

Demokurs

Modul 32741 Vertiefung der Wirtschafts-
mathematik und Statistik

Teil Statistik

15. Juli 2010

gegeben, wobei die Stichproben(ko)varianzen durch

$$S(X, Y) = \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X})(Y_n - \bar{Y}) \quad (4.3)$$

$$S(X)^2 = \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X})^2 = S(X, X) \quad (4.4)$$

$$S(Y)^2 = \frac{1}{N-1} \sum_{n=1}^N (Y_n - \bar{Y})^2 = S(Y, Y) \quad (4.5)$$

definiert sind.

Explizit kann man auch

**Korrelations-
koeffizient
(Stichprobe)**

$$R = \frac{\sum_{n=1}^N (X_n - \bar{X})(Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^N (X_n - \bar{X})^2 \sum_{n=1}^N (Y_n - \bar{Y})^2}} \quad (4.6)$$

schreiben.

4.1.1 Signifikanztest $\rho = 0$

Bei bivariat normalverteilten Merkmalen (Abb. 4.1)

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho \\ \sigma_x \sigma_y \rho & \sigma_y^2 \end{bmatrix} \right). \quad (4.7)$$

ist unter

$$H_0 : \rho = 0 \quad (4.8)$$

$$H_1 : \rho \neq 0 \quad (4.9)$$

**Test des Korrelati-
onskoeffizienten**

die Teststatistik

$$T = \sqrt{N-2} \frac{R}{\sqrt{1-R^2}} \sim t(N-2). \quad (4.10)$$

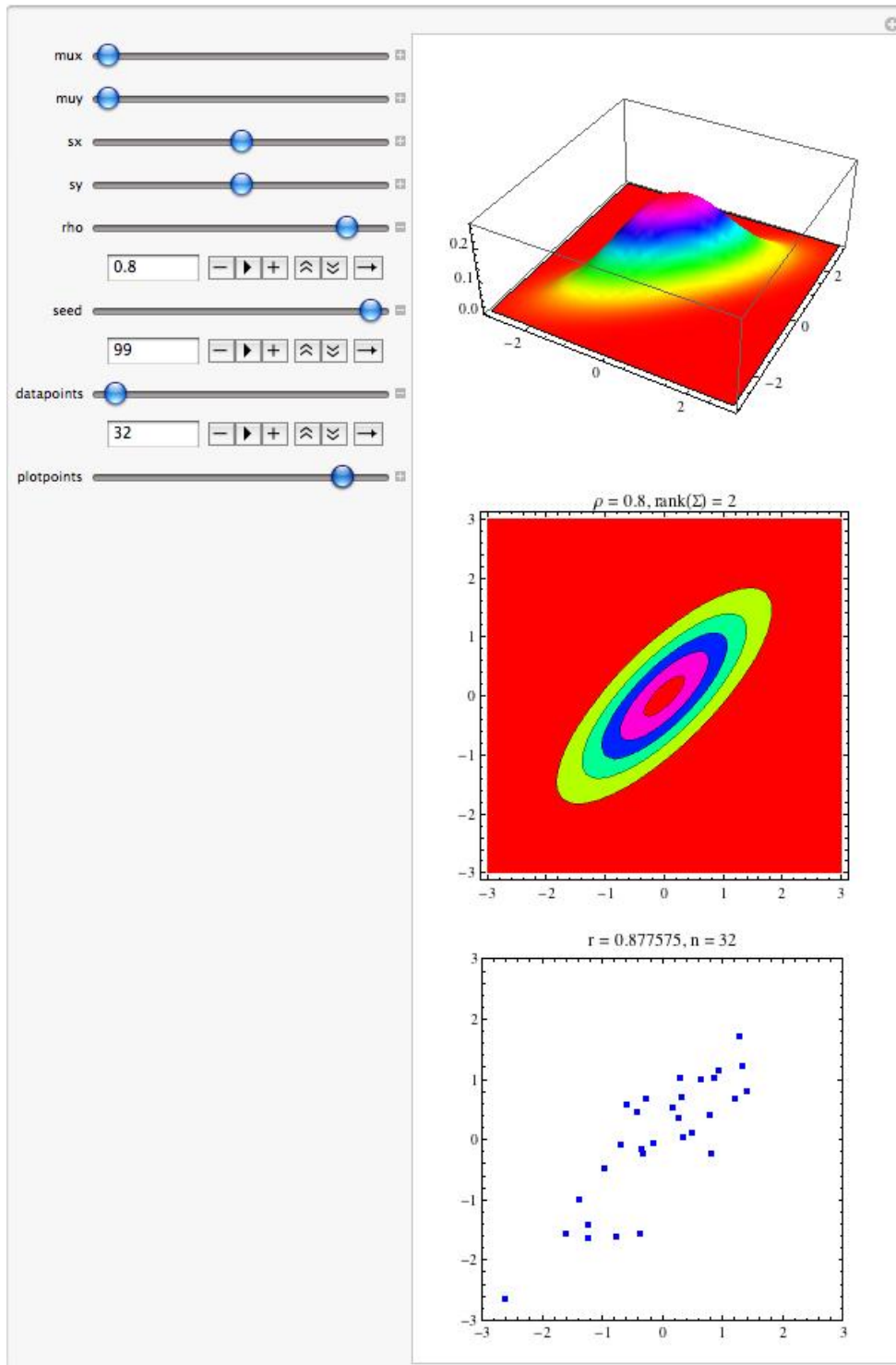


Abbildung 4.1: Bivariate Normalverteilung mit $\rho = 0.8$ und simulierten Daten (Streudiagramm; $r = .877$, $N = 32$, $seed = 99$).

