Largest eigenvalues and eigenvectors in multivariate analysis

Iain Johnstone, Statistics, Stanford

imj@stanford.edu

Yale, May, 2009

What's this?

Eigenvalues: Theme

- **Many problems of classical multivariate statistics involve eigenvalues** $x_1 > x_2 > \ldots > x_p$
- **Asymptotics in** $p \Rightarrow$ **useful information, [even** p **small]**
- **Illustrate for null distributions for** x_1

Eigenvalues: Theme

- **Many problems of classical multivariate statistics involve eigenvalues** $x_1 > x_2 > \ldots > x_p$
- **Asymptotics in** $p \Rightarrow$ **useful information, [espec.** p **small]**
- **Illustrate for null distributions for** x_1

Example: (multiple) Regression.

$$\underset{n\times p}{y} = \underset{n\times q}{X}\ \underset{q\times p}{\beta} + \underset{n\times p}{\epsilon} \qquad \epsilon \sim N(0, I_n \otimes \Sigma_p).$$

Tests of $H_0 : \beta = 0$ **use roots of** $\det[A + x_i(A + B)] = 0$.

$$A = SS_H = \hat{\beta}^T X^T X \hat{\beta} \qquad \text{“Hypothesis”}$$

$$B = SS_E = y^T (I - P_X) y \qquad \text{“Error”}$$

Double Wishart Setting

$$A \sim W_p(n_1, I)$$
$$B \sim W_p(n_2, I)$$

2 independent Wisharts, $p \leq n_1, n_2$

"null hypothesis" setting

Common feature: roots $:= (x_i)_{i=1}^p$ **of generalized eigenproblem:**

$$\det[x(A+B) - A] = 0$$

Single Wishart

- **Principal Component analysis**
- **Factor analysis**
- **Multidimensional scaling**

Double Wishart

- **Canonical correlation analysis**
- **Multivariate Analysis of Variance (MANOVA)**
- **Multivarate regression analysis**
- **Discriminant analysis**
- **Tests of equality of covariance matrices**

Joint density of eigenvalues

Single Wishart: $\det[A - x_i I] = 0.$

Double Wishart: $\det[A - x_i(A+B)] = 0.$

In each case, (Fisher, Girshick, Hsu, Mood, Roy, (1939)):

$$f(x_1, \ldots, x_p) = c \prod_i w^{1/2}(x_i) \prod_{i<j} (x_i - x_j) \qquad x_1 \geq \ldots \geq x_p$$

Single Wishart: $w(x) = x^{n-p} e^{-x},$ **(Laguerre)**

Double Wishart: $w(x) = x^{p-q-1}(1-x)^{n-p-q-1}.$ **(Jacobi)**

Outline

I. Setting

   1. Single & Double Wishart, examples

II. Largest Eigenvalue: Limiting Law (H_0)

   1. Gaussian
   2. Laguerre/Single W.
   3. Jacobi/Double W.

III. Largest Eigenvalue: Concentration

   1. Single W.
   2. Double W.

IV. Largest Eigenvector(s)

Symmetric Gaussian matrices

A **– symmetric** $N \times N$ **matrix drawn from (**$\beta = 1$ **for GOE)**

$$f(A) = c \exp\{-\tfrac{\beta}{2} \mathbf{tr}(A^T A)\}.$$

[for GUE: complex Hermitian]. Largest eigenvalue:

$$\theta_N = \lambda_{max}(A)$$

Centering and scaling constants:

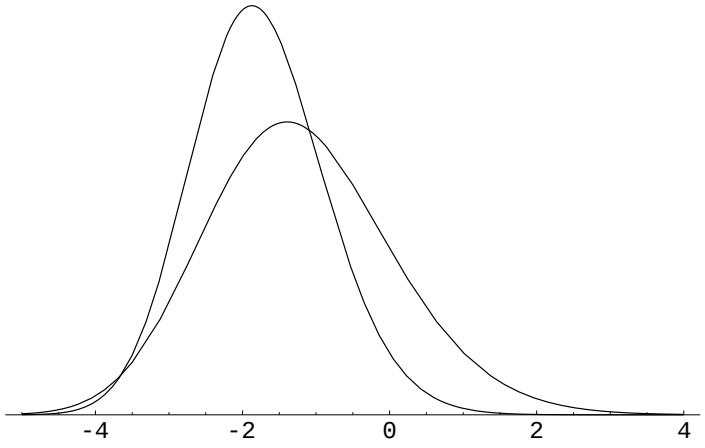
$$\mu_N = \sqrt{2N}, \qquad \sigma_N = 2^{-1/2} N^{-1/6}$$

ratio $N^{-2/3}$**!**

Tracy-Widom limit: TW (1994, 1996) For $\beta = 1, 2, 4 \quad (\mathbb{R}, \mathbb{C}, \mathbb{Q}),$

$$\frac{\theta_N - \mu_N}{\sigma_N} \stackrel{\mathcal{D}}{\Rightarrow} F_\beta.$$

