3. Основные понятия латентного анализа

Латентный анализ был развит П. Лазарсфельдом во второй половине 40-х годов XX в. в процессе изучения социальных установок американских солдат. Метод впервые был изложен в четвертом томе серии «Исследования по социальной психологии во второй мировой войне» [14].

Существо метода заключается в следующем. Предполагается, как и в теории тестов, что исследуемая социальная установка представляет собой в числовом отношении некоторый гипотетический (латентный) континуум. Индивиды будут как-то располагаться на этом континууме в соответствии с определенным значением своей социальной установки. Индивидам задаются

[14] *Lazarsfeld P. F.* The logical and mathematical foundation of latent structure analysis.-In: Measurement and Prediction. N. Y., 1950.

вопросы, и ответы на вопросы выражают как бы внешнюю эмпирическую структуру исследуемого социального явления.

Задача метода — в установлении внутренней латентной структуры, которая обусловливает именно данный характер ответов. Первоначально для простоты будем считать вопросы дихотомическими, т. е. ответы на них альтернативны, типа «да — нет». Вообще говоря, метод не связан с этим ограничением. На исследуемом континууме мы не можем ввести единицу измерения и начало отсчета. Поэтому в лучшем случае мы будем получать ординальную шкалу измерения. При исследовании данной социальной установки можно давать различные наборы вопросов. Вполне понятно, что вовсе необязательно при каждой эмпирической структуре (она, естественно, будет различна) индивид будет обладать одной и той же латентной структурой, т. е. быть в той же самой точке континуума. Не существует детерминистского проецирования эмпирической структуры (ответов) на латентную структуру, а можно попытаться определить только вероятность, с какой данная структура ответов соответствует определенной точке латентного континуума.

Вводится так называемая функция *i-го* вопроса $f_i(x)$. Это вероятность положительного ответа индивида на i-й вопрос, при условии, если индивид находится в точке x латентного континуума. Функция вопроса (в английской транскрипции — traceline) введена Лазарсфельдом по аналогии с операционной характеристикой теории тестов и является вероятностной характеристикой вопроса. Можно выделить три типа вопроса по виду их функций (рис. 16).

Тип I—это такие вопросы, когда с увеличением значений латентной переменной вероятность ответить на него положительно увеличивается, с уменьшением — уменьшается. Пока мы не обращаем внимания на форму кривой.

Тип II — зависимость обратная: с увеличением исследуемой переменной вероятность положительного ответа уменьшается.

Тип III — вопросы таковы, что наибольшая вероятность ответить на них положительно при среднем значении переменной; вероятность уменьшается при увеличении и уменьшении исследуемой переменной.

Далее вводится так называемый маргинал *i-го* вопроса — p_i. Это число лиц, которые положительно ответили на *i-й* вопрос.

Наконец, поскольку задача вероятностная, необходимо найти закон распределения лиц на континууме, т. е. плотность вероятности $\varphi(x)$.

Таким образом, нам даны n вопросов (дихотомических), введены величины:

$f_i(x)$ —функции вопросов;
p_i — маргиналы вопросов;
$\varphi(x)$ —закон распределения лиц на латентном континууме;
$\varphi(x)dx$ — число лиц в интервале x и $x + dx$;
$f_i(x)\,\varphi(x)dx$ — число лиц в интервале x и $x + dx$, которые положительно ответили на i-й вопрос;

$\int_{-\infty}^{\infty} f_i = (x)\varphi(x)dx$ — число лиц на всем континууме, которые положительно ответили на i-й вопрос, т. е. это число равно маргиналу p_i — известной величине.

Отсюда основное расчетное уравнение латентного анализа:

$$p_i = \int_{-\infty}^{\infty} f_i = (x)\varphi(x)dx \qquad (1)$$

Слева — эмпирические переменные (которые мы получаем в опыте), справа — латентные переменные, которые нам неизвестны. Цель исследования — нахождение функции $\varphi\{x)$.

