

# *Multimodality of the Likelihood Function for the Behrens-Fisher Problem*

Max Buot (Xavier University)

D.R. (SAMSI and Penn State University)

# The Behrens-Fisher Problem

---

Independent, multivariate normal populations,  $N_p(\mu_1, \Sigma_1)$  and  $N_p(\mu_2, \Sigma_2)$

Popn. 1: Industrial characteristics of factory A

Popn. 2: Industrial characteristics of factory B

Random samples:  $X_1, \dots, X_{N_1}$  from  $N_p(\mu_1, \Sigma_1)$ , and  $Y_1, \dots, Y_{N_2}$  from  $N_p(\mu_2, \Sigma_2)$

Problem: Carry out statistical inference for  $\mu_1 - \mu_2$

Test  $H_0 : \mu_1 - \mu_2 = 0$  vs.  $H_a : \mu_1 - \mu_2 \neq 0$

Under  $H_0$ ,  $\mu$  denotes the common value of  $\mu_1$  and  $\mu_2$

## Fisher's Iris Data

---

The Iris dataset was introduced by R. A. Fisher as an example in discriminant analysis

The data report four characteristics (sepal width, sepal length, petal width and petal length) on three species of Irises: setosa, versicolor, and virginica

A scatterplot of the sepal lengths

Another scatterplot of the petal lengths

A 3-D scatterplot of the Iris Data

These scatterplots motivate the Behrens-Fisher problem for comparing versicolor and virginica

The usual formula for the density function of  $X \sim N_p(\mu, \Sigma)$

$$f(x; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right)$$

The likelihood function under  $H_0$ :

$$L(\mu, \Sigma_1, \Sigma_2) = \prod_{j=1}^{N_1} f(x_j; \mu, \Sigma_1) \prod_{j=1}^{N_2} f(y_j; \mu, \Sigma_2)$$

Fisher's brilliant maximum likelihood idea

Treat the data  $\{X_j\}$  and  $\{Y_j\}$  as constants

Treat the parameters  $(\mu, \Sigma_1, \Sigma_2)$  as variables

Maximize the likelihood function  $L$  w.r.t.  $(\mu, \Sigma_1, \Sigma_2)$

The maximum likelihood estimator: The value of  $(\mu, \Sigma_1, \Sigma_2)$  which maximizes  $L$

The MLE is a function of the data  $\{X_j\}$  and  $\{Y_j\}$

The Behrens-Fisher likelihood function:

$$L = \prod_{j=1}^{N_1} (2\pi)^{-p/2} |\Sigma_1|^{-1/2} \exp \left( -\frac{1}{2} (x_j - \mu)' \Sigma_1^{-1} (x_j - \mu) \right) \\ \cdot \prod_{j=1}^{N_2} (2\pi)^{-p/2} |\Sigma_2|^{-1/2} \exp \left( -\frac{1}{2} (y_j - \mu)' \Sigma_2^{-1} (y_j - \mu) \right)$$

Sample means:

$$\bar{X} = N_1^{-1} \sum_{j=1}^{N_1} X_j, \quad \bar{Y} = N_2^{-1} \sum_{j=1}^{N_2} Y_j$$

Standard calculations lead to the likelihood equations:

$$\hat{\Sigma}_1 = N_1^{-1} \sum_{j=1}^{N_1} (X_j - \hat{\mu})(X_j - \hat{\mu})',$$

$$\hat{\Sigma}_2 = N_2^{-1} \sum_{j=1}^{N_2} (Y_j - \hat{\mu})(Y_j - \hat{\mu})'$$

$$(N_1 \hat{\Sigma}_1^{-1} + N_2 \hat{\Sigma}_2^{-1}) \hat{\mu} = N_1 \hat{\Sigma}_1^{-1} \bar{X} + N_2 \hat{\Sigma}_2^{-1} \bar{Y}$$

At first glance, a system of  $p(p + 2)$  equations

The equations cannot be solved algebraically

A numerical value or formula for  $\hat{\mu}$  leads to the same for  $\hat{\Sigma}_1, \hat{\Sigma}_2$

