

# The Statistical Analysis of Two-step Monotone Incomplete Multivariate Normal Data

Megan M. Romer, Donald St. P. Richards

Pennsylvania State University

## Introduction

In all areas of research, incomplete data sets are ubiquitous. We consider problems arising in the statistical analysis of monotone incomplete data drawn from a multivariate normal population. Draw a random sample of  $N$  units from the population and measure  $d$  variables on each unit. Denote these vectors by  $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_N$ .

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{d-2} \\ Z_{d-1} \\ Z_d \end{pmatrix}, \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{d-2} \\ Z_{d-1} \\ * \end{pmatrix}, \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ * \\ * \\ * \end{pmatrix}, \dots, \begin{pmatrix} Z_1 \\ * \\ \vdots \\ * \\ * \\ * \end{pmatrix}$$

## Two-step Notation

Partition  $N$  mutually independent observations into a block of complete data, of dimension  $d = p + q$ , and a block of incomplete data, of dimension  $p$ :

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{Y}_1 \end{pmatrix} \begin{pmatrix} \mathbf{X}_2 \\ \mathbf{Y}_2 \end{pmatrix} \dots \begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \mathbf{X}_{n+1} \mathbf{X}_{n+2} \dots \mathbf{X}_N$$

The first  $n$  vectors are observations from  $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and the last  $N - n$  vectors are observations on the first  $p$  variables of the same population. Define  $\tau = n/N$  and  $\bar{\tau} = 1 - \tau$ . The sample means:

$$\bar{\mathbf{X}}_1 = \frac{1}{n} \sum_{j=1}^n \mathbf{X}_j, \quad \bar{\mathbf{X}}_2 = \frac{1}{N-n} \sum_{j=n+1}^N \mathbf{X}_j$$

$$\bar{\mathbf{Y}} = \frac{1}{n} \sum_{j=1}^n \mathbf{Y}_j, \quad \bar{\mathbf{X}} = \frac{1}{N} \sum_{j=1}^N \mathbf{X}_j,$$

The corresponding matrices of sums of squares and products:

$$\mathbf{A}_{11,n} = \sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}}_1)(\mathbf{X}_j - \bar{\mathbf{X}}_1)'$$

$$\mathbf{A}_{22} = \sum_{j=1}^n (\mathbf{Y}_j - \bar{\mathbf{Y}})(\mathbf{Y}_j - \bar{\mathbf{Y}})'$$

$$\mathbf{A}_{12} = \mathbf{A}'_{21} = \sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}}_1)(\mathbf{Y}_j - \bar{\mathbf{Y}})'$$

$$\mathbf{A}_{11,N} = \sum_{j=1}^N (\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'$$

Partition  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  into  $p$  and  $q$  rows and columns. The maximum likelihood estimators, (MLEs), for  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  [Anderson, 1957]:

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{X}} \\ \bar{\mathbf{Y}} - \bar{\tau} \mathbf{A}_{21} \mathbf{A}_{11,n}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_{11} = \frac{1}{N} \mathbf{A}_{11,N}, \quad \hat{\boldsymbol{\Sigma}}_{12} = \frac{1}{N} \mathbf{A}_{11,N} \mathbf{A}_{11,n}^{-1} \mathbf{A}_{12} = \hat{\boldsymbol{\Sigma}}'_{21}$$

$$\hat{\boldsymbol{\Sigma}}_{22} = \frac{1}{n} \mathbf{A}_{22} + \frac{1}{N} \mathbf{A}_{21} \mathbf{A}_{11,n}^{-1} \mathbf{A}_{11,N} \mathbf{A}_{11,n}^{-1} \mathbf{A}_{12}$$

## References

- T. W. Anderson. Maximum likelihood estimators for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.*, 52:200-203, 1957.
- W. Y. Chang and D. St. P. Richards. Finite-sample inference with monotone incomplete multivariate normal data I. *J. of Mult. Anal.*, to appear, 2009.
- M. M. Romer. The Statistical Analysis of Monotone Incomplete Multivariate Normal Data. Doctoral Dissertation, Penn State University, 2009.