Beispiel 5.3 (BIP 2007 und Inflationsrate (Konfidenzintervall))

Im Beispiel war

$$\begin{aligned}\hat{\alpha} &= -0.827 \\ \hat{\beta} &= 0.601 \\ \hat{\sigma}^2 &= .774 \\ \hat{\sigma} &= \sqrt{.774} = .88 \\ \sum (x_n - \bar{x})^2 &= \sum x_n^2 - N\bar{x}^2 = 107.121.\end{aligned}$$

Daraus findet man den Streuungsterm

$$\begin{aligned}\sqrt{\widehat{\text{Var}}(\hat{E})} &= \hat{\sigma} \sqrt{\frac{1}{N} + \frac{(x - \bar{x})^2}{\sum_n (x_n - \bar{x})^2}} \\ &= 0.88 \sqrt{\frac{1}{32} + \frac{(x - 5.5275)^2}{107.121}}.\end{aligned}$$

Das 95%-KI ist damit

$$(-0.827 + 0.601 x) \pm 1.7972 \sqrt{1/32 + 0.00934(-5.5275 + x)^2}$$

mit dem Quantil $t(.975, 30) = 2.04227$. An der Stelle $x = \bar{x} = 5.5275$ ergibt sich der minimale Wert

$$2.497 \pm 0.318.$$

Abb. 5.5 zeigt das Konfidenz-Intervall als Konfidenz-Band für alle x -Werte zwischen 0 und 11.

■

**5.3.2 Prognoseintervall für individuelles Y_0
(Fall 2)**

Der individuelle zufällige Wert $Y_0 = \alpha + \beta X_0 + \epsilon_0 = E[Y_0|X_0] + \epsilon_0$ kann mit Hilfe von $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$ geschätzt werden. Der dabei zu erwartende quadrierte Prognose-Fehler ist (vgl. 5.91; alles folgende bedingt auf X_0)

$$E[Y_0 - \hat{Y}_0]^2 = \sigma^2 \left(1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum_n (X_n - \bar{X})^2} \right). \quad (5.95)$$

