Multivariate Datenanalyse

MSc Psychologie WiSe 2022/23

Prof. Dr. Dirk Ostwald

(6) Multivariate Normalverteilungen

Anwendungsfälle

Allgemeines Lineares Modell

- T-Tests, ANOVA, einfache und multiple Regression

Multivariates Allgemeines Lineares Modell

- T^2-Tests, MANOVA, einfache und multiple multivariate Regression

Hierarchische Modelle

- Linear Mixed Models, Bayesianische Regression, Diskriminanzanalyse

Lineare Normalverteilungsmodelle

- Probabilistische Hauptkomponentenanalyse, Faktorenanalyse

Kalman Filter, Gaussian Processes, Gaussian Random Fields, ...

Konstruktion und Definition

Marginale Normalverteilungen

Gemeinsame Normalverteilungen

Bedingte Normalverteilungen

Selbstkontrollfragen

Definition (Normalverteilte Zufallsvariable)

ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1}$$

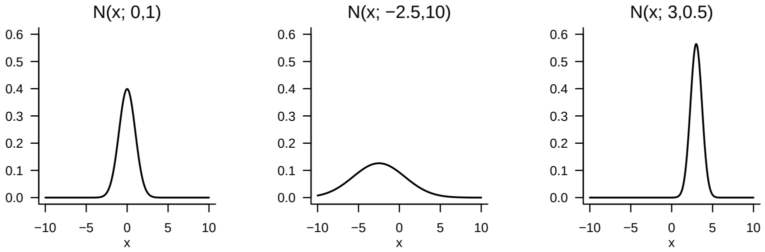
Dann sagen wir, dass ξ einer *Normalverteilung (oder Gauß-Verteilung) mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$* unterliegt und nennen ξ eine *normalverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim N(\mu, \sigma^2)$ ab. Die WDF einer normalverteilten Zufallsvariable bezeichnen wir mit

$$N\left(x; \mu, \sigma^2\right) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{2}$$

Bemerkungen

- Es gelten $\mathbb{E}(\xi) = \mu$ und $\mathbb{V}(\xi) = \sigma^2$.
- Der Parameter μ entspricht dem Wert höchster Wahrscheinlichkeitsdichte.
- Der Parameter σ^2 spezifiziert die Breite der WDF.
- $\xi \sim N(0, 1)$ heißt auch *standardnormalverteilt*.

Visualisierung univariater Normalverteilungsdichtefunktionen



Konstruktion und Definition

Theorem (Konstruktion bivariater Normalverteilungen)

$\zeta_1 \sim N(0,1)$ und $\zeta_2 \sim N(0,1)$ seien zwei unabhängige standardnormalverteilte Zufallsvariablen. Weiterhin seien $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$ und $\rho \in]-1,1[$. Schließlich seien

$$\begin{aligned} \xi_1 &:= \sigma_1\zeta_1 + \mu_1 \\ \xi_2 &:= \sigma_2\left(\rho\zeta_1 + (1-\rho^2)^{1/2}\zeta_2\right) + \mu_2. \end{aligned} \tag{3}$$

Dann hat die WDF des Zufallsvektors $\xi := (\xi_1, \xi_2)^T$, also der gemeinsamen Verteilung von ξ_1 und ξ_2, die Form

$$p : \mathbb{R}^2 \to \mathbb{R}_{>0},\ x \mapsto p(x) := (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \tag{4}$$

wobei

$$m := 2, \mu := \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} \text{ und } \Sigma := \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2\end{pmatrix} \tag{5}$$

Bemerkungen

- Für einen Beweis siehe DeGroot and Schervish (2012), S. 338-339.
- Man nennt die gemeinsame Verteilung von ξ_1 und ξ_2 *bivariate Normalverteilung*.

