Large Covariance Matrices

Wald Lecture III.

Harold Hotelling and Abraham Wald

Orientation

- Multivariate statistics is long-established field:
  - null Wishart, Canonical Correlation root distributions date from 1930's
  - classical distribution theory got 'stuck'
- Random matrix theory
  - nuclear physics 1950's, now many areas of math, including probability
  - e.g. Gaussian, Laguerre, Jacobi ensembles
- Contemporary multivariate statistics – large p, with or without large n
  - Is there a payoff to statistics from RMT?
  - expand arsenal of math tools for thinking about multivariate data analysis

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues [v. brief]
- Ex. 2: CCA etc - eigenvalues [main]

  [Joint with Peter Forrester]
- Ex. 3: sparse PCA - eigenvectors [brief]
- Some related problem areas [mention]

Gaussian data matrices

p variables

$$Z = (z_{ik}) = \begin{bmatrix} \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \vdots \\ \cdots\cdots\cdots \end{bmatrix} \begin{matrix} n \\ \text{cases} \end{matrix} = \begin{bmatrix} \vec{z}_1^{\,T} \\ \vdots \\ \vec{z}_n^{\,T} \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \end{bmatrix}$$

Independent rows: $\vec{z}_i \sim N_p(0, \Sigma), \qquad i = 1, \ldots n$

or: $Z \sim N(0, I_n \otimes \Sigma_p)$

Zero mean $\Rightarrow$ no centering in **sample covariance matrix:**

$$S = (S_{kk'}), \qquad S = \frac{1}{n} Z^T Z, \qquad S_{kk'} = \frac{1}{n} \sum_{i=1}^{n} z_{ik} z_{ik'}$$

$$nS \sim W_p(n, \Sigma)$$

Growing Gaussian: $p = p(n) \nearrow$ with n

A less developed theory

Nonparametric estimation of sparse means

$Y_i \stackrel{ind}{\sim} N(\mu_i, 1)$ under $H_0 : \mu_i \equiv 0$

$$P\{\max Y_i > \sqrt{2 \log n}\} \to 0 \qquad n \text{ large}$$

(and associated extreme value theory) – well understood

vs.

Nonparametric estimation for Covariances

$Z^T Z \sim W_p(n, \Sigma)$ under $H_0 : \Sigma = I$ Eigenvalues l_i

Until recently

$$P\{\max l_i >?\} \to 0 \qquad (n, p) \text{ large}$$

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues [v. brief]
- Ex. 2: CCA etc - eigenvalues [main]

  [Joint with Peter Forrester]
- Ex. 3: sparse PCA - eigenvectors [brief]
- Some related problem areas [mention]

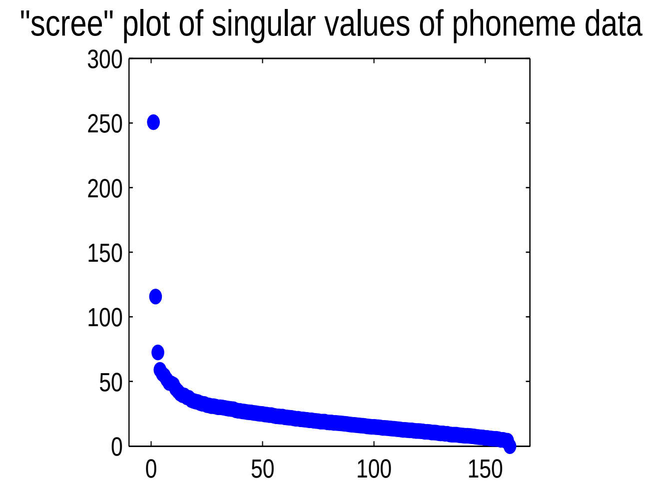
Ex. 1: Principal Components Analysis

- $n \times p$ data matrix Z: n cases, p variables
- spectral decomp: $Z^T Z = U \text{diag}\{l_1, \ldots, l_p\} U^T$

Ex. 1: Principal Components Analysis

- $n \times p$ data matrix Z: n cases, p variables
- spectral decomp: $Z^T Z = U \mathrm{diag}\{l_1, \ldots, l_p\} U^T$
- how many dimensions of "significant" variation? typically, graphical methods:

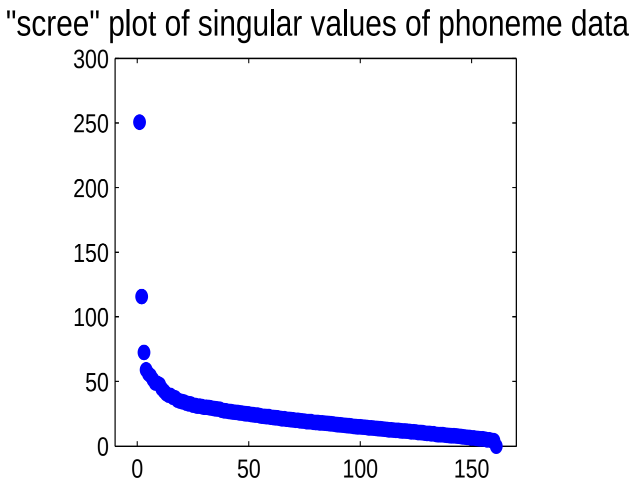
plot l_k versus k,
find "elbow"



