DOI 10.31483/r-33128

Сизых Дмитрий Сергеевич

Сизых Наталья Васильевна

ОСОБЕННОСТИ ПРЕПОДАВАНИЯ МЕТОДОВ ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ ПОДГОТОВКИ СПЕЦИАЛИСТОВ ПО НАПРАВЛЕНИЮ «БИЗНЕС-ИНФОРМАТИКА»

***Аннотация**: рассмотрены особенности преподавания методов дискриминантного анализа для подготовки специалистов по направлению «бизнес-информатика», то есть специалистов, для которых особое значение имеет понимание модели, алгоритма ее практической реализации и подходов к оценке и анализу полученных результатов. А также – понимание того, с какой целью, где, когда и как можно и необходимо использовать данный метод. Все это обусловливает практическую направленность процесса преподавания.*

***Ключевые слова**: дискриминантный анализ, алгоритм, методика преподавания, практика проведения дискриминантного анализа, расчёт с помощью SPSS.*

***Abstract**: teaching peculiarities of the discriminant analysis methods for training specialists in the «business informatics» field are considered in the study. For these specialists it is important to understand the essence of the model, algorithm of its practical implementation, and approaches for assessing and analyzing the obtained results. Also it is important to understand, why, where, when and how it is possible to apply this method. All this determines the practical orientation of the teaching process.*

***Keywords**: discriminant analysis, algorithm, teaching methods, discriminant analysis practice implementation, application of the SPSS software for the data analysis.*

Введение

Не вызывает сомнения тот факт, что подготовка различных специалистов требует разного подхода к преподаванию одного и того же учебного материала. В настоящее время в связи с расширением процессов автоматизации, с цифровизацией экономики и управления все более востребованным и популярным

становится подготовка специалистов по направлению «бизнес-информатика». В соответствии с будущим направлением деятельности данных специалистов их подготовка включает учебный материал по математике, управлению, программированию, моделированию. Однако методика преподавания дисциплин, включающих данный материал, должна иметь свои особенности преподавания, которые заключаются в практической направленности как изложения материала, так и проведения учебных занятий.

В данном разделе монографии рассмотрены особенности преподавания методов дискриминантного анализа для подготовки специалистов по направлению «бизнес-информатика». Для данных специалистов особое значение имеет понимание модели, алгоритма ее практической реализации и подходов к оценке и анализу полученных результатов. А также – понимание того, с какой целью, где, когда и как можно и необходимо использовать данный метод. На решение данных вопросов и должен быть направлен процесс изложения учебного материала и методика преподавания.

Рассмотрим на отдельном конкретном примере методику изложения материала по изучению особенностей применения метода дискриминантного анализа. Во-первых, данный анализ проводится с помощью специального программного обеспечения (в данном случае пакет SPSS) и поэтому проведение дискриминантного анализа рассматривается на базе основных этапов работы данной программы. Во-вторых, анализ основных этапов работы пакета SPSS сопровождается изучением, используемого данной программой алгоритма. В данном случае вручную просчитываются основные этапы дискриминантного анализа. И, в-третьих, на каждом этапе анализа, при получении результатов, проводится их объяснение.

Общая информация по методу дискриминантного анализа

Дискриминантный анализ относится к статистическому классификационному анализу (дискриминация) объектов исследования по совокупности признаков [1–7; 11]. Области практического применения дискриминантного анализа различны: социология, общественные науки, экономика, финансовые исследования, инвестиции и рыночные исследования, маркетинг, управление

предприятием, позиционирование и пр. Классическим примером использования методов дискриминантного анализа являются скоринг системы в банковских и кредитных организациях, которые позволяют перед принятием решения о выдаче кредитов классифицировать своих клиентов по ряду признаков на надежных и ненадежных. В маркетинге выявляют отличительные характеристики потребителей товаров, на фондовых рынках выявляют предпочтения и характер поведения инвесторов и пр.

Модель анализа относится к моделям «классификации с учителем», то есть анализ проводится по показателям максимального сходства объектов относительно заранее заданных обучающих выборок. В основе данного анализа лежит возможность использовать полученную ранее информацию (или опыт) для решения (предсказания) вопросов классификации новых объектов за счет полученных прогностических переменных (дискриминирующих переменных, предикторов) или функций. В процессе выполнения дискриминантного анализа выявляются различия заранее заданных групп объектов исследования, поиск переменных и функций, разделяющих эти группы и позволяющих относить новые наблюдаемые объекты в одну или несколько заданных групп, а также выполняется классификация объектов. Таким образом, дискриминантный анализ включает совокупность нескольких тесно связанных статистических методов.

Дискриминантный анализ, также как и кластерный и факторный, относится к классификационным видам анализа. По сравнению с кластерным анализом в дискриминантном проводится классификация с обучением (имеются эталонные группы) и формулируется правило, по которому новые объекты относятся к одному из уже существующих классов. А в кластерном анализе объекты классифицируются на основе их различия без какой-либо предварительной информации о составе классов. Таким образом, в кластерном анализе образуются новые кластеры, а в дискриминантном анализе новые кластеры не образуются, а лишь отдельные новые объекты классифицируются по уже существующим группам. Кроме того, дискриминантный анализ по вычислительным процедурам похож на дисперсионный анализ и на множественный регрессионный. Например, в

дискриминантном анализе, так же, как и в многомерном дисперсионном анализе MANOVA, с целью определения наличия значимых различий между группами (с точки зрения всех признаков), сравниваются матрицы ковариаций с помощью многомерного *F*-критерия. В дискриминантном анализе, как и во множественном регрессионном, составляется уравнение регрессии, с помощью которого решаются задачи предсказания и прогнозирования. Но в дискриминантном анализе используется номинальная зависимая переменная, в отличие от количественной переменной в регрессионном анализе. Логистическая регрессия также, как и дискриминантный анализ, используется при необходимости сегментирования, но поскольку последний является более универсальным, то его применение более предпочтительно, а логистическую регрессию рекомендуют применять в тех случаях, если есть сомнения в результатах дискриминантного анализа.

В общем, цель дискриминантного анализа состоит в классификации новых объектов по заранее заданным группам (классам, обучающим выборкам). Дискриминантный анализ направлен на решение следующих задач:

– выбор и статистическая оценка (анализ и моделирование) признаков (дискриминационных переменных), которые наилучшим образом различают (дискриминируют) формирующиеся совокупности между собой;

– построение дискриминантной модели для классификации;

– классификация новых объектов на основе дискриминантной модели;

– прогнозирование поведения новых объектов относительно объектов, входящих в обучающие группы;

– оценивание точности и качества прогнозов на основе полученной дискриминантной модели;

– сопоставление и уточнение результатов классификации объектов кластерного анализа;

– использование дискриминантных моделей для различных скоринговых систем и пр.

Все перечисленные задачи относятся к следующим четырем типам:

– определение функции дискриминации, позволяющей распределять новые объекты по заданным заранее группам;

– классификация объектов;

– восстановление потерянных или недостающих признаков принадлежности объекта к той или иной группе;

– прогнозирование (предсказание) будущих событий на основании имеющихся данных.

Практическое применение дискриминантного анализа связано с рядом ограничений и рекомендаций (чаще практического характера):

– шкала значений признаков должна быть числовой (интервальной или относительной), а если применяется порядковая шкала, то рекомендуемое число градаций должно быть не менее пяти;

– многомерная нормальность закона распределения дискриминантных признаков для каждого класса;

– условием использования линейных дискриминантных функций является равенство матриц ковариаций в разных классах, но при малых объемах выборок даже в случае наличия не слишком больших различий матриц ковариаций можно также применять линейные функции дискриминации, а во всех остальных случаях рекомендованы квадратичные дискриминаторы;

– в отношении наличия выбросов в исходных данных: рекомендовано проводить дополнительные исследования и, возможно, исключать выбросы;

– отдельное внимание требует анализ сильно коррелированных признаков и возможного их влияния на дискриминантную функцию: имеется рекомендация о линейной независимости между дискриминантными признаками (отсутствия мультиколлинеарности), поскольку коррелированные признаки не несут новой информации, но, по данным практиков, это практически не искажает полученные результаты дискриминантного анализа;

– по количеству объектов и признаков имеются разные рекомендации: в общем случае количество объектов должно превышать количество

дискриминантных признаков на две и более единиц, должно быть два или более эталонных класса и не менее двух объектов в каждом классе;

– дискриминантная функция должна включать небольшое количество признаков (не более 10), поскольку увеличение количества признаков в ней практически всегда связано со снижением качества статистической модели классификации, ее точности и надежности.