Each  $\hat{\Sigma}_i$  is a sum of rank-one matrices, e.g.,

$$\hat{\Sigma}_1 = N_1^{-1} \sum_{j=1}^{N_1} (X_j - \hat{\mu})(X_j - \hat{\mu})'$$

If  $N_1, N_2 > p$  then, a.s., each  $\hat{\Sigma}_i$  is positive definite

A crucial observation:

$$\hat{\Sigma}_1 = \tilde{S}_1 + (\bar{X} - \hat{\mu})(\bar{X} - \hat{\mu})'$$

where

$$\tilde{S}_1 = N_1^{-1} \sum_{j=1}^{N_1} (X_j - \bar{X})(X_j - \bar{X})'$$

$N_1 \tilde{S}_1 \sim W_p(N_1 - 1, \Sigma_1)$ , the Wishart distribution

If  $N_1 > p$  then  $\tilde{S}_1$  is positive definite, a.s.

Similar results hold for  $\hat{\Sigma}_2$

This leads to an algorithm for studying the system of equations

## An iterative algorithm

(1) Start with initial estimates  $\hat{\Sigma}_{i,0} = \tilde{S}_i$ ,  $i = 1, 2$ , where

$$\tilde{S}_1 = N_1^{-1} \sum (X_j - \bar{X})(X_j - \bar{X})'$$

$$\tilde{S}_2 = N_2^{-1} \sum (Y_j - \bar{Y})(Y_j - \bar{Y})'$$

(2) Calculate the corresponding initial estimate of  $\mu$ :

$$\hat{\mu}_0 = (N_1 \hat{\Sigma}_{1,0}^{-1} + N_2 \hat{\Sigma}_{2,0}^{-1})^{-1} (N_1 \hat{\Sigma}_{1,0}^{-1} \bar{X} + N_2 \hat{\Sigma}_{2,0}^{-1} \bar{Y})$$

(3) Use  $\hat{\mu}_0$  from Step 2 to calculate  $\hat{\Sigma}_{i,1}$  using the formulas

$$\hat{\Sigma}_{1,1} = \tilde{S}_1 + (\bar{X} - \hat{\mu}_0)(\bar{X} - \hat{\mu}_0)'$$

$$\hat{\Sigma}_{2,1} = \tilde{S}_2 + (\bar{Y} - \hat{\mu}_0)(\bar{Y} - \hat{\mu}_0)'$$

(4) Return to Step 2 until both  $\hat{\Sigma}_{1,j}$  and  $\hat{\Sigma}_{2,j}$  converge

## Potential problems with this algorithm

---

Do the choice of initial values of the  $\hat{\Sigma}_i$  lead to convergence?

Will the algorithm converge to a global maximum of  $\log L$ ?

How many stationary points does  $L$  have?

How many solutions are there to the system of equations?

The number of equations is quadratic in  $p$

Can we reduce the number of equations from  $p(p + 2)$  to  $p$ ?

## Woodbury's theorem

---

Let  $M$  be a  $p \times p$  nonsingular matrix and  $v \in \mathbb{R}^p$ . Then

$$(M + vv')^{-1} = M^{-1} - \frac{M^{-1}vv'M^{-1}}{1 + v'M^{-1}v}$$

Multiply this equation on the right by  $v$ , and simplify:

$$\begin{aligned}(M + vv')^{-1}v &= M^{-1}v - \frac{M^{-1}vv'M^{-1}v}{1 + v'M^{-1}v} \\ &= \frac{(1 + v'M^{-1}v)M^{-1}v - (M^{-1}v)(v'M^{-1}v)}{1 + v'M^{-1}v} \\ &= \frac{M^{-1}v}{1 + v'M^{-1}v}\end{aligned}$$

Set  $M = \tilde{S}_1$ ,  $v = \bar{X} - \hat{\mu}$ :

$$\begin{aligned}\hat{\Sigma}_1^{-1}(\bar{X} - \hat{\mu}) &\equiv (\tilde{S}_1 + (\bar{X} - \hat{\mu})(\bar{X} - \hat{\mu})')^{-1}(\bar{X} - \hat{\mu}) \\ &= \frac{\tilde{S}_1^{-1}(\bar{X} - \hat{\mu})}{1 + (\bar{X} - \hat{\mu})'\tilde{S}_1^{-1}(\bar{X} - \hat{\mu})}\end{aligned}$$