Tracy-Widom distributions (1994,96)



$$q'' = sq + 2q^3 \quad \textbf{(Painlevé II)}$$

$$q(s) \sim \textbf{Ai}(s) \textbf{ as } s \to \infty$$

$$F_2(s) = e^{-\int_s^\infty (x-s)^2 q(x)dx}$$

$$F_1(s)^2 = F_2(s) e^{-\int_s^\infty q(x)dx}.$$

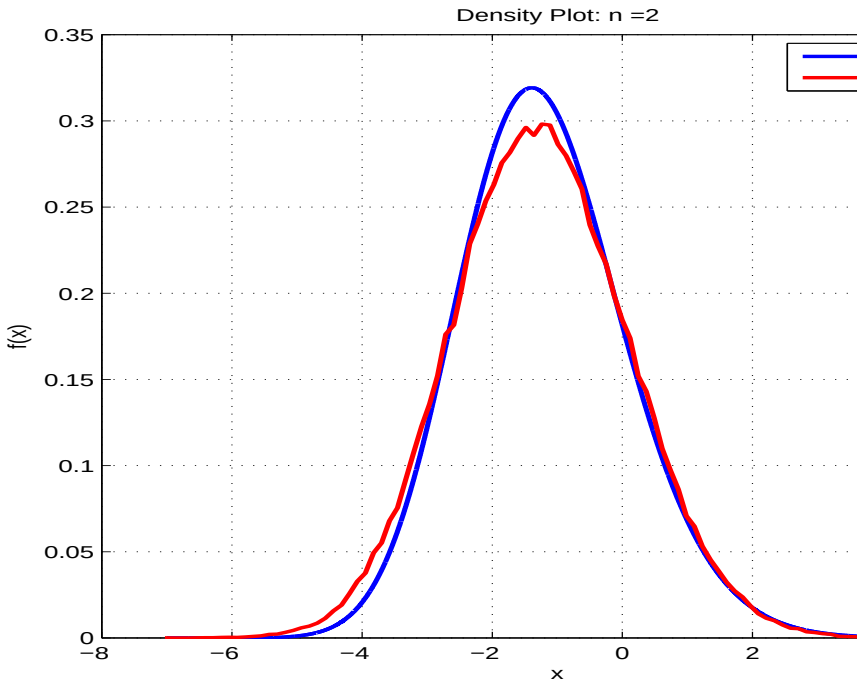
Right tail decay: $1 - F_\beta(s) \approx e^{-(2/3)s^{3/2}}$.

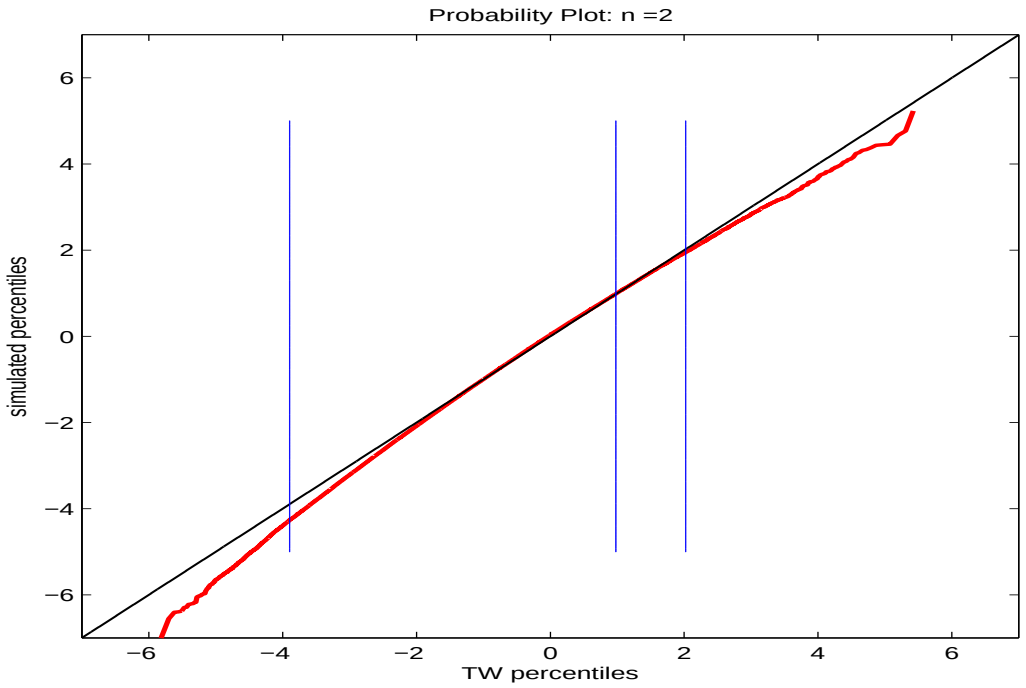
Rate of convergence: **First order** $N^{-1/3}$, **Choup (2006,07)**

"Second-order": J + Ma (2008) For $\beta = 1, 2$, **&** $\mu_N \overset{(\beta=1)}{=} \sqrt{2N-1}$,

$$|P\{\lambda_{max}(A) \le \mu_N + \sigma_N s\} - F_\beta(s)| \le C N^{-2/3} e^{-s/2}.$$