Konstruktion bivariater Normalverteilungen

```
# Parameterdefinitionen
mu_1   = 5.0                                                # \mu_1
mu_2   = 4.0                                                # \mu_2
sig_1  = 1.5                                                # \sigma_1
sig_2  = 1.0                                                # \sigma_2
rho    = 0.9                                                # \rho

# Realisierungen der standardnormalverteilten ZVen
n      = 100                                                # Anzahl Realisierungen
zeta_1 = rnorm(n)                                           # \zeta_1 \sim N(0,1)
zeta_2 = rnorm(n)                                           # \zeta_1 \sim N(0,1)

# Evaluation von Realisierungen von \xi_1 und \xi_2
xi_1   = sig_1*zeta_1 + mu_1                                # Realsierungen von zeta_1
xi_2   = sig_2*(rho*zeta_1 + sqrt(1-rho^2)*zeta_2) + mu_2 # Realsierungen von zeta_2


# Parameter der gemeinsamen Verteilung von \xi_1 und \xi_2
mu     = matrix(c(mu_1,                                     # \mu \in \mathbb{R}^2
                  mu_2),
                 nrow = 2, byrow = TRUE)
Sigma  = matrix(c(sig_1^2        , rho*sig_1*sig_2,       # \Sigma \in \mathbb{R}^{2 x 2}
                 rho*sig_1*sig_2, sig_2^2),
                 nrow = 2, byrow = TRUE)
print(mu)
```

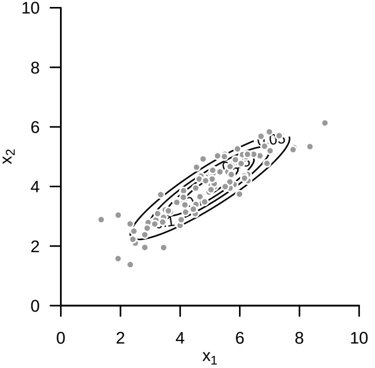
```
>      [,1]
> [1,]    5
> [2,]    4
```

```
print(Sigma)
```

```
>      [,1] [,2]
> [1,] 2.25 1.35
> [2,] 1.35 1.00
```

Konstruktion bivariater Normalverteilungen

- Realisierungen von $\xi = (\xi_1, \xi_2)$, $-$ Isokonturen von p



Konstruktion und Definition

Definition (Multivariate Normalverteilung)

ξ sei ein m-dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^m$ und WDF

$$p : \mathbb{R}^m \to \mathbb{R}_{>0},\ x \mapsto p(x) := (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \quad (6)$$

Dann sagen wird, dass ξ einer *multivariaten (oder m-dimensionalen) Normalverteilung* mit *Erwartungswertparameter* $\mu \in \mathbb{R}^m$ und positive-definitem *Kovarianzmatrixparameter* $\Sigma \in \mathbb{R}^{m \times m}$ unterliegt und nennen ξ einen *(multivariat) normalverteilten Zufallsvektor*. Wir kürzen dies mit $\xi \sim N(\mu, \Sigma)$ ab. Die WDF eines multivariat normalverteilten Zufallsvektors bezeichnen wir mit

$$N(x; \mu, \Sigma) := (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \quad (7)$$

Bemerkungen

- Es gelten $\mathbb{E}(\xi) = \mu$ und $\mathbb{C}(\xi) = \Sigma$.
- Der Parameter $\mu \in \mathbb{R}^m$ entspricht dem Wert höchster Wahrscheinlichkeitsdichte
- Die Diagonalelemente von Σ spezifizieren die Breite der WDF bezüglich $\xi_1, ..., \xi_m$.
- Das i, jte Element von Σ spezifiziert die Kovarianz on ξ_i und ξ_j.
- Der Term $(2\pi)^{-m/2}|\Sigma|^{-1/2}$ ist die Normalisierungskonstante für den Exponentialfunktionsterm.

