# *Statistical Inference with Monotone Incomplete Multivariate Normal Data*

This talk is based on joint work with my wonderful co-authors:

Wan-Ying Chang (US Census Bureau)

Megan Romer (Penn State University)

Tomoya Yamada (Sapporo Gakuin University)

## Background

We have a *population* of “patients”

We draw a *random sample* of $N$ patients, and measure $m$ variables on each patient:

| | |
|---|---|
| $1$ | Visual acuity |
| $2$ | LDL (low-density lipoprotein) cholesterol |
| $3$ | Systolic blood pressure |
| $4$ | Glucose intolerance |
| $5$ | Insulin response to oral glucose |
| $6$ | Actual weight $\div$ Expected weight |
| $\vdots$ | $\vdots$ |
| $m$ | White blood cell count |

We obtain data:

$$\text{Patient} \quad \begin{matrix} 1 & 2 & 3 & \cdots & N \\ \begin{pmatrix} v_{1,1} \\ v_{1,2} \\ \vdots \\ v_{1,m} \end{pmatrix} & \begin{pmatrix} v_{2,1} \\ v_{2,2} \\ \vdots \\ v_{2,m} \end{pmatrix} & \begin{pmatrix} v_{3,1} \\ v_{3,2} \\ \vdots \\ v_{3,m} \end{pmatrix} & \cdots & \begin{pmatrix} v_{N,1} \\ v_{N,2} \\ \vdots \\ v_{N,m} \end{pmatrix} \end{matrix}$$

Vector notation: $V_1, V_2, \ldots, V_N$

$V_1$: The measurements on patient $1$, stacked into a column

etc.

# Classical multivariate analysis

Statistical analysis of $N$ $m$-dimensional data vectors

Common assumption: The population has a multivariate normal distribution

$V$: The vector of measurements on a randomly chosen patient

Multivariate normal populations are characterized by:

$\mu$: The population mean vector

$\Sigma$: The population covariance matrix

For a given data set, $\mu$ and $\Sigma$ are unknown

We wish to perform inference about $\mu$ and $\Sigma$

Construct confidence regions for, and test hypotheses about, $\mu$ and $\Sigma$

Anderson (2003). *An Introduction to Multivariate Statistical Analysis*

Eaton (1984). *Multivariate Statistics: A Vector-Space Approach*

Johnson and Wichern (2002). *Applied Multivariate Statistical Analysis*

Muirhead (1982). *Aspects of Multivariate Statistical Theory*

Standard notation: $V \sim N_p(\mu, \Sigma)$

The probability density function of $V$: For $v \in \mathbb{R}^m$,

$$f(v) = (2\pi)^{-m/2}|\Sigma|^{-1/2} \exp\big( -\tfrac{1}{2}(v-\mu)'\Sigma^{-1}(v-\mu)\big)$$

$V_1, V_2, \ldots, V_N$: Measurements on $N$ randomly chosen patients

Estimate $\mu$ and $\Sigma$ using Fisher's maximum likelihood principle

Likelihood function: $L(\mu, \Sigma) = \prod_{j=1}^{N} f(v_j)$

Maximum likelihood estimator: The value of $(\mu, \Sigma)$ that maximizes $L$

$\widehat{\mu} = \frac{1}{N}\sum_{j=1}^{N} V_j$: The sample mean and MLE of $\mu$

$\widehat{\Sigma} = \frac{1}{N}\sum_{j=1}^{n}(V_j - \bar{V})(V_j - \bar{V})'$: The MLE of $\Sigma$

What are the probability distributions of $\widehat{\mu}$ and $\widehat{\Sigma}$?

$$\widehat{\mu} \sim N_p(\mu, \tfrac{1}{N}\Sigma)$$

Law of Large Numbers: $\widehat{\mu} \to \mu$, a.s., as $N \to \infty$

$N\widehat{\Sigma}$ has a Wishart distribution, a generalization of the $\chi^2$

$\widehat{\mu}$ and $\widehat{\Sigma}$ also are mutually independent

## Monotone incomplete data

Some patients were not measured completely

The resulting data set, with $*$ denoting a missing observation

$$\begin{pmatrix} v_{1,1} \\ v_{1,2} \\ v_{1,3} \\ \vdots \\ v_{1,m} \end{pmatrix} \quad \begin{pmatrix} * \\ v_{2,2} \\ v_{2,3} \\ \vdots \\ v_{2,m} \end{pmatrix} \quad \begin{pmatrix} * \\ * \\ v_{3,2} \\ \vdots \\ v_{3,m} \end{pmatrix} \quad \cdots \quad \begin{pmatrix} * \\ * \\ * \\ \vdots \\ v_{N,m} \end{pmatrix}$$