Approximations at $N = 2$

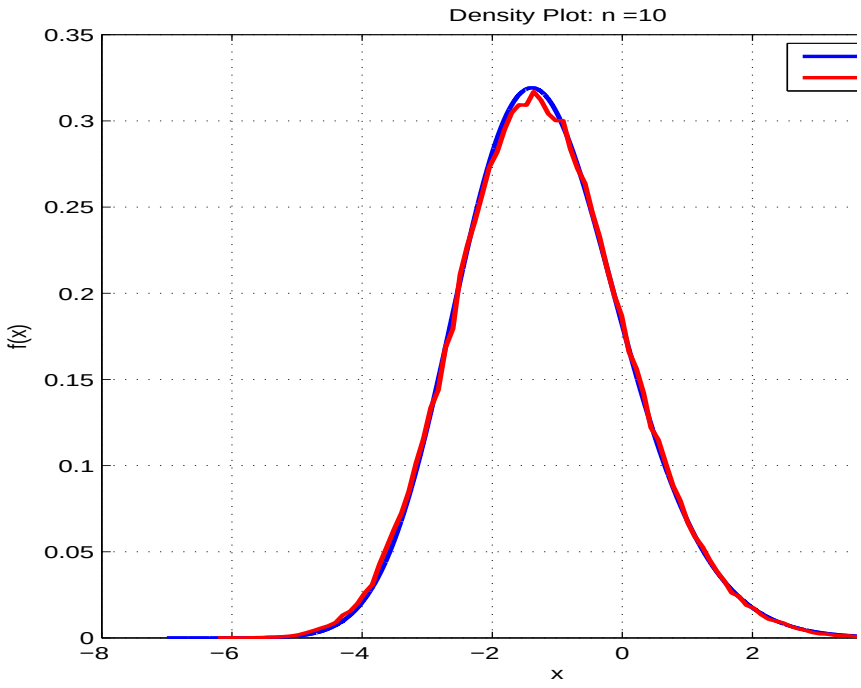


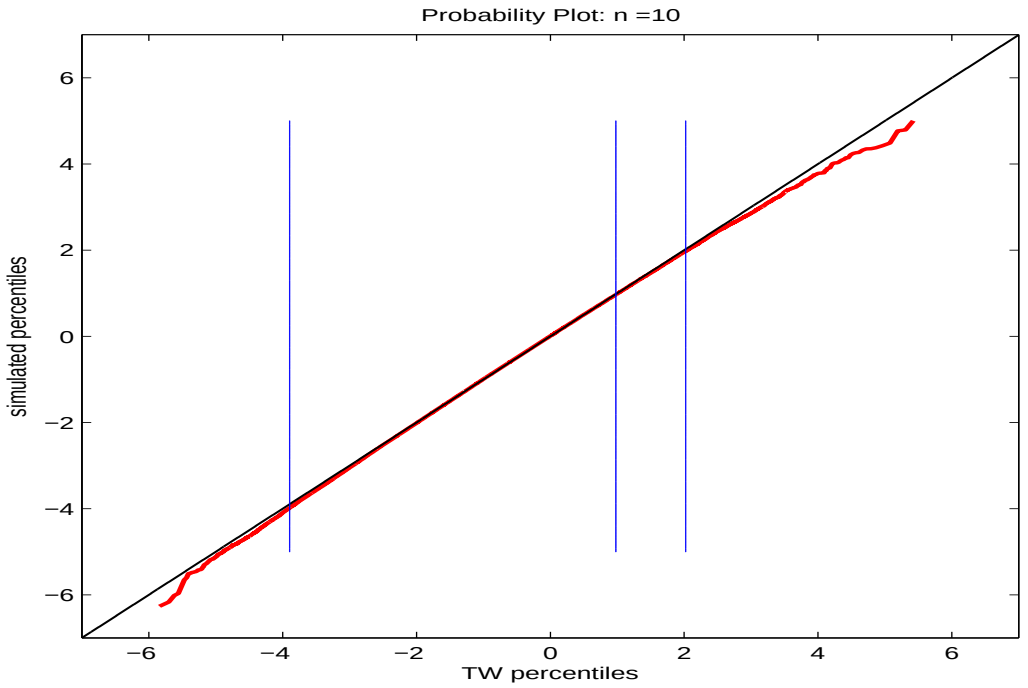


– Better approximation in right tail : location of turning point of Airy function

– additional $O\left(N^{-4/3}\right)$ mean correction included

Approximations at $N = 10$





Outline

I. Setting

   1. Single & Double Wishart, examples

II. Largest Eigenvalue: Limiting Law (H_0)

   1. Gaussian
   2. Laguerre/Single W.
   3. Jacobi/Double W.

III. Largest Eigenvalue: Concentration

   1. Single W.
   2. Double W.

IV. Largest Eigenvector(s)

Tracy Widom Limits

Theorem For {real, complex}, {single, double} Wishart matrices, if $n/p \to \gamma$, [or $(n_1/p, n_2/p) \to (\gamma_1, \gamma_2)$,] then

$$P\{n\hat{\ell}_1 \le \mu_{np} + \sigma_{np} s | H_0\} \to F_\beta(s)$$

– "universality" in RMT (Deift, 06 ICM)

Single Wishart: $\beta = 2$ (Johansson, 00), $\beta = 1$ (J, 01)

"Microarray case" $n, p \to \infty, \quad n/p \to 0, \infty.$ (El Karoui, 03)

Non Gaussian: (Soshnikov, 02, 06; S. + Fyodorov, 06, Péché, 07)

- subGaussian: $\Rightarrow$ TW limit; heavy tails $\Rightarrow$ Poisson

Second order accuracy

For {real, complex}, {single, double } Wishart matrices, if

$n/p \to \gamma,$ **[or** $(n_1/p, n_2/p) \to (\gamma_1, \gamma_2),$**]** **then**

$$|P\{n\hat{\ell}_1 \le \mu_{np} + \sigma_{np} s | H_0\} - F_\beta(s)| \le C e^{-cs} \mathbf{p^{-2/3}}.$$

Single Wishart **Complex: El Karoui (2006).**

Real: Ma (2008, poster here)

$$\mu_{np} = \left(\sqrt{n-\tfrac{1}{2}} + \sqrt{p-\tfrac{1}{2}}\right)^2$$

$$\sigma_{np} = \left(\sqrt{n-\tfrac{1}{2}} + \sqrt{p-\tfrac{1}{2}}\right)\left(\frac{1}{\sqrt{n-\frac{1}{2}}} + \frac{1}{\sqrt{p-\frac{1}{2}}}\right)^{1/3}.$$

Beyond the "Null Hypothesis"

Classical RMT ensembles (e.g. $W_p(n, I)$)

$\leftrightarrow$ "null hypothesis", symmetry, no structure.

