Канонический корреляционный анализ

1 Постановка задачи

Обычные коэффициенты корреляции используются для выявления линейной зависимости между двумя признаками X и Y. Если нужно выявить зависимость между признаком X_0 и набором признаков $X_1, \ldots, X_p$, то в качестве характеристики этой зависимости рассматривается множественный коэффициент корреляции, равный коэффициенту корреляции между X_0 и его наилучшим линейным предсказанием

$$\widehat{X_0} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

Этот принцип можно распространить на выявление зависимости между двумя наборами признаков $X_1, \ldots, X_r$ и $X_{r+1}, \ldots, X_{r+s}$. Математическая задача здесь состоит в том, чтобы найти такие нормированные линейные комбинации

$$U_1 = \beta_{10} + \beta_{11} X_1 + \ldots + \beta_{1r} X_r \quad \text{и} \quad V_1 = \alpha_{10} + \alpha_{11} X_{r+1} + \ldots + \alpha_{1s} X_{r+s},$$

чтобы корреляция $R_1 = cor(U_1, V_1)$ между U_1 и V_1 была максимальной. Случайные величины U_1 и V_1, а также их коэффициенты называются каноническими.

2 Метод нахождения канонических коэффициентов

Не умаляя общности, будем считать центрированными компоненты случайного вектора

$$X = (X_1, \ldots, X_r, X_{r+1}, \ldots, X_{r+s})^T$$

$EX_i = 0,\ i = 1, 2, \ldots, r+s$, $U_1 = (X_1, \ldots, X_r)^T, U_2 = (X_{r+1}, \ldots, X_{r+s})^T$. Обозначим через $\Sigma_{11} = EU_1U_1^T$, $\Sigma_{22} = EU_2U_2^T$ и $\Sigma_{12} = EU_1U_2^T$, $\Sigma_{21} = EU_2U_1^T$, $\Sigma_{12} = \Sigma_{21}^T$ ковариационные матрицы подвекторов, образующие общую ковариационную матрицу

$$\Sigma = EXX^T = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Канонические случайные величины обозначим через $V_1 = L^T U_1$, $V_2 = M^T U_2$, где $L = (l_1, \ldots, l_r)^T$, $M = (m_1, \ldots, m_s)^T$ векторы канонических коэффициентов. Поскольку

$$\begin{aligned} DV_1 = EV_1^2 = EL^T U_1 U_1^T L = L^T \Sigma_{11} L = 1, \\ DV_2 = EV_2^2 = EM^T U_2 U_2^T M = M^T \Sigma_{22} M = 1, \\ cov(V_1, V_2) = EV_1 V_2 = EL^T U_1 U_2^T M = L^T \Sigma_{12} M = M^T \Sigma_{21} L, \end{aligned}$$

то экстремальная задача нахождения L и M таких, что

$$cor(V_1, V_2) = cov(V_1, V_2) \ \rightarrow \ max \quad \text{при} \quad DV_1 = DV_2 = 1 \tag{1}$$

сводится к задаче

$$L^T \Sigma_{12} M \ \rightarrow \ max \quad \text{при} \quad L^T \Sigma_{11} L = 1 \text{ и } M^T \Sigma_{22} M = 1. \tag{2}$$

Для решения этой задачи используем метод множителей Лагранжа. Введем функцию Лагранжа

$$\mathcal{L}(L, M) = L^T \Sigma_{12} M - \frac{\lambda_1}{2} \left(L^T \Sigma_{11} L - 1 \right) - \frac{\lambda_2}{2} \left(M^T \Sigma_{22} M - 1 \right)$$

и построим систему нормальных уравнений

$$\begin{cases} \Sigma_{12} M - \lambda_1 \Sigma_{11} L = 0 \\ \Sigma_{21} L - \lambda_2 \Sigma_{22} M = 0 \\ L^T \Sigma_{11} L = 1 \\ M^T \Sigma_{22} M = 1 \end{cases}$$