*Monotone data*: Each $*$ is followed by $*$'s only

We may need to renumber patients to display the data in monotone form

# Physical Fitness Data

A well-known data set from a SAS manual on missing data

Patients: Men taking a physical fitness course at NCSU

Three variables were measured:

Oxygen intake rate (ml. per kg. body weight per minute)

RunTime (time taken, in minutes, to run 1.5 miles)

RunPulse (heart rate while running)

| Oxygen | RunTime | RunPulse | | Oxygen | RunTime | RunPulse |
|---|---|---|---|---|---|---|
| 44.609 | 11.37 | 178 | \| | 39.407 | 12.63 | 174 |
| 45.313 | 10.07 | 185 | \| | 46.080 | 11.17 | 156 |
| 54.297 | 8.65 | 156 | \| | 45.441 | 9.63 | 164 |
| 51.855 | 10.33 | 166 | \| | 54.625 | 8.92 | 146 |
| 49.156 | 8.95 | 180 | \| | 39.442 | 13.08 | 174 |
| 40.836 | 10.95 | 168 | \| | 60.055 | 8.63 | 170 |
| 44.811 | 11.63 | 176 | \| | 37.388 | 14.03 | 186 |
| 45.681 | 11.95 | 176 | \| | 44.754 | 11.12 | 176 |
| 39.203 | 12.88 | 168 | \| | 46.672 | 10.00 | * |
| 45.790 | 10.47 | 186 | \| | 46.774 | 10.25 | * |
| 50.545 | 9.93 | 148 | \| | 45.118 | 11.08 | * |
| 48.673 | 9.40 | 186 | \| | 49.874 | 9.22 | * |
| 47.920 | 11.50 | 170 | \| | 49.091 | 10.85 | * |
| 47.467 | 10.50 | 170 | \| | 59.571 | * | * |
| 50.388 | 10.08 | 168 | \| | 50.541 | * | * |
| | | | \| | 47.273 | * | * |

Monotone data have a staircase pattern; we will consider the two-step pattern

Partition $V$ into an incomplete part of dimension $p$ and a complete part of dimension $q$

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}, \begin{pmatrix} * \\ Y_{n+1} \end{pmatrix}, \begin{pmatrix} * \\ Y_{n+2} \end{pmatrix}, \ldots, \begin{pmatrix} * \\ Y_N \end{pmatrix}$$

Assume that the individual vectors are independent and are drawn from $N_m(\mu, \Sigma)$

Goal: Maximum likelihood inference for $\mu$ and $\Sigma$, with analytical results as extensive and as explicit as in the classical setting

# Where do monotone incomplete data arise?

Panel survey data (Census Bureau, Bureau of Labor Statistics)

Astrophysics

Early detection of diseases

Wildlife survey research

Covert communications

Mental health research

Climate and atmospheric studies

⋮

We have $n$ observations on $\binom{X}{Y}$ and $N-n$ additional observations on $Y$

Difficulty: The likelihood function is more complicated

$$
\begin{aligned}
L &= \prod_{i=1}^{n} f_{X,Y}(x_i, y_i) \cdot \prod_{i=n+1}^{N} f_Y(y_i) \\
&= \prod_{i=1}^{n} f_Y(y_i) f_{X|Y}(x_i) \cdot \prod_{i=n+1}^{N} f_Y(y_i) \\
&= \prod_{i=1}^{N} f_Y(y_i) \cdot \prod_{i=1}^{n} f_{X|Y}(x_i)
\end{aligned}
$$

Partition $\mu$ and $\Sigma$ similarly:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Let

$$\mu_{1\cdot 2} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(Y - \mu_2), \quad \Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$Y \sim N_q(\mu_2, \Sigma_{22}), \quad X|Y \sim N_p(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$$

$\widehat{\mu}$ and $\widehat{\Sigma}$: Wilks, Anderson, Morrison, Olkin, Jinadasa, Tracy, ...