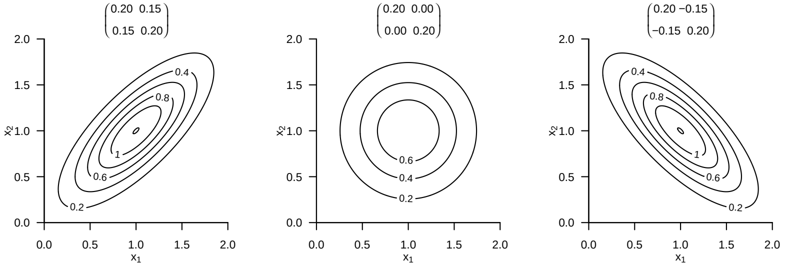
Konstruktion und Definition

Visualisierung bivariater Normalverteilungsdichtefunktionen

```
# multivariate Normalverteilungstools
# install.packages("mvtnorm")
library(mvtnorm)


# Ergebnisraumdefintion
x_min  = 0                                          # x_i Minimum
x_max  = 2                                          # x_i Maxim
x_res  = 1e3                                        # x_i Auflösung
x_1    = seq(x_min, x_max, length.out = x_res)      # x_1 Raum
x_2    = seq(x_min, x_max, length.out = x_res)      # x_2 Raum
X      = expand.grid(x_1,x_2)                       # X = (x_1,x_2)^T Raum

# Parameterdefinition
mu     = c(1,1)                                     # \mu \in \mathbb{R}^2
S      = list(matrix(c(0.2,  0.15,  0.15, 0.2), 2), # \Sigma in \mathbb{R}^{2 \times 2}
              matrix(c(0.2,  0.00,  0.00, 0.2), 2), # \Sigma in \mathbb{R}^{2 \times 2}
              matrix(c(0.2, -0.15, -0.15, 0.2), 2)) # \Sigma in \mathbb{R}^{2 \times 2}

# Kovarianzparametervariantenschleife
for (Sigma in S){

  # Wahrscheinlichkeitsdichtefunktionauswertung
  p      = matrix(                                  # Matrixkonversion des von
                dmvnorm(as.matrix(X), mu, Sigma),   # dmvnorm() ausgegebenen Vektors
                nrow = x_res)

  # Visualisierung
  contour(
  x_1,
  x_2,
  p,
  xlim      =  c(x_min,x_max),
  ylim      =  c(x_min,x_max),
  nlevels   = 5)}
```

Konstruktion und Definition

Visualisierung bivariater Normalverteilungsdichtefunktionen



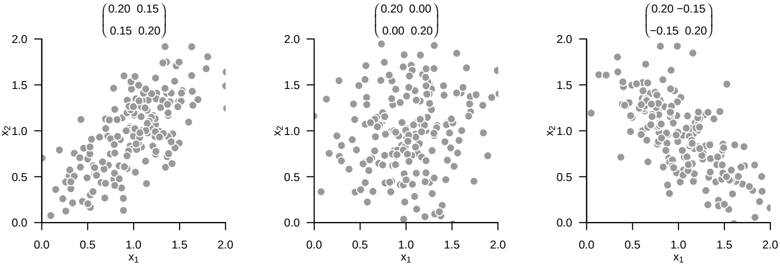
Konstruktion und Definition

Realisierung bivariater normalverteilter Zufallsvektoren

```r
# R Paket für multivariate Normalverteilungen
library(mvtnorm)

# Parameterdefinition
mu     = c(1,1)                                # \mu \in \mathbb{R}^2
Sigma  = matrix(c(0.2,  0.15,  0.15, 0.2), 2)  # \Sigma in \mathbb{R}^{2 \times 2}

# Zufallsvektorrealisierungen
rmvnorm(n = 10, mu, Sigma)
```

```
>        [,1]  [,2]
>  [1,] 1.552 0.975
>  [2,] 0.795 1.238
>  [3,] 0.172 0.242
>  [4,] 1.644 1.972
>  [5,] 0.449 0.221
>  [6,] 0.875 0.730
>  [7,] 0.639 0.682
>  [8,] 1.655 1.502
>  [9,] 0.887 0.381
> [10,] 1.405 1.255
```