Некоторые замечания к ограничениям дискриминантного анализа:

1. Предполагается, что анализируемые признаки представляют выборку из многомерного нормального распределения. Для проверки данного условия можно воспользоваться специальными критериями нормальности и графиками. Данная возможность имеется в любом программном обеспечении для дискриминантного анализа. Однако, следует отметить, что при проведении дискриминантного анализа допускается пренебрежение условием нормальности распределения признаков.

2. Предполагается, что матрицы ковариаций (дисперсий) признаков должны быть однородными. Для проверки данного положения можно построить матричную диаграмму рассеяния, а также воспользоваться многочисленными критериями и способами проверки нарушения данного положения в имеющихся данных. Однако, считается, что малые отклонения от однородности матриц ковариаций (дисперсий) не важны. При этом для принятия окончательного решения анализируются внутригрупповые матрицы дисперсий и корреляций (по матрице рассеяния), поскольку условие об их статистическом равенстве является относительно важным. Если произвольно принять, что ковариационные матрицы статистически неразличимы, то могут быть исключены признаки, имеющие большое значение для хорошей дискриминации. Проверка на однородность проводится с помощью М критерия Бокса, который чувствителен к отклонению от многомерной нормальности, и поэтому можно пренебрегать его показателем. В общем случае критерий М Бокса используется для проверки многомерной нормальности распределения по показателю равенства ковариационных матриц.

3. Наличие возможной зависимости между средними значениями признаков по классам и дисперсиями (или стандартными отклонениями) между собой

может вносить существенные погрешности в корректность критериев значимости. Наличие больших значений средних и больших значений изменчивости указывает на ненадежность значений средних. При этом критерий значимости может ошибочно указывать на статистическую значимость. Данная погрешность может быть вызвана и наличием нескольких выбросов, которые влияют на средние значения и увеличивают их изменчивость. Рекомендуется провести анализ описательных статистик по признакам, проанализировать диаграммы рассеяния. Для выявления такого случая следует изучить описательные статистики, то есть средние и стандартные отклонения или дисперсии.

4. Другое предположение в дискриминантном анализе заключается в том, что признаки, используемые для дискриминации между совокупностями, не являются полностью избыточными. Избыточность признаков по отношению друг к другу может привести к получению плохо обусловленных матриц ковариаций (дисперсий), что не позволит выполнить необходимое для дискриминантного анализа обращение матриц. Например, если какой-то признак является суммой нескольких других признаков, то можем получить плохо обусловленные матрицы. Для решения данной проблемы для каждого дискриминантного признака вычисляется и отслеживается показатель толерантности. Значение толерантности вычисляется как $1\text{-}R^2$, где R^2 – коэффициент множественной корреляции для соответствующего признака со всеми другими признаками в используемой модели. Таким образом, толерантность определяет долю дисперсии, относящейся к соответствующему признаку. Когда признак почти полностью избыточный, значение толерантности для данного признака будет приближаться к нулю и поэтому матрица ковариации является плохо обусловленной.

Алгоритм метода дискриминантного анализа и объяснение полученных данных: иллюстрация на практическом примере

Рассмотрим простейшие примеры процесса дискриминации новых объектов по эталонным классам. По результатам рассмотрения данных примеров проанализируем алгоритм решения задач дискриминантного анализа и особенности использования специальных автоматизированных программ (на примере пакета SPSS) для проведения дискриминантного анализа [8–10; 12–16].

Расчетные примеры иллюстрируют простейшие варианты использования дискриминантного анализа, без оптимизации и учета некоторых дополнительных условий. Примеры показывают особенности получения дискриминантных канонических функций и процесс принятия решения по ним. Данные функции не оптимальны, но позволяют принять правильные решения в простейших случаях. Расчет проводится в Excel и сравнивается с результатами данного анализа, проведенного с помощью SPSS.

Проведем анализ полученных результатов на примерах дискриминации стран ЕС по инновационности их предприятий. Данные взяты из базы Евростат. Для примера возьмем два варианта заданий по анализу шести стран ЕС относительно инновационности их предприятий. В первом задании (вариант 1), страны разделены на два эталонных класса, а в варианте 2 имеем три эталонных класса. Таким образом, в качестве объектов возьмем страны: Дания, Франция, Финляндия, Швеция, Австрия, Португалия, а предикторы (признаки) – укрупненные факторы Инновационное производство и Инновационное управление и маркетинг, которые были получены по результатам факторного анализа.

Таблица 1

Исходные данные

Объекты		Признаки	
Обозначение	Страна	Инновационное производство	Инновационное управление и маркетинг
1	Дания	0,19805	0,45875
2	Франция	0,27982	0,47377
3	Финляндия	0,9568	0,04895
4	Швеция	1,02114	0,07407
5	Австрия	0,39664	0,72057
6	Португалия	0,36248	0,83139
Страны для определения класса (группы) по инновационным показателям (дискриминируемые объекты)			
7	Словения	0,1142	0,08966
8	Италия	0,68678	0,63041
9	Ирландия	0,69931	0,48736

Рассмотрим новые три страны (дискриминируемые объекты), которые необходимо с помощью метода дискриминантного анализа отнести к одному из уже известных классов. Для дискриминируемых объектов имеются показатели по такому же набору признаков, как и для эталонных объектов.

Задача дискриминации сводится к определению дискриминирующих признаков, определению дискриминантной функции и самому процессу определения к каким классам относятся новые страны по показателям инновационности их предприятий.

Таким образом, рассмотрим следующие два варианта дискриминантного анализа:

Вариант 1: разделим рассматриваемые страны на два эталонных класса и определим дискриминантную функцию (одну) для этих классов:

– первый класс (Финляндия и Швеция) – это страны с высоким уровнем инновационного производства, но низким уровнем инновационного управления и маркетинга, т.е. основные инновационные доходы страны получают за счет новых технологий, оборудования и пр.;

– второй класс (Португалия и Австрия, Дания и Франция) – это страны с высоким уровнем инновационного управления и маркетинга и более низким уровнем инновационного производства, т.е. большее внимание на предприятиях данных стран уделяется развитию и применению инновационных методов управления и маркетинга, а инновационные технологии развиваются на более низком уровне.

Вариант 2: разделим рассматриваемые страны на три класса и определим две дискриминантные функции для этих классов:

– первый класс (Финляндия и Швеция);

– второй класс (Дания и Франция, класс с минимальным показателем инновационного производства) – показатель инновационного управления выше показателя инновационного производства, при этом имеем среднее значение по инновационному управлению и минимальное значение по инновационному производству;

– третий класс (Португалия и Австрия, класс с максимальным показателем инновационного управления) – показатель инновационного управления выше показателя инновационного производства, при этом имеем максимальное значение по инновационному управлению и среднее значение по инновационному производству.

Вариант 1.

Имеем два эталонных класса и три новых страны, которые надо распределить по эталонным классам.

1. Исходные данные представим в виде матриц *X, Y и Z.*

Таблица 2

Исходные данные для варианта 1

Показатели / Эталонные классы и новые объекты	Страны	Инновационное производство	Управление инновационными предприятиями
Технологичные страны	Финляндия	0,9568	0,04895
	Швеция	1,02114	0,07407
Остальные	Дания	0,19805	0,45875
	Франция	0,27982	0,47377
	Австрия	0,39664	0,72057
	Португалия	0,36248	0,83139
Новые страны для разделения	Словения	0,1142	0,08966
	Италия	0,68678	0,63041
	Ирландия	0,69931	0,48736