## Exact Stochastic Representation for $\hat{\boldsymbol{\mu}}$

**Theorem 1**[Chang and Richards, 2009; Romer, 2009]: Let  $\mathbf{V}_1 \sim N_{p+q}(\mathbf{0}, \frac{1}{N} \boldsymbol{\Sigma} + \frac{\bar{\tau}}{n} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22.1} \end{pmatrix})$ ,  $Q_1 \sim \chi_{n-p}^2$ ,  $Q_2 \sim \chi_p^2$ ,  $\mathbf{V}_2 \sim N_q(\mathbf{0}, \mathbf{I}_q)$ , where  $\mathbf{V}_1, \mathbf{V}_2, Q_1$ , and  $Q_2$  are mutually independent. Then,

$$\hat{\boldsymbol{\mu}} \stackrel{\mathcal{L}}{=} \boldsymbol{\mu} + \mathbf{V}_1 + \left( \frac{\bar{\tau} Q_2}{n Q_1} \right)^{1/2} \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Sigma}_{22.1}^{1/2} \mathbf{V}_2 \end{pmatrix}.$$

Explicit formulas for  $\mathbf{V}_1, \mathbf{V}_2, Q_1$ , and  $Q_2$  in terms of the data:

$$Q_1 = \frac{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)}{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \mathbf{A}_{11,n}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)}, \quad Q_2 = n \bar{\tau} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

$$\mathbf{V}_1 = \begin{pmatrix} \bar{\mathbf{X}} \\ \bar{\mathbf{Y}} \end{pmatrix}, \quad \mathbf{V}_2 = - \frac{\sum_{j=1}^n \mathbf{Y}_j (\mathbf{X}_j - \bar{\mathbf{X}})' \mathbf{A}_{11,n}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)}{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \mathbf{A}_{11,n}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)}$$

## A Generalization of Hotelling's $T^2$ -statistic

Let  $T^2 = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})' \widehat{\text{Cov}}(\hat{\boldsymbol{\mu}})^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ , where  $\widehat{\text{Cov}}(\hat{\boldsymbol{\mu}})$  is the MLE of  $\text{Cov}(\hat{\boldsymbol{\mu}})$ . Define  $T_1^2 = n(\bar{\mathbf{Y}} - \mathbf{A}_{21} \mathbf{A}_{11,n}^{-1} \bar{\mathbf{X}}_1)' \mathbf{A}_{22.1,n}^{-1} (\bar{\mathbf{Y}} - \mathbf{A}_{21} \mathbf{A}_{11,n}^{-1} \bar{\mathbf{X}}_1)$  and  $T_2^2 = N \bar{\mathbf{X}}' (\mathbf{A}_{11,N})^{-1} \bar{\mathbf{X}}$ . Then  $\frac{1}{N} T^2 = \gamma^{-1} T_1^2 + T_2^2$ .

**Proposition**[Romer, 2009]: Let  $\boldsymbol{\Lambda}_{11}$  and  $\boldsymbol{\Lambda}_{22}$  be  $p \times p$  and  $q \times q$  positive definite matrices, respectively,  $\boldsymbol{\Lambda}_{21}$  be  $q \times p$ ,  $\boldsymbol{\nu}_1 \in \mathbb{R}^p$ , and  $\boldsymbol{\nu}_2 \in \mathbb{R}^q$ . The statistics  $T_1^2$  and  $T_2^2$ , and consequently  $T^2$ , are algebraically invariant under the transformation  $\begin{pmatrix} \mathbf{X}_j \\ \mathbf{Y}_j \end{pmatrix} \mapsto \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \boldsymbol{\Lambda}_{21} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{X}_j \\ \mathbf{Y}_j \end{pmatrix} + \begin{pmatrix} \boldsymbol{\nu}_1 \\ \boldsymbol{\nu}_2 \end{pmatrix}$ .