Умножая первое уравнение на L^T, а второе на M^T, получаем, что

$$\begin{cases} L^T \Sigma_{12} M - \lambda_1 L^T \Sigma_{11} L = 0 \\ M^T \Sigma_{21} L - \lambda_2 M^T \Sigma_{22} M = 0 \\ L^T \Sigma_{11} L = 1 \\ M^T \Sigma_{22} M = 1 \end{cases} \iff \quad \lambda_1 = \lambda_2 = cor(V_1, V_2) = \varrho \,.$$

Итак, имеем систему уравнений:

$$\begin{cases} \Sigma_{12} M - \varrho \Sigma_{11} L = 0 \\ \Sigma_{21} L - \varrho \Sigma_{22} M = 0 \end{cases} . \tag{3}$$

Умножим первое уравнение на $\Sigma_{21} \Sigma_{11}^{-1}$, а второе на ϱ, и сложим уравнения:

$$\begin{cases} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} M - \varrho \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{11} L = 0 \\ \varrho \Sigma_{21} L - \varrho^2 \Sigma_{22} M = 0 \end{cases}$$
$$\implies \quad (\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \varrho^2 \Sigma_{22}) M = 0,$$

откуда получаем, что ϱ^2 является собственным корнем, а M собственным вектором уравнения

$$|\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \varrho^2 \Sigma_{22}| = 0. \tag{4}$$

Аналогично в случае другого множества величин

$$(\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \varrho^2 \Sigma_{11}) L = 0,$$

и ϱ^2 является собственным корнем, а L собственным вектором уравнения

$$|\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \varrho^2\Sigma_{11}| = 0. \tag{5}$$

Количество ненулевых корней равно рангу матрицы Σ_{12} В первом случае количество корней равно рангу матрицы Σ_{22}, то есть s, во втором рангу матрицы Σ_{11}, то есть r. Количество ненулевых корней равно m рангу матрцы Σ_{12}, а кратности нулевых корней различны. Таким образом, имеем $m \leq min(r, s)$ канонических случайных величин $V_{1i} = L_i^T U_1$ и $V_{2i} = M_i^T U_2$, упорядоченных в соответствии с корнями $\varrho_1^2 \geq \varrho_2^2 \geq \ldots \geq \varrho_m^2$. Далее для удобства будем считать канонические корреляции ϱ_i положительными, иначе можно поменять знаки канонических коэффициентов M или L.

3 Свойства канонических случайных величин

$$1)\ cov(V_{1i}, V_{1j}) = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad cov(V_{2i}, V_{2j}) = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}.$$

$$2)\ \ cov(V_{1i}, V_{2j}) = \begin{cases} \varrho_i, & \text{при } i = j \leq m, \\ 0, & \text{при } i \neq j \text{ или } i = j > m \end{cases}$$

Обозначим через A симметричную матрицу $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, M_i и M_j собственные векторы, соответствующие разным собственным числам ϱ_i^2 и ϱ_j^2. В системе

$$\begin{cases} (A - \varrho_i^2\Sigma_{22})M_i = 0 \\ (A - \varrho_j^2\Sigma_{22})M_j = 0 \end{cases}$$

первое уравнение умножим на M_j^T, второе на M_i^T, и уравнения сложим:

$$\begin{cases} M_j^T A M_i - \varrho_i^2 M_j^T \Sigma_{22} M_i = 0 \\ M_i^T A M_j - \varrho_j^2 M_i^T \Sigma_{22} M_j = 0 \end{cases} \quad \Longrightarrow \quad (\varrho_i^2 - \varrho_j^2) M_i^T \Sigma_{22} M_j = 0,$$

откуда $cov(V_{2i}, V_{2j}) = EV_{2i}V_{2j} = EM_i^T U_2 U_2^T M_j = M_i^T \Sigma_{22} M_j = 0$.