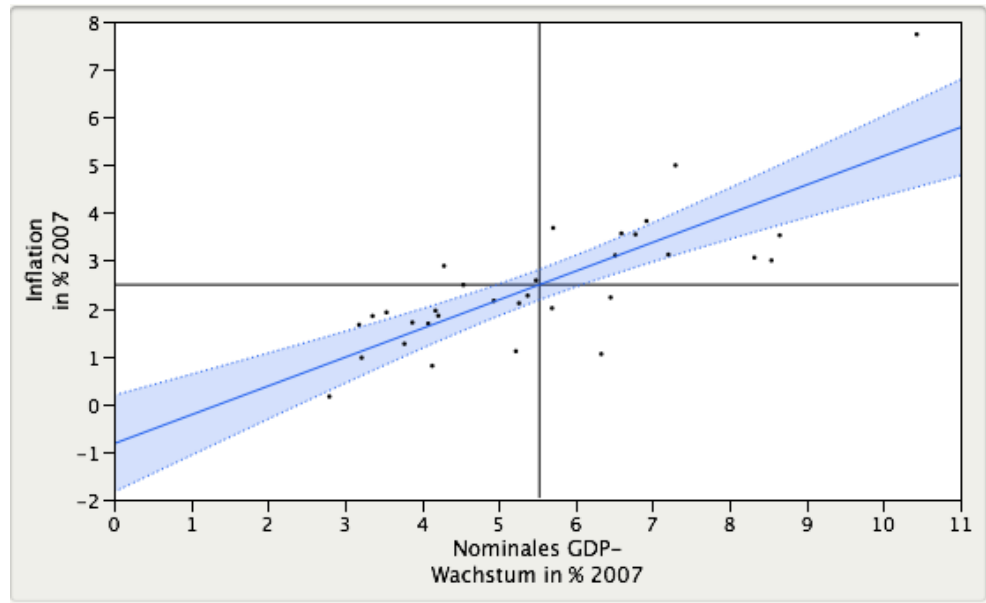


Abbildung 5.5: Geschätzte Gerade mit Konfidenzband $\hat{E}[Y|X] \pm t(1 - \alpha/2, N - 2)\sqrt{\widehat{\text{Var}}(\hat{E})}$.

Man erhält also im Vergleich zu Glg. 5.91 einen zusätzlichen Term σ^2 , der durch den Gleichungsfehler ϵ_0 im stochastischen $Y_0 = \alpha + \beta X_0 + \epsilon_0$ erzeugt wird.

Dies kann wie folgt gezeigt werden (bedingt auf X_0):

Da $E[Y_0] = \alpha + \beta X_0$ und $E[\hat{Y}_0] = \alpha + \beta X_0$ (Erwartungstreue der KQ-Schätzer) gilt

$$E[Y_0 - \hat{Y}_0]^2 = \text{Var}(Y_0 - \hat{Y}_0) = \text{Var}(Y_0) + \text{Var}(\hat{Y}_0). \quad (5.96)$$

Hierbei wurde $\text{Cov}(Y_0, \hat{Y}_0) = \text{Cov}(\alpha + \beta X_0 + \epsilon_0, \hat{\alpha} + \hat{\beta} X_0) = 0$ ausgenutzt, da die KQ-Schätzer unabhängig vom Gleichungsfehler ϵ_0 sind. Außerdem gilt ganz allgemein für Zufallsvariablen $E[Z^2] = \text{Var}(Z)$ für $E[Z] = 0$.

Setzt man noch $\text{Var}(Y_0) = \sigma^2$ und $\text{Var}(\hat{Y}_0) = \sigma^2 \left(\frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum_n (X_n - \bar{X})^2} \right)$ (Fall 1) ein und ersetzt wieder $\sigma \rightarrow \hat{\sigma}$, so ergibt sich das gesuchte Prognoseintervall

Das F -Quantil $f(.95, 1, 30) = 4.17088$ ist wesentlich kleiner als die Teststatistik, sodaß die Nullhypothese verworfen wird. Dies stimmt mit dem Resultat des t -Tests überein (Bsp. 5.2). Die T -Statistik war 7.07611. Quadriert man diese, so ergibt sich $T^2 = 50.0713$, was mit der F -Statistik übereinstimmt. Dies ist kein Zufall, sondern folgt aus dem Zusammenhang $T(N - 2)^2 = N(0, 1)^2/\chi^2(N - 2) = F(1, N - 2)$ der T , χ^2 und F -Statistik (vgl. Abs. 1.3.4.2, Nummer 4).