Realisierung bivariater normalverteilter Zufallsvektoren



Konstruktion und Definition

Marginale Normalverteilungen

Gemeinsame Normalverteilungen

Bedingte Normalverteilungen

Selbstkontrollfragen

Marginale Normalverteilungen

Theorem (Marginale Normalverteilungen)

Es sei $m := k + l$ und $\xi = (\xi_1, ..., \xi_m)^T$ sei ein m-dimensionaler normalverteilter Zufallsvektor mit Erwartungswertparameter

$$\mu = \begin{pmatrix} \mu_\upsilon \\ \mu_\zeta \end{pmatrix} \in \mathbb{R}^m, \tag{8}$$

mit $\mu_\upsilon \in \mathbb{R}^k$ and $\mu_\zeta \in \mathbb{R}^l$ und Kovarianzmatrixparameter

$$\Sigma = \begin{pmatrix} \Sigma_{\upsilon\upsilon} & \Sigma_{\upsilon\zeta} \\ \Sigma_{\zeta\upsilon} & \Sigma_{\zeta\zeta} \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{9}$$

mit $\Sigma_{\upsilon\upsilon} \in \mathbb{R}^{k \times k}$, $\Sigma_{\upsilon\zeta} \in \mathbb{R}^{k \times l}$, $\Sigma_{\zeta\upsilon} \in \mathbb{R}^{l \times k}$, und $\Sigma_{\zeta\zeta} \in \mathbb{R}^{l \times l}$. Dann sind $\upsilon := (\xi_1, ..., \xi_k)^T$ und $\zeta := (\xi_{k+1}, ..., \xi_m)^T$ k- und l-dimensionale normalverteilte Zufallsvektoren, respektive, und es gilt

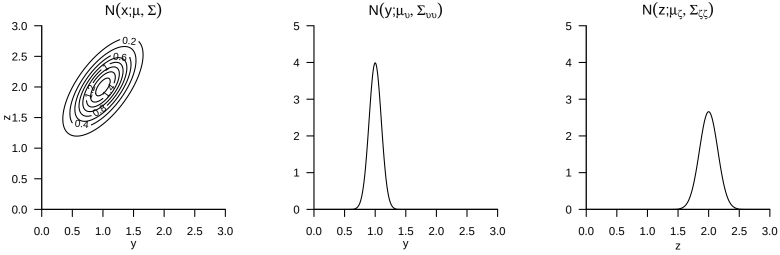
$$\upsilon \sim N(\mu_\upsilon, \Sigma_{\upsilon\upsilon}) \text{ and } \zeta \sim N(\mu_\zeta, \Sigma_{\xi\xi}), \tag{10}$$

Bemerkungen

- Die Marginalverteilungen einer multivariaten Normalverteilung sind auch Normalverteilungen.
- Die Parameter der Marginalverteilungen ergeben sich aus den Parametern der gemeinsamen Verteilung.
- Für Beweise, siehe z.B. Mardia, Kent, and Bibby (1979), Kapitel 3 oder Anderson (2003), Kapitel.

Marginale Normalverteilungen

$$m := 2, k = 1, l = 1, \mu := \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \Sigma := \begin{pmatrix} 0.10 & 0.08 \\ 0.08 & 0.15 \end{pmatrix}$$



Konstruktion und Definition

Marginale Normalverteilungen

Gemeinsame Normalverteilungen

Bedingte Normalverteilungen

Selbstkontrollfragen

Theorem (Gemeinsame Normalverteilungen)

ξ sei ein m-dimensionaler normalverteilter Zufallsvektor mit WDF

$$p_\xi : \mathbb{R}^m \to \mathbb{R}_{>0},\ x \mapsto p_\xi(x) := N(x; \mu_\xi, \Sigma_{\xi\xi}) \text{ mit } \mu_\xi \in \mathbb{R}^m, \Sigma_{\xi\xi} \in \mathbb{R}^{m\times m}, \tag{11}$$