For $W_p(n, \Sigma)$: For what conditions on Σ does

$$P\{\hat{\ell}_1 \leq \mu_{np}(\Sigma) + \sigma_{np}(\Sigma)s\} \to F_\beta(s) \qquad ??$$

***Some* answers:**

- **sufficiently many $\ell_k(\Sigma)$ accumulate near $\ell_1(\Sigma)$ (data in $\mathbb{C}$)**

  El Karoui, 2007

- **small number of (not too!) isolated $\ell_i(\Sigma)$**

  Baik-Ben Arous-Péché, 2005, Baik-Silverstein 2006, Paul 2007

  Harding (2008, Economics Letters)

Outline

I. Setting

   1. Single & Double Wishart, examples

II. Largest Eigenvalue: Limiting Law (H_0)

   1. Gaussian
   2. Laguerre/Single W.
   3. Jacobi/Double W.

III. Largest Eigenvalue: Concentration

   1. Single W.
   2. Double W.

IV. Largest Eigenvector(s)

Roy's Greatest Root Test: Package Output

SAS: (From Gledhill et. al.: NOAA Fisheries Reef Fish Video Surveys.)

Table 18. Multivariate statistics and F approximations for the CANDISC procedure ran with all mincount and habitat data.

S=5 M=2 N=48.5

Statistic	*Value*	*F Value*	*Num DF*	*Den DF*	*Pr > F*
Wilks' Lambda	0.58109196	1.15	50	454.87	0.2325
Pillai's Trace	0.50115530	1.15	50	515	0.2342
Hotelling-Lawley Trace	0.59036501	1.15	50	305.82	0.2355
Roy's Greatest Root	0.26759820	2.76	10	103	0.0047