Аналогично $cov(V_{1i}, V_{1j}) = EV_{1i}V_{1j} = EL_i^T U_1 U_1^T L_j = L_i^T \Sigma_{11} L_j = 0$.

$$cov(V_{1i}, V_{2j}) = EL_1^T U_1 U_2^T M_j = L_i^T \Sigma_{12} M_j$$

Рассмотрим первое уравнение из системы (3) в случае $\varrho_i \neq 0$ и умножим его обе части на L_i^T:

$$\Sigma_{12} M_i = \varrho_i \Sigma_{11} L_i \quad \Longleftrightarrow \quad L_i^T \Sigma_{12} M_i = \varrho_i L_i^T \Sigma_{11} L_i = \varrho_i,$$

при умножении на L_j^T получим $L_j^T \Sigma_{12} M_i = \varrho_i L_j^T \Sigma_{11} L_i = 0$. Если $\varrho_i = 0$, тогда $\Sigma_{12} M_i = 0 \iff L_i^T \Sigma_{12} M_i = 0$.

Бартлетту принадлежит критерий, основанный на распределении логарифма s сомножителей $\prod_{j=1}^{s}(1-\varrho_j^2)$. Для определенности считаем, что $r \leq s$, иначе группы просто меняются местами. Если предположить, что k канонических корреляций не равны нулю, то статистика критерия для проверки того, что остальные равны нулю, такова:

$$-\left(n-k-1-\frac{r+s+1}{2}+\sum_{j=1}^{r}\varrho_j^2\right)\ln(\prod_{j=k+1}^{r}(1-\varrho_j^2)) \sim \chi^{\mathbf{2}}((r-k)(s-k)). \qquad (6)$$

4 Пример канонических корреляций из исследования факторов, влияющих на антропологию ребенка

4.1 Селекция признаков по матери и ребенку

Для примера рассмотрим данные об антропологических характеристиках новорожденных младенцев у матерей, страдающих избыточным весом.

```
> library(xtable)
> library("CCA")
> data<- read.csv("~/Documents/share/R/lessions/A_canon_corr/dataNegrusha3.csv")
```

	X	Group	ChildWeight	ChildHeight	MotherWeightBefore
1	1	0	3750	52	76.30
2	2	1	3500	54	84.00
3	3	1	3840	54	90.00
4	4	1	4120	54	82.50
5	5	1	3750	53	86.00
6	6	1	3550	53	76.00

	MotherWeightAfter	Skinfold	Matake	LogHDL	LogLDL
1	77	19	29.40	0.78	1.19
2	92	37	33.70	0.98	1.27
3	91	24	37.84	0.72	1.05
4	80	34	43.93	0.03	0.85
5	80	30	26.11	0.11	0.90
6	88	53	39.28	0.06	0.77