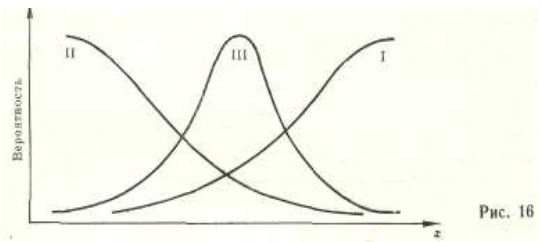


Рис. 16

Вводится основное математическое допущение, «условие локальной независимости». Оно заключается в том, что если взяты два вопроса, то для индивида в точке X вероятность положительно ответить одновременно на оба вопроса, которую обозначим $f_{ij}(x)$, равна произведению вероятностей положительного ответа на каждый вопрос;

$$\boldsymbol{f_{ij}(x)=f_i(x)f_j(x).} \qquad (2)$$

В общем виде, если взято k вопросов, уравнение (2) принимает вид

$$f_\sigma(x)=f_i f_j \ldots f_l=\prod_{i=1}^{k} f_i(x) \qquad (3)$$

В случае уравнения (1) мы для *n* вопросов получим следующую систему уравнений:

$$\int_{-\infty}^{\infty} f_i = (x)\varphi(x)dx = \mathrm{p}_\sigma \qquad (4)$$

где а — все наборы индексов *i, j* ...

Общего решения эта система уравнений не имеет. В зависимости от условий, налагаемых на функции, получаются те или иные модификации основного расчетного уравнения, которые называются моделями латентного анализа.

Некоторые модели допускают решение и в настоящее время все интенсивнее проникают в практику социологического измерения.

Рассмотрим различные варианты соотношения эмпирических и латентных переменных. Существуют следующие важные комбинации:

Тип I — это наиболее общая и сильная модель латентного анализа. Она может получиться в том случае, если на входе будут стоять качественные эмпирические переменные, а на выходе— количественные латентные переменные, т. е. из данных, обладающих весьма малой информацией, мы получаем весьма богатую информацию. Грубо говоря, мы задаем дихотомические вопросы (номинальная шкала измерения) респондентам в отношении удовлетворенности жизнью, а получаем по меньшей мере интервальную шкалу удовлетворенности.

Тип II — качественные эмпирические и качественные латентные переменные; наиболее разработанный тип моделей — модели так называемых латентных классов, когда все респонденты расположены не непрерывно на латентном континууме, а в отдельных точках, классах. Эти модели наиболее разработаны, во-первых, для дихотомических вопросов, во-вторых, для ограниченного числа вопросов и классов. Под классами понимается простая классификация или номинальная шкала измерения. Делаются в настоящее время попытки получить модель упорядоченных классов.

Тип III — количественные эмпирические и количественные латентные переменные. Эта модель латентного анализа имеет определенный аналог с факторным анализом.

Тип IV — количественные эмпирические и качественные латентные переменные. Это так называемая модель латентно-профильного анализа, разработанного Гибсоном.

Лазарсфельд предложил обобщить латентный анализ на случай многомерного латентного континуума. Для большей нагляд-

ности приведем следующий пример. Когда мы исследуем удовлетворенность жизнью, то задаем определенные вопросы и пытаемся решить соответствующее расчетное уравнение латентного анализа, считая, что удовлетворенность жизнью представляет собой некоторую одномерную величину. Это понятие можно уточнить, если считать, что она — результат, к примеру, удовлетворенности работой и удовлетворенности личной жизнью. Тогда наша первоначальная латентная переменная заменяется двумя тоже латентными переменными, которые мы и будем искать.

В этом случае мы имеем не одномерный континуум—линию, на которой мы строили функции вопросов и функции распределения лиц, а двумерный континуум — плоскость. На ней будут уже поверхности — двумерные функции вопросов и двумерные функции распределения лиц.

Если обозначить одну латентную переменную *x*, а другую — *y*, то основное расчетное уравнение (4) для двумерного случая перейдет в

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_\sigma(x, y)\varphi(x, y)\,dxdy = p_\sigma \qquad (5)$$

где σ — набор индексов *i, j* ...

В последнее время делаются попытки применить латентный анализ к исследованию процессов. В частности, предложена модель применения метода латентных классов к простейшему марковскому процессу повторного поведения.

Существо модели латентных классов заключается в том, что латентная переменная считается прерывной [15]. Это означает, что все респонденты расположены в дискретных точках — классах. Будем считать, что задано *n* дихотомических вопросов, а респонденты расположены в *m* латентных классах. Для этого случая преобразуем основное уравнение (4).

Вместо непрерывной функции плотности будем иметь *m* частот, которые соответствуют относительным объемам латентных классов.