■

Der F -Bruch läßt sich auch mit Hilfe des Korrelationskoeffizienten ausdrücken, da

$$\frac{SQE}{SQR/(N - 2)} = \frac{SQE}{SQT - SQE}(N - 2) \quad (5.122)$$

$$= \frac{SQE/SQT}{1 - SQE/SQT}(N - 2). \quad (5.123)$$

Somit gilt

$$F = \frac{R_{xy}^2}{1 - R_{xy}^2}(N - 2). \quad (5.124)$$

Große (betragsmäßige) Korrelationen führen also zu großen F -Statistiken.

Im Beispiel ist $r_{xy}^2 = 0.625$ und somit $F = \frac{0.625}{1 - 0.625}30 = 50.07$.

5.4.4 Residualanalyse

Diagnose

Nach dem Schätzen der Parameter und dem Testen des Modells sollte auch eine Analyse der Residuen vorgenommen werden (Diagnose). Hiermit wird überprüft, ob die Annahmen des Modells (vgl. Abs. 5.1.2.2) zumindest approximativ erfüllt sind oder ob grobe Abweichungen vorliegen.

Beispielsweise sollten die Residuen $\hat{\epsilon}_n$ unsystematisch streuen und keine Abhängigkeit von den Regressoren X_n aufweisen. Dies zeigt sich im Streudiagramm Abb. 5.7. Die eingezeichnete Regressionslinie hat nur eine sehr kleine Steigung. Man hat allerdings den Eindruck, daß für große

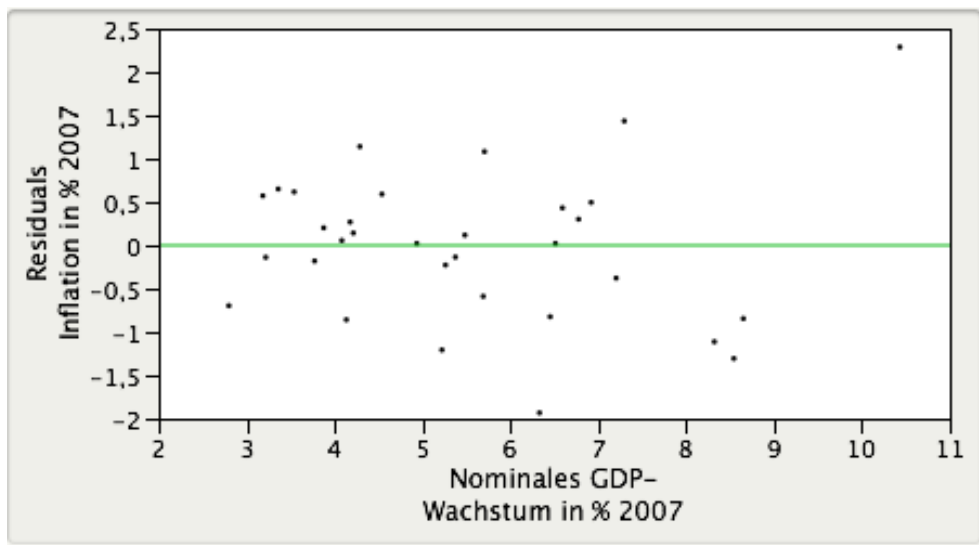


Abbildung 5.7: Streudiagramm der Residuen $\hat{\epsilon}_n$ mit den Regressoren X_n . Eingezeichnet ist auch eine geschätzte Regressionslinie.