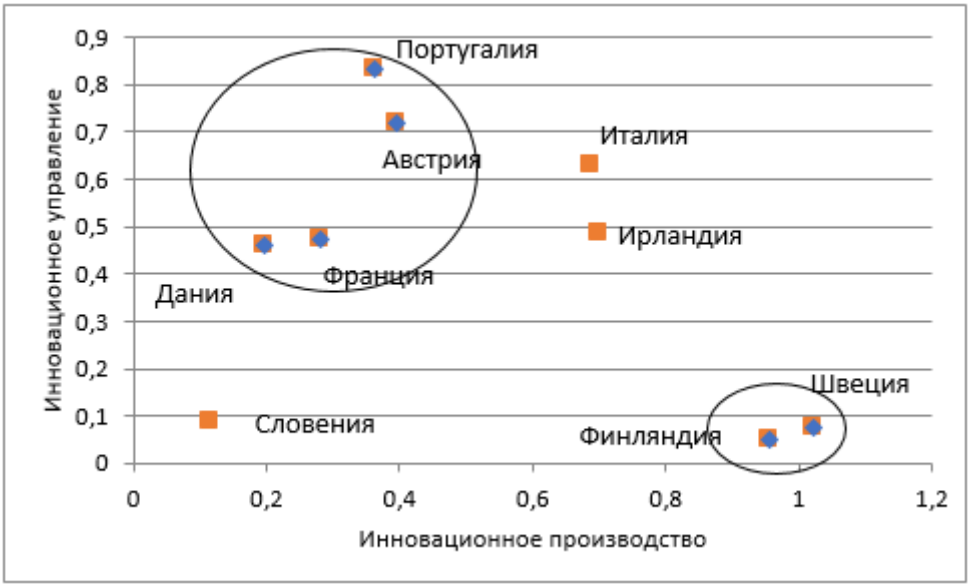

Рис. 1. Расположения стран в двумерном пространстве признаков примера

Имеем три матрицы: X и Y по эталонным классам (обучающие подмножества), а Z по новым объектам для распределения:

$$X = \begin{pmatrix} 0,9568 & 0,04895 \\ 1,02114 & 0,07407 \end{pmatrix};\ Y = \begin{pmatrix} 0,19805 & 0,45875 \\ 0,27982 & 0,47377 \\ 0,39664 & 0,72057 \\ 0,36248 & 0,83139 \end{pmatrix};\ Z = \begin{pmatrix} 0,1142 & 0,08966 \\ 0,68678 & 0,63041 \\ 0,69931 & 0,48736 \end{pmatrix}$$

2. Рассчитаем средние значения и получим векторы средних по матрицам эталонных классов:

Определим средние значения по каждому признаку для каждого из 2-х эталонных классов объектов отдельно. Среднее значение для каждого признака в каждом классе считаем по формуле:

$$\overline{x}_{jl} = \frac{1}{n_l}\sum_{i=1}^{n_l} x_{ij},$$

где x_{ij} – значение j-того признака для i-того объекта в рассматриваемом классе;

$\overline{x}_{jl}$ – среднее значение j-того признака для данного класса l;

n_l – общее количество объектов в классе l;

$i = 1,2,\ldots,n_l$ – количество объектов в классах: в рассматриваемом примере в 1-м классе 2 объекта $n_x = 2$, а во 2-м классе 4 объекта $n_y = 4$;

$j = 1,2,\ldots,m$ – количество признаков в рассматриваемом примере $m = 2$.

Результаты расчетов по каждому классу представим в виде векто-столбцов:

$$\overline{X} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \ldots \\ \overline{x}_m \end{pmatrix},$$ где m – количество признаков, в нашем примере $m = 2$

Таким образом, имеем:

$$\overline{X} = \begin{pmatrix} 0,98897 \\ 0,06151 \end{pmatrix},\ \overline{Y} = \begin{pmatrix} 0,30925 \\ 0,62112 \end{pmatrix}$$

5. Для каждого эталонного класса рассчитаем ковариационные матрицы.

Ковариационные матрицы будут квадратные размерности m * m. Поскольку в рассматриваемом примере используется два признака, то будем иметь матрицы размером 2 * 2.

Рассчитаем центрированные матрицы по следующей общей формуле:

$$X_{центр} = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \dots & x_{1m} - \overline{x}_m \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \dots & x_{1m} - \overline{x}_m \\ \dots & \dots & \dots & \dots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \dots & x_{nm} - \overline{x}_m \end{pmatrix}$$

Имеем следующие две центрированные матрицы для эталонных классов:

$$X_{центр} = \begin{pmatrix} -0,03217 & -0,01256 \\ 0,03217 & 0,01256 \end{pmatrix}$$

$$Y_{центр} = \begin{pmatrix} -0,11120 & -0,16237 \\ -0,02943 & -0,14735 \\ 0,08739 & 0,09945 \\ 0,05323 & 0,21027 \end{pmatrix}$$

Ковариационные матрицы рассчитываем по формуле:

$$S_x = \frac{1}{n_x - 1}(X_{центр}{}^T * X_{центр}) = \frac{1}{1}(\begin{pmatrix} -0,03217 & 0,03217 \\ -0,01256 & 0,01256 \end{pmatrix} * \begin{pmatrix} -0,03217 & -0,01256 \\ 0,03217 & 0,01256 \end{pmatrix}) =$$

$$= \begin{pmatrix} 0,002070 & 0,000808 \\ 0,000808 & 0,000316 \end{pmatrix} = \begin{pmatrix} 0,002 & 0,001 \\ 0,001 & 0 \end{pmatrix}$$

$$S_y = \frac{1}{n_y - 1}(\mathrm{Y}_{центр}{}^T * \mathrm{Y}_{центр}) = \frac{1}{3}(\begin{pmatrix} -0,11120 & -0,02943 & 0,08740 & 0,05323 \\ -0,16237 & -0,14735 & 0,09945 & 0,21027 \end{pmatrix} *$$

$$* \begin{pmatrix} -0,11120 & -0,16237 \\ -0,02943 & -0,14735 \\ 0,08740 & 0,09945 \\ 0,05323 & 0,21027 \end{pmatrix}) = \frac{1}{3}\begin{pmatrix} 0,023702 & 0,042276 \\ 0,042276 & 0,102180 \end{pmatrix} = \begin{pmatrix} 0,008 & 0,014 \\ 0,014 & 0,034 \end{pmatrix}$$

Рассчитаем общую классификационную матрицу для всех объектов в эталонных классах:

$$S_{общая} = \frac{1}{n_x + n_y - 1}\left(\begin{pmatrix} 0,4210 & 0,4850 & -0,338 & -0,256 & -0,139 & -0,173 \\ -0,386 & -0,361 & 0,0242 & 0,0392 & 0,2860 & 0,3968 \end{pmatrix} * \right.$$

$$\left. * \begin{pmatrix} 0,4210 & -0,386 \\ 0,4853 & -0,361 \\ -0,338 & 0,0242 \\ -0,256 & 0,0392 \\ -0,139 & 0,2860 \\ -0,173 & 0,3968 \end{pmatrix}\right) = \frac{1}{5}\left(\begin{pmatrix} 0,642 & -0,464 \\ -0,464 & 0,520 \end{pmatrix}\right) = \begin{pmatrix} 0,128 & -0,093 \\ -0,093 & 0,104 \end{pmatrix}$$

Полученные результаты полностью совпадают с данными SPSS (см. таблицу 21).

Таблица 3

Ковариационные матрицы по результатам расчета в SPSS

Группа		Инновационное производство	Инновационное управление
1	Инновационное производство	,002	,001
	Инновационное управление	,001	,000
2	Инновационное производство	,008	,014
	Инновационное управление	,014	,034
Всего	Инновационное производство	,128	-,093
	Инновационное управление	-,093	,104
a. Количество степеней свободы итоговой ковариационной матрицы – 5.			

6. Рассчитаем объединенную ковариационную матрицу для двух эталонных классов.

Воспользуемся следующей формулой для расчета объединенной ковариационной матрицы $S_{x,y}$:

$$S_{x,y} = \frac{1}{n_x + n_y - 2}((n_x - 1) * S_x + (n_y - 1) * S_y) =$$

$$= \frac{1}{2 + 4 - 2}\left(1 * \begin{pmatrix} 0,002 & 0,001 \\ 0,001 & 0,00 \end{pmatrix} + 3 * \begin{pmatrix} 0,008 & 0,014 \\ 0,014 & 0,034 \end{pmatrix}\right) =$$

$$= \frac{1}{4}\begin{pmatrix} 0,0258 & 0,0431 \\ 0,0431 & 0,1025 \end{pmatrix} = \begin{pmatrix} 0,006 & 0,011 \\ 0,011 & 0,026 \end{pmatrix}$$

Результаты совпадают с данными SPSS (см. таблицу 4).

Таблица 4

Объединенные внутригрупповые матрицы по результатам расчета в SPSS

		Инновационное производство	Инновационное управление
Ковариация	Инновационное производство	,006	,011
	Инновационное управление	,011	,026
Корреляция	Инновационное производство	1,000	,838
	Инновационное управление	,838	1,000

Количество степеней свободы ковариационной матрицы – 4.

Кроме ковариации в SPSS, была подсчитана и взаимная корреляция по признакам эталонных классов объектов.