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

R: (`car, heplots` packages, Fox et. al.)

```
Multivariate Tests:
                    Df test stat approx F   num Df   den Df    Pr(>F)
Pillai          5.0000  0.417938 1.845226  15.0000 171.0000 0.0320861 *
Wilks           5.0000  0.623582 1.893613  15.0000 152.2322 0.0276949 *
Hotelling-Lawley 5.0000  0.538651 1.927175  15.0000 161.0000 0.0239619 *
Roy             5.0000  0.384649 4.384997   5.0000  57.0000 0.0019053 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Tracy-Widom approximation

Let $W = \textbf{logit}(x_1) = \log(x_1/(1-x_1))$.

Result: [J,08] **As** $p \propto n_1, n_2 \to \infty$, **(and with** $O(p^{-2/3})$ **error):**

$$\frac{W - \mu_p}{\sigma_p} \stackrel{\mathcal{D}}{\Rightarrow} W_\infty \sim F_1.$$

In other words:

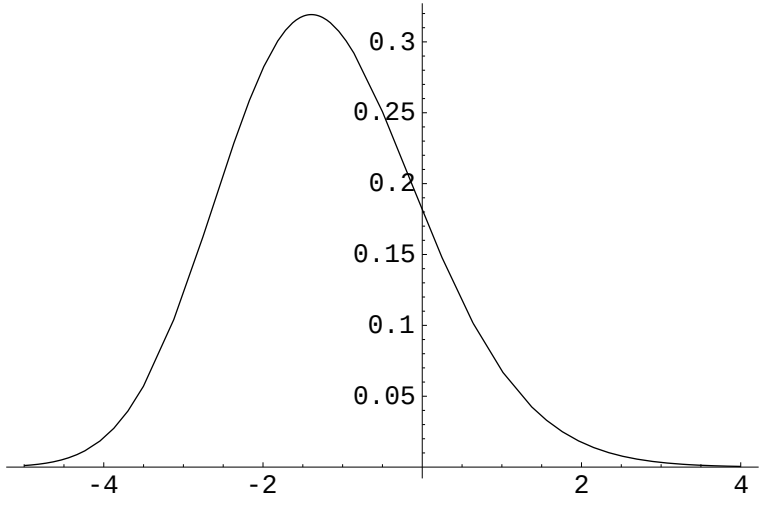
$$x_1 \approx \frac{e^{\mu+\sigma W_\infty}}{1+e^{\mu+\sigma W_\infty}}.$$

Percentiles f_α**:**

$$f_{.90} = 0.4501$$

$$f_{.95} = 0.9793$$

$$f_{.99} = 2.0234$$



Centering and Scaling Constants

- **set** $N = n_1 + n_2 - 1$

- **define** γ, ϕ **from**

$$\sin^2(\gamma/2) = (p - \tfrac{1}{2})/N, \qquad \sin^2(\phi/2) = (n_1 - \tfrac{1}{2})/N$$

- **Then** μ **and** σ **are given by**

$$\mu_p = 2 \log \tan\left(\frac{\phi + \gamma}{2}\right), \qquad \sigma_p^3 = \frac{16}{N^2} \frac{1}{\sin^2(\phi + \gamma) \sin\phi \sin\gamma}.$$

- **(given a table of** F_1**), straightforward to code:**

`papptw`, `qapptw`, `rapptw`.

Accuracy of approximate α^{th} percentile

Approximate percentile using Tracy-Widom percentile f_α:

$$x_\alpha = x_\alpha^{TW}(p, n_1, n_2) = e^{\mu_p + f_\alpha \sigma_p} / (1 + e^{\mu_p + f_\alpha \sigma_p}).$$

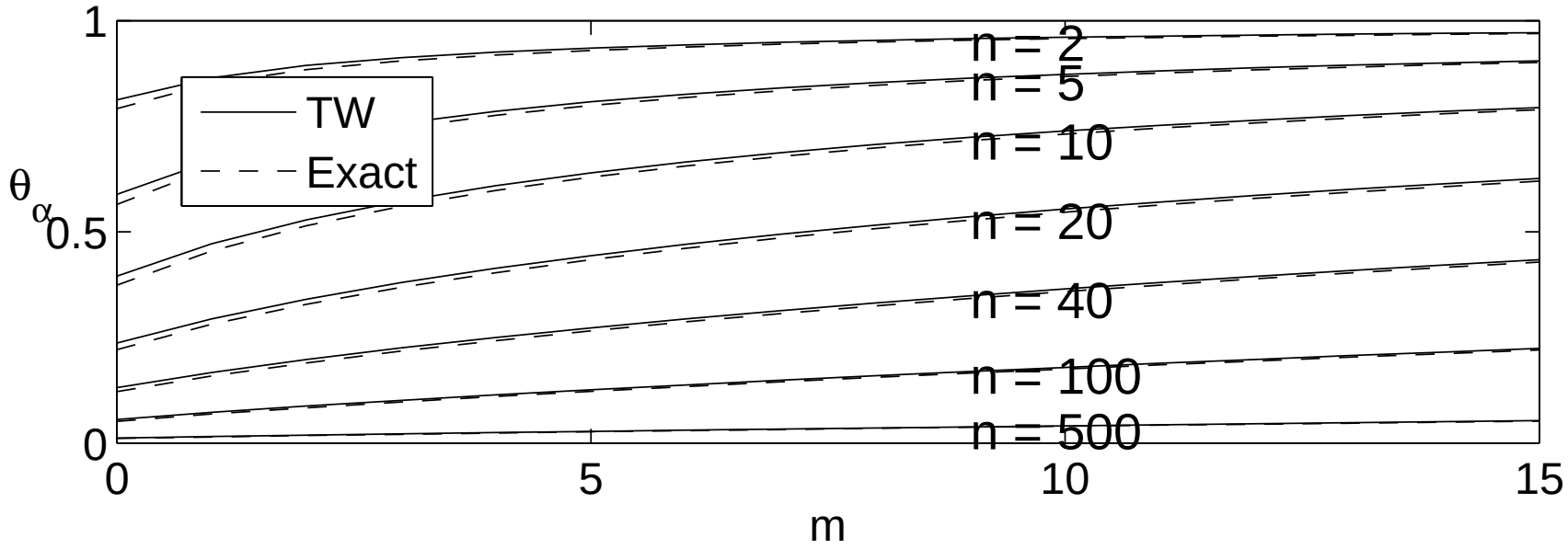
William Chen's tables: (2003, 2004a, 2002, 2004b) $\Rightarrow$ can compare

- **'Exact'** $x_\alpha(p, n_1, n_2)$ **with**
- **T-W approx** $x_\alpha^{TW}(p, n_1, n_2)$**.**

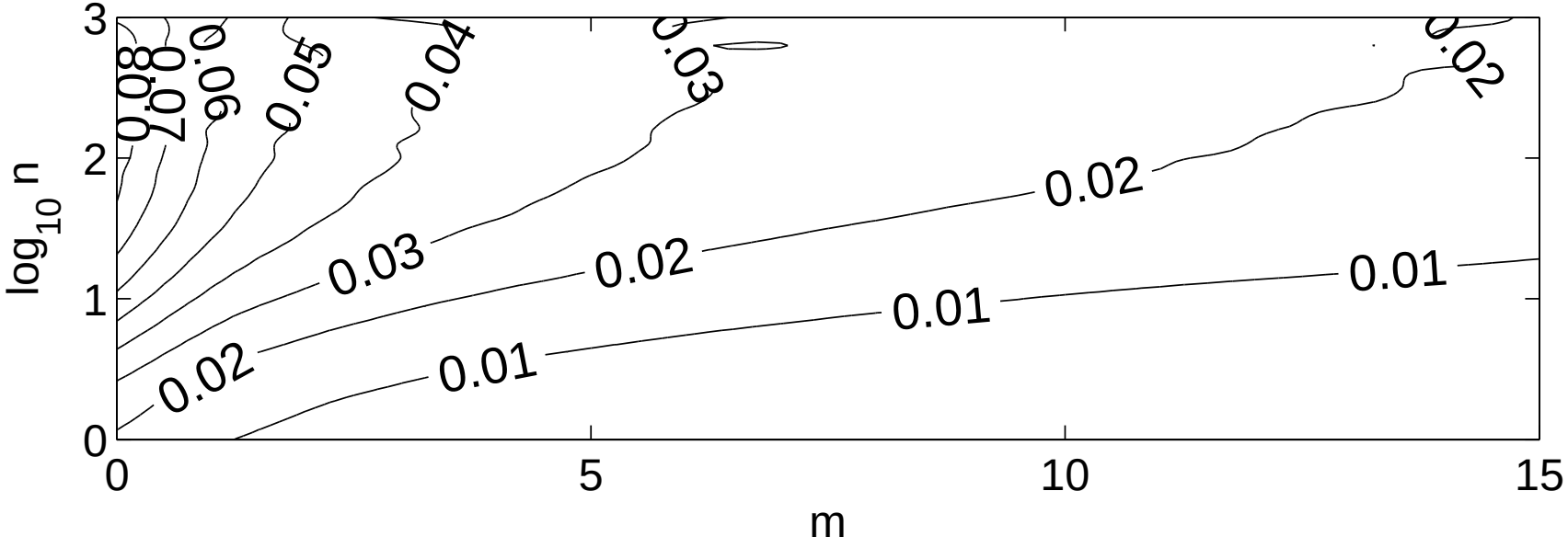
Relative error: $r = (x_\alpha^{TW} / x_\alpha) - 1.$

Tracy-Widom based on $p \to \infty$, but try $\boxed{p = 2}$ as n_1, n_2 vary