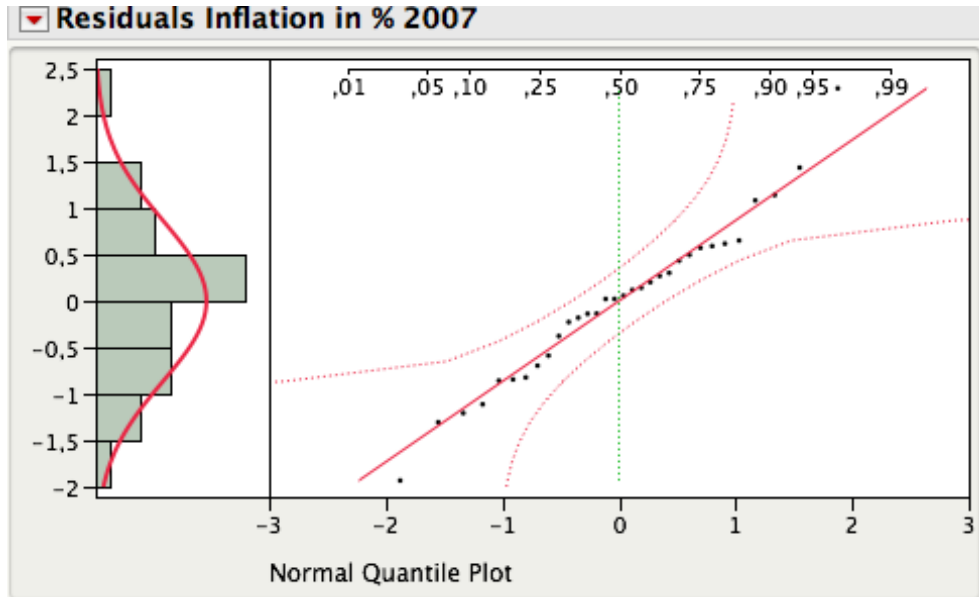


Abbildung 5.8: Histogramm und Normal-Quantil-Plot der Residuen $\hat{\epsilon}_n$. Es sind keine groben Abweichungen von der Normalverteilung zu erkennen.