7. Рассчитаем обратную матрицу к объединенной ковариационной матрице $S_{x,y}$.:

$$S^{-1}{}_{x,y} = \begin{pmatrix} 742,86 & -314,28 \\ -314,28 & 171,43 \end{pmatrix}$$

Найдем вектор оценок коэффициентов дискриминации B (вектор дискриминантных множителей). В общем виде имеем следующую дискриминантную функцию:

$$F = b_1 * x_{i1} + b_2 * x_{i2} + \ldots + b_j * x_{ij} + \ldots + b_m * x_{im},$$

Значения дискриминантных функций определяются для объектов эталонных классов. Если эталонных классов более двух, то строится несколько (более одной) дискриминантных функций. Коэффициенты нескольких дискриминантных функций (например, канонических) выбираются по следующему алгоритму:

– для первой дискриминантной функции коэффициенты выбираются таким образом, чтобы центроиды различных эталонных классов как можно больше отличались друг от друга;

– для второй дискриминантной функции коэффициенты выбираются также, но при этом должно выполняться дополнительное условие, состоящее в том, что значения второй функции должны быть не коррелированы со значениями первой;

– аналогично определяются коэффициенты и для других дискриминантных функций.

Коэффициенты дискриминантных функций определяются из условия обеспечения наибольшего различия между дискриминантными функциями по методу наименьших квадратов. Наибольшее различие между эталонными классами достигается при минимальной внутригрупповой вариации и максимальной межгрупповой вариации признаков.

Поскольку для определения коэффициентов дискриминантных функций используются сложные расчеты, то приведем расчет по упрощенному алгоритму. Полученный вариант дискриминантной функции позволяет провести классификацию новых объектов. Вектор оценок коэффициентов дискриминации *B* определим по следующему соотношению:

$$B = S_{x,y}^{-1}(\bar{X} - \bar{Y}) = \begin{pmatrix} 742,86 & -314,28 \\ -314,28 & 171,43 \end{pmatrix} * \left(\begin{pmatrix} 0,98897 \\ 0,06151 \end{pmatrix} - \begin{pmatrix} 0,30925 \\ 0,62112 \end{pmatrix} \right) =$$

$$= \begin{pmatrix} 680,81 \\ -309,56 \end{pmatrix}$$

Для рассматриваемого примера b_1 = 680,81, b_2 = -309,56. При этом дискриминантная функция имеет следующий общий вид:

$$F = 680,81 * x_{i1} + (-309,56) * x_{i2}$$

8. Просчитаем значения дискриминантной функции для двух эталонных классов X и Y:

– для класса X:

F_{x1} = 680,81 * 0,9568 + (-309,56) * 0,04895 = 636,25

F_{x2} = 680,81 * 1,02114 + (-309,56) * 0,07407 = 672,27

– для класса Y:

F_{y1} = 680,81 * 0,19805 + (-309,56) * 0,45875 = -7,18

F_{y2} = 680,81 * 0,27982 + (-309,56) * 0,47377 = 43,84

F_{y3} = 680,81 * 0,39664 + (-309,56) * 0,72057 = 46,98

F_{y4} = 680,81 * 0,36248 + (-309,56) * 0,83139 = -10,58

9. В каждом эталонном классе рассчитаем среднее арифметическое значение дискриминантной функции по сумме всех полученных значений:

$$\bar{F}_x = 654,26$$

$$\bar{F}_y = 18,26$$

10. Определяем общее среднее арифметическое значение для средних значений дискриминантных функций двух эталонных классов:

$$\bar{F} = \frac{1}{2}(\bar{F}_x + \bar{F}_y) = 336,26$$

Данное значение является константой дискриминации, относительно значения которого, решается вопрос по зачислению новых объектов в эталонные классы.

11. Рассчитаем значения дискриминантных функций для новых объектов, находящихся в группе новых объектов Z.

$$F_{z1} = 680,81 * 0,1142 + (-309,56) * 0,08966 = 49,99$$

$$F_{z2} = 680,81 * 0,68678 + (-309,56) * 0,6304 = 272,42$$

$$F_{z3} = 680,81 * 0,6993 + (-309,56) * 0,4874 = 325,23$$

12. Выполним распределение трех новых объектов по эталонным классам X и Y.

Распределять объекты будем по следующему правилу: сравниваем значение дискриминантной функции нового объекта с общим средним значением дискриминантной функции по двум эталонным классам и учитываем при этом соотношение средних значений дискриминантной функции по эталонным классам X и Y. Имеем следующие варианты:

– если $F_z > \bar{F}$, то при $\bar{F}_x > \bar{F}_y$ объект относится к классу X, а при $\bar{F}_x < \bar{F}_y$ к классу Y;

– если $F_z < \bar{F}$, то при $\bar{F}_x > \bar{F}_y$ объект относится к классу Y, а при $\bar{F}_x < \bar{F}_y$ к классу X.

Для рассматриваемого примера имеем:

– $F_{z1} < \bar{F}$, а $\bar{F}_x > \bar{F}_y$, то новый объект 1 (Словения) относится к классу Y;

– $F_{z2} < \bar{F}$, а $\bar{F}_x > \bar{F}_y$, то новый объект 1 (Италия) относится к классу Y;

– $F_{z3} < \bar{F}$, а $\bar{F}_x > \bar{F}_y$, то новый объект 1 (Ирландия) относится к классу Y.

Таким образом, Словения, Италия и Ирландия относятся к классу Y, стран с высоким уровнем инновационного управления и маркетинга и более низким уровнем инновационного производства.

Вариант 2

Имеем три эталонных класса и три новых страны, которые надо распределит по эталонным классам.

1. Исходные данные представим в виде матриц эталонных классов *X*, *Y*, *Z*, и матрицы новых объектов *D*.

Таблица 5

Исходные данные для варианта 2

Показатели / Эталонные классы И новые объекты	Страны	Инновационное производство	Управление инновационными предприятиями
Технологичные страны	Финляндия	0,9568	0,04895
	Швеция	1,02114	0,07407
Минимальный показатель инновационного производства	Дания	0,19805	0,45875
	Франция	0,27982	0,47377
Максимальный показатель инновационного управления	Австрия	0,39664	0,72057
	Португалия	0,36248	0,83139
Новые страны для разделения	Словения	0,1142	0,08966
	Италия	0,68678	0,63041
	Ирландия	0,69931	0,48736

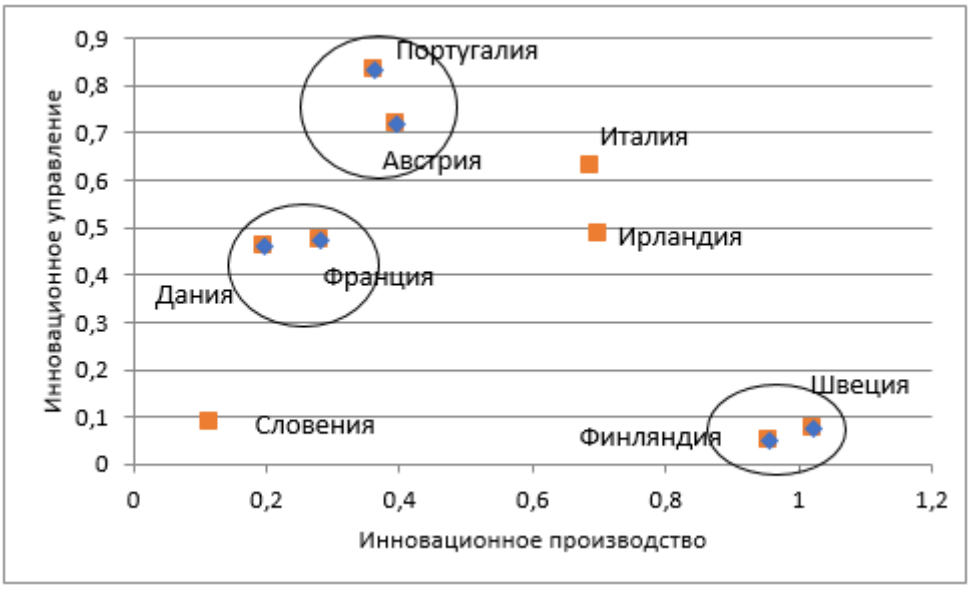

Рис. 2. Расположения стран в двумерном пространстве признаков

Имеем четыре матрицы:

$$X = \begin{pmatrix} 0,9568 & 0,04895 \\ 1,02114 & 0,07407 \end{pmatrix}; \; Y = \begin{pmatrix} 0,19805 & 0,45875 \\ 0,27982 & 0,47377 \end{pmatrix}; \; Z = \begin{pmatrix} 0,39664 & 0,72057 \\ 0,36248 & 0,83139 \end{pmatrix}$$

$$D = \begin{pmatrix} 0,1142 & 0,08966 \\ 0,68678 & 0,63041 \\ 0,69931 & 0,48736 \end{pmatrix}$$

2. Рассчитаем средние значения и получим векторы средних по построенным матрицам для эталонных классов X, Y, Z:

Таким образом, имеем:

$$\bar{X} = \begin{pmatrix} 0,98897 \\ 0,06151 \end{pmatrix}, \; \bar{Y} = \begin{pmatrix} 0,2389 \\ 0,4663 \end{pmatrix}, \; \bar{Z} = \begin{pmatrix} 0,3796 \\ 0,7760 \end{pmatrix}$$

3. Для каждого эталонного класса рассчитаем ковариационные матрицы.