Ex. 1: Principal Components Analysis

- $n \times p$ data matrix Z: n cases, p variables
- spectral decomp: $Z^T Z = U \text{diag}\{l_1, \ldots, l_p\} U^T$
- how many dimensions of "significant" variation?
  typically, graphical methods:

  plot l_k versus k,
  find "elbow"



- testing for sphericity: $H_0 : \Sigma = I$, e.g. using l_1
- Any guidance from distribution theory?
- Until recently: $\exists$ tables, but no simple approximations or asymptotics

RMT and largest root l_1

- RMT language: "spectrum edge" of "Laguerre ensemble"
- if $n/p \to c$ then (IMJ, 01)

$$\frac{l_1 - \mu_{np}}{\sigma_{np}} \to F_1 \qquad \text{(Tracy-Widom)}$$

- good approximations for (n, p) not so large
  - especially using second-order corrections to μ_{np}, σ_{np}
- leans heavily on RMT:
  - Tracy-Widom distribution
  - (Fredholm) determinant representations
  - (non-standard) asymptotics of orthogonal polynomials

Recent results for l_1

More general p dependence: $n \to \infty, p \to \infty$, (even $p \gg n$) (El Karoui)

Rate of convergence under $n/p \to c$ can be $O(p^{-2/3})$ (El Karoui)

Progress under alternative hypotheses:

$\Sigma =$ diag$(\sigma_1^2, \ldots, \sigma_r^2, 1, \ldots, 1)$, r fixed:

- complex Gaussian: limit distributions, phase transitions (Baik - Ben Arous - Péché)
- fourth moments: strong law behavior (Baik - Silverstein)

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

Ex. 2: Canonical Correlation Analysis (CCA)

$$X = [\, x_1 \;\; x_2 \;\; \cdots \;\; x_p \,] \qquad\qquad Y = [\, y_1 \;\; y_2 \;\; \cdots \;\; y_q \,]$$

$$\begin{matrix} | & | & & | \end{matrix} \qquad\qquad \begin{matrix} | & | & & | \end{matrix}$$

Goal: find $a^T x$ most correlated with $b^T y$: $\quad \rightarrow \quad$ maximize

$$\mathrm{Corr}(a^T x, b^T y) = \frac{a^T S_{xy} b}{\sqrt{a^T S_{xx} a}\sqrt{b^T S_{yy} b}} \qquad \begin{pmatrix} S_{xy} = X^T Y \\ S_{xx} = X^T X \\ \ldots\ldots \end{pmatrix}$$

i.e.

$$r_k = \max \left\{ a^T S_{xy} b \;\; : \;\; \begin{matrix} a^T S_{xx} a = b^T S_{yy} b = 1 & \\ a^T S_{xx} a_j = b^T S_{yy} b_j = 0 & j < k \end{matrix} \right\}$$

ctd.

$$r_k = \max \left\{ a^T S_{xy} b \ : \ \begin{array}{ll} a^T S_{xx} a = b^T S_{yy} b = 1 & \\ a^T S_{xx} a_j = b^T S_{yy} b_j = 0 & j < k \end{array} \right\}$$