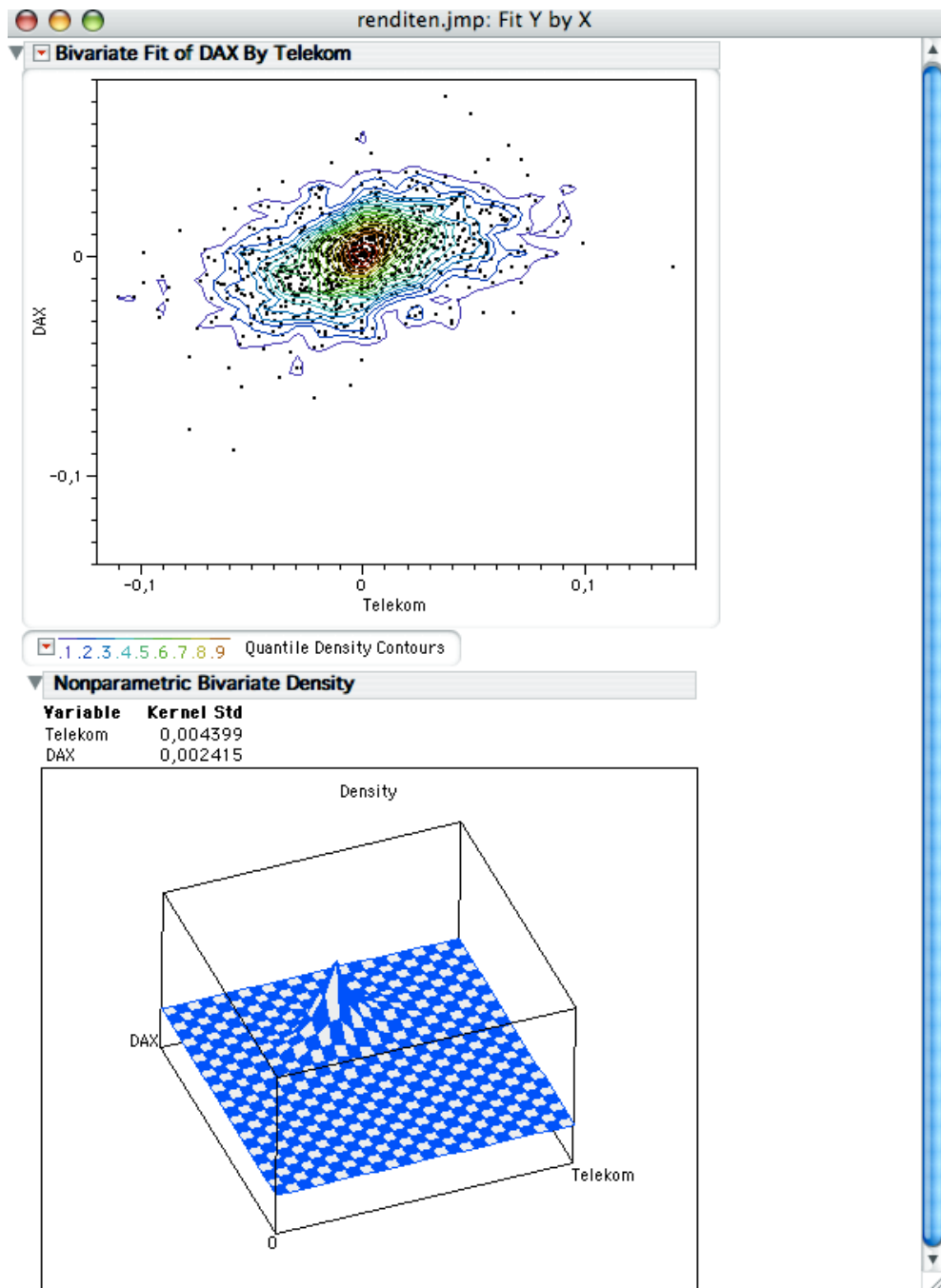


Abbildung 12.20: Streudiagramm, Dichte-Graphik (Höhenlinien) (oben) und 3-D-Darstellung der bivariaten empirischen Dichte (unten) von Dax und Telekom-Rendite (SAS/JMP).

$P = [\psi_1, \dots, \psi_p] : p \times p$, $M = \text{Diag}(\mu_1, \dots, \mu_p) : p \times p$ (Diagonalmatrix).

Die Summendarstellung von Σ wird als **Eigenwertzerlegung** oder **Spektral-Darstellung** bezeichnet. Man spricht auch von **Diagonalisierung** ($P'\Sigma P = M$).

**Eigenwert-
zerlegung**

Die Wichtigkeit dieser Formeln kann gar nicht überschätzt werden. Sie erlauben, eine Matrix als Überlagerung von Projektionen $\psi_i\psi_i'$ auf eindimensionale Unterräume darzustellen, mit den Eigenwerten (Spektrum) als Gewicht.

Ganz allgemein gilt für die Spur (= trace) der Matrix

Spur

$$\sum_i \sigma_{ii} := \text{tr}(\Sigma) = \sum_i \mu_i = \text{tr}(M), \quad (14.318)$$

da $\text{tr}(\Sigma) = \text{tr}(PMP') = \text{tr}(MP'P) = \text{tr}(M)$.

Übung: Beweisen Sie die zyklische Eigenschaft $\text{tr}(AB) = \text{tr}(BA)$ der Spur.

Beispiel 14.9 (Eigenwerte einer Korrelationsmatrix)

Für die (theoretische) Korrelationsmatrix

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad (14.319)$$

ergeben sich die Eigenwerte aus

$$\det \begin{pmatrix} 1 - \mu & \rho \\ \rho & 1 - \mu \end{pmatrix} = 0 = (1 - \mu)^2 - \rho^2 \quad (14.320)$$

$$\mu_{1,2} = 1 \pm \rho \quad (14.321)$$

Die Summe der Eigenwerte ist also $2 = \text{tr}(R) =$ Summe der Diagonale := Spur = trace. Ganz allgemein gilt

$$\sum_i R_{ii} := \text{tr}(R) = \sum_i \mu_i = p. \quad (14.322)$$

Die Eigenvektoren ergeben sich aus den Bedingungen

$$(R - \mu_1 I_2)\psi_1 = \begin{bmatrix} -\rho & \rho \\ \rho & -\rho \end{bmatrix} \begin{bmatrix} \psi_{11} \\ \psi_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (14.323)$$

$$(R - \mu_2 I_2)\psi_2 = \begin{bmatrix} \rho & \rho \\ \rho & \rho \end{bmatrix} \begin{bmatrix} \psi_{21} \\ \psi_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (14.324)$$

Etwa löst

$$\begin{bmatrix} \psi_{11} \\ \psi_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (14.325)$$

$$\begin{bmatrix} \psi_{21} \\ \psi_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (14.326)$$

$$(14.327)$$

obige Gleichungen. Das Betrags-Quadrat der Vektoren ist $[1, 1][1, 1]' = 2$, $[1, -1][1, -1]' = 2$, sodaß man

$$\psi_1 = \begin{bmatrix} \psi_{11} \\ \psi_{12} \end{bmatrix} = 1/\sqrt{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (14.328)$$

$$\psi_2 = \begin{bmatrix} \psi_{21} \\ \psi_{22} \end{bmatrix} = 1/\sqrt{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (14.329)$$

als orthonormierte Eigenvektoren findet.