Anderson and Olkin (1985): An elegant derivation of $\widehat{\Sigma}$

Sample means:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j, \qquad \bar{Y}_1 = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

$$\bar{Y}_2 = \frac{1}{N-n}\sum_{j=n+1}^{N} Y_j, \quad \bar{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_j$$

Sample covariance matrices:

$$A_{11} = \sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})', \quad A_{12} = \sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y}_1)'$$

$$A_{22,n} = \sum_{j=1}^{n}(Y_j - \bar{Y}_1)(Y_j - \bar{Y}_1)', \quad A_{22,N} = \sum_{j=1}^{N}(Y_j - \bar{Y})(Y_j - \bar{Y})'$$

## The MLE's of $\mu$ and $\Sigma$

Notation: $\tau = n/N$, $\bar{\tau} = 1 - \tau$

$$\widehat{\mu}_1 = \bar{X} - \bar{\tau} A_{12} A_{22,n}^{-1} (\bar{Y}_1 - \bar{Y}_2), \qquad \widehat{\mu}_2 = \bar{Y}$$

$\widehat{\mu}_1$ is called the *regression estimator of* $\mu_1$

In sample surveys, extra observations on a subset of variables are used to improve estimation of a parameter

$\widehat{\Sigma}$ is more complicated:

$$\begin{aligned} \widehat{\Sigma}_{11} &= \frac{1}{n}(A_{11} - A_{12} A_{22,n}^{-1} A_{21}) + \frac{1}{N} A_{12} A_{22,n}^{-1} A_{22,N} A_{22,n}^{-1} A_{21} \\ \widehat{\Sigma}_{12} &= \frac{1}{N} A_{12} A_{22,n}^{-1} A_{22,N} \\ \widehat{\Sigma}_{22} &= \frac{1}{N} A_{22,N} \end{aligned}$$

## Seventy year-old unsolved problems

Explicit confidence levels for elliptical confidence regions for $\mu$

In testing hypotheses on $\mu$ or $\Sigma$, are the LRT statistics unbiased?

Calculate the higher moments of the components of $\widehat{\mu}$

Determine the asymptotic behavior of $\widehat{\mu}$ as $n$ or $N \to \infty$

The Stein phenomenon for $\widehat{\mu}$?

The crucial obstacle: The exact distribution of $\widehat{\mu}$

## The exact distribution of $\widehat{\mu}$

Chang and D.R. (J. Multivariate Analysis, 2009): For $n > p + q$,

$$\widehat{\mu} \stackrel{\mathcal{L}}{=} \mu + V_1 + \left(\tfrac{1}{n} - \tfrac{1}{N}\right)^{1/2} \left(\tfrac{Q_2}{Q_1}\right)^{1/2} \begin{pmatrix} V_2 \\ \mathbf{0} \end{pmatrix},$$

where $V_1$, $V_2$, $Q_1$, and $Q_2$ are independent;

$$V_1 \sim N_{p+q}(\mathbf{0}, \Omega), \quad V_2 \sim N_p(\mathbf{0}, \Sigma_{11 \cdot 2}), \quad Q_1 \sim \chi^2_{n-q}, \quad Q_2 \sim \chi^2_q;$$

$$\Omega = \tfrac{1}{N}\Sigma + \left(\tfrac{1}{n} - \tfrac{1}{N}\right) \begin{pmatrix} \Sigma_{11 \cdot 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Consequences: $\widehat{\mu}$ is an unbiased estimator of $\mu$. Also, $\widehat{\mu}_1$ and $\widehat{\mu}_2$ are independent iff $\Sigma_{12} = \mathbf{0}$.

Romer and D.R. (2009): Explicit formulas for the $V$'s and $Q$'s

Computation of the higher moments of $\widehat{\mu}$ now is straightforward

Due to the term $1/Q_1$, even moments exist only up to order $n-q$

The covariance matrix of $\widehat{\mu}$:

$$\mathrm{Cov}(\widehat{\mu}) = \frac{1}{N}\Sigma + \frac{(n-2)\bar{\tau}}{n(n-q-2)} \begin{pmatrix} \Sigma_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Asymptotics for $\widehat{\mu}$: If $n, N \to \infty$ with $N/n \to \delta \geq 1$ then

$$\sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{\mathcal{L}} N_{p+q}\left(\mathbf{0}, \Sigma + (\delta - 1) \begin{pmatrix} \Sigma_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right)$$