По-отдельности рассмотрим признаки, относящиеся к детям и к их матерям, и применим процудуру канонического корреляционного анализа.

```
> library("CCA")
> X<-data[,c(3:4)]
> Y<-data[,c(5:10)]
> CC<-cc(X,Y)
> CC[[2]]$Xnames
```

```
[1] "ChildWeight" "ChildHeight"
```

```
> CC[[2]]$Ynames
```

```
[1] "MotherWeightBefore" "MotherWeightAfter"
[3] "Skinfold"           "Matake"
[5] "LogHDL"             "LogLDL"
```

```
> B<-cbind(X,Y)
> str(B)
```

```
'data.frame':        36 obs. of  8 variables:
 $ ChildWeight       : int  3750 3500 3840 4120 3750 3550 3500 3600 3000 3850 ...
 $ ChildHeight       : int  52 54 54 54 53 53 54 52 49 53 ...
 $ MotherWeightBefore: num  76.3 84 90 82.5 86 76 91 98.2 94 80.7 ...
 $ MotherWeightAfter : int  77 92 91 80 80 88 82 98 94 76 ...
 $ Skinfold          : int  19 37 24 34 30 53 26 34 45 29 ...
 $ Matake            : num  29.4 33.7 37.8 43.9 26.1 ...
 $ LogHDL            : num  0.7839 0.9821 0.7178 0.0296 0.1133 ...
 $ LogLDL            : num  1.191 1.267 1.051 0.85 0.904 ...
```

```
> B<-as.matrix(B)
> (B-mean(B))/sd(B)
```

```
     ChildWeight ChildHeight MotherWeightBefore
[1,]    2.815398  -0.3676131         -0.3466972
[2,]    2.600213  -0.3658917         -0.3400695
[3,]    2.892864  -0.3658917         -0.3349051
[4,]    3.133871  -0.3658917         -0.3413606
[5,]    2.815398  -0.3667524         -0.3383480
[6,]    2.643250  -0.3667524         -0.3469554
[7,]    2.600213  -0.3658917         -0.3340443
[8,]    2.686287  -0.3676131         -0.3278470
```

```
      [,1]        [,2]       [,3]
 [9,] 2.169844 -0.3701953 -0.3314621
[10,] 2.901472 -0.3667524 -0.3429099
[11,] 2.858435 -0.3667524 -0.3273306
[12,] 1.989089 -0.3710561 -0.3245762
[13,] 2.385028 -0.3684739 -0.3344747
[14,] 2.686287 -0.3676131 -0.3383480
[15,] 2.901472 -0.3641702 -0.3397252
[16,] 2.075163 -0.3701953 -0.3417910
[17,] 2.255918 -0.3693346 -0.3383480
[18,] 2.453888 -0.3684739 -0.3451479
[19,] 2.341992 -0.3676131 -0.3340443
[20,] 2.505532 -0.3684739 -0.3495376
[21,] 2.600213 -0.3684739 -0.3161410
[22,] 3.831069 -0.3624487 -0.3245762
[23,] 3.245767 -0.3667524 -0.3301710
[24,] 2.453888 -0.3667524 -0.3245762
[25,] 2.600213 -0.3667524 -0.3187232
[26,] 2.290347 -0.3693346 -0.3082222
[27,] 2.471102 -0.3684739 -0.3286217
[28,] 2.677680 -0.3667524 -0.3073614
[29,] 2.522747 -0.3676131 -0.3228547
[30,] 2.410851 -0.3676131 -0.3314621
[31,] 2.531354 -0.3667524 -0.3194118
[32,] 3.202730 -0.3650309 -0.3503984
[33,] 2.402243 -0.3676131 -0.3686460
[34,] 2.307562 -0.3693346 -0.3482465
[35,] 2.514139 -0.3676131 -0.3614158
[36,] 2.531354 -0.3667524 -0.3615880
      MotherWeightAfter   Skinfold     Matake      LogHDL
 [1,]        -0.3460947 -0.3960175 -0.3870658 -0.4116968
 [2,]        -0.3331836 -0.3805242 -0.3833665 -0.4115262
 [3,]        -0.3340443 -0.3917138 -0.3797995 -0.4117537
 [4,]        -0.3435125 -0.3831064 -0.3745587 -0.4123461
 [5,]        -0.3435125 -0.3865494 -0.3898964 -0.4122740
 [6,]        -0.3366265 -0.3667524 -0.3785630 -0.4123214
 [7,]        -0.3417910 -0.3899923 -0.3827991 -0.4116509
 [8,]        -0.3280192 -0.3831064 -0.3565063 -0.4121933
 [9,]        -0.3314621 -0.3736383 -0.3653009 -0.4121457
[10,]        -0.3469554 -0.3874101 -0.3731919 -0.4122895
[11,]        -0.3280192 -0.3822457 -0.3638406 -0.4119778
[12,]        -0.3271584 -0.3908531 -0.3712258 -0.4122740
[13,]        -0.3538413 -0.3917138 -0.3840778 -0.4119301
[14,]        -0.3202725 -0.3839672 -0.3792663 -0.4120283
[15,]        -0.3323229 -0.3788027 -0.3647062 -0.4118145
[16,]        -0.3417910 -0.3762205 -0.3850241 -0.4119353
```