Similarly,

$$\hat{\Sigma}_2^{-1}(\bar{Y} - \hat{\mu}) = \frac{\tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})}{1 + (\bar{Y} - \hat{\mu})'\tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})}$$

The third set of likelihood equations:

$$(N_1 \hat{\Sigma}_1^{-1} + N_2 \hat{\Sigma}_2^{-1}) \hat{\mu} = N_1 \hat{\Sigma}_1^{-1} \bar{X} + N_2 \hat{\Sigma}_2^{-1} \bar{Y}$$

Rewrite this as

$$N_1 \hat{\Sigma}_1^{-1} (\bar{X} - \hat{\mu}) + N_2 \hat{\Sigma}_2^{-1} (\bar{Y} - \hat{\mu}) = 0$$

Substitute the consequence of Woodbury's theorem:

$$\frac{N_1 \tilde{S}_1^{-1} (\bar{X} - \hat{\mu})}{1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu})} + \frac{N_2 \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu})}{1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu})} = 0$$

We can clear denominators since  $\hat{\mu} \in \mathbb{R}^p$  and  $N_1, N_2 > p$

$$N_1(1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1}(\bar{Y} - \hat{\mu})) \tilde{S}_1^{-1}(\bar{X} - \hat{\mu}) \\ + N_2(1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1}(\bar{X} - \hat{\mu})) \tilde{S}_2^{-1}(\bar{Y} - \hat{\mu}) = 0$$

A system of  $p$  equations in the components of  $\mu$

Each equation is of degree 3

Algebraic geometry: Solving systems of polynomial equations

Bézout's theorem: The number of ( $\mathbb{R}$  or  $\mathbb{C}$ ) solutions of our system is at most  $3^p$

“at most”?

## Historical remarks on polynomial systems

---

1-variable (The Fundamental Theorem of Algebra): A polynomial of degree  $n$  has exactly  $n$  roots

2-variables: Minding (1838)

$p$  variables: Bernstein- Khovanskiĭ-Kouchnirenko (1975) – BKK

BKK's revolutionary result: A formula for the exact number of roots in terms of the monomials appearing in the system of equations

B. Sturmfels (1998). Polynomial equations and convex polytopes. *Amer. Math. Monthly*, **105**, 907–922.

Cox-Little-O'Shea (1998). *Using Algebraic Geometry*. Springer, New York.

## Application of BKK to Behrens-Fisher

---

Carrier set: The set of monomials which appear in a polynomial

We work with monomials or with vectors of corresponding exponents

Example: Identify  $x^3y^2z^{14}$  with  $(3, 2, 14)$

A carrier set is a collection of vectors with nonnegative integral components

$C_1, C_2$ : Carrier sets corresponding to polynomials  $h_1, h_2$

Minkowski sum:  $C_1 + C_2 = \{u + v : u \in C_1, v \in C_2\}$

$C_1 + C_2$  is the carrier set of  $h_1h_2$

The Minkowski sum is commutative and associative

## Behrens-Fisher meets BKK

---

$e_1, \dots, e_p$ : The standard basis for the vector space  $\mathbb{R}^p$

The Behrens-Fisher equations:

$$\begin{aligned} N_1(1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu})) \tilde{S}_1^{-1} (\bar{X} - \hat{\mu}) \\ + N_2(1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu})) \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu}) = 0 \end{aligned}$$

The quadratic term  $1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu})$  has carrier set  $\{0, e_1, \dots, e_p\} + \{0, e_1, \dots, e_p\}$

The linear term  $\bar{Y} - \hat{\mu}$  has carrier set  $\{0, e_1, \dots, e_p\}$

The carrier set of each Behrens-Fisher equation is the Minkowski sum

$$\Pi := \{0, e_1, \dots, e_p\} + \{0, e_1, \dots, e_p\} + \{0, e_1, \dots, e_p\}$$

Fact:

$$\Pi = \left\{ \sum_{j=1}^p n_j e_j : n_j \in \mathbb{N}_0, 1 \leq j \leq p, \sum_{j=1}^p n_j \leq 3 \right\}$$