Имеем следующие три центрированные матрицы для эталонных классов:

$$X_{центр} = \begin{pmatrix} -0,03217 & -0,01256 \\ 0,03217 & 0,01256 \end{pmatrix}, \; Y_{центр} = \begin{pmatrix} -0,0409 & -0,0075 \\ 0,0409 & 0,0075 \end{pmatrix},$$

$$Z_{центр} = \begin{pmatrix} 0,0171 & -0,0554 \\ -0,0171 & 0,0554 \end{pmatrix}$$

Ковариационные матрицы рассчитываем по формуле:

$$S_x = \frac{1}{n_x - 1}(X_{центр}{}^T * X_{центр}) = \begin{pmatrix} 0,002 & 0,0008 \\ 0,0008 & 0,0003 \end{pmatrix}$$

$$S_y = \begin{pmatrix} 0,003 & 0,001 \\ 0,001 & 0,000 \end{pmatrix}$$

$$S_z = \begin{pmatrix} 0,001 & -0,002 \\ -0,002 & 0,006 \end{pmatrix}$$

4. Рассчитаем объединенную ковариационную матрицу для трех эталонных классов.

Воспользуемся следующей формулой для расчета объединенной ковариационной матрицы $S_{x,y,z}$:

$$S_{x,y,z} = \frac{1}{n_x + n_y + n_z - 3}((\mathrm{n}_x - 1) * S_x + (n_y - 1) * S_y + (n_z - 1) * S_z) = \begin{pmatrix} 0,002 & 0,000 \\ 0,000 & 0,002 \end{pmatrix}$$

Поскольку в данном примере имеем три эталонных класса, то канонических дискриминантных функций будет две: первая будет разделять 1-й класс X и объединенный 2-й и 3-й (Y и Z), а вторая – будет разделять 2-й Y и 3-й Z классы. Расчеты и анализ проводим согласно предыдущему примеру.

5. Расчет и анализ для первой дискриминантной функции. Данный расчет был проведен в предыдущем примере, поэтому воспользуемся только его результатами: все три новые объекты отнесены к объединенному 2-му и 3-му классам (Y и Z).

Первая дискриминантная каноническая функция имеет вид: $F1$ = 680,81 * x_{i1} + (-309,56) * $x_{i2.}$

6. Расчет и анализ для второй дискриминантной функции, которая позволит разделить объекты классов Y и Z.

Рассчитываем объединенную ковариационную матрицу для классов Y и Z:

$$S_{\mathrm{y,z}} = \frac{1}{n_y + n_z - 2}((n_y - 1) * S_y + (n_z - 1) * S_z) =$$

$$= \frac{1}{2}\left(\begin{pmatrix} 0,004 & -0,001 \\ -0,001 & 0,006 \end{pmatrix}\right) = \begin{pmatrix} 0,002 & -0,0005 \\ -0,0005 & 0,003 \end{pmatrix}$$

Рассчитаем обратную матрицу к объединенной ковариационной матрице $S_{y,z}$.

$$S^{-1}{}_{\mathrm{y,z}} = \begin{pmatrix} 521,74 & 86,96 \\ 86,96 & 347,83 \end{pmatrix}$$

Найдем вектор оценок коэффициентов дискриминации B второй дискриминантной функции (вектор дискриминантных множителей).

$$B = S_{\mathrm{y,z}}{}^{-1}(\bar{Y} - \bar{Z}) = \begin{pmatrix} -100,3 \\ -119,96 \end{pmatrix}$$

В общем виде имеем вторую дискриминантную функцию, которая разделяет второй и третий класс объектов:

$$F2 = -100,3 * x_{i1} - 119,96 * x_{i2},$$

значения, которой определяются для объектов классов Y и Z. Для рассматриваемого примера b_1 = -100,3, b_2 = -119,96.

Просчитаем значения дискриминантной функции для разделения двух эталонных классов Y и Z:

– для класса Y:

$$F_{y1} = -74,9$$

$$F_{y2} = -84,9$$

– для класса Z:

$$F_{z1} = -126,2$$

$$F_{z2} = -136,09.$$

В каждом эталонном классе рассчитаем среднее арифметическое значение дискриминантной функции:

$$\bar{F}_y = -79,9$$

$$\bar{F}_z = -131,2$$

Определяем общее среднее арифметическое значение для средних значений дискриминантных функций эталонных классов Y и Z:

$$\bar{F} = \frac{1}{2}(\bar{F}_y + \bar{F}_z) = -105,53$$

Данное значение является константой дискриминации, относительно значения которого, решается вопрос по зачислению новых объектов в эталонные классы Y и Z.

Рассчитаем значения дискриминантных функций для новых объектов, находящихся в группе новых объектов D.

$$F_{d1} = -22,2$$

$$F_{d2} = -144,5$$

$$F_{d3} = -128,6$$

7. Выполним распределение трех новых объектов по эталонным классам Y и Z.

Для рассматриваемого примера имеем:

$F_{d1} > \bar{F}$, а $\bar{F}_y > \bar{F}_z$, то новый объект 1 (Словения) относится к классу Y;

$F_{d2} < \bar{F}$, а $\bar{F}_y > \bar{F}_z$, то новый объект 2 (Италия) относится к классу Z;

$F_{d3} < \bar{F}$, а $\bar{F}_y > \bar{F}_z$, то новый объект 3 (Ирландия) относится к классу Z.

Таким образом, Словения относится ко второму классу (Дания и Франция, страны с минимальным показателем инновационного производства), а Италия и Ирландия относятся к третьему классу (Австрия и Португалия, страны с максимальным показателем инновационного управления). При этом ни один из трех новых объектов не попал в первый класс.

Обобщенный алгоритм проведения дискриминантного анализа включает следующие этапы:

1. Определение цели анализа и исходных данных. В зависимости от цели исследования – выбрать группирующую переменную (зависимая переменная) с взаимоисключающими значениями, поскольку объекты могут относиться только к одному классу.

Определить набор признаков (независимые переменные) – очень важный этап для успешного проведения дискриминантного анализа, поскольку только действительные объекты используются при расчетах для построения дискриминантной функции. Проанализировать пропущенные значения признаков и принять соответствующие решения (в SPSS данные показатели иллюстрируются в таблице «Анализ обработанных наблюдений»). Если число признаков велико, то сложно или даже невозможно провести дискриминантный анализ со всеми признаками одновременно, поэтому необходимо предварительно рассмотреть и проанализировать важность всех признаков для проведения анализа. Если нет предварительных решений по сокращению признаков, то в дальнейшем необходимо использовать пошаговый метод отбора признаков для дискриминантного анализа. При этом необходимо помнить, что по умолчанию SPSS реализует стандартный метод дискриминантного анализа с использованием всех признаков в дискриминантной функции.

2. Анализ признаков. Необходимо проверить, удовлетворяют ли имеющиеся данные всем предположениям, необходимым для проведения дискриминантного

анализа. Выделить выбросы, принять решение по отбрасыванию признаков, которые заведомо не могут быть хорошими дискриминаторами анализируемых объектов. Программные пакеты позволяют получить таблицу с описательными статистиками для каждого признака, гистограмму распределения, что позволит провести быстрый визуальный контроль нарушений нормальности. Можно визуально проанализировать диаграмму рассеяния для любой пары признаков. Можно использовать большой набор различных описательных графиков, категоризованных гистограмм и пр.

Формируется таблица «Статистика группы» с данными о средних значениях дискриминационных признаков в каждой из исследуемых групп объектов, стандартное отклонение и валидное количество действительных наблюдений. Визуально можно оценить имеются ли различия средних значений признаков в выделенных группах

Результаты теста на равенство средних значений в эталонных классах находятся в таблице «Критерии равенства групповых средних»: рассчитывается значение показателя лямбда Уилкса, F-критерия и уровня значимости. На основании этих данных можно определить значимые признаки для дискриминации объектов.