Обозначим их ν^α, $\alpha=1, 2, \ldots, m$. Вместо непрерывного графика (-го вопроса получатся отдельные вероятности для каждого класса, которые обозначим p_i^α Это вероятность положительного ответа на *i-й* вопрос в классе $\boldsymbol{\alpha}$. Условие локальной независимости (3) будет иметь вид

$$\rho^\alpha = \prod_\sigma \rho_\sigma^\alpha \qquad (6)$$

[15] *Lazarsfeld P. F., Henry N. W.* Latent Structure Analysis. N. Y., 1968.

Основные уравнения примут вид

$$p_\sigma = \sum_{a=1}^{m} p_\sigma^a \nu^a; \ \alpha=1, \dots, m, \qquad (7)$$

где σ — наборы индексов.

Важная сторона модели латентных классов — число эмпирических данных и число латентных (неизвестных) переменных. Как известно, необходимым условием существования решения системы латентных уравнений является тот факт, что число неизвестных должно быть не больше числа уравнений. Число уравнений 2".

Имеем

$1=\upsilon^1+\dots+\upsilon^m;$

$p_i= p_i^1\upsilon^1+\dots p_i^m+\upsilon^m;$

$p_{ij}= p_{ij}'\upsilon'+\dots p_{ji}^m+\upsilon^m;$

В 1-й строке –1 уравнение ($C_n^0=1$)

В 2-й строке –n уравнение ($C_n'=1$)

В 3-й строке $-\frac{n(n-1)}{2}$ уравнений ($C^2_n= \frac{n(n-1)}{2}$)

В *i-й* строке — $C_n\underline{i}$ уравнений. Всего *n* строк, и, следовательно, общее число уравнений равно сумме биноминальных коэффициентов:

$1+ C^1_n+\dots+ C^n_n=2_n$

Число неизвестных латентных параметров равно $m(n+1)$, поскольку ***mn*** — число латентных вероятностей и ***m*** — число латентных частот в классах.

Таким образом, необходимое (но недостаточное) условие разрешимости модели латентных классов соблюдено —

$$M(n+1) \leq 2_n \qquad (8)$$

Если окажется, что $m(n+1) <2^n$, то необходимы такие дополнительные условия, налагаемые на эмпирические переменные, чтобы

$$m(n+1)=2^n. \qquad (9)$$

Только в этом случае модель имеет решение. Условия, налагаемые на эмпирические данные, называются условиями редуцируемости.

Из нескольких других оснований, связанных с решением расчетных уравнений, можно получить, что

$n \geq 2m-1$ (8')

Объединяя условия (8) и (8'), получаем выражение, которое дает значение наименьшего числа вопросов:

$n \geq 2\log_2 m+1$ (8'')

Очевидно, что модель латентных классов может иметь практическое значение только при небольшом числе вопросов. Дело здесь даже не в том, что это приведет к огромной вычислительной работе. Можно легко увидеть, уравнение (9) выполняется для $m = 2$ и n=3. Проведем вычисления по всем этапам латентного анализа для этого случая.

Основные уравнения (7) примут вид

$$p_i = p_i^1 v^1 + p^2 v^2, \qquad i=1,2,3;$$
$$p_{ij} = p_i^1 p^{j1} v^1 + p_i^2 p_j^2 v^2 \qquad i, j=1,2,3; \qquad (10)$$
$$p_{ijk} = p_i^1 p_j^1 p_k^1 v^1 + p_i^2 p_j^2 p_k^2 v^2, \qquad i, j, k= 1, 2, 3.$$

Или в развернутом виде:

$$p_{1\,2} = p_2^1 p_2^1 v^1 + p_1^2 p_2^2 v^2; \qquad p_1 = p_1^1 v^1 + p_1^2 v^2;$$
$$p_{1\,3} = p_1^1 p_3^1 v^1 + p_1^2 p_3^2 v^2; \qquad p_2 = p_2^1 v^1 + p_2^2 v^2;$$
$$p_{2\,3} = p_2^1 p_3^1 v^1 + p_2^2 p_3^2 v^2; \qquad p_3 = p_3^1 v^1 + p_3^2 v^2;$$
$$p_{1\,2\,3} = p_1^1 p_2^1 p_3^1 v^1 + p_1^2 p_2^2 p_3^2 v^2;$$