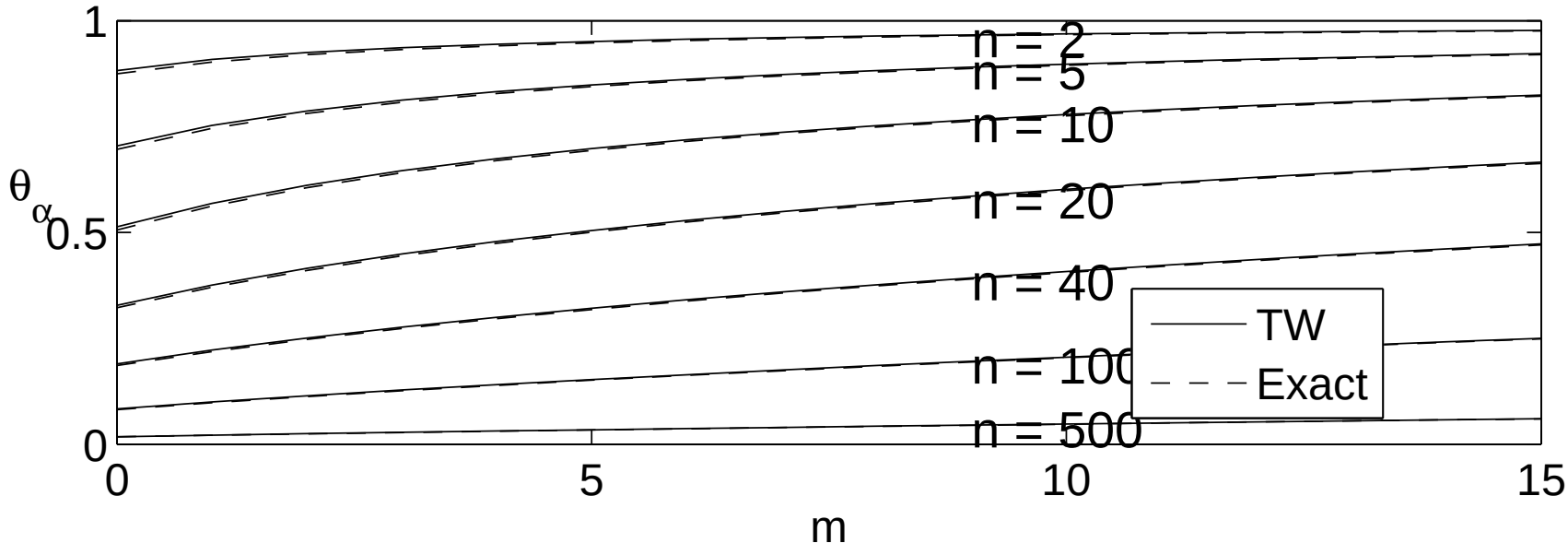
$p = 2$ at 95th percentile



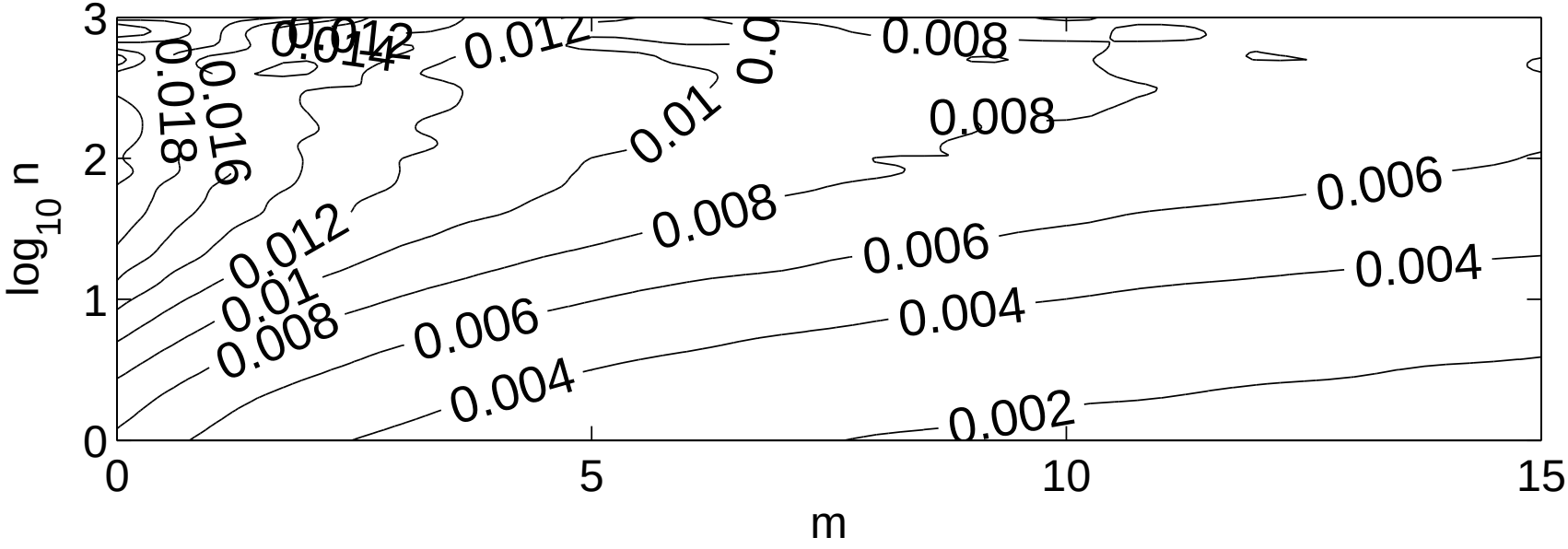
Contours of relative error $r = (\theta_\alpha^{TW}/\theta_\alpha)-1$



p = 4 at 90th percentile



90th %tile, S=4, Contours of relative error r = $(\theta_\alpha^{TW}/\theta_\alpha)-1$



Recap, for Double Wishart

T-W approximation to null distribution of Roy's largest root:

- **Conventional percentiles: generally accurate to $< 10\%$ relative error**
- **[Rough $p-$value assessments: qualitatively o.k. over many orders of magnitude]**
- **Similar $O(N^{-2/3})$ approximations for Wishart and Gaussian.**

Recommend: Tracy-Widom approximation replace $F-$ lower bound in default package printouts

Software: RMTstat

In development, for R, MATLAB (Z. Ma, P. Perry, M. Shahram, IMJ)

[Preliminary R version now at CRAN]

- **Tracy-Widom distribution: `ptw, qtw, dtw, rtw`**
  (Prähofer-Spohn tables + interpolation)
- **Sampling from {Gaussian, Laguerre, Jabobi}, {Unitary, Orthogonal} ensembles**
- **TW approximation to extreme eigenvalues**
  - **null distributions**
  - **'spiked' models**
- **common tests based on largest root**

Outline

I. Setting

   1. Single & Double Wishart, examples

II. Largest Eigenvalue: Limiting Law (H_0)

   1. Gaussian
   2. Laguerre/Single W.
   3. Jacobi/Double W.

III. Largest Eigenvalue: Concentration

   1. Single W.
   2. Double W.

IV. Largest Eigenvector(s)

Wishart: Concentration for l_1

$$A \sim W_p(n, I), \qquad A = X^T X \qquad X \sim N(0, I_n \otimes I_p)$$

Theorem (e.g. Davidson-Szarek, 2001) With $l_1(A) = \sigma_1^2(X)$**,**

$$P\{\sigma_1(X) > E\sigma_1(X) + \sqrt{n}t\} \leq e^{-nt^2/2},$$

and

$$E\sigma_1(X) \leq \sqrt{n} + \sqrt{p}.$$

from Standard Result: If $X \sim N_m(0, I)$ **and** $f : \mathbb{R}^m \to \mathbb{R}$ **is Lipschitz-1 (implied by** $|\nabla f| \leq 1$**), then**

$$P\{f(X) > Ef(X) + s\} \leq e^{-s^2/2}.$$

'Small' vs. 'Large' Deviations

Rescale $X \sim N(0, I_n \otimes I_p)/\sqrt{n}$ **so that** $E\sigma_1(X) \approx 1 + \sqrt{\gamma}$.