## Exact Stochastic Representation for the $T^2$ -statistic

**Theorem 2**[Romer, 2009]: Let  $\cos^2 \theta \sim \text{Beta}(\frac{1}{2}, \frac{1}{2}(p-1))$ ,  $Q_1 \sim \chi_p^2$ ,  $Q_2 \sim \chi_p^2$ ,  $Q_3 \sim \chi_{n-p-q}^2$ ,  $Q_4 \sim \chi_q^2$ ,  $\mathbf{W} \sim \text{W}_2(N-p, \mathbf{I}_2)$ , and  $\boldsymbol{\beta} \sim \text{Beta}(\frac{1}{2}(n-p-2), \frac{1}{2}(N-n-1))$  be mutually independent. Then,

$$T^2 \stackrel{\mathcal{L}}{=} \frac{N Q_4}{\gamma Q_3} \left( 1 + Q_1 \boldsymbol{\beta}^{-1} \mathbf{e}'_1 \mathbf{W}^{-1} \mathbf{e}_1 \right) + \mathbf{u}' \mathbf{W}^{-1} \mathbf{u} - \frac{v}{1 + v \mathbf{e}'_1 \mathbf{W}^{-1} \mathbf{e}_1} (\mathbf{e}'_1 \mathbf{W}^{-1} \mathbf{u})^2,$$

where  $\gamma = 1 + \frac{(n-2)N\bar{\tau}}{n(n-p-2)}$ ,  $\mathbf{u} = (u_1, u_2)' = N^{1/2}((\sqrt{\tau} Q_1^{1/2} + \sqrt{\bar{\tau}} Q_2^{1/2} \cos \theta), \sqrt{\bar{\tau}} Q_2^{1/2} \sin \theta)'$ , and  $v = \bar{\tau} Q_1 + \tau Q_2 - 2(\bar{\tau} Q_1 \tau Q_2)^{1/2} \cos \theta$ .