и мы имеем уравнение частот:

$$v^1 + v^2 = 1.$$

Всего восемь уравнений и восемь неизвестных; тем самым можно найти все восемь неизвестных параметров:

$$v^1, v^2, p_1^1, p_2^1, p_3^1, \quad p_1^2, p_2^2, p_3^2.$$

Весьма важной задачей латентного анализа является вычисление условных вероятностей. Последняя означает вероятность того, что индивид с данным вариантом ответа попадает в *i-й* класс:

$$P(1|s) = \frac{p_s^1 v^1}{p_s^1 v^1 + p_s^2 v^2};$$

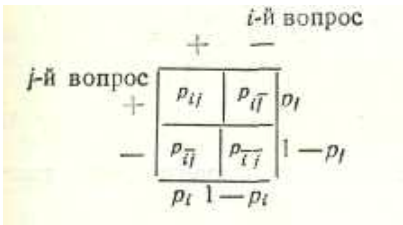
$$P(2|s) = 1 - P(1|s);$$

из общей формулы Бейесса

$$P(x|s) = \frac{P(s|x)\,P(x)}{P(s)}$$

Лица тех вариантов ответов, у которых $p(x) \geq \frac{1}{2}$ попадают в один класс, а у которых $p(x) \geq \frac{1}{2}$, в другой класс (в случае двух классов). Эта ситуация сходна с операцией отнесения к факторам в факторном анализе.

Для решения уравнений модели латентных классов Лазарсфельд развил специальную алгебру, так называемую алгебру дихотомических систем. Основная идея решения вытекает из рассмотрения четырехклеточной таблицы.



где p_{ij} — относительное число лиц, которые положительно ответили на *i-й* и *j-й* вопросы; p_{ij} — число лиц, которые положительно ответили на *j*-й вопрос и трицательно — на *i*-й; $p_{i\bar{j}}$ — число лиц, которые положительно ответили на *i*-й вопрос и отрицательно — на *j*-й; $p_{\bar{i}\bar{j}}$ — число лиц, отрицательно ответивших на оба вопроса.

Рассмотрим определитель

$$[ij] = \begin{vmatrix} p_{ij} & p_{i\bar{j}} \\ p_{\bar{i}j} & p_{\bar{i}\bar{j}} \end{vmatrix} = p_{ij}p_{\bar{i}\bar{j}} - p_{\bar{i}j}p_{i\bar{j}}.$$

Поскольку из таблицы

$$p_{ij} + p_{\bar{i}j} = p_j; \quad p_{ij} + p_{i\bar{j}} = p_i; \quad p_{\bar{i}\bar{j}} = 1 - p_i - p_{\bar{i}j},$$

то имеем

$$p_{\bar{i}j} = p_j - p_{ij}; \quad p_{i\bar{j}} = p_i - p_{ij}; \quad p_{\bar{i}\bar{j}} = 1 - p_i - p_j + p_{ij};$$
$$[ij] = p_{ij}p_{\bar{i}\bar{j}} - p_{i\bar{j}} - p_{\bar{i}j} = p_{ij}(1 - p_i - p_j - p_{ij}) =$$
$$= (p_j - p_{ij})(p_i - p_{ij}) = p_{ij} - p_i p_j.$$

Назовем определитель *[ij]* произведением двух вопросов — *i*-го и *j*-го. На этом определителе основываются три меры связи между

дихотомическими вопросами четырехпольной таблицы:

$$\varphi = \frac{[ij]}{\sqrt{p_i p_{\bar{j}} p_j p_{\bar{j}}}},\ \chi^2 = \varphi;\ f_{ij} = \frac{[ij]}{p_i p_{\bar{i}}}$$

Для тех вопросов – *i, j, k* – введем понятие условного

произведения $[ij;k] = [ij]\, p^{1}_{k}\, p^{2}_{k}$

Выразим неизвестные параметры системы через определители, значения которых известны на основе эмпирических данных.

Имеем

$$[ij] = \begin{vmatrix} p_{ij} & p_{i\bar{j}} \\ p_{\bar{i}j} & p_{\bar{i}\bar{j}} \end{vmatrix} = p_{ij} - p_i p_j = \begin{vmatrix} p_{ij} & p_i \\ p_j & 1 \end{vmatrix} = \begin{vmatrix} 1 & p_i \\ p_j & p_{ij} \end{vmatrix} =$$

$$= \begin{vmatrix} v^1 + v^2 & v^1 p_i^1 + v^2 p_i^2 \\ v^1 p_j^1 + v^2 p_j^2 & v^1 p_i^1 p_j^1 + v^2 p_i^2 p_j^2 \end{vmatrix}.$$

Представим последний определитель как произведение таких определителей:

$$[ij] = \begin{vmatrix} v^1 & v^2 \\ v^1 p_j^1 & v^2 p_j^2 \end{vmatrix} \begin{vmatrix} 1 & p_i^1 \\ 1 & p_i^2 \end{vmatrix} = (p_i^2 - p_i^1)(p_j^2 - p_j^1) v^1 v^2.$$

Следует отметить, что, по крайней мере, один определитель *[ij]* (*ij* = 1, 2, 3) не равен нулю; в противном случае все три вопроса независимы и не имеют никакого отношения к исследуемому явлению.