Формируются и предъявляются объединенные внутригрупповые матрицы ковариации и корреляции, ковариационные матрицы отдельно по каждому классу и общая ковариационная матрица по всем объектам.

Проверяется показатель эквивалентности ковариационных матриц и выводится критерий Бокса по классам объектов.

Если на данном этапе, после анализа признаков, принимается решение изменения состава дискриминационных признаков, то следует заново сформировать задание на проведение дискриминантного анализа с измененным списком признаков (независимых переменных).

3. Пошаговый дискриминантный анализ. Позволяет автоматически выбирать дискриминантные признаки по заданным критериям. Критерии либо задаются априорно, либо выбираются те, что по умолчанию встроены в SPSS:

– F-критерий для включения признаков в уравнение регрессии (по умолчанию таким критерием является F≥3,84) и F-критерий для исключения предикторов из уравнения регрессии (по умолчанию F≤2,71). Если необходимо включить большее количество признаков, то необходимо понизить верхнее критическое значение и задать уровень априорно, а чтобы исключить большее количество признаков из модели необходимо увеличить нижнее критическое значение и задать его априорно;

– задать уровень значимости для Fкритерия (по умолчанию введены значения 0,05 и 0,01;

– задать априорную вероятность распределения объектов по классам: можно априорно задать «все группы равны» и «вычислить по размерам групп» (по умолчанию одинаковая вероятность по всем классам).

Кроме критериев, задается пошаговый вывод результатов (по умолчанию); для расстояний Махалонобиса выводится F попарных расстояний.

4. Построение дискриминантной модели. Рассчитываются и анализируются коэффициенты канонической дискриминантной функции: коэффициенты стандартизованной канонической дискриминантной функции, не стандартизованные коэффициенты, не стандартизованные канонические дискриминантные функции, вычисленные в групповых средних. Построенная дискриминантная модель должна максимально четко разделять исследуемые группы, поэтому проводится расчет показателей качества дискриминантной модели и предоставляются следующие показатели: объединенные внутригрупповые корреляции между дискриминирующими переменными и стандартизованными каноническими дискриминантными функциями (матрица структуры); Собственное значение (% дисперсии и каноническая корреляция), Лямбда Уилкса (Хи-квадрат и показатель значимости).

5. Построение функций классификации. Рассчитываются линейные дискриминантные функции Фишера с предоставленными данными по учтенным априорным вероятностям. Имеется возможность построения территориальной карты. Статистика по наблюдениям позволяет получить информацию по

предсказанным группам (с указанием возможного неправильного предсказания), апостериорную вероятность предсказания, квадрат расстояния Махаланобиса до центроида, дискриминантные баллы. Матрица классификации позволяет проанализировать статистику по сгруппированным и не сгруппированным объектам.

6. Анализ полученных результатов. Проводится изучение и анализ полученных результатов классификации, принимается решение по поводу возможного расширения (или наоборот сужения) набора признаков. Отдельно рассматриваются неверно классифицированные наблюдения, которые могут образовывать незамеченную ранее группу и принимается при необходимости проведение повторного дискриминантного анализа по внесенным необходимым коррективам;

7. Проверка качества. Заключительным этапом является проверка качества построенного решающего правила.

Анализ полученных результатов процесса дискриминации продемонстрируем на примерах, рассмотренных в этом разделе (дискриминация стран ЕС по показателю инновационности их предприятий). Дискриминантный анализ данных примеров был проведен с помощью пакета SPSS. Проанализируем полученные результаты.

Вариант 1. Дискриминантный анализ по двум эталонным классам.

Используем стандартный метод дискриминации (не пошаговый). Априорные вероятности не задаем. Все показатели стандартные – по умолчанию. Проводим анализ согласно приведенному алгоритму.

1. Определение цели анализа и исходных данных. Имеем шесть стран ЕС (объекты), которые характеризуются признаками (независимые переменные, имеем два признака), отражающими показатели инновационности предприятий этих стран. Исходная информация взята из базы Евростат.

Группирующая переменная (зависимая переменная) имеет две градации: значение 1 – страны с высоким уровнем инновационного производства и значение 2 – страны с высоким уровнем инновационного управления и маркетинга. Имеем два эталонных класса: в 1-м два объекта, во 2-м 4 объекта. Минимальные условия по количеству классов, объектов и признаков соблюдены.

Группирующая переменная измерена в номинальной шкале, а признаки – в количественной. Пропущенных значений признаков нет.

Таблица 6

Анализ сводки обработки наблюдений

Невзвешенные наблюдения		N	Проценты
Валидные		6	66,7
Исключено	Отсутствующие или выходящие за пределы диапазона коды групп	3	33,3
	По крайней мере одна дискриминирующая переменная	0	0,0
	И отсутствующие или выходящие за пределы диапазоны коды групп, и по крайней мере одна дискриминирующая переменная	0	0,0
	Всего	3	33,3
Всего		9	100,0

В данной таблице: валидных объектов, т.е. объектов в эталонных классах 6 (66.7%); исключено 3 объекта (33,3%), которые надо дискриминировать по двум эталонным группам. Таким образом, имеем 9 объектов, что составляет все 100%.

2. Анализ признаков. Поскольку в данном исследовании используется два независимых признака, которые получены после факторного анализа данных базы Евростат, то признаки не требуют проверки на взаимную корреляцию и являются линейно независимыми. Поскольку объектов небольшое количество, то имеется возможность применять линейные функции дискриминации и не проводится проверка на нормальность закона распределения дискриминантных признаков. Нет выбросов.

Таблица 7

Статистика группы

Группа		Среднее	Средне кв. отклонение	N валидных (по списку)	
				Невзвешенные	Взвешенные
1	Инновационное производство	,98897	,0455	2	2
	Инновационное управление	,06151	,0178	2	2
2	Инновационное производство	,30925	,0889	4	4

	Инновационное управление	,62112	,1846	4	4
Всего	Инновационное производство	,53582	,3583	6	6
	Инновационное управление	,43458	,3225	6	6

Приведена статистика по двум эталонным классам (всего 6 объектов). Средние значения признаков двух классов имеют существенные отличия и незначительные среднеквадратические отклонения, а это указывает на то, что объекты по классам сгруппированы компактно и имеется существенная разница между признаками классов. При разнице между средними показателями признаков по классам имеем практически равные общие средние и одинаковые среднеквадратические отклонения, а это указывает, что эталонные два класса имеют противоположные значения по рассматриваемым двум признакам, т.е. в первом классе выше значение первого признака и ниже значение второго, а во втором классе – наоборот. Известно, что без учета закона статистического распределения признаков сопоставление средних будет некорректным, но в данном случае имеем небольшое количество анализируемых объектов, и учесть закон статистического распределения признаков не имеется возможности.

Таблица 8

Критерии равенства групповых средних

	Лямбда Уилкса	F	ст.св.1	ст.св.2	Значимость
Инновационное производство	,040	95,62	1	4	,001
Инновационное управление	,197	16,3	1	4	,016

Показатель статистической значимости дискриминации по F статистике Фишера: наибольший вклад в определение принадлежности объекта к классу имеет признак «Инновационное производство» 95,6% с высоким уровнем значимости 0,001 (99,9%, значительно выше допустимого 95%), а второй признак способствует определению принадлежности к классу на 16,3% и имеет достаточный уровень значимости 0,016 (98,4%). Расчетное значение *F* сравнивалось с

табличным значением $F_{\alpha,k-1,n-k}$ при выбранном уровне значимости α = 0,05 и числе степеней свободы k-1 = 2 – 1 = 1 и n-k = 6 – 2 = 4. Согласно показателям уровней значимости делаем вывод, что различия между классами являются случайными и имеются значимые различия между признаками классов. Доля дисперсии, которая не обусловлена различиями между группами (относительный вклад остаточной дисперсии) для обоих признаков минимальна: для первого признака Λ-статистики Уилкса = 0,04 (4%), а для второго 0,197 (19,7%).

Обе переменные корректно дискриминируют классы.

Проверять показатель эквивалентности ковариационных матриц по критерию Бокса по классам объектов не будем, поскольку имеем всего 6 объектов.

Проанализируем объединенные внутригрупповые матрицы и ковариационные матрицы (общую по всем объектам и отдельно по группам). Данные в таблицах 9 и 10.