Übung: Zeigen Sie, daß ψ_1, ψ_2 orthonormiert sind.

Es ist wichtig, daß die Eigenvektoren gar nicht von der Korrelation ρ abhängen. Sie zeigen in Richtung der Winkelhalbierenden der Quadranten. Abb. 14.29 zeigt simulierte Daten aus einer bivariaten Normalverteilung

$$N\left(0, R = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right). \quad (14.330)$$

Die Eigenwerte von R sind $1 \pm 0.9 = 1.9, 0.1$ und die orthogonale Matrix der Eigenvektoren lautet

$$P = 2^{-1/2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (14.331)$$

$$PP' = P'P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2. \quad (14.332)$$

Im gedrehten Koordinatensystem gilt daher

$$\mathbf{y} = P'\mathbf{x} = \begin{bmatrix} \psi'_1 \mathbf{x} \\ \psi'_2 \mathbf{x} \end{bmatrix} = 2^{-1/2} \begin{bmatrix} x_1 + x_2 \\ x_1 - x_2 \end{bmatrix} \quad (14.333)$$

und $\text{Cov}(\mathbf{y}) = P'RP = M = \text{diag}(1.9, 0.1)$.

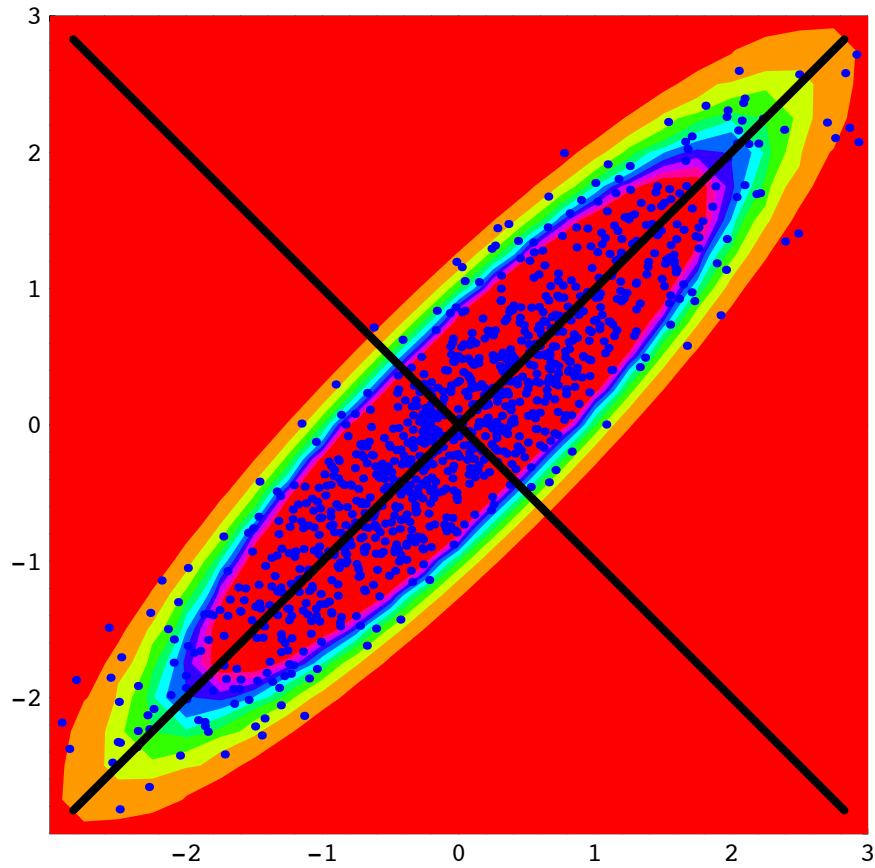


Abbildung 14.29: Simulierte normalverteilte Daten $\mathbf{x}_n, n = 1, \dots, N = 1000$ mit Kovarianz-Matrix $R = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$. Die Hauptachsen zeigen in Richtung der Winkelhalbierenden.