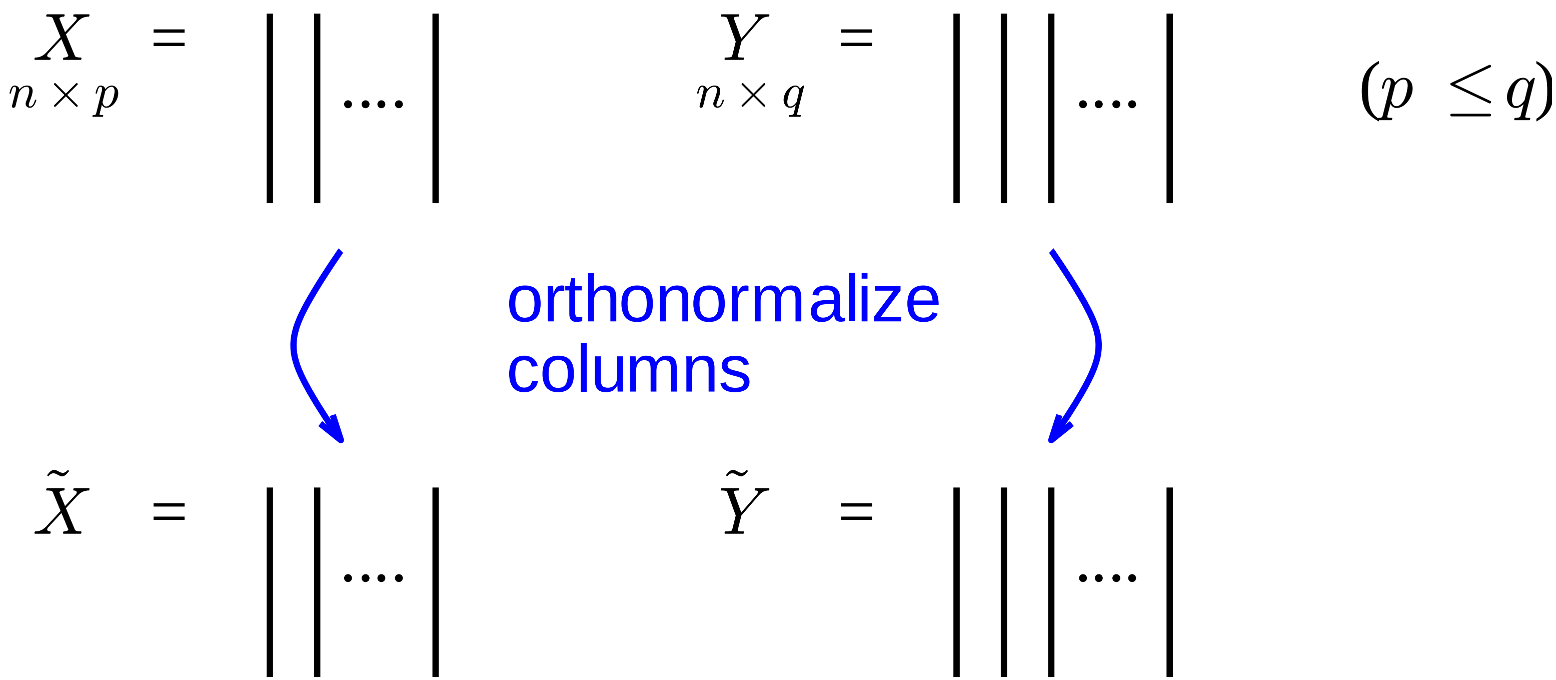
$\longrightarrow$ determinantal equation

$$\det(S_{xy} S_{yy}^{-1} S_{yx} - r^2 S_{xx}) = 0$$

$\longrightarrow$ $r_1^2 \geq r_2^2 \geq \cdots \geq r_p^2$ $\quad \begin{pmatrix} \text{and} & a_1, \ldots, a_p \\ & b_1, \ldots, b_p \end{pmatrix}$

$\longrightarrow$ how many r_k^2 are “significant”?

The SVD view of CCA



SVD : $$\tilde{X}^T \tilde{Y} = U \begin{pmatrix} \mathbf{r_1} & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \mathbf{r_p} & 0 & \cdots & 0 \end{pmatrix} V^T$$

Useful for comparison theorems (and computation)

Example: First use of CCA

"Regressions between Sets of Variables", F.V. Waugh, *Econometrica*, 1942

$\mathbf{X}$ = **"wheat characteristics"**	$\mathbf{Y}$ = **"flour characteristics"**
x_1 = kernel texture	y_1 = wheat per bbl. of flour
x_2 = test weight	y_2 = ash in flour
x_3 = damaged kernels	y_3 = crude protein in flour
x_4 = foreign material	y_4 = gluten quality index
x_5 = crude protein in wheat	

$a^T x$ index of wheat quality

$b^T y$ index of flour quality

GOAL: highly correlated grading of raw & finished products

$$p = 5 \qquad q = 4 \qquad n = 138$$

Example

Barnett & Preisendorfer, (1987), *Monthly Weather Review*

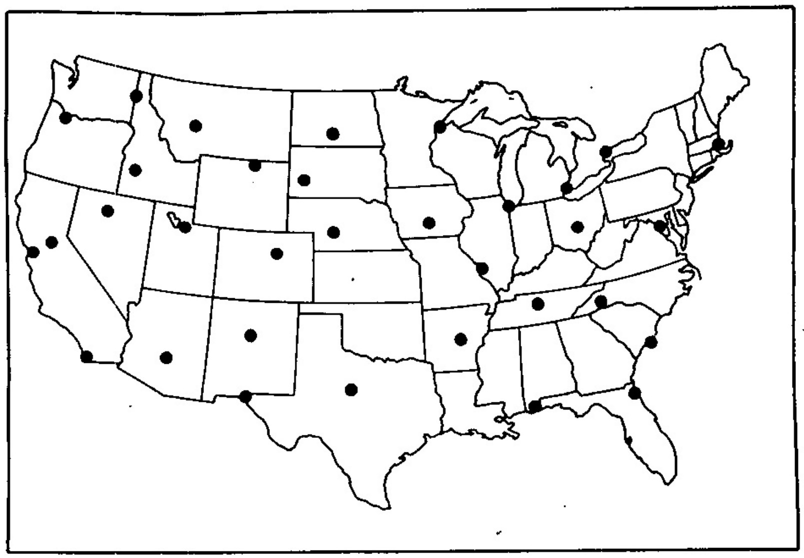


FIG. 1. Locations of stations/districts providing surface air temperature predictand data.