Tracy-Widom limit suggests for $t \approx sn^{-2/3}$,

$$P\{\sigma_1 \geq E\sigma_1 + ct\} \approx e^{-(2/3)nt^{3/2}}.$$

Lipschitz concentration of measure $\Rightarrow \forall t > 0$

$$P\{\sigma_1 \geq E\sigma_1 + t\} \leq e^{-nt^2/2}.$$

- **Not optimal (though simple) for 'small' deviations** $t \ll 1$

  (recent: Majumdar-Vargassola)

- **Effective for 'large' deviations** $t > 1$

Double Wishart

Motivations:

- **analog of Davidson-Szarek**
- **arises in Candes-Tao sparse linear regression,** $Y = A\beta + \epsilon$**,**
  restricted orthogonality constants $\theta_{S,S'}$**, and analogs** $\gamma_{S,S'}$**:**

$$\langle A_T v, A_{T'} v' \rangle \leq \theta_{S,S'} \|v\| \|v'\| \qquad (\leq \gamma_{S,S'} \|A_T v\| \|A_{T'} v'\|).$$

Use CCA form: $X \sim N(0, I_n \otimes I_p) \perp\!\!\!\perp Y \sim N(0, I_n \otimes I_q)$

$$R_{p,q;n}(X, Y) = \max\{\textbf{Corr}(Xu, Yv) : \|u\| = \|v\| = 1\}$$

[not Lipschitz in X, Y**!]**

R_{max} as singular value

Use SVD: $X = U_X D_X V_X^T, \qquad Y = U_Y D_Y V_Y^T$

$$\Rightarrow \quad R(X,Y) = \sigma_1(U_X^T U_Y), \qquad \textbf{with}$$

$U_X \in \mathbb{V}_{n,p}$ **– Stiefel manifold of orthonormal** $p-$**frames, etc**

Claim

$$P\{\sigma_1 \geq E\sigma_1 + t\} \leq e^{-(n-2)t^2/8}.$$

$\Rightarrow$ **other double Wishart cases (regression, MANOVA, etc).**

Concentration for Riemannian manifolds

Theorem (Ledoux, 1992) **Assume:**

(1. manifold) (X, g) **compact, connected, smooth Riemannian manifold, dimension** ≥ 2

(2. volume) $d\mu$ **normalized Riemannian volume element**

(3. Lipschitz) F $1-$**Lipschitz on** X**, i.e.** $|\nabla F| \leq 1$

(4. curvature) $c(X)$ $= \inf\{$**Ric**$(\nabla f, \nabla f) : |\nabla f| = 1\}$

Ricci curvature

Then

$$\mu\{F \geq \int F d\mu + t\} \leq e^{-ct^2/2}.$$

Need to interpret (1) - (4) for $F = \sigma_1(U_X^T U_Y) = R(X, Y)$**.**

Outline

I. Setting

   1. Single & Double Wishart, examples

II. Largest Eigenvalue: Limiting Law (H_0)

   1. Gaussian
   2. Laguerre/Single W.
   3. Jacobi/Double W.

III. Largest Eigenvalue: Concentration