Введем обозначение:

$p_i^2 - p_i^1 = d_i;\ i = 1, 2, 3$. Соберем вместе все имеющиеся уравнения для нашего случая трех вопросов и двух латентных классов:

$$[1\,2] = d_1 d_2 v^1 v^2; \quad \text{(I)}$$

$$[1\,3] = d_1 d_3 v^1 v^2; \quad \text{(II)}$$

$$[2\,3] = d_2 d_3 v^1 v^2; \quad \text{(III)}$$

$$[1\,2;\,3] = [1\,2]\, p_3^1 p_3^2. \quad \text{(IV)}$$

Рассмотрим величину

$$[1\,2;\,\bar{3}] = [1\,2]\, \bar{p}_3^1 \bar{p}_3^2 = [1\,2]\ (1 - p_3^1)(1 - p_3^2)$$

или

$$1 - p_3^1 - p_3^2 + p_3^1 p_3^2 = \frac{[12;\bar{3}]}{[12]}.$$

Но из (IV) $p_3^1 \cdot p_3^2 = \frac{[12;3]}{[12]}$.

Отсюда

$$p^1{}_3+p^2{}_3=1-\frac{\left[12;\bar{3}\right]}{[12]}+\frac{[12;3]}{[12]}$$

Следовательно $p^1{}_2$ и $p^2{}_3$ являются корнями некоторого квадратного уравнения

$$Z^2-(1+\frac{[12;3]}{[12]}-\frac{\left[12;\bar{3}\right]}{[12]})z+\frac{[12;3]}{[12]}=0$$

Мы положили, что $[12]\neq 0$ и ищем параметры для третьего вопроса (в случае, если [12] = 0, то мы будем искать параметры такого вопроса, где определитель других двух не равен нулю).

Как только $p^1{}_3$ и $p^2{}_3$ найдены, все остальные параметры можно найти без труда.

Имеем, по определению

$$,p_3=v_1p_3+v^2p_3{}^2,\ v^1+v^2=1 \qquad (12)$$

Получаем v^1 и v^2

Далее мы имеем две системы линейных уравнений:

$p_{13} = v^1p_1{}^1p_3{}^1+v^2p_1{}^2p_3{}^2;$

$p_{23} =v^1p^1{}_2p^1{}_3+v^2p^2{}_2p^2{}_3;$

$p_1 =v^1p^1{}_1+v^2p_1{}^2$

$p_2 =v^1p^1{}_2+v^2p_2{}^3$

из которых получаем $p_1{}^1, p_1{}^2, p_2{}^1, p_2{}^2$.

Проводя вычисления уравнений (11) — (14), получаем значения маргиналов для классов, т. е.

$p_1{}^1, p_2{}^1, p_3{}^1\ p_1{}^2, p_2{}^2, p_3{}^2$

Зная эти величины, можно получить частоты вариантов ответов для классов. Например, если берем ответный вариант — + —, то его частота в классе 1 равна

$v^1q_1{}^1p_2{}^1q_3{}^1$, где $q_1{}^1 = 1—p_1{}^1;$ $q_3{}^1=1—p_3{}^1$ а для класса 2 соответственно равна

$v^2q_1{}^2p_2{}^2q_3{}^2$, *где* $q_l{}^2=1—p_1{}^2;\ q_3{}^2=1—p_3{}^2$

Таким образом последовательно получаем все частоты вариантов ответов.