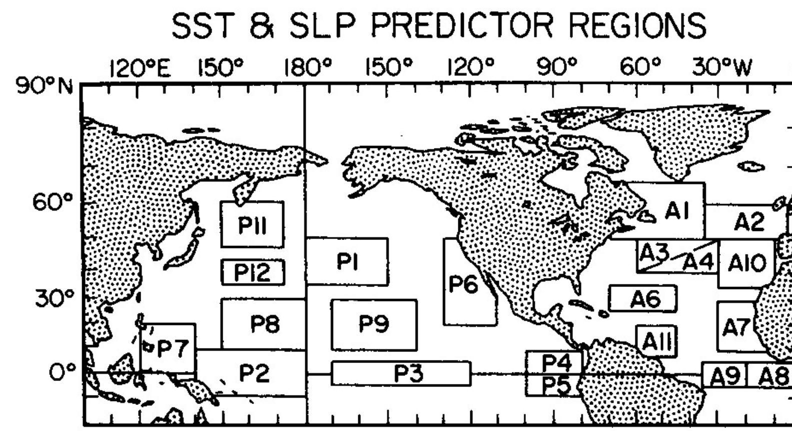


FIG. 2. SST from the large averaging areas shown above were used as predictor information. SLP predictor data came from the region 20°–70°N, 140°E to the Greenwich Meridian.

Y variables: surface air temperatures at 33 U.S. locations; monthly data, 1931-1980

X variables: sea surface temp(SST) in 21 regions for 3 prior months in 2 seasons.

$$p = 126 \qquad q = 33 \qquad n = 600$$

Neglect of CCA in STAT?

Title/Abstract/Keyword search:

articles published in 15 mos. in 2002-03:

[**ISI Web of Science**]

Keyword	**Stat/Prob Journals**	**Other Journals**	**Total**
Canonical Correlation	7	116	123
Gibbs Sampler	49	52	101

Recent Variants

Functional CCA Leurgans, Moyeed, Silverman 94

Curve data $\{X_i(t), Y_i(t), i = 1, \ldots, n\}$ $t \in T$

$p = q = \#$ discretization points t_k, maybe large

$\rightarrow$ regularized CCA: $\max \frac{(a^T \Sigma_{XY} b)^2}{a^T(\Sigma_{XX} + \lambda D^4)a \quad b^T(\Sigma_Y Y + \lambda D^4)b}$

Kernel ICA Bach, Jordan '02

From $y^i = Ax^i$, $\quad i = 1, \ldots, n$ estimate A. If A is 2×2 (here), set

$$X = \begin{bmatrix} \Phi(x_1^1) \\ \vdots \\ \Phi(x_1^n) \end{bmatrix} \qquad Y = \begin{bmatrix} \Phi(x_2^1) \\ \vdots \\ \Phi(x_2^n) \end{bmatrix}$$

$\Phi : \mathbb{R} \rightarrow \mathcal{F}$ feature sp.

$p = q = \dim(\mathcal{F})$ large

$\rightarrow$ regularized CCA on X, Y (...)

[cf. Renyi (59) $\rho^*(X_1, X_2) = \max_{\theta,\phi} \text{corr}[\theta(X_1), \phi(X_2)]$)

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

Roots of a Determinantal Equation

The equation $|X^T Y (Y^T Y)^{-1} Y^T X - r^2 X^T X| = 0$ becomes

$$|A - r^2(A + B)| = 0$$

where

$$A = X^T P X \qquad P = Y(Y^T Y)^{-1} Y^T$$
$$B = X^T P^{\perp} X \qquad P^{\perp} = I - P$$

Roots of a Determinantal Equation

The equation $|X^TY(Y^TY)^{-1}Y^TX - r^2X^TX| = 0$ becomes

$$|A - r^2(A+B)| = 0$$

where

$$A = X^TPX \qquad P = Y(Y^TY)^{-1}Y^T$$
$$B = X^TP^{\perp}X \qquad P^{\perp} = I - P$$

Stochastic model: $\underset{n\times(p+q)}{[X:Y]} \sim N(0, I_n \otimes \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix})$

Null distribution: $\Sigma_{XY} = 0 \quad \leftrightarrow \quad \rho_1 = \ldots = \rho_p = 0$

In this case (transforming to $\Sigma_{XX} = I$)

$$A \sim W_p(q, I)$$
$$B \sim W_p(n-q, I)$$

independent, Wishart

Basic Setting

$A \sim W_p(q, I)$
$B \sim W_p(n - q, I)$

2 **independent** Wisharts $p \leq q, n - q$

[Recall: if Z has rows $z_i^T \overset{ind}{\sim} N_p(0, \Sigma)$, then

$$Z^T Z = \sum_{i=1}^n z_i z_i^T \sim W_p(n, \Sigma)$$]

Multivariate Beta roots $:= (u_i)_{i=1}^p$

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

$$\Updownarrow$$

Multivariate F roots $:= (w_i)_{i=1}^p$

$$\det[wB - A] = 0$$

Largest Root test: based on $u_1 (\geq u_2 \geq \ldots u_p)$

Related Classical Problems

2 Wishart setting central to classical multivariate analysis:

- **Multiple Response Linear Model**

$$\underset{n\times \mathbf{p}}{Y} = \underset{n\times q}{X} \underset{q\times \mathbf{p}}{\beta} + \underset{n\times \mathbf{p}}{E}, \qquad E \sim N(0, I_n \otimes \boldsymbol{\Sigma})$$

Largest root test of $H_0 : \beta = 0$ uses u_1.