   1. Single W.
   2. Double W.

IV. Largest Eigenvector(s)

Estimation of Eigenvectors

$S \sim W_p(n, \Sigma), \qquad \Sigma = \sigma^2 I + \sum_{\nu=1}^{M} \lambda_\nu \theta_\nu \theta_\nu^T$

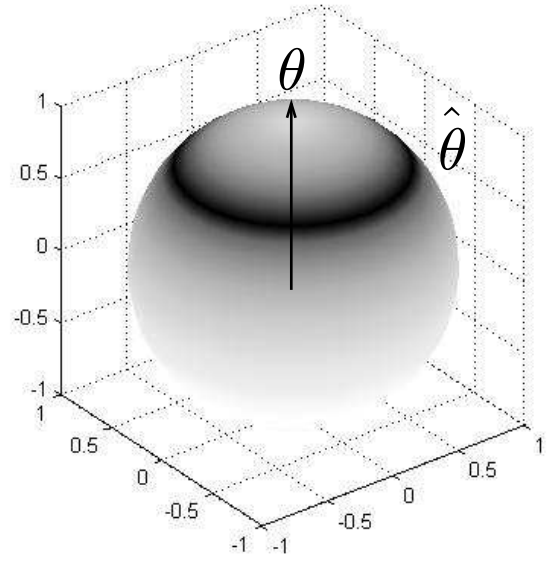
Estimation and inference for θ_ν??

Classical: p **fixed,** n **large:** $\sqrt{n}(\hat{\theta}_\nu - \theta_\nu) \to N_p(0, \Gamma_\nu)$

BUT: inconsistency when $p/n \to \gamma > 0$**:**

Reimann, v.d.Broeck, Bex, Hoyle, Rattray; Paul, Baik, Silverstein

$$\langle \hat{\theta}_\nu, \theta_\nu \rangle \to \begin{cases} 0 & \lambda_\nu \in [0, \sqrt{\gamma}] \\ \frac{1-\gamma/\lambda_\nu^2}{1+\gamma/\lambda_\nu} & \lambda_\nu > \sqrt{\gamma} \end{cases}$$



Eigenvectors: Elements of an Estimation Theory

- **Assume $\exists$ a basis with sparse representation:**

$$\theta \in \Theta_q(C): \quad \textbf{e.g.} \quad |\theta_{\nu,(\mu)}| \le C|\mu|^{-1/q} \qquad q < 2$$

$\Longrightarrow$ **near sharp upper & lower bounds for minimax risk:**

$$\inf_{\hat{\theta}} \sup_{\theta_\nu \in \Theta_q(C)} E\|\hat{\theta}_\nu - \theta_\nu\|^2. \qquad \textbf{(Paul)}$$

- **Goal: Approximate by "signal in Gaussian noise" model**

LEMMA **($M = 1$) Let $\hat{C} = \langle \hat{\theta}, \theta \rangle$ and $\hat{\theta}^\perp = \hat{\theta} - \hat{C}\theta$. Then**

$$\hat{\theta} = \hat{C}\theta + \hat{S}U \qquad \textbf{(Paul)}$$

  - $U = \hat{\theta}^\perp / \|\hat{\theta}^\perp\|$ **is uniform on "S^{p-2}" $\Rightarrow$ nearly Gaussian.**
  - **move from eigenvectors to sparse mean estimation.**

References

THANK YOU!

IMJ, "High Dimensional Statistical Inference and Random Matrices", *Proc ICM 2006*, Vol 1, 307–333.

IMJ, "Multivariate Analysis and Jacobi Ensembles: Largest eigenvalue, Tracy-Widom Limits and Rates of Convergence", *Ann. Statist.*, 2008

IMJ, "Approximate null distribution of the largest root in multivariate analysis" *Ann. Applied Stat.* to appear.

D. Paul and IMJ, "Sparse principal component analysis for high dimensional data", in preparation.

Webpage: `www-stat.stanford.edu/~imj`