$A \in \mathbb{R}^{n\times m}$ sei eine Matrix, $b \in \mathbb{R}^n$ sei ein Vektor und υ sei ein n-dimensionaler bedingt normalverteilter Zufallsvektor mit bedingter WDF

$$p_{\upsilon|\xi}(\cdot|x) : \mathbb{R}^n \to \mathbb{R}_{>0},\ y \mapsto p_{\upsilon|\xi}(y|x) := N(y; A\xi + b, \Sigma_{\upsilon\upsilon}) \text{ mit } \Sigma_{\upsilon\upsilon} \in \mathbb{R}^{n\times n}. \tag{12}$$

Dann ist der $m+n$-dimensionale Zufallsvektor $(\xi, \upsilon)^T$ normalverteilt mit (gemeinsamer) WDF

$$p_{\xi,\upsilon} : \mathbb{R}^{m+n} \to \mathbb{R}_{>0},\ \begin{pmatrix} x \\ y \end{pmatrix} \mapsto p_{\xi,\upsilon}\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \mu_{\xi,\upsilon}, \Sigma_{\xi,\upsilon}\right), \tag{13}$$

mit $\mu_{\xi,\upsilon} \in \mathbb{R}^{m+n}$ and $\Sigma_{\xi,\upsilon} \in \mathbb{R}^{m+n\times m+n}$ und insbesondere

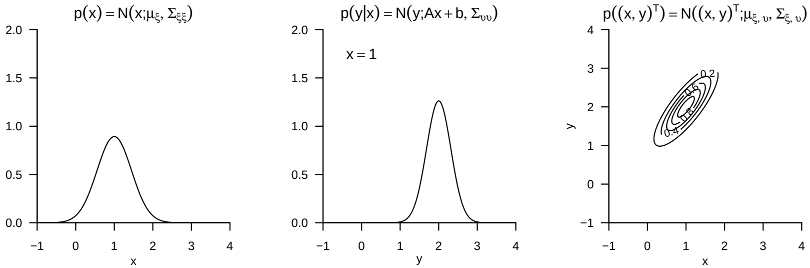
$$\mu_{\xi,\upsilon} = \begin{pmatrix} \mu_\xi \\ A\mu_\xi + b \end{pmatrix} \text{ und } \Sigma_{\xi,\upsilon} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\xi}A^T \\ A\Sigma_{\xi\xi} & \Sigma_{\upsilon\upsilon} + A\Sigma_{\xi\xi}A^T \end{pmatrix}. \tag{14}$$

Bemerkungen

- Eine marginale und eine bedingte multivariate Normalverteilung induzieren eine gemeinsame Normalverteilung.
- Die Parameter der gemeinsamen Verteilungen ergeben sich als linear-affine Transformation der Parameter der induzierenden Verteilungen.

Gemeinsame Normalverteilungen

$$m := 1, n := 1, \mu_\xi := 1, \Sigma_{\xi\xi} := 0.2, A := 1, b := 1, \Sigma_{\upsilon\upsilon} := 0.1$$



Konstruktion und Definition

Marginale Normalverteilungen

Gemeinsame Normalverteilungen

Bedingte Normalverteilungen

Selbstkontrollfragen

Theorem (Bedingte Normalverteilungen)