Таблица 9

Объединенные внутригрупповые матрицы

		Инновационное производство	Инновационное управление
Ковариация	Инновационное производство	,006	,011
	Инновационное управление	,011	,026
Корреляция	Инновационное производство	1,000	,838
	Инновационное управление	,838	1,000

Количество степеней свободы ковариационной матрицы – 4.

Таблица 10

Ковариационные матрицы

Группа		Инновационное производство	Инновационное управление
1	Инновационное производство	,002	,001
	Инновационное управление	,001	,000
2	Инновационное производство	,008	,014
	Инновационное управление	,014	,034
Всего	Инновационное производство	,128	-,093
	Инновационное управление	-,093	,104

Количество степеней свободы итоговой ковариационной матрицы – 5.

Имеются небольшие различия в ковариационных матрицах по классам, но поскольку объектов мало, то данные показатели можно проигнорировать. Корреляционная матрица показала достаточно большую взаимную корреляцию между имеющимися двумя признаками, но поскольку на предыдущем этапе было установлено, что оба признака являются значимыми, то данный показатель тоже можно проигнорировать.

3. Пошаговый дискриминантный анализ. Изначально была задана стандартная процедура дискриминации, при которой в дискриминантную функцию включаются все имеющиеся признаки, поэтому пошаговый дискриминантный анализ в данном примере не рассматривается.

4. Построение дискриминантной модели. Проанализируем сводку показателей по построенной канонической дискриминантной функции. Поскольку эталонных классов 2, то имеем одну дискриминантную функцию. Не стандартизованные коэффициенты канонической дискриминантной функции приведены в таблице 11.

Таблица 11

Коэффициенты канонической дискриминантной функции

	Функция
	1
Инновационное производство	22,536
Инновационное управление	-10,503
(Константа)	-7,511

Поскольку получены не стандартизированные коэффициенты, то имеем информацию об абсолютном вкладе данного признака в значение дискриминантной функции. Таким образом, имеем каноническую дискриминантную функцию вида:

$$F = 22{,}536 * x_{i1} - 10{,}503 * x_{i2} - 7{,}511$$

Данная функция является оптимальной и отличается от той, которая вычислялась «вручную» в примере, приведенном выше в данном разделе. Однако

решения о принадлежности новых объектов эталонным классам принимаются на основании данной функции так же, как и показано в приведенном выше примере:

– значения дискриминантной функции для двух эталонных классов X и Y равны:

$$F_{x1} = 13{,}538;\ F_{x2} = 14{,}724;$$

$$F_{y1} = -7{,}866;\ F_{y2} = -6{,}181;\ F_{y3} = -6{,}14;\ F_{y4} = -8{,}074.$$

Расчётные показатели полностью совпадают с табличными данными (см. табл. 21):

– среднее арифметическое значение дискриминантной функции отдельно по эталонным классам и общее среднее значение (константой дискриминации):

$$\bar{F}_x = 14{,}1303;$$

$$\bar{F}_y = -7{,}065.$$

Эти данные полностью совпадают с табличными (см. табл. 16 Функции в центроидах групп)

Константа дискриминации $\bar{F} = 3{,}5325$;

– значения дискриминантной функции для новых объектов:

$$F_{z1} = -5{,}879;\ F_{z2} = 1{,}345;\ F_{z3} = 3{,}13;$$

Расчётные показатели полностью совпадают с табличными данными (см. табл. 21);

– распределение трех новых объектов по эталонным классам X и Y:

поскольку, если $F_z < \bar{F}$, то при $\bar{F}_x > \bar{F}_y$ объект относится к классу Y, значит, все три новые объекты относятся ко второму классу Y.

Таким образом, результаты классификации с помощью SPSS подтвердили полученные ранее расчеты.

Коэффициенты стандартизованной канонической дискриминантной функции, которые показывают относительный вклад признака в значение дискриминантной функции, приведены в таблице 12.

Таблица 12

Коэффициенты стандартизованной канонической дискриминантной функции

	Функция
	1
Инновационное производство	1,809
Инновационное управление	-1,681

Поскольку коэффициенты по абсолютной величине для обоих признаков практически равны и достаточно большие, то на включение новых объектов в тот или иной класс данные признаки оказывают практически одинаковое влияние и ни один признак нельзя сократить.

Структурные коэффициенты, которые являются коэффициентами взаимной корреляции между отдельными признаками и дискриминантной функцией, приведены в таблице 13.

Таблица 13

Матрица структуры

	Функция
	1
Инновационное производство	,400
Инновационное управление	-,165

Структурные коэффициенты по абсолютной величине не сильно различаются, то это еще раз подтверждает то, что вся информация о дискриминантной функции заключена в обоих коэффициентах. Поскольку структурные коэффициенты предоставляют более точную информацию о влиянии данного признака на дискриминацию объектов посредством данной дискриминантной функции, то можно сделать вывод, что влияние первого признака на дискриминантную функцию больше второго немногим более, чем в 2 раза.

Для оценки качества полученной дискриминантной функции используются показатели, приведенные в таблицах 14 и 15.

Таблица 14

Собственные значения

Функция	Собственное значение	% дисперсии	Суммарный %	Каноническая корреляция
1	149,759	100,0	100,0	,997

Межгрупповая дисперсия превышает внутригрупповую дисперсию (изменчивость признаков между классами превышает изменчивость признаков внутри классов) в 149,759 раз, что объясняется также и разностью между средними значениями. Значит, полученная функция имеет высокую точность и хорошо дискриминирует классы. При этом дискриминантная функция учитывает 100% дисперсии признаков (дискриминантная функция всегда вычисляется для равной 100% дисперсии зависимой переменной).

Поскольку каноническая корреляция 0,997, то изменчивость дискриминантной функции объясняется разницей между классами на 99,7%, что указывает на высокую ее разделительную способность.

Таблица 15

Лямбда Уилкса

Критерий для функций	Лямбда Уилкса	Хи-квадрат	ст. св.	Значимость
1	,007	15,047	2	,001

Для имеющейся дискриминантной функции доля дисперсии, которая не обусловлена различиями между группами (относительный вклад остаточной дисперсии), очень мала 0,7%, что указывает на высокую значимость функции. Это подтверждается показателем по критерию Пирсона $\chi^2 = 15,047$, который указывает на вероятность того, что различия между классами являются случайными. Расчетное значение χ^2 больше табличного для степени свободы $\nu = (2 – 0) * (2 – 0\text{-}1) = 2$ и дискриминантная функция значима на уровне 0,001 (что намного меньше критического показателя 0,05), а, значит, имеет смысл использовать ее для дальнейшей классификации объектов.

Значения функции, вычисленные для центроидов, приведены в таблице 11. С помощью данных значений определяется принадлежность новых объектов к эталонным классам.

Таблица 16

Функции в центроидах групп (не стандартизованные канонические дискриминантные функции, вычисленные в групповых средних)

Группа	Функция
	1
1	14,131
2	-7,065

Полученные данные совпадают с расчетными.

5. Построение функций классификации. В SPSS в разделе Статистика классификаций приводятся данные по функциям Фишера и результатам распределения объектов по классам.

В таблице 17 приводятся данные по всем классифицированным объектам: всего объектов 9, из них 3 объекта новых, которые надо было дискриминировать по эталонным классам.

Таблица 17

Сводка обработки классификаций

Обработано		9
Исключено	Отсутствующие или выходящие за пределы диапазона коды групп	0
	По крайней мере одна дискриминирующая переменная	0
Использовано в выводе		9

Согласно таблице 17 обработаны все 9 объектов и исключенных переменных (признаков) нет.

При вычислении функций Фишера используются формулы апостериорной вероятности Байеса и поэтому учитываются априорные вероятности, которые могут быть заданы исследователем. Поскольку в данном исследовании априорно вероятность не задавалась, то вероятности распределения объектов в любой их

двух эталонных классов будут одинаковы (по умолчанию) и, значит, равны по 0,5. Суммарная вероятность равна 1. Эти данные находятся в таблице 18.

Таблица 18

Априорные вероятности для групп

Группа	Априорная	Используемые в анализе наблюдения	
		Невзвешенные	Взвешенные
1	,500	2	2
2	,500	4	4
Всего	1,000	6	6

Используя априорную вероятность, были построены линейные дискриминантные функции Фишера (см. таблицу 19), количество которых равно количеству эталонных классов (две функции).