Основное расчетное уравнение допускает возможность решения при определенных ограничениях, наложенных не на $\varphi>(x)$, а на функцию $f_\sigma(x)$. Допустим, что функции вопросов выражаются

некоторыми полиномами

$f_i(x)=a_i + b_ix + c_ix^2 + ...$

В общем случае — степенью k. Для простоты рассмотрим только случаи $k = 1$ и $k = 2$, т. е. когда функции вопросов — прямые и параболы. Прежде всего возьмем случай $k=1$:

$f_i = a_i + b_ix;$

из (1)

$$p_i = \int_{-\infty}^{\infty} (a_i + b_i x)\,\varphi(x)\,dx;$$

$$p_i = \int_{-\infty}^{\infty} [a_i\varphi(x) + b_i x\varphi(x)]\,dx = a_i \int_{-\infty}^{\infty} \varphi(x)\,dx + b_i \int_{-\infty}^{\infty} x\varphi(x)\,dx.$$

Интегралы суть моменты функции $\varphi(x)$:

$p_i = a_i + b_i M^{(1)}$.

Далее, условия локальной независимости:

$$f_{ij}(x) = f_i(x) f_j(x) = (a_i + b_i x)(a_j + b_j x) = a_{ij} + (a_i b_j + a_i b_j)\,x + b_i b_j x^2;$$

$$p_{ij} = \int_{-\infty}^{\infty} [a_i a_j + (a_i b_j + a_i b_j)\,x + b_i b_j x^2]\,\varphi(x)\,dx = a_{ij} + (a_i b_j + a_i b_j)\,M^{(1)} + b_i b_j M^{(2)}.$$

Можно заметить, что для двух вопросов будет шесть неизвестных (a_i, a_j ,b_i, b_j, $M^{(1)}$,$M^{(2)}$) и три уравнения; для трех вопросов—восемь неизвестных и семь уравнений; для четырех зопросов — десять неизвестных и 16 уравнений.

Аналогичные выкладки можно произвести для случая квадратной функции вопросов:

$f_i(x)=a_i+b_ix+c_ix^2$

$f_{ij}(x)=(a_i+b_ix+c_ix^2)(a_i+b_ix+c_ix^2)=$

$=a_ia_j+(a_ib_i+a_ib_i)x+(a_ic_i+b_ib_i+a_ic_i)x^2+(b_ic_i+b_ic_i)x^3+c_ic_ix^4;$

$$p_i = \int_{-\infty}^{\infty} (a_i + b_i x + c_i x^2)\,\varphi(x)\,dx = a_i + b_i M^{(1)} + c_i M^{(2)};$$

$$p_{ij} = a_i a_j + (a_i b_j + a_j b_i)\,M^{(1)} + (a_i c_j + b_i b_j + a_j c_i)\,M^{(2)} + (b_j c_i + b_i c_j)\,M^{(3)} + c_i c_j M^{(4)}.$$

Имеем $[ij] = p_{ij} - p_i p_j$.

Оказывается, что

$$[ij] = b_i b_j \{M^{(2)} - M^{(1)2}\} = b_i b_j \sigma^2, \text{ где } \sigma^2 = M^{(2)} - M^{(1)2}.$$

Аналогично

$$[i\kappa] = b_i b_k \sigma^2;\ [j\kappa] = b_j b_k \sigma^2,$$

Введем величину

$$S_i = \sqrt{\frac{[ij]\,[ik]}{[jk]}}.$$

Тогда можно выразить коэффициенты линейной функции вопроса $f_i\,(x)$ на основании эмпирических данных p_i и S_i:

$$a_i = p_i - \frac{M^{(1)} S_i}{\sigma};\qquad b_i = \frac{S_i}{\sigma}.$$

Два первых момента — средняя и дисперсия — не определяются. Полагаем их равными соответственно нулю и единице. В гаком случае можно легко определить третий момент функции φ(x):

$$M^{(3)} = \sigma^3 k + M^{(1)}(M^{(2)} + 3\sigma^2),$$

где

$$K = \frac{p_{123} - p_1 p_2 p_3}{S_1 S_2 S_3} - \left(\frac{p_1}{S_1} + \frac{p_2}{S_2} + \frac{p_3}{S_3}\right).$$

Зная функции вопросов, можно получить все последующие моменты $\varphi(x)$. Например, с помощью f_{ijk} имеем выражение

$$p_{ijk} = a_i a_{jk} + (a_k a_j b_i + a_i a_k b_j + a_j a_i b_k)\,M^1 + (a_i b_j b_k + a_j b_k b_i + a_k b_i b_j)\,M^2 + b_i b_j b_k M^{(3)},$$

из которого легко определяется $M^{(3)}$. Добавляя уравнения для других совместных частот, получим моменты высших порядков, и таким образом $\boldsymbol{\varphi(x)}$ будет определена.