- **Multiple Discrimination**

q populations; n observations on p variables.
A and B: between and within class covariance matrices.

- **Testing Equality of Two Covariance Matrices**

From CCA to Other Settings

Basic setting: u_1 largest root of $\det[u(A+B) - A] = 0$

CCA	$[X\ Y] \sim N_{p+q}(0, I_n \otimes \Sigma)$ $H_0 : \Sigma_{XY} = 0$	p	q	$n-q$
Multivariate Linear Model	$\underset{n\times p}{Y} = X \underset{p\times q}{\beta} + \underset{q\times p}{E}$ $H_0 : \underset{g\times q}{C}\ \underset{q\times p}{\beta}\ \underset{p\times r}{M} = 0$	r $\uparrow$ dimen	g $\uparrow$ hypoth. d.f.	$n-q$ $\uparrow$ error d.f
Equality of Covariance	$n_i \hat{\Sigma}_i \sim W_p(n_i, \Sigma_i)$ $H_0 : \Sigma_1 = \Sigma_2$	p	n_1	n_2
Mult. Discrim.	n_i obs on q pops $N_p(\mu_i, \Sigma)$ $i = 1, \ldots, q$	p	$q-1$	$n-q$

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

Why non-standard asymptotics?

Large literature on $\det[A - u(A+B)] = 0$.
Books: e.g. Anderson(58,02); Muirhead(82)

Exact distributions are complex; $\exists$ much asymptotics with p, q **fixed**, $n \to \infty$

BUT, some results not “numerically available”:
e.g. null distribution of largest root u_1

$$nu_1 \to \quad \text{largest root of } \; W_p(q, I) \qquad \text{[finite LOE]}$$

$\Rightarrow$ here, aim for simple approximations from large $(p, q(p), n(p))$ asymptotics as $p \to \infty$

Limiting Empirical Spectrum

Note: $u_i = r_i^2$ **squared** correlation scale

Assume (p, q, n) large, such that

$$0 < \sin^2 \frac{\gamma_0}{2} \leftarrow \frac{p}{n} \leq \frac{q}{n} \rightarrow \sin^2 \frac{\phi_0}{2}$$

Limiting Empirical Spectrum

Note: $u_i = r_i^2$ **squared** correlation scale

Assume (p, q, n) large, such that

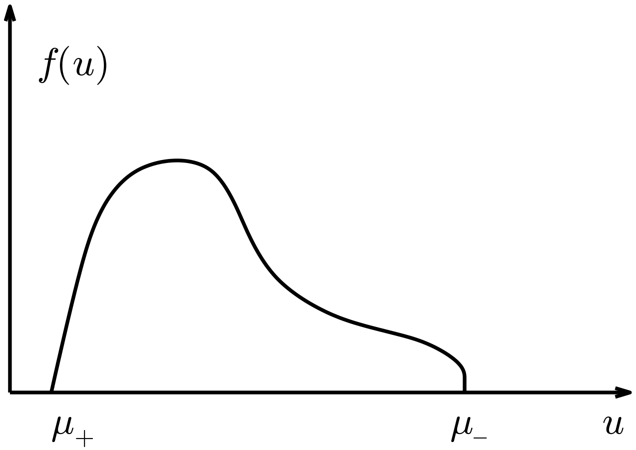
$$0 < \sin^2 \frac{\gamma_0}{2} \leftarrow \frac{p}{n} \leq \frac{q}{n} \rightarrow \sin^2 \frac{\phi_0}{2}$$

Then (Wachter, 1980)

$$F_p(u) = p^{-1} \#\{i : u_i \leq u\} \rightarrow \int_0^u f(u')du'$$

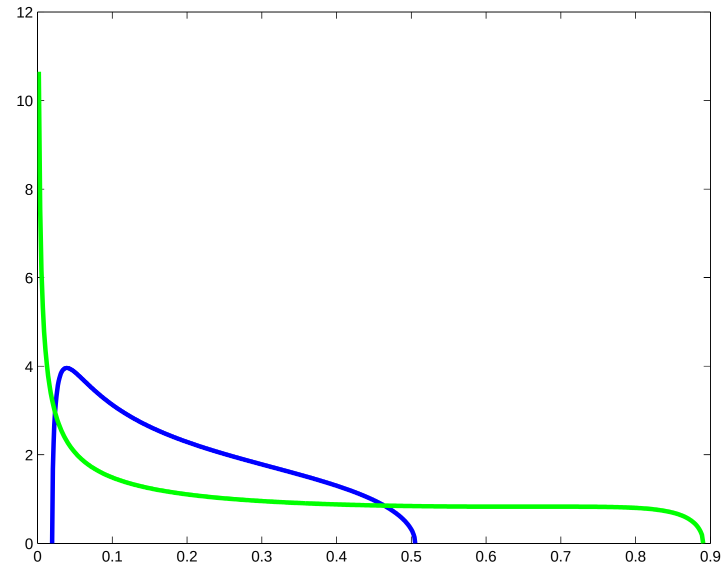
$$f(u) = \frac{c_u \sqrt{(\mu_+ - u)(u - \mu_-)}}{u(1-u)} \qquad c_u = 2\pi \sin^2 \gamma_0 / 2$$

$$\mu_\pm = \cos^2\left(\frac{\pi}{2} - \frac{\phi_0 \pm \gamma_0}{2}\right)$$



Examples

p	q	n	$\gamma_0/2$	$\phi_0/2$	μ_-	μ_+
10	20	100	.325	.466	.020	.505
50	50	150	.616	.617	.000	.890
4	5	137	.182	.210	.0004	.140