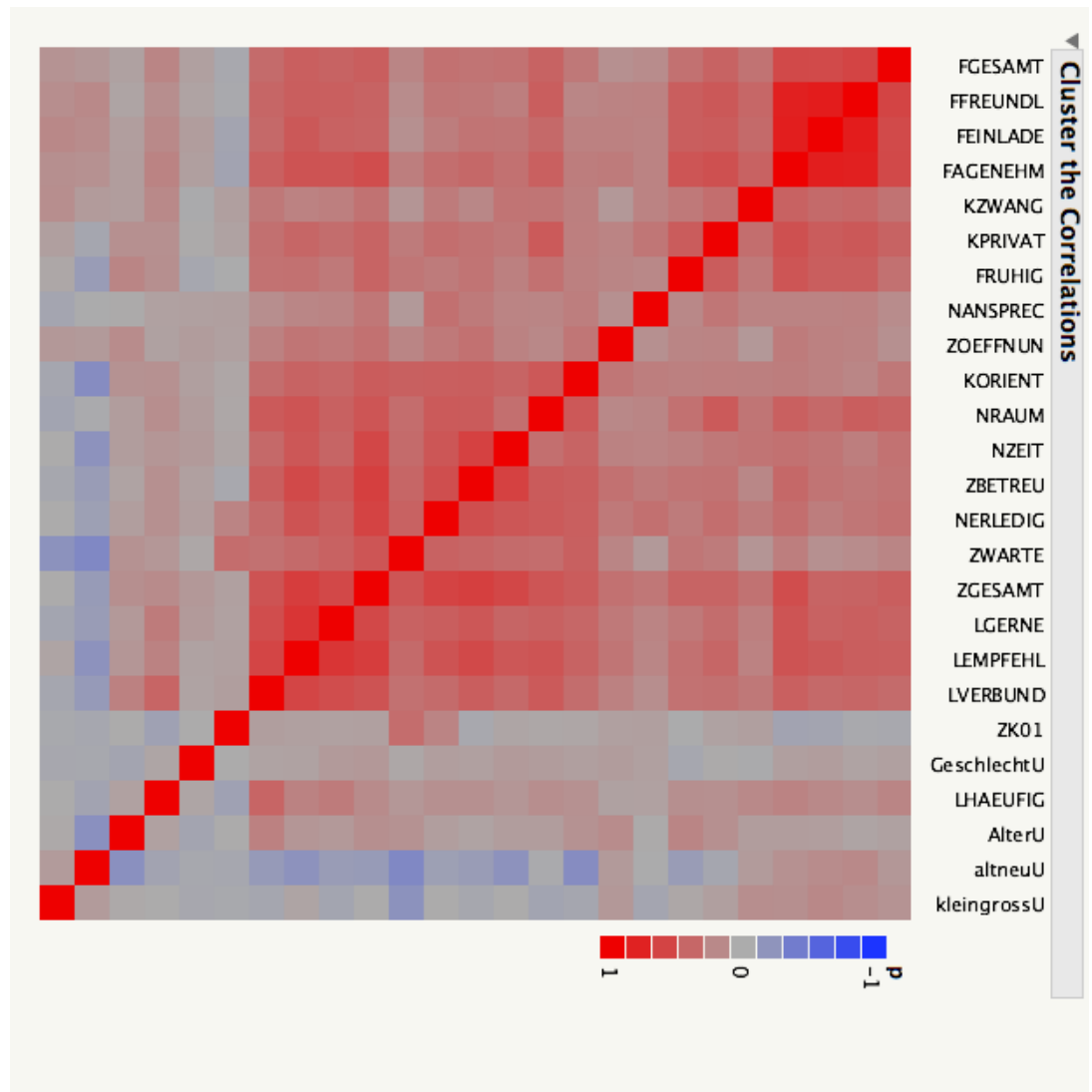


Abbildung 15.12: JMP: Korrelationsmatrix aller Variablen (Cluster der Korrelationen). Die Stärke der Korrelation ist durch die Farbe markiert (rot: $r > 0$, blau: $r < 0$). Die items der Konstrukte bilden einen positiv korrelierten Block.