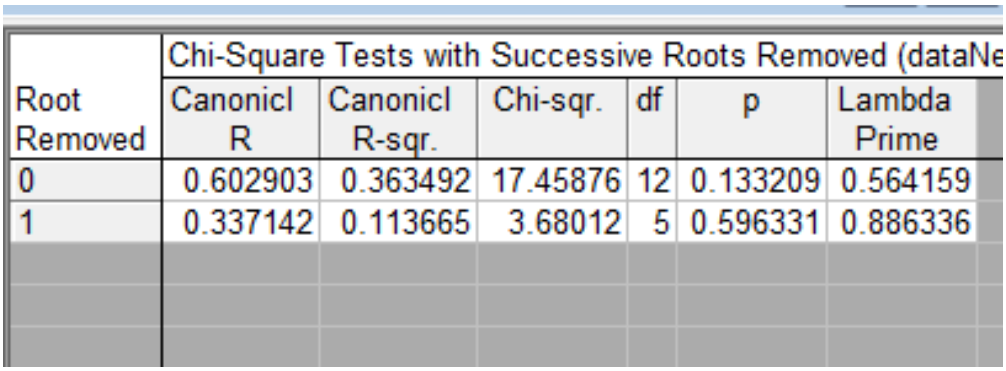
```
[17,]       -0.3349051 -0.3848279 -0.3739512 -0.4119563
[18,]       -0.3426517 -0.3770813 -0.3768909 -0.4117007
[19,]       -0.3366265 -0.3839672 -0.3730856 -0.4118704
[20,]       -0.3529806 -0.3891316 -0.3782432 -0.4119999
[21,]       -0.3090829 -0.3624487 -0.3654232 -0.4119405
[22,]       -0.3219940 -0.3856886 -0.3757815 -0.4118898
[23,]       -0.3280192 -0.3839672 -0.3791224 -0.4115624
[24,]       -0.3245762 -0.3899923 -0.3746066 -0.4118898
[25,]       -0.3219940 -0.3719168 -0.3664158 -0.4116038
[26,]       -0.3065007 -0.3658917 -0.3572292 -0.4119943
[27,]       -0.3219940 -0.3865494 -0.3839180 -0.4115591
[28,]       -0.3176903 -0.3736383 -0.3515241 -0.4119670
[29,]       -0.3142473 -0.3744990 -0.3626513 -0.4122740
[30,]       -0.3133866 -0.3684739 -0.3664718 -0.4123053
[31,]       -0.3159688 -0.3744990 -0.3628943 -0.4119833
[32,]       -0.3435125 -0.3917138 -0.3893529 -0.4119510
[33,]       -0.3693346 -0.4046249 -0.4020132 -0.4120758
[34,]       -0.3392088 -0.3839672 -0.3909355 -0.4122512
[35,]       -0.3607272 -0.3934353 -0.3999950 -0.4117127
[36,]       -0.3650309 -0.3968782 -0.3944362 -0.4117007
          LogLDL
 [1,] -0.4113465
 [2,] -0.4112810
 [3,] -0.4114670
 [4,] -0.4116398
 [5,] -0.4115932
 [6,] -0.4117127
 [7,] -0.4110124
 [8,] -0.4115932
 [9,] -0.4113413
[10,] -0.4112786
[11,] -0.4110580
[12,] -0.4111740
[13,] -0.4114404
[14,] -0.4118011
[15,] -0.4112090
[16,] -0.4110543
[17,] -0.4112001
[18,] -0.4112202
[19,] -0.4111120
[20,] -0.4115359
[21,] -0.4115392
[22,] -0.4116546
[23,] -0.4117453
[24,] -0.4113207
```

```
[25,] -0.4112157
[26,] -0.4114375
[27,] -0.4111826
[28,] -0.4110885
[29,] -0.4114731
[30,] -0.4114375
[31,] -0.4111120
[32,] -0.4116812
[33,] -0.4117007
[34,] -0.4115828
[35,] -0.4110250
[36,] -0.4114761
```