What this might mean in practice

A (hopefully hypothetical) clinical trial:

- $\mathbf{n = 100}$ (randomly chosen) patients
- X variables: $\mathbf{p = 20}$ physiologic measurements:
  blood pressure, heart rate, BMI , serum albumin, ...
- Y variables: $\mathbf{q = 10}$ financial variables:
  income, assets, tax, ...
- *then* ... someone fakes the financial data, .. **and yet,**

$$\boxed{\mu_+^2 \approx .7^2}$$

i.e. Some linear physiologic feature $(a^T X)$ and some linear financial feature $(b^T Y)$ have *observed* correlation **about 0.7**

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

Approximate Law for Largest Root

Assume 2 Wishart Setting with $p, q(p), n(p) \to \infty$.

$$\frac{\gamma_p}{2} = \sin^{-1}\sqrt{\frac{p}{n}}, \qquad \frac{\phi_p}{2} = \sin^{-1}\sqrt{\frac{q}{n}}.$$

$$\mu_\pm = \cos^2\left(\frac{\pi}{2} - \frac{\phi_p \pm \gamma_p}{2}\right), \quad \sigma_{p+}^3 = \frac{1}{(2n)^2}\frac{\sin^4(\phi_p + \gamma_p)}{\sin\phi_p \sin\gamma_p}.$$

Approximate Law for Largest Root

Assume 2 Wishart Setting with $p, q(p), n(p) \to \infty$.

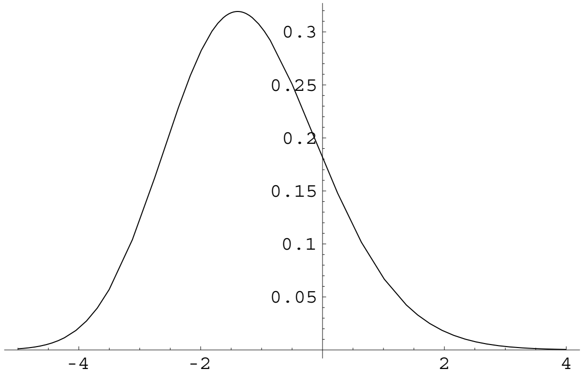
$$\frac{\gamma_p}{2} = \sin^{-1}\sqrt{\frac{p}{n}}, \qquad \frac{\phi_p}{2} = \sin^{-1}\sqrt{\frac{q}{n}}.$$

$$\mu_\pm = \cos^2\left(\frac{\pi}{2} - \frac{\phi_p \pm \gamma_p}{2}\right), \quad \sigma_{p+}^3 = \frac{1}{(2n)^2}\frac{\sin^4(\phi_p + \gamma_p)}{\sin\phi_p \sin\gamma_p}.$$

Theorem (IMJ + Peter Forrester)

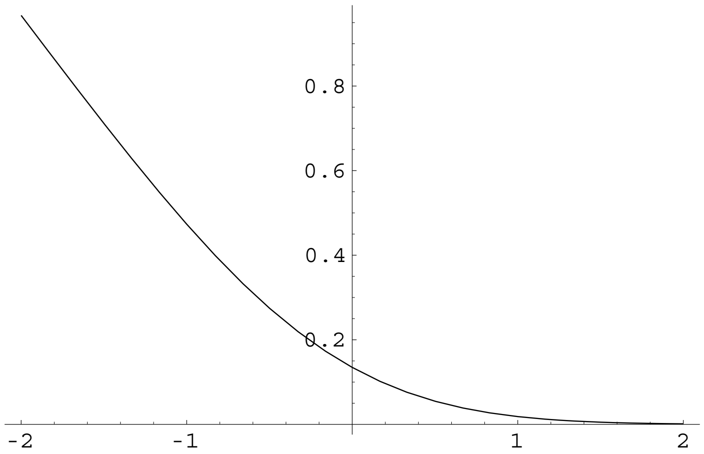
$$P\{u_1 \le s\} = P\{\mu_+ + \sigma_+ W_1 \le s\} + o(1)$$

W_1 follows the *Tracy-Widom* F_1 distribution.