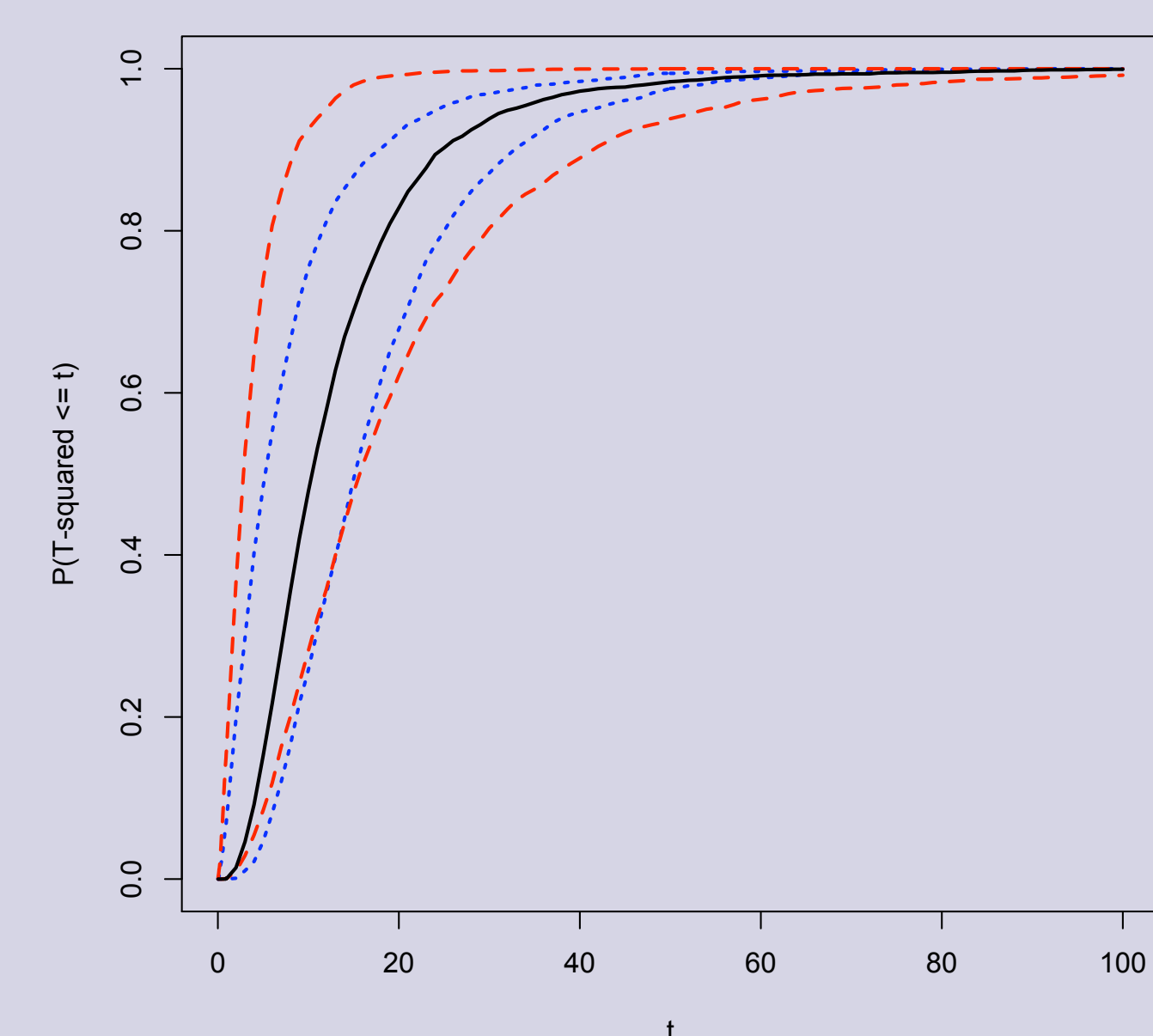
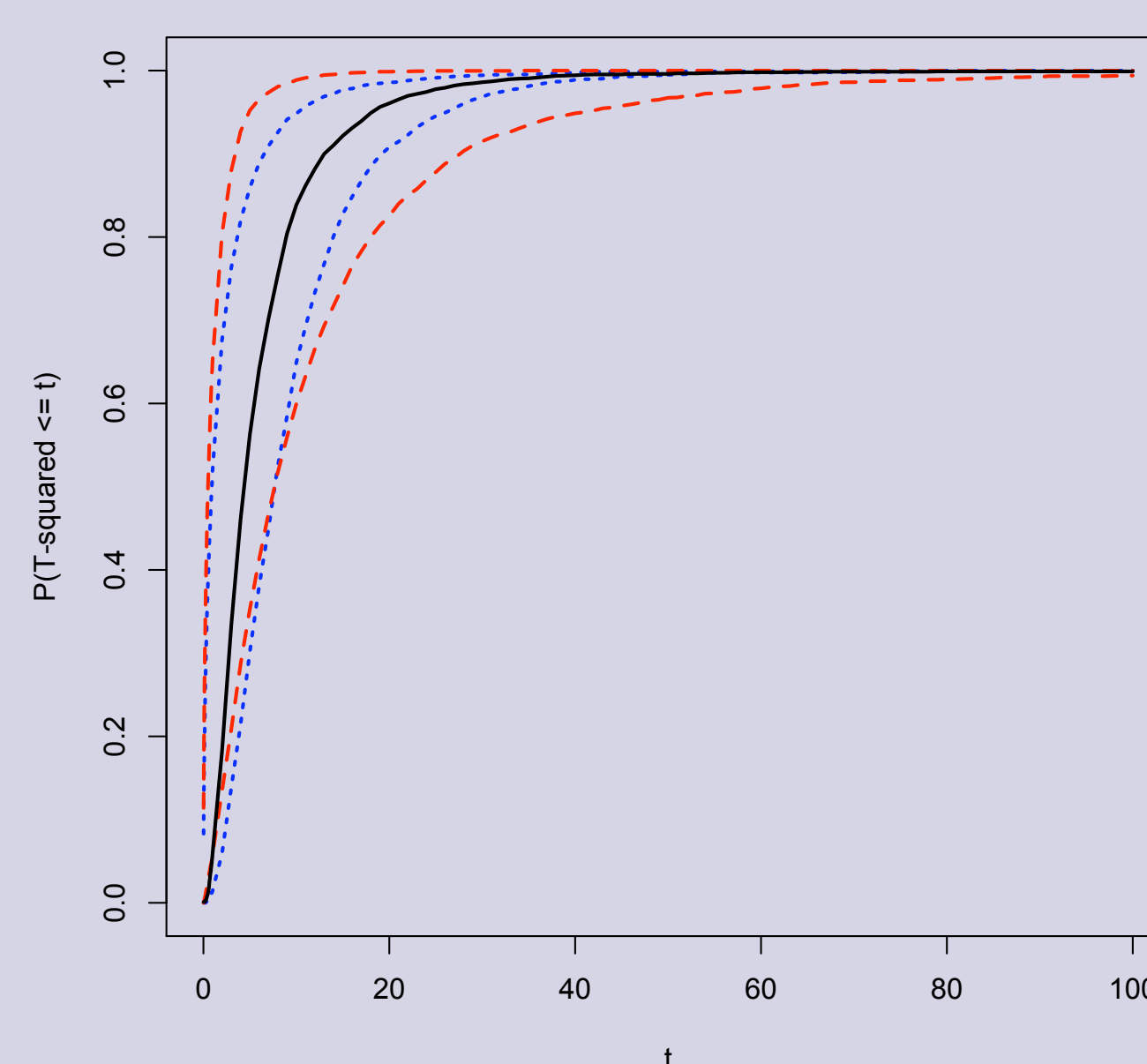
## Applications of the $T^2$ -statistic