Each  $n_j$  is the number of times that  $e_j$  appears in formation of the Minkowski sum

Construct the convex hull of  $\Pi$

$$\begin{aligned}\mathcal{C}(\Pi) &= \left\{ \sum_{j=1}^p u_j e_j : u_j \geq 0, 1 \leq j \leq p, \sum_{j=1}^p u_j \leq 3 \right\} \\ &\equiv \left\{ (u_1, \dots, u_p) : u_j \geq 0, 1 \leq j \leq p, \sum_{j=1}^p u_j \leq 3 \right\}.\end{aligned}$$

BKK's Theorem: The exact number of isolated solutions of the system is  $p! \cdot \text{Vol}(\mathcal{C}(\Pi))$

$$\text{Vol}(\mathcal{C}(\Pi)) = \int_{\mathcal{C}(\Pi)} du_1 \cdots du_p$$

This is well-known to fans of the Dirichlet distributions

$$\text{Vol}(\mathcal{C}(\Pi)) = \frac{3^p}{\Gamma(p+1)} = \frac{3^p}{p!}$$

Theorem: The Behrens-Fisher problem has exactly  $3^p$  solutions

Almost surely, every solution is isolated

Roots appear in complex conjugate pairs, so there is at least one real root

$p = 1$ : Sugiura and Gupta (1987)

## How many real roots are there?

---

Even in the 1-D case, this is difficult to say

Consider the case in which  $p = 2$

We “randomly” generated  $(N_1, N_2, \mu, \Sigma_1, \Sigma_2)$

We simulated a sample of size  $N_j$  from  $N_p(\mu, \Sigma_j)$ ,  $j = 1, 2$

Verschelde (1999): PHCpack, a numerical continuation program for solving systems of polynomial equations

We applied PHCpack to solve the system of equations

Repeat the process thousands of times

It is possible to find situations in which multiple root occur

Example:  $N_1 = 11, N_2 = 5$

$$\bar{X} = \begin{pmatrix} -1.5516 \\ -9.4713 \end{pmatrix}, \quad \tilde{S}_1 = \begin{pmatrix} 0.3998 & -0.1026 \\ -0.1026 & 0.2378 \end{pmatrix}$$

$$\bar{Y} = \begin{pmatrix} -1.9175 \\ -10.4805 \end{pmatrix}, \quad \tilde{S}_2 = \begin{pmatrix} 0.4193 & 0.0792 \\ 0.0792 & 0.0334 \end{pmatrix}$$

The real solutions for  $\hat{\mu}$  are:

$$\begin{pmatrix} -1.3570 \\ -10.2957 \end{pmatrix}, \quad \begin{pmatrix} -1.2478 \\ -9.9902 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} -1.4451 \\ -9.6333 \end{pmatrix}$$

This example seems to be a rare exception. Our simulated bivariate Behrens-Fisher equations had:

1 real solution about 99.5% of the time

3 real solutions about 0.5% of the time

We found no cases having 5 or more real solutions

We wanted to test for differences between small-sample and large-sample cases

We simulated Behrens-Fisher problems with  $N_1, N_2 \leq 15$

Simulations with  $3 \leq N_1, N_2 \leq 15$

Number of solutions	Frequency	Percentage
1	4450	99.29%
3	32	0.71%

No simulation returned more than 3 real solutions

For larger sample sizes, our simulations resulted in:

Simulations with  $15 \leq N_1, N_2 \leq 60$

Number of solutions	Frequency	Percentage
1	4404	99.46%
3	24	0.54%

Again, no simulation resulted in more than 3 real solutions

There seems to be little chance that a randomly generated, two-dimensional Behrens-Fisher problem will have 3 or more real solutions, and there is a high chance that it will have a unique real solution

Is there an explanation for this phenomenon?

It can be shown that

$$\begin{aligned} L(\hat{\mu}, \hat{\Sigma}_1, \hat{\Sigma}_2) &= (2\pi e)^{-(N_1+N_2)p/2} |\tilde{S}_1|^{-N_1/2} |\tilde{S}_2|^{-N_2/2} \\ &\quad \times (1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu}))^{-N_1/2} \\ &\quad \times (1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu}))^{-N_2/2} \end{aligned}$$

We can maximize  $L$  by evaluating

$$(1 + (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu}))^{-N_1/2} \cdot (1 + (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu}))^{-N_2/2}$$

at each real solution of the system of score equations

This can be done computationally using PHCpack

Can we express  $\hat{\mu}$  entirely in terms of the data?

