ , в то время как в формуле (7) она находится по массиву  $\{X_{i,t}^j\}_{i=1,\dots,n}$  (при каждом отдельном значении  $t$ ). В результате, при «новом» способе расчета весов  $w^{(1)}, \dots, w^{(k)}$  с помощью формул (2) и (3) веса  $w^{(1)}, \dots, w^{(k)}$  перестают быть переменными во времени. По-видимому, отказ от концепции переменных весов является той необходимой «платой», чтобы получить свойство динамической сопоставимости.

С целью сопоставления «старого» и «нового» способов расчета композитных индексов рассмотрим следующий

### Искусственный пример 1.

Будем строить композитный индекс продолжительности жизни  $I$  на основе двух показателей: ожидаемой продолжительности жизни мужчин (показатель  $X^{(1)}$ ) и ожидаемой продолжительности жизни женщин (показатель  $X^{(2)}$ ).

## Искусственный пример 1 (продолжение).

Предположим, что в 2005 г. ожидаемая продолжительность жизни как мужчин, так и женщин в нашем гипотетическом примере совпадает с той реальной ожидаемой продолжительностью жизни, которая была в регионах РФ в 2005 г. В рамках нашего примера предположим также, что в последующие годы с 2006 по 2013 г. ожидаемая продолжительность жизни мужчин с каждым годом росла на один год, а ожидаемая продолжительность жизни женщин за тот же период не менялась, оставаясь на уровне 2005 г. Таким образом, мы имеем следующую динамику показателей  $X^{(1)}$  и  $X^{(2)}$ :

$$\begin{cases} X_{i,t}^{(1)} = X_{i,2005}^{(1)} + (t - 2005), \\ X_{i,t}^{(2)} = X_{i,2005}^{(2)}, \end{cases} \quad i = 1, \dots, 79, \quad t = 2006, \dots, 2013.$$



## Искусственный пример 1 (продолжение).

Используя в качестве данных искусственно созданные показатели  $X^{(1)}$  и  $X^{(2)}$ , мы рассчитали композитный индекс продолжительности жизни «старым» и «новым» способами.

В результате, для каждого из регионов композитный индекс продолжительности жизни, рассчитанный «новым» способом, является строго возрастающей функцией по переменной  $t$ .

Что касается «старого» способа расчета, то, например, для Чукотского автономного округа имеет место хотя и незначительное, но строгое снижение композитного индекса продолжительности жизни (см. рис. ниже).

## Искусственный пример 1 (продолжение).

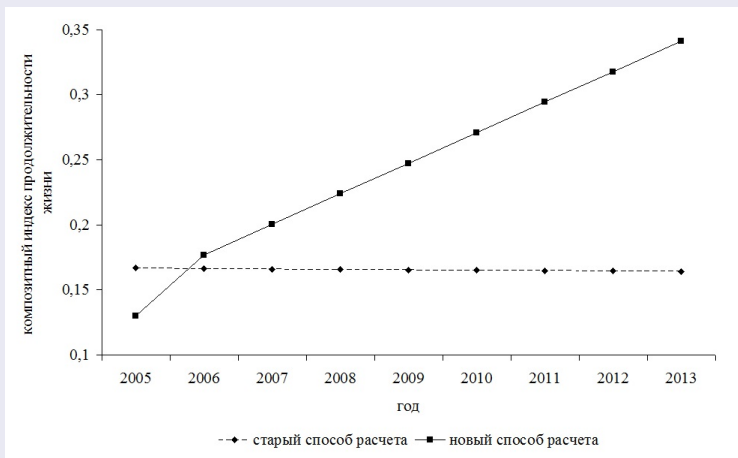


Рис.: Гипотетическая динамика композитного индекса продолжительности жизни для Чукотского автономного округа.

## Искусственный пример 1 (продолжение).

Мы получили, что для Чукотского автономного округа, имеют место неравенства  $X_{i,2005}^1 \leq X_{i,2013}^1$  и  $X_{i,2005}^2 \leq X_{i,2013}^2$ , но при этом для «старого» способа расчета индексов  $I_{i,2005} > I_{i,2013}$ . Стало быть, «старый» способ расчета композитных индексов не обладает свойством динамической сопоставимости.