## The analog of Hotelling's $T^2$-statistic

$$T^2 = (\widehat{\mu} - \mu)' \widehat{\mathrm{Cov}}(\widehat{\mu})^{-1} (\widehat{\mu} - \mu)$$

where

$$\widehat{\mathrm{Cov}}(\widehat{\mu}) = \frac{1}{N}\widehat{\Sigma} + \frac{(n-2)\bar{\tau}}{n(n-q-2)} \begin{pmatrix} \widehat{\Sigma}_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

An obvious ellipsoidal confidence region for $\mu$ is

$$\left\{ \nu \in \mathbb{R}^{p+q} : (\widehat{\mu} - \nu)' \widehat{\mathrm{Cov}}(\widehat{\mu})^{-1} (\widehat{\mu} - \nu) \leq c \right\}$$

What is the corresponding confidence level?

Theorem: For $t \geq 0$, $P(T^2 \leq t)$ is bounded above by

$$P\Big(F_{q,N-q} \leq (q^{-1} - N^{-1})t\Big)$$

and bounded below by

$$P\Big(\frac{N^2 Q_2}{nQ_1}\big(1 + \frac{qQ_3}{Q_5}\big) + \frac{Nq}{Q_5}\big(\tau^{1/2} Q_3^{1/2} + \bar{\tau}^{1/2} Q_4^{1/2}\big)^2 \leq t\Big),$$

where $Q_1 \sim \chi^2_{n-p-q}$, $Q_2 \sim \chi^2_p$, $Q_3 \sim \chi^2_q$, $Q_4 \sim \chi^2_q$, $Q_5 \sim \chi^2_2$, and $Q_1, \ldots, Q_5$ are mutually independent.

Romer (2009) has now derived the exact distribution of $T^2$

Shrinkage estimation for $\mu$ when $\Sigma$ is unknown

# A decomposition of $\widehat{\Sigma}$

Notation: $A_{11\cdot 2,n} := A_{11} - A_{12}A_{22,n}^{-1}A_{21}$

$$
\begin{aligned}
n\widehat{\Sigma} &= \tau \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22,n} \end{pmatrix} + \bar{\tau} \begin{pmatrix} A_{11\cdot 2,n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\
&\quad + \tau \begin{pmatrix} A_{12}A_{22,n}^{-1} & \mathbf{0} \\ \mathbf{0} & I_q \end{pmatrix} \begin{pmatrix} B & B \\ B & B \end{pmatrix} \begin{pmatrix} A_{22,n}^{-1}A_{21} & \mathbf{0} \\ \mathbf{0} & I_q \end{pmatrix}
\end{aligned}
$$

where

$$
\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22,n} \end{pmatrix} \sim \mathrm{W}_{p+q}(n-1, \Sigma) \quad \text{and} \quad B \sim \mathrm{W}_q(N-n, \Sigma_{22})
$$

are independent. Also, $N\widehat{\Sigma}_{22} \sim \mathrm{W}_q(N-1, \Sigma_{22})$

$$A_{22,N} = \sum_{j=1}^{n}(Y_j - \bar{Y}_1 + \bar{Y}_1 - \bar{Y})(Y_j - \bar{Y}_1 + \bar{Y}_1 - \bar{Y})' + \sum_{j=n+1}^{N}(Y_j - \bar{Y}_2 + \bar{Y}_2 - \bar{Y})(Y_j - \bar{Y}_2 + \bar{Y}_2 - \bar{Y})'$$

$$A_{22,N} = A_{22,n} + B$$

$$B = \sum_{j=n+1}^{N}(Y_j - \bar{Y}_2)(Y_j - \bar{Y}_2)' + \frac{n(N-n)}{N}(\bar{Y}_1 - \bar{Y}_2)(\bar{Y}_1 - \bar{Y}_2)'$$

Verify that the terms in the decomposition of $\widehat{\Sigma}$ are independent

The marginal distribution of $\widehat{\Sigma}_{11}$ is non-trivial

If $\Sigma_{12} = \mathbf{0}$ then $A_{22,n}$, $B$, $A_{11\cdot 2,n}$, $A_{12}A_{22,n}^{-1}A_{21}$, $\bar{X}$, $\bar{Y}_1$, and $\bar{Y}_2$ are independent

Matrix $F$-distribution: $F_{a,b}^{(q)} = W_2^{-1/2} W_1 W_2^{-1/2}$

where $W_1 \sim \mathrm{W}_q(a, \Sigma_{22})$ and $W_2 \sim \mathrm{W}_q(b, \Sigma_{22})$