(ξ, υ) sei ein $m+n$-dimensionaler normalverteilter Zufallsvektor mit WDF

$$p_{\xi,\upsilon} : \mathbb{R}^{m+n} \to \mathbb{R}_{>0}, (x, y) \mapsto p_{\xi,\upsilon}(x, y) := N\left((x, y); \mu_{\xi,\upsilon}, \Sigma_{\xi,\upsilon}\right), \tag{15}$$

mit

$$\mu_{\xi,\upsilon} = \begin{pmatrix} \mu_\xi \\ \mu_\upsilon \end{pmatrix}, \Sigma_{\xi,\upsilon} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\upsilon} \\ \Sigma_{\upsilon\xi} & \Sigma_{\upsilon\upsilon} \end{pmatrix}, \tag{16}$$

mit $x, \mu_\xi \in \mathbb{R}^m, y, \mu_\upsilon \in \mathbb{R}^n$ and $\Sigma_{\xi\xi} \in \mathbb{R}^{m\times m}, \Sigma_{\xi\upsilon} \in \mathbb{R}^{m\times n}, \Sigma_{\upsilon\upsilon} \in \mathbb{R}^{n\times n}$. Dann ist die bedingte Verteilung von ξ gegeben υ eine m-dimensionale Normalverteilung mit bedingter WDF

$$p_{\xi|\upsilon}(\cdot|y) : \mathbb{R}^m \to \mathbb{R}_{>0}, x \mapsto p_{\xi|\upsilon}(x|y) := N(x; \mu_{\xi|\upsilon}, \Sigma_{\xi|\upsilon}) \tag{17}$$

mit

$$\mu_{\xi|\upsilon} = \mu_\xi + \Sigma_{\xi\upsilon}\Sigma_{\upsilon\upsilon}^{-1}(\upsilon - \mu_\upsilon) \in \mathbb{R}^m \tag{18}$$

und

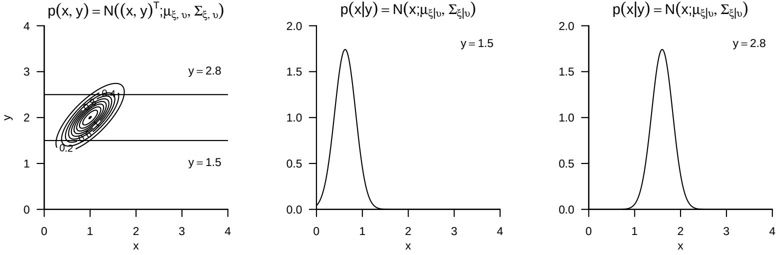
$$\Sigma_{\xi|\upsilon} = \Sigma_{\xi\xi} - \Sigma_{\xi\upsilon}\Sigma_{\upsilon\upsilon}^{-1}\Sigma_{\upsilon\xi} \in \mathbb{R}^{m\times m}. \tag{19}$$

Bemerkungen

- Die Parameter einer bedingten (multivariaten) Normalverteilung ergeben sich aus den Parametern einer gemeinsamen multivariaten Normalverteilung. Im Zusammenspiel mit den vorherigen Theoremen können die Parameter bedingter und marginale Normalverteilungen aus den Parametern der komplementären bedingten und marginalen Normalverteilungen bestimmt werden.

Bedingte Normalverteilungen

$$m := 2, n := 1, \mu := \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \Sigma := \begin{pmatrix} 0.12 & 0.09 \\ 0.09 & 0.12 \end{pmatrix}$$



Konstruktion und Definition

Marginale Normalverteilungen

Gemeinsame Normalverteilungen

Bedingte Normalverteilungen

Selbstkontrollfragen

Selbstkontrollfragen

1. Definieren Sie die WDF einer univariaten normalverteilten Zufallsvariable.
2. Geben Sie das Theorem zur Konstruktion bivariater Normalverteilungen wieder.
3. Definieren Sie die WDF eines multivariaten normalverteilten Zufallsvektors.
4. Geben Sie das Theorem zu Marginalen Normalverteilungen wieder.
5. Geben Sie das Theorem zu Gemeinsamen Normalverteilungen wieder.
6. Geben Sie das Theorem zu Bedingten Normalverteilungen wieder.

Referenzen

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley-Interscience.

DeGroot, Morris H., and Mark J. Schervish. 2012. *Probability and Statistics*. 4th ed. Boston: Addison-Wesley.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. Probability and Mathematical Statistics. London ; New York: Academic Press.