*Первой главной компонентой* показателей  $Z^{(1)}, \dots, Z^{(k)}$  называется их линейная комбинация

$$I_{\text{МГК}} = v^{(1)}Z^{(1)} + \dots + v^{(k)}Z^{(k)} \quad (9)$$

с наибольшей дисперсией, коэффициенты в которой определяются как решение следующей экстремальной задачи

$$\begin{cases} D(v^{(1)}Z^{(1)} + \dots + v^{(k)}Z^{(k)}) \rightarrow \max_{v^{(1)}, \dots, v^{(k)} \in \mathbb{R}}, \\ (v^{(1)})^2 + \dots + (v^{(k)})^2 = 1. \end{cases} \quad (10)$$

Показатели  $Z^{(1)}, \dots, Z^{(k)}$  в формулах (9) и (10) предполагаются уже нормализованными.

Два распространённых способа нормировки:

- «стандартизация» — вычитание среднего и деление на стандартное отклонение (по умолчанию используется в статистических пакетах),
- сведение к единой шкале с заданными наибольшим и наименьшим значением (методология см. [4, Айвазян, 2003]).

Далее первый способ будем называть «МГК с корреляционной матрицей», а второй — «МГК с ковариационной матрицей».

- МГК с корреляционной матрицей вообще не учитывает разброс показателей — веса определяются только корреляциями.
- МГК с ковариационной матрицей измеряет разброс как концентрацию распределения у крайних значений (см. рис. ниже).

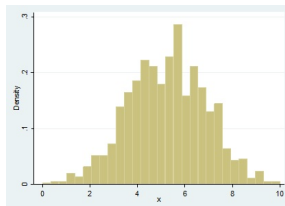


Рис.: Низкий разброс.

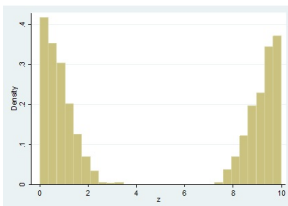


Рис.: Высокий разброс.

## Искусственный пример 2.

Рассмотрим динамику двух показателей  $Z^{(1)}$  и  $Z^{(2)}$  за семь лет:

$t$	$Z^{(1)}$	$Z^{(2)}$
1	1	0
2	0.71	0.16
3	0.42	0.33
4	0.14	0.49
5	0.69	0.8
6	0.28	0.83
7	0	1



## Искусственный пример 2 (продолжение).

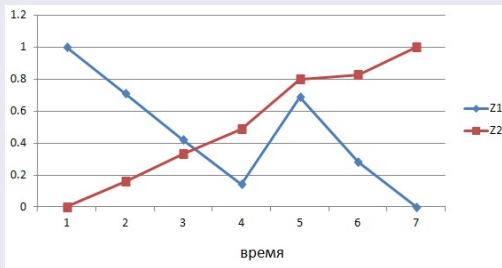


Рис.: Динамика показателей  $Z^{(1)}$  и  $Z^{(2)}$ .

Первая главная компонента:

$$I_{\text{МГК, cov}} = -0.68Z^{(1)} + 0.73Z^{(2)},$$

$$I_{\text{МГК, corr}} = -0.71Z^{(1)} + 0.71Z^{(2)}.$$

## Искусственный пример 2 (продолжение).

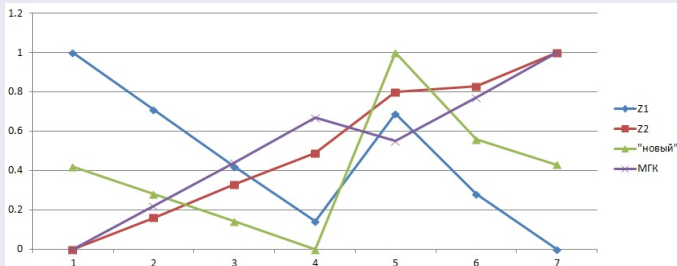


Рис.: Динамика индексов в искусственном примере 2.

**Нелогичность поведения МГК:** на промежутке  $t \in [4; 5]$  оба показателя  $Z^{(1)}$  и  $Z^{(2)}$  растут, а индекс МГК — падает!

**Причина:** отрицательная корреляция между показателями  $Z^{(1)}$  и  $Z^{(2)}$ .

# Сопоставление с модифицированным МГК — 1

Возможное решение проблемы отрицательных весов и динамической несопоставимости — возвести веса МГК в квадрат [4, Айвазян, 2003].

Индекс  $I_{\text{ММГК}}$  модифицированного метода главных компонент (ММГК) определяется следующим образом.