Theorem: Suppose that $\Sigma_{12} = \mathbf{0}$. Then

$$\Sigma_{11}^{-1/2}\widehat{\Sigma}_{11}\Sigma_{11}^{-1/2} \stackrel{\mathcal{L}}{=} \frac{1}{n}W_1 + \frac{1}{N}W_2^{1/2}\big(I_p + F\big)W_2^{1/2}$$

where $W_1$, $W_2$, and $F$ are independent, and

$$W_1 \sim \mathrm{W}_p(n-q-1, I_p), \quad W_2 \sim \mathrm{W}_p(q, I_p), \text{ and}$$

$$F \sim F^{(p)}_{N-n,n-q+p-1}$$

$$N\Sigma_{11}^{-1/2}\widehat{\Sigma}_{11}\Sigma_{11}^{-1/2} \stackrel{\mathcal{L}}{=} \frac{N}{n}\Sigma_{11}^{-1/2}A_{11\cdot 2,n}\Sigma_{11}^{-1/2} + \Sigma_{11}^{-1/2}A_{12}A_{22,n}^{-1}(A_{22,n} + B)A_{22,n}^{-1}A_{21}\Sigma_{11}^{-1/2}$$

Theorem: With no assumptions on $\Sigma_{12}$,

$$\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1} \stackrel{\mathcal{L}}{=} \Sigma_{12}\Sigma_{22}^{-1} + \Sigma_{11\cdot 2}^{1/2} W^{-1/2} K \Sigma_{22}^{-1/2}$$

where $W$ and $K$ are independent, and

$$W \sim \mathrm{W}_p(n-q+p-1, I_p), \quad K \sim N_{pq}(\mathbf{0}, I_p \otimes I_q)$$

In particular, $\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}$ is an unbiased estimator of $\Sigma_{12}\Sigma_{22}^{-1}$

The general distribution of $\widehat{\Sigma}$ requires the hypergeometric functions of matrix argument

Saddlepoint approximations

The distribution of $|\widehat{\Sigma}|$ is much simpler:

$$|\widehat{\Sigma}| = |\widehat{\Sigma}_{11\cdot 2}| \cdot |\widehat{\Sigma}_{22}|$$

$|\widehat{\Sigma}_{11\cdot 2}|$ and $|\widehat{\Sigma}_{22}|$ are independent; each is a product of independent $\chi^2$ variables

Hao and Krishnamoorthy (2001):

$$|\widehat{\Sigma}| \stackrel{\mathcal{L}}{=} n^{-p} N^{-q} \; |\Sigma| \cdot \prod_{j=1}^{p} \chi^2_{n-q-j} \cdot \prod_{j=1}^{q} \chi^2_{N-j}$$

It now is plausible that tests of hypothesis on $\Sigma$ are unbiased

## Testing $\Sigma = \Sigma_0$

Data: Two-step, monotone incomplete sample

$\Sigma_0$: A given, positive definite matrix

Test $H_0 : \Sigma = \Sigma_0$ *vs.* $H_a : \Sigma \neq \Sigma_0$ (WLOG, $\Sigma_0 = I_{p+q}$)

Hao and Krishnamoorthy (2001): The LRT statistic is

$$\begin{aligned} \lambda_1 \quad \propto \quad & |A_{22,N}|^{N/2} \exp\big(-\tfrac{1}{2}\text{tr}\, A_{22,N}\big) \\ & \times |A_{11\cdot 2,n}|^{n/2} \exp\big(-\tfrac{1}{2}\text{tr}\, A_{11\cdot 2,n}\big) \\ & \times \exp\big(-\tfrac{1}{2}\text{tr}\, A_{12}A_{22,n}^{-1}A_{21})\big). \end{aligned}$$

Is the LRT unbiased? If $C$ is a critical region of size $\alpha$, is

$$P(\lambda_1 \in C | H_a) \geq P(\lambda_1 \in C | H_0)?$$

Pitman (1939): Even with $1$-d data, $\lambda_1$ is not unbiased

Bartlett: $\lambda_1$ becomes unbiased if sample sizes are replaced by degrees of freedom

With two-step monotone data, perhaps a similarly modified statistic, $\lambda_2$, is unbiased?

Answer: Still unknown.