Собираем данные в один вектор, нормируем и центрируем данные и вычисляем ковариационную матрицу. Вычисляем собственные числа и собственные векторы характеристического уравнений (4) и (5).

```
> Step1<-function(X,Y)
+ {
+   if(ncol(X)>ncol(Y)){Z<-X;X<-Y;Y<-Z}
+   r<-ncol(X)
+   s<-ncol(Y)
+ U<-apply(cbind(X,Y),2,function(x)(x-mean(x))/sd(x))
+ n<-nrow(U)
+
+ U1<-U[,seq(r)]
+ U2<-U[,r+seq(s)]
+
+ SS<-cov(U)
+ S11<-SS[seq(r),seq(r)]
+ S12<-SS[seq(r),seq(s)+r]
+ S21<-SS[seq(s)+r,seq(r)]
+ S22<-SS[seq(s)+r,seq(s)+r]
+ A<-solve(S11)%*%S12%*%solve(S22)%*%S21
+ B<-solve(S22)%*%S21%*%solve(S11)%*%S12
+ return(list(param=c(n,r,s),U1=U1,U2=U2,A=A,B=B))
+ }
> st1<-Step1(X,Y)
> r<-st1$param[2]
> s<-st1$param[3]
```

Проверяем значимость коэффициентов согласно (6).



Root Removed	Chi-Square Tests with Successive Roots Removed (dataNe					
	Canonicl R	Canonicl R-sqr.	Chi-sqr.	df	p	Lambda Prime
0	0.602903	0.363492	17.45876	12	0.133209	0.564159
1	0.337142	0.113665	3.68012	5	0.596331	0.886336

Рис. 1: Результаты канонического анализа, выполненного в пакете Statistica.

	r	r2	chi	df	p.value
1	0.60	0.36	17.46	12.00	0.13
2	0.34	0.11	3.60	5.00	0.61

Таблица 1: Проверка значимости канонических корреляций

4.2 Интерпретация канонических величин

Для интерпретации новых канонических величин удобнее всего использовать факторные нагрузки – коэффициенты корреляции между ними с исходными признаками ($CC\$scores\$corr.X.xscores$, $CC\$scores\$corr.Y.yscores$).

Информация о величине и значимости канонических коэффициентов корреляции находится во вкладке

Canonical factos- Chi square tests,

факторные нагрузки

Factor structures - Factor structures and redundancies,

канонические коэффициенты в

Canonical scores - Left and right set canonical weights.

Анализируя зависимость между каноническими величинами и исходными признаками, обнаруживаем линейную зависимость между уменьшением ЛПНП при сроке 25 недель и увеличением массы новорожденного, а также между увеличением роста новорожденного при повышении ЛПВП. При увеличении объема выборки эти корреляции могли бы оказаться значимыми.

4.3 Значения канонических величин

Нормированные и центрированные линейные комбинации канонических весов с признаками совпадают с каноническими величинами $CC\$scores\$xscores$ и $CC\$scores\$yscores$, вычисленными в *cc* (рис.3).

	V1.1	V1.2
ChildWeight	0.92	0.40
ChildHeight	0.57	0.82

	1	2
ChildWeight	-0.92	-0.40
ChildHeight	-0.57	-0.82

Таблица 2: Факторные нагрузки канонической величины левого множества, вычисленные в cc и вручную.

	V2.1	V2.2
MotherWeightBefore	0.07	-0.23
MotherWeightAfter	0.10	-0.18
Skinfold	-0.17	-0.29
Matake	-0.01	-0.10
LogHDL	0.04	0.78
LogLDL	-0.83	0.32

	1	2
MotherWeightBefore	-0.07	0.23
MotherWeightAfter	-0.10	0.18
Skinfold	0.17	0.29
Matake	0.01	0.10
LogHDL	-0.04	-0.78
LogLDL	0.83	-0.32

Таблица 3: Факторные нагрузки правой канонической величины, вычисленные в cc и вручную.