Probably not, but the formula for  $L(\hat{\mu})$  provides us with information about the existence of multiple solutions of the likelihood equations

If  $\bar{X} = \bar{Y}$  then the formula for  $L(\hat{\mu}, \hat{\Sigma}_1, \hat{\Sigma}_2)$  shows that the unique solution of the likelihood equations is  $\hat{\mu} = \bar{X} = \bar{Y}$

If the distance between  $\bar{X}$  and  $\bar{Y}$  is small (in the sense that  $(\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu})$  and  $(\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu})$  are close) then the likelihood equations are likely to have a unique real solution

Conclusion: Behrens-Fisher problems arising in practical data analysis are likely to have a unique real solution, because inference about the difference between the population means usually is performed only if that difference, *viz.*,  $\mu$ , is small

We now devise a new iterative algorithm for solving Behrens-Fisher problems

Easy algebraic identity for vectors  $u, v \in \mathbb{R}^p$ :

$$u'u - v'v = (u - v)'(u + v)$$

Set  $u = \tilde{S}_1^{-1/2}(\bar{X} - \hat{\mu})$ ,  $v = \tilde{S}_2^{-1/2}(\bar{Y} - \hat{\mu})$ :

$$\begin{aligned} & (\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu}) - (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu}) \\ &= (\tilde{S}_1^{-1/2} \bar{X} - \tilde{S}_2^{-1/2} \bar{Y} - (\tilde{S}_1^{-1/2} \hat{\mu} - \tilde{S}_2^{-1/2}) \hat{\mu})' \\ & \quad \times (\tilde{S}_1^{-1/2} \bar{X} + \tilde{S}_2^{-1/2} \bar{Y} - (\tilde{S}_1^{-1/2} + \tilde{S}_2^{-1/2}) \hat{\mu}). \end{aligned}$$

To make  $(\bar{X} - \hat{\mu})' \tilde{S}_1^{-1} (\bar{X} - \hat{\mu}) \approx (\bar{Y} - \hat{\mu})' \tilde{S}_2^{-1} (\bar{Y} - \hat{\mu})$ , we choose  $\hat{\mu}$  so that  $\tilde{S}_1^{-1/2} \bar{X} + \tilde{S}_2^{-1/2} \bar{Y} - (\tilde{S}_1^{-1/2} + \tilde{S}_2^{-1/2}) \hat{\mu} \approx 0$

## A new iterative algorithm for Behrens-Fisher problems

---

(1) Start with the initial value for  $\hat{\mu}$ :

$$\hat{\mu}_0 = (\tilde{S}_1^{-1/2} + \tilde{S}_2^{-1/2})^{-1} (\tilde{S}_1^{-1/2} \bar{X} + \tilde{S}_2^{-1/2} \bar{Y})$$

(2) Calculate corresponding estimates of the  $\hat{\Sigma}_i$ ,  $i = 1, 2$ :

$$\hat{\Sigma}_{1,0} = \tilde{S}_1 + (\bar{X} - \hat{\mu}_0)(\bar{X} - \hat{\mu}_0)'$$

$$\hat{\Sigma}_{2,0} = \tilde{S}_2 + (\bar{Y} - \hat{\mu}_0)(\bar{Y} - \hat{\mu}_0)'$$

(3) Update  $\hat{\mu}_0$ , using the likelihood equation for  $\mu$ :

$$\hat{\mu}_1 = (N_1 \hat{\Sigma}_{1,0}^{-1} + N_2 \hat{\Sigma}_{2,0}^{-1})^{-1} (N_1 \hat{\Sigma}_{1,0}^{-1} \bar{X} + N_2 \hat{\Sigma}_{2,0}^{-1} \bar{Y})$$

(4) Return to Step 2 until both  $\hat{\Sigma}_{1,j}$  and  $\hat{\Sigma}_{2,j}$  converge

It is a challenging problem to ascertain the convergence properties of this algorithm