Theorem: If $|\Sigma_{11}| < 1$ then $\lambda_2$ is unbiased

With monotone incomplete data, further modification is needed

Theorem: The modified LRT,

$$\begin{aligned}\lambda_3 \quad \propto \quad & |A_{22,N}|^{(N-1)/2} \exp\left(-\tfrac{1}{2}\text{tr}\, A_{22,N}\right) \\ & \times |A_{11\cdot 2,n}|^{(n-q-1)/2} \exp\left(-\tfrac{1}{2}\text{tr}\, A_{11\cdot 2,n}\right) \\ & \times |A_{12}A_{22,n}^{-1}A_{21}|^{q/2} \exp\left(-\tfrac{1}{2}\text{tr}\, A_{12}A_{22,n}^{-1}A_{21}\right),\end{aligned}$$

is unbiased. Also, $\lambda_1$ is not unbiased

For diagonal $\Sigma = \text{diag}(\sigma_{jj})$, the power function of $\lambda_3$ increases monotonically as any $|\sigma_{jj} - 1|$ increases, $j = 1, \ldots, p + q$.

With monotone two-step data, test

$$H_0 : (\mu, \Sigma) = (\mu_0, \Sigma_0) \;\; vs. \;\; H_a : (\mu, \Sigma) \neq (\mu_0, \Sigma_0)$$

where $\mu_0$ and $\Sigma_0$ are given. The LRT statistic is

$$\lambda_4 = \lambda_1 \exp\big( - \tfrac{1}{2}(n\bar{X}'\bar{X} + N\bar{Y}'\bar{Y})\big)$$

Remarkably, $\lambda_4$ is unbiased

The sphericity test, $H_0 : \Sigma \propto I_{p+q}$ *vs.* $H_a : \not\propto I_{p+q}$

The unbiasedness of the LRT statistic is an open problem

## The Stein phenomenon for $\widehat{\mu}$

$\widehat{\mu}$: The mean of a complete sample from $N_m(\mu, I_m)$

Quadratic loss function: $L(\widehat{\mu}, \mu) = \|\widehat{\mu} - \mu\|^2$

Risk function: $R(\widehat{\mu}) = E\, L(\widehat{\mu}, \mu)$

C. Stein: $\widehat{\mu}$ is inadmissible for $m \geq 3$

James-Stein estimator for shrinking $\widehat{\mu}$ to $\nu \in \mathbb{R}^m$:

$$\widehat{\mu}_c = \left(1 - \frac{c}{\|\widehat{\mu} - \nu\|^2}\right)(\widehat{\mu} - \nu) + \nu$$

Baranchik's positive-part shrinkage estimator:

$$\widehat{\mu}_c^+ = \left(1 - \frac{c}{\|\widehat{\mu} - \nu\|^2}\right)_+ (\widehat{\mu} - \nu) + \nu$$

We collect a monotone incomplete sample from $N_{p+q}(\mu, \Sigma)$

Does the Stein phenomenon hold for $\widehat{\mu}$, the MLE of $\mu$?

The phenomenon seems almost universal: It holds for many loss functions, inference problems, and distributions

Various results available on shrinkage estimation of $\Sigma$ with incomplete data, but no such results available for $\mu$

The crucial impediment: The distribution of $\widehat{\mu}$ was unknown

Theorem (Yamada and D.R.): For $p \geq 2$, $n \geq q+3$, and $\Sigma = I_{p+q}$, both $\widehat{\mu}$ and $\widehat{\mu}_c$ are inadmissible:

$$R(\widehat{\mu}) > R(\widehat{\mu}_c) > R(\widehat{\mu}_c^+)$$

for all $\nu \in \mathbb{R}^{p+q}$ and all $c \in (0, 2c^*)$, where

$$c^* = \frac{p-2}{n} + \frac{q}{N}.$$

Non-radial loss functions

Replace $\|\widehat{\mu} - \nu\|^2$ by non-radial functions of $\widehat{\mu} - \nu$

Shrinkage to a random vector $\nu$, calculated from the data

# Kurtosis tests for multivariate normality

$m$-dimensional complete, random sample: $V_1, \ldots, V_N$

Extensive literature on testing for multivariate normality

Mardia's statistic for testing for kurtosis:

$$b_{2,m} = \sum_{j=1}^{N} \left[ (V_j - \bar{V})' S^{-1} (V_j - \bar{V}) \right]^2$$