Painlevé II and Tracy-Widom

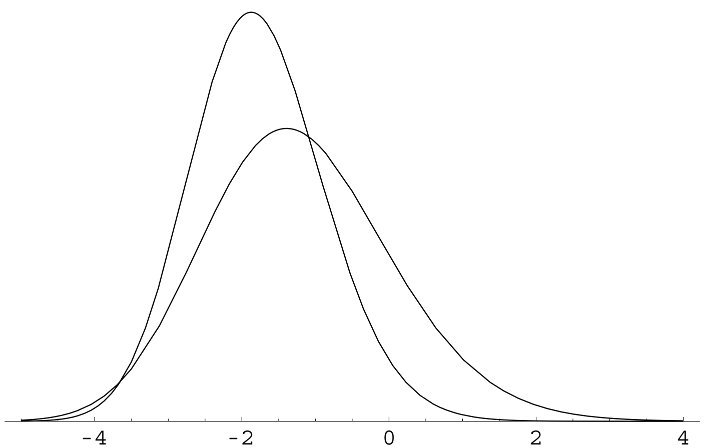
Painlevé II:



$$q'' = xq + 2q^3$$
$$q(x) \sim \mathrm{Ai}(x) \qquad \text{as} \quad x \to \infty$$

$$q^2(x) \sim \begin{cases} x/2 & \text{at } -\infty \\ e^{-(4/3)x^{3/2}} & \text{at } \infty \end{cases}$$

Tracy-Widom distributions:



$$f_j = F_j'$$
$$(\log F_2)'' = -q^2$$
$$(\log \mathbf{F_1})' = -\tfrac{1}{2}(q' + q^2)$$

Approximate Law for Largest Root

Assume 2 Wishart Setting with $p, q(p), n(p) \to \infty$.

$$\frac{\gamma_p}{2} = \sin^{-1}\sqrt{\frac{p - \mathbf{.5}}{n - \mathbf{1}}}, \qquad \frac{\phi_p}{2} = \sin^{-1}\sqrt{\frac{q - \mathbf{.5}}{n - \mathbf{1}}}.$$

$$\mu_\pm = \cos^2\Big(\frac{\pi}{2} - \frac{\phi_p \pm \gamma_p}{2}\Big), \qquad \sigma^3_{p+} = \frac{1}{(2n - \mathbf{2})^2}\frac{\sin^4(\phi_p + \gamma_p)}{\sin\phi_p \sin\gamma_p}.$$

Main 'Result' (IMJ + Peter Forrester)

$$P\{u_1 \le s\} = P\{\mu_+ + \sigma_+ W_1 \le s\} + O(p^{-2/3}).$$

- small corrections $(\mathbf{.5}, \mathbf{1}, \mathbf{2})$ greatly improve approximation for p, q small, so
- error is $O(p^{-2/3})$ [instead of $O(p^{-1/3})$]

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

$TW(p,q,n)$ approximation

Use $\mu_+(p,q,n) + \sigma_+(p,q,n)F_{1,\alpha}$
(with $F_{1,\alpha} = \alpha^{th}$ percentile of F_1)
to approximate α^{th} percentile of u_1

Claim: As a rough guide, the TW-approximation

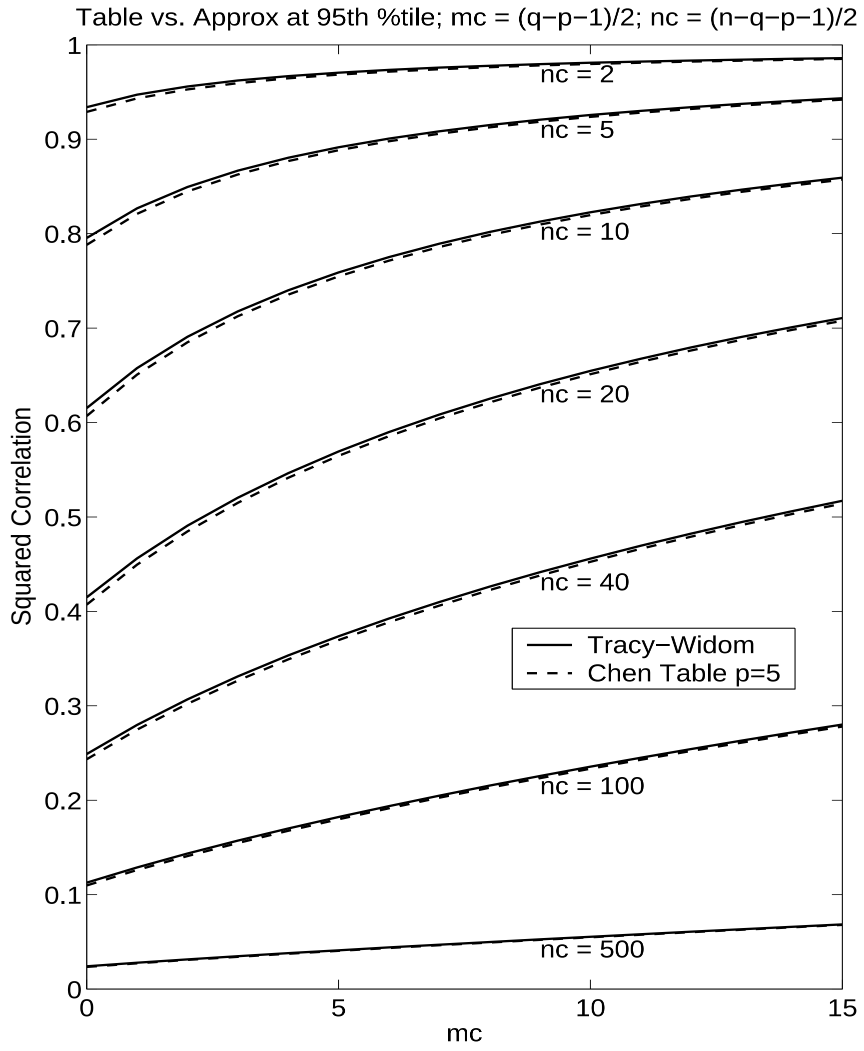
- mostly mimics tables where extant
- extends tables if not
- improves on some software (S/SPLUS/R; SAS)