Сначала рассчитывают коэффициенты  $v^{(1)}, \dots, v^{(k)}$  тем же способом, что и при МГК:

$$\begin{cases} D(v^{(1)}Z^{(1)} + \dots + v^{(k)}Z^{(k)}) \rightarrow \max_{v^{(1)}, \dots, v^{(k)} \in \mathbb{R}}, \\ (v^{(1)})^2 + \dots + (v^{(k)})^2 = 1. \end{cases}$$

а потом по определению полагают

$$I_{\text{ММГК}} := (v^{(1)})^2 Z^{(1)} + \dots + (v^{(k)})^2 Z^{(k)}.$$

Обращаем внимание на то, что ММГК веса не могут быть отрицательными, но **могут быть нулевыми**.

Пусть признаки  $Z^{(1)}$  и  $Z^{(2)}$  независимы, причем  $D(Z^{(2)}) > D(Z^{(1)})$ .

Индекс, построенный МГК и ММГК с ковариационной матрицей:

$$I_{\text{МГК, cov}} = I_{\text{ММГК, cov}} = Z^{(2)}.$$

Индекс, построенный ММГК с корреляционной матрицей:

$$I_{\text{ММГК, corr}} = 0.5Z^{(1)} + 0.5Z^{(2)}.$$

**Вывод:** при применении ММГК с ковариационной матрицей индекс не учитывает показатель с меньшей дисперсией.

Несложно придумать схожий пример и для ММГК с корреляционной матрицей.

Пусть признаки  $Z^{(1)}$ ,  $Z^{(2)}$  и  $Z^{(3)}$  имеют корреляционную матрицу:

$$\begin{pmatrix} 1 & 0.4 & -0.4 \\ 0.4 & 1 & 0.7 \\ -0.4 & 0.7 & 1 \end{pmatrix}$$

Тогда первая модифицированная главная компонента равна:

$$I_{\text{ММГК, corr}} = 0.5Z^{(2)} + 0.5Z^{(3)}.$$

**Обратим внимание**, что первый признак  $Z^{(1)}$  не входит в данную компоненту.

**Причина:**  $\text{corr}(Z^{(1)}, Z^{(2)}) = -\text{corr}(Z^{(1)}, Z^{(3)})$ .

Этот пример может показаться слишком экзотическим, но можно придумать и более естественный случай.

Пусть признаки  $Z^{(1)}$ ,  $Z^{(2)}$  и  $Z^{(3)}$  имеют корреляционную матрицу:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.7 \\ 0 & 0.7 & 1 \end{pmatrix}$$

Тогда первая модифицированная главная компонента все равно равна:

$$I_{\text{ММГК, corr}} = 0.5Z^{(2)} + 0.5Z^{(3)}.$$

**Вывод:** МГК и ММГК с корреляционной матрицей придают нулевой вес признаку, некоррелированному с остальными признаками.

## МГК с корреляционной матрицей:

- не учитывает разброс показателей,
- учитывает корреляции,
- не гарантирует неотрицательность весов и динамическую сопоставимость,
- может придавать нулевой вес показателю, несмотря на его вариабельность.

## МГК с ковариационной матрицей:

- учитывает разброс как сосредоточенность у крайних значений,
- учитывает корреляции,
- не гарантирует неотрицательность весов и динамическую сопоставимость,
- может придавать нулевой вес показателю, несмотря на его вариабельность.

### ММГК с корреляционной матрицей:

- не учитывает разброс показателей,
- учитывает корреляции,
- гарантирует неотрицательность весов и динамическую сопоставимость,
- может придавать нулевой вес показателю, несмотря на его вариабельность.

### ММГК с ковариационной матрицей:

- учитывает разброс как сосредоточенность у крайних значений,
- учитывает корреляции,
- гарантирует неотрицательность весов и динамическую сопоставимость,
- может придавать нулевой вес показателю, несмотря на его вариабельность.







### «Новый» способ:

- учитывает разброс показателей,
- не учитывает корреляции,
- гарантирует неотрицательность весов и динамическую сопоставимость,
- придаёт положительный вес изменчивым показателям.

### Возникают вопросы:

- нужно ли учитывать корреляции?
- нужно ли учитывать разброс?

Спасибо за внимание!

-  Борzych Д. А. Количественный анализ динамики уровня здоровья населения РФ // Вестник НГУЭУ. 2016. № 1. С. 133-143.
-  Борzych Д. А., Фурманов К. К., Чернышева И. К. О способе построения динамически сопоставимых композитных индексов // Вестник НГУЭУ. 2016. № 4. С. 67-83.
-  Регионы России. Социально-экономические показатели. 2013 : стат. сб. / Росстат. – М., 2013.
-  Айвазян С. А. К методологии измерения синтетических категорий качества жизни населения // Экономика и математические методы. Т. 39. №2. 2003, а.