Invariance under nonsingular affine transformations of the data

Asymptotic distribution of $b_{2,m}$

Monotone incomplete data, i.i.d., unknown population:

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}, \begin{pmatrix} * \\ Y_{n+1} \end{pmatrix}, \begin{pmatrix} * \\ Y_{n+2} \end{pmatrix}, \ldots, \begin{pmatrix} * \\ Y_N \end{pmatrix}$$

A generalization of Mardia's statistic:

$$\begin{aligned} \hat{\beta} &= \sum_{j=1}^{n} \left[ \left( \begin{pmatrix} X_j \\ Y_j \end{pmatrix} - \widehat{\mu} \right)' \widehat{\Sigma}^{-1} \left( \begin{pmatrix} X_j \\ Y_j \end{pmatrix} - \widehat{\mu} \right) \right]^2 \\ &\quad + \sum_{j=n+1}^{N} \left[ (Y_j - \widehat{\mu}_2)' \widehat{\Sigma}_{22}^{-1} (Y_j - \widehat{\mu}_2) \right]^2 \end{aligned}$$

# An alternative to $\hat{\beta}$

Impute each missing $X_j$ using linear regression:

$$\widehat{X}_j = \begin{cases} X_j, & 1 \le j \le n \\ \widehat{\mu}_1 + \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}(Y_j - \widehat{\mu}_2), & n+1 \le j \le N \end{cases}$$

Construct

$$\widehat{\beta}_* = \sum_{j=1}^{N} \left[ \left( \widehat{\begin{pmatrix} X_j \\ Y_j \end{pmatrix}} - \widehat{\mu} \right)' \widehat{\Sigma}^{-1} \left( \widehat{\begin{pmatrix} X_j \\ Y_j \end{pmatrix}} - \widehat{\mu} \right) \right]^2$$

A remarkable result: $\widehat{\beta} \equiv \widehat{\beta}_*$

$\widehat{\beta}$ is invariant under nonsingular affine transformations of the data

Yamada, Romer, and D.R. (2010): Under certain regularity conditions,

$$(\widehat{\beta} - c_1)/c_2 \xrightarrow{\mathcal{L}} N(0,1)$$

as $n, N \to \infty$

The constants $c_1, c_2$ depend on $n, N$ and the underlying population distribution

In the normal case, $c_1, c_2$ depend only on $n, N, p, q$

## References

Chang and D.R. (2009). Finite-sample inference with monotone incomplete multivariate normal data, I. J. Multivariate Analysis.

Chang and D.R. (2009). Finite-sample inference with monotone incomplete multivariate normal data, II. J. Multivariate Analysis.

D.R. and Yamada (2010). The Stein phenomenon for monotone incomplete multivariate normal data. J. Multivariate Analysis.

Yamada (2009). The asymptotic expansion of the distribution of the canonical correlations with monotone incomplete multivariate normal data. Preprint, Sapporo Gakuin University.

Romer (2009). The Statistical Analysis of Monotone Incomplete Multivariate Normal Data. Doctoral Dissertation, Penn State University.

Romer and D.R. (2010). Maximum likelihood estimation of the mean of a multivariate normal population with monotone incomplete data. Preprint, Penn State University.

Yamada, Romer, and D.R. (2010). Kurtosis tests for multivariate normality with monotone incomplete data. Preprint, Penn State University.

# Astrostatistics research problems

K. R. Lang, Astrophysical Formulae, Vol. II: Space, Time, Matter and Cosmology, 3rd. ed., Springer, 2006

Numerous monotone incomplete data sets

Is it true that astrophysicists often discard incomplete data?

Incomplete longitudinal data (light curves, luminosity data)

Incomplete time series

Small-sample distributions of test statistics, e.g., Mardia's statistic, often are unexplored even with complete data

How to relax the MCAR assumption to MAR?

A. Isenman, “Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning,” Springer, 2008

COMBO-17 Survey

Apply classical multivariate statistical procedures (principal components, MANOVA, ...) to the COMBO-17 survey

I hear that some variables in the survey “are not of scientific interest,” e.g., the absence of high-redshift (i.e. distant) high-absolute-magnitude (i.e. faint) galaxies, the dropoff in flux with redshift, the dropoff in image size with redshift, ...

Carry out a statistical analysis of the variables which “are not of scientific interest”

Discover amazing results that put you in the NY Times and

Get an invitation to Stockholm

Send me 10% of the prize money (I’m a reasonable guy)