**Theorem 3**[Romer, 2009]: Let  $Q_1 \sim \chi_p^2$ ,  $Q_2 \sim \chi_p^2$ ,  $Q_3 \sim \chi_{n-p-q}^2$ ,  $Q_4 \sim \chi_q^2$ ,  $Q_5 \sim \chi_{N-p}^2$ ,  $Q_6 \sim \chi_{N-p}^2$ ,  $Q_7 \sim \chi_1^2$ ,  $Q_8 \sim \chi_{N-p-1}^2$ , and  $\boldsymbol{\beta} \sim \text{Beta}(\frac{1}{2}(n-p-2), \frac{1}{2}(N-n-1))$  be mutually independent. For  $t \geq 0$ ,

$$P(T^2 \leq t) \leq P\left( \frac{N Q_4}{\gamma Q_3} \left( 1 + \frac{Q_1}{Q_8 \boldsymbol{\beta}} \right) \leq t \right)$$

$$P(T^2 \leq t) \geq P\left[ \frac{N Q_4}{\gamma Q_3} \left( 1 + \frac{Q_1 (Q_6 + Q_7)}{Q_5 Q_6 \boldsymbol{\beta}} \right) + \max\left( \frac{N \bar{\tau} Q_2 (Q_6 + Q_7)}{Q_5 Q_6}, \frac{N \bar{\tau} Q_2 Q_7}{Q_6 + Q_7} \right) + \frac{N(\tau Q_1 + \sqrt{\tau \bar{\tau}} Q_1 Q_2)(Q_6 + Q_7)}{Q_5 Q_6} + \frac{N(2\sqrt{\tau \bar{\tau}} Q_1 Q_2 + \bar{\tau} Q_2)}{Q_6} \leq t \right]$$

**Comparison of Bounds:** Chang & Richards (Red); Theorem 2 (Black); Theorem 3 (Blue)  
( $p=2; q=1; n=10; N=15$ ) ( $p=3; q=3; n=15; N=20$ )



**Theorem 4**[Romer, 2009]: Let  $T_\alpha^2$  be a  $100(1 - \alpha)\%$  percentage point for  $T^2$ . Then, with confidence at least  $1 - \alpha$ ,  $\boldsymbol{\mu}$  satisfies for all  $\mathbf{v}$  the simultaneous inequalities:  $\mathbf{v}' \boldsymbol{\mu} \in \left( \mathbf{v}' \hat{\boldsymbol{\mu}} \pm \sqrt{T_\alpha^2} \cdot \sqrt{\mathbf{v}' \widehat{\text{Cov}}(\hat{\boldsymbol{\mu}}) \mathbf{v}} \right)$ .

## Funding

This research was supported in part by the National Science Foundation grant DMS-0705210.