Таблица 19

Коэффициенты функции классификации

	Группа	
	1	2
Инновационное производство	502,825	25,144
Инновационное управление	-208,961	13,671
(Константа)	-242,906	-8,827

Для объектов первого эталонного класса имеем следующую функцию Фишера:

$$d_1 = -242,906 + 502,825 * x_1 - 208,961 * x_2$$

Для второго эталонного класса функция Фишера имеет вид:

$$d_2 = -8,827 + 25,144 * x_1 + 13,671 * x_2$$

Решение о классификации объекта по функциям Фишера принимается следующим образом: рассчитываются все функции Фишера для каждого из объектов и по наибольшему значению принимается решение о принадлежности данного объекта к данному классу. Сделаем расчеты, данные занесем в таблицу, сравним полученную классификацию объектов с результатами SPSS.

Таблица 20

Расчетные показатели по функциям Фишера

Объекты		Функция Фишера для класса 1	Функция Фишера для класса 2	Результат классификации
1	Финляндия	227,96832	15,9	1
2	Швеция	255,07098	17,8612	1
3	Дания	-239,18237	2,42434	2
4	Франция	-201,20496	4,6857	2
5	Австрия	-194,03652	10,997	2
6	Португалия	-234,37008	11,6531	2
7	Словения	-204,21883	-4,72981	2
8	Италия	-29,306951	17,0597	2
9	Ирландия	6,8853178	15,4191	2

Полученные данные по классификации полностью совпадают с расчетами в SPSS.

Статистика по наблюдением приведена в таблице 21.

Таблица 21

Статистика по наблюдениям

Номер наблюдения		Фактическая группа	Наивысшая группа					Вторая по высоте группа			Дискриминантные баллы
			Предсказанная группа	P(D>d | G = g)		P(G = g | D = d)	Квадрат расстояния Махаланобиса до центроида	Группа	P(G = g | D = d)	Квадрат расстояния Махаланобиса до центроида	Функция 1
				PM	ст.св.						
Исходный	1	1	1	,553	1	1,000	,352	2	0,000	424,488	13,538
	2	1	1	,553	1	1,000	,352	2	0,000	474,771	14,724
	3	2	2	,423	1	1,000	,641	1	0,000	483,855	-7,866
	4	2	2	,376	1	1,000	,782	1	0,000	412,565	-6,181
	5	2	2	,355	1	1,000	,855	1	0,000	410,924	-6,140
	6	2	2	,313	1	1,000	1,018	1	0,000	493,065	-8,074
	7	не сгруппировано	2	,235	1	1,000	1,408	1	0,000	400,387	-5,879
	8	не сгруппировано	2	,000	1	1,000	70,738	1	0,000	163,472	1,345
	9	не сгруппировано	2	,000	1	1,000	103,95	1	0,000	121,016	3,130

В таблице приводятся номера фактического класса (группы) для объектов эталонных классов и прогнозируемого класса (предсказанная группа) для всех объектов, включая новые. Если прогнозируемая группа у каких-то эталонных объектов не совпадает с фактической, то это указывается двумя звездочками (**). Данные несовпадения отражаются и в таблице классификации (см. табл. 22) в процентах вероятности сделанного прогноза. При появлении объектов со звездочками рекомендуется данные объекты из эталонных классов изъять, провести повторный дискриминантный анализ с указанием изъятых объектов в качестве новых для дискриминации.

В таблице 21 приведена статистика по определению прогнозируемого класса по апостериорной вероятности Байеса и по квадрату расстояния Махалонибиса до центроида эталонных классов. Данные приведены по двум группам по уровню прогноза: с наиболее вероятным уровнем прогноза и последующая вторая по уровню прогноза группа. Кроме этого, в наивысшей группе приведены данные по условной вероятности принадлежности объекта к группе (G) при данной величине дискриминантной функции (D), то есть значение $P(D > d \mid G = g)$. С учетом априорной вероятности (по умолчанию составляет 0,5 для каждого эталонного класса), получаем апостериорную вероятность $P(G = g \mid D = d)$ наблюдаемого значения дискриминантной функции (D), если задана принадлежность объекта к группе (G). Показатель степени свободы определяется как количество групп, входящих в состав дискриминантного уравнения минус 1. В данном примере дискриминантное уравнение распределяет объекты по двум группам, значит степень свободы равна $2 – 1 = 1$.

Показатели по квадрату расстояния Махалонибиса до центроида эталонных классов также подтверждает правильность классификации (объект относится к тому классу, расстояние до центроида которого меньше).

В последнем столбце таблицы 21 приведены дискриминантные баллы, которые показывают величину канонической дискриминантной функции для соответствующего объекта. Данные значения рассчитаны по канонической

дискриминантной функции и полностью соответствуют значениям, рассчитанным по данным таблицы 11.

Результаты классификации приведены в классификационной матрице в таблице 22.

Таблица 22

Результаты классификации

Группа			Предсказанная принадлежность к группе		Всего
			1	2	
Исходный	Количество	1	2	0	2
		2	0	4	4
		Несгруппированные наблюдения	0	3	3
	%	1	100,0	0,0	100,0
		2	0,0	100,0	100,0
		Несгруппированные наблюдения	0,0	100,0	100,0

В таблице показано, что 100,0% исходных сгруппированных объектов классифицированы правильно. Фактические и прогнозные классы совпали полностью.

6. Анализ полученных результатов. Полученные результаты полностью соответствуют поставленной цели, расширение набора признаков не требуется, неверно классифицированных наблюдений нет.

7. Проверка качества. Для проверки качества дискриминации можно использовать повторный дискриминантный анализ на аналогичной группе объектов или использовать иные подходы к проверке качества. Кроме того, можно оценить показатель эффективности τ:

$$\tau = \frac{6-3}{6-3} = 1, \text{ то есть } 100\%.$$

Таким образом, применение дискриминантных функций позволяет получить 100%-ную эффективность классификации.

Список литературы

1. Айвазян С.А. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков [и др.]. – М.: Финансы и статистика, 1989. – 607 с.

2. Афифи А. Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзенс; пер. с англ. – М.: Мир, 1982. – 488 с.

3. Болч Б. Многомерные статистические методы для экономики / Б. Болч, К.Дж. Хуань; пер. с англ. – М.: Статистика, 1979. – 317 с.

4. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб.: Питер, 2001. – 752 с.

5. Каримов Р.Н. Обработка экспериментальной информации: уч. пособие. Ч. 3: Многомерный анализ. – Саратов, 2000. – 108 с.

6. Кендалл М.Дж. Многомерный статистический анализ и временные ряды / М.Дж. Кендалл, А. Стьюарт; пер. с англ. – М.: Наука; Гл. ред. физ.-мат. лит., 1976. – 736 с.

7. Ким Дж.О. Факторный, дискриминантный и кластерный анализ / Дж.О. Ким, Ч.У. Мьюллер, У.Р. Клекка [и др.]. – М.: Финансы и статистика, 1989. – 215 с.

8. Плис А.И. Практикум по прикладной статистике в среде SPSS: учеб. пособие. В 2-х ч. Ч. 1: Классические процедуры статистики / А.И. Плис, Н.А. Слиеина. – М.: Финансы и статистика, 2004.

9. Статистические методы для ЭВМ / пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1986. – 464 с.

10. Таганов Д.Н. SPSS: Статистический анализ в маркетинговых исследованиях. – СПб.: Питер, 2005.

11. Факторный, дискриминантный и кластерный анализ / Дж. Ким, Ч.У. Мюллер [и др.]; пер. с англ. – М.: Финансы и статистика, 1989. – 215 с.

12. BiihlAchim, Zdfel Peter. SPSS 11. Einfuhrung in die moderne Datenanalyse unter Windows. Munchen: Pearson Studium, 2002.

13. Brosius Felix, SPSS 11. Fundierte Einfuhrung in SPSS und Statik. Bonn: mitp-Verlag, 2002.

14. Jassen Jiirgen, Laatz Wilfried. Statistische Datenanalyse mit SPSS fur Windows. Eine anwendungsorientierte E nfuhrung in das Basissystem und das Modul exakte Tests. Berlin: Springer, 2003.

15. Schmalen Helmut. Grundlagen und Probleme der Betribswirtschaf. 12 Auflage. Stuttgard: Schaffer-Poeschel Verlag, 2002.

16. Wittenberg Reinhard. Datenanalyse mit SPSS fur Windows. Stuttgart: Lucius& Lucius, 2003.

Сизых Дмитрий Сергеевич – канд. техн. наук, доцент кафедры управления информационными системами и цифровой инфраструктурой, Национальный исследовательский университет «Высшая школа экономики», Россия, Москва

Сизых Наталья Васильевна – канд. техн. наук, доцент кафедры управления информационными системами и цифровой инфраструктурой, Национальный исследовательский университет «Высшая школа экономики», Россия, Москва