4.4 Селекция признаков по временному фактору

Изменим структуру левого и правого множеств.

```
> X<-data[,c(3:4,6)]
> Y<-data[,c(5,7:10)]
> CC<-cc(X,Y)
```

	1	2	3
ChildWeight	-0.05	-0.93	-0.37
ChildHeight	0.03	-0.61	-0.79
MotherWeightAfter	-1.00	0.04	-0.05

Таблица 4: Факторные нагрузки левого множества

Первая значимая каноническая корреляция указывает на зависимость между весом матери в 25 недель и после родов. Вторая указывает на факт снижения массы новорожденного в зависимости от повышения уровня ЛПНП. Третья практически не значима, но указывает на повышение роста новорожденного в зависимости от повышения уровня ЛПВП.

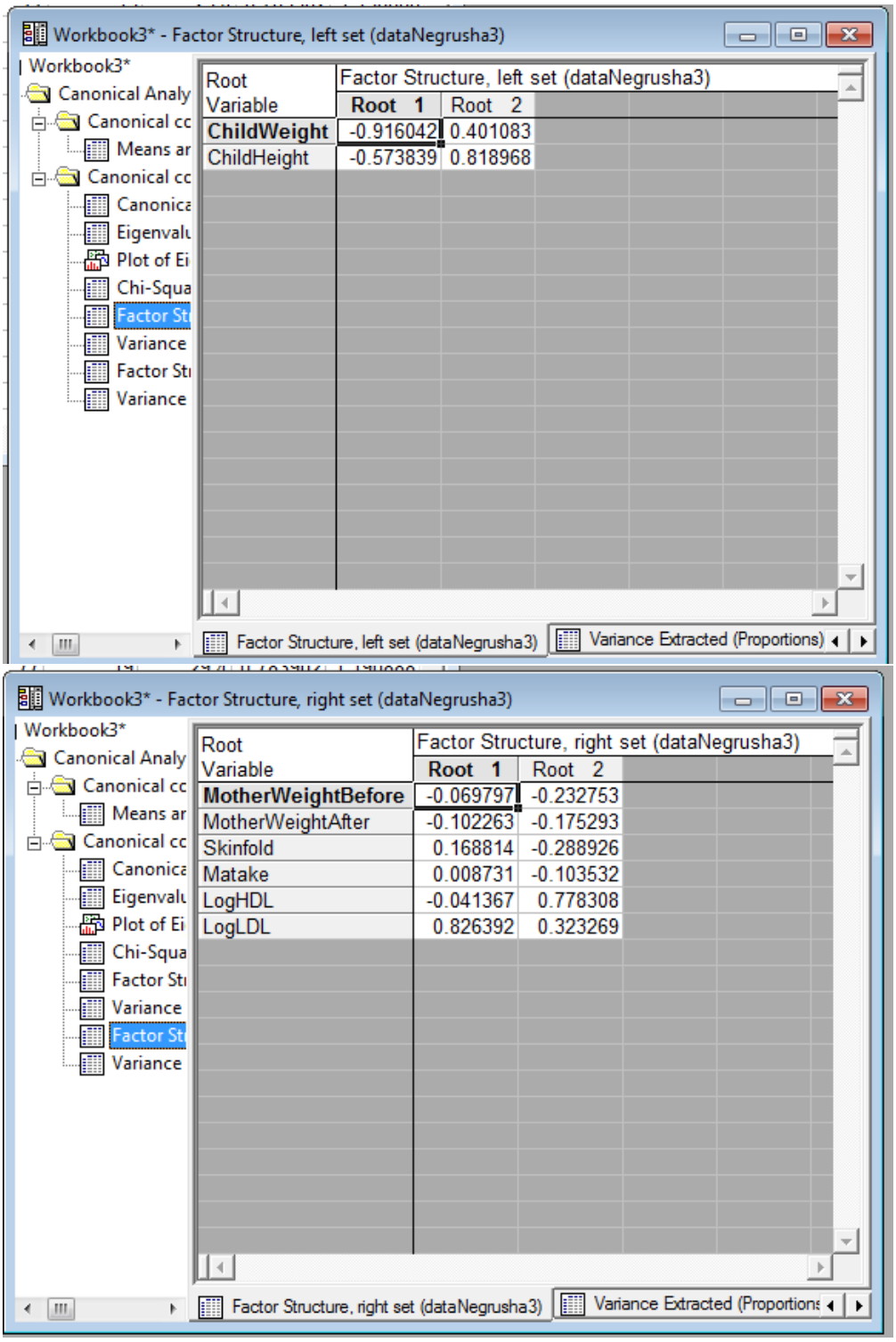


Рис. 2: Анализ весов канонических коэффициентов в пакете Statistica.

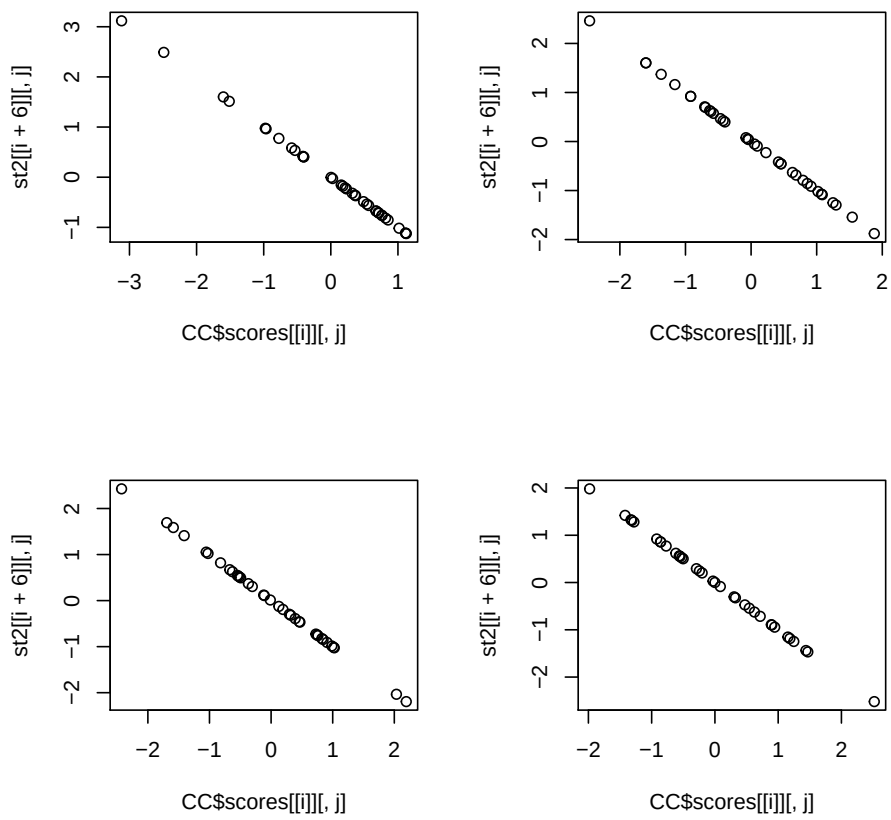


Рис. 3: Идентичность значений канонических величин, вычисленных непосредственно и автоматически.

	1	2	3
MotherWeightBefore	-0.95	0.08	-0.07
Skinfold	-0.79	0.30	0.03
Matake	-0.83	0.14	-0.18
LogHDL	0.11	-0.07	-0.81
LogLDL	-0.02	0.83	-0.43

Таблица 5: Факторные нагрузки правого множества

	r	r2	chi	df	p.value
1	0.93	0.86	76.42	15.00	0.00
2	0.60	0.36	16.52	8.00	0.04
3	0.31	0.10	3.03	3.00	0.39

Таблица 6: Значимость канонических корреляций