Approximation vs. Tables for $p = 5$

Tables: William Chen, IRS, (2002)

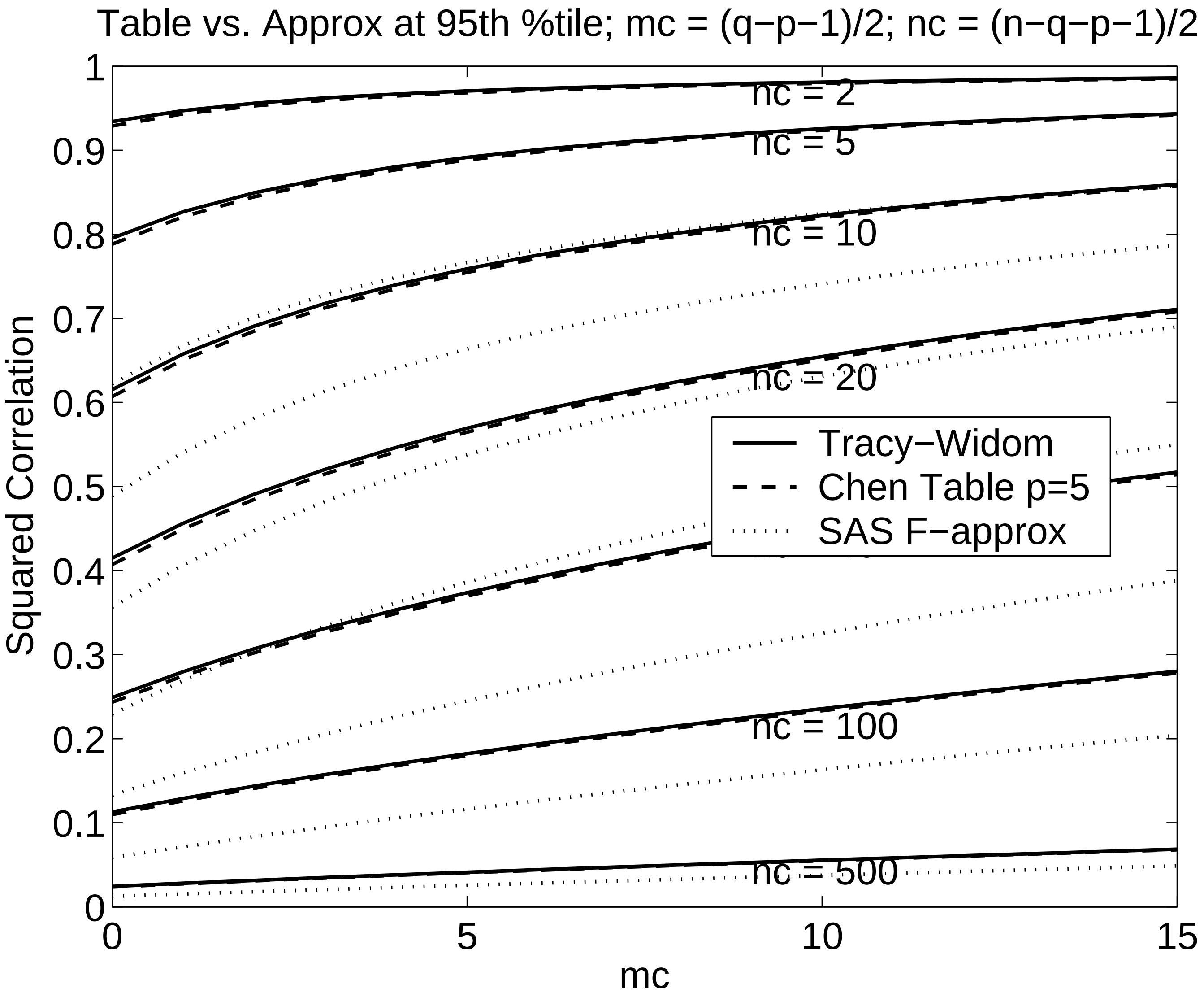
$$m_c = \frac{q-p-1}{2} \in [0, 15],$$

$$n_c = \frac{n-q-p-1}{2} \in [1, 1000]$$



Upper Bound in SAS

Approximate $\frac{n-q}{q}\frac{u_1}{1-u_1}$ by $F_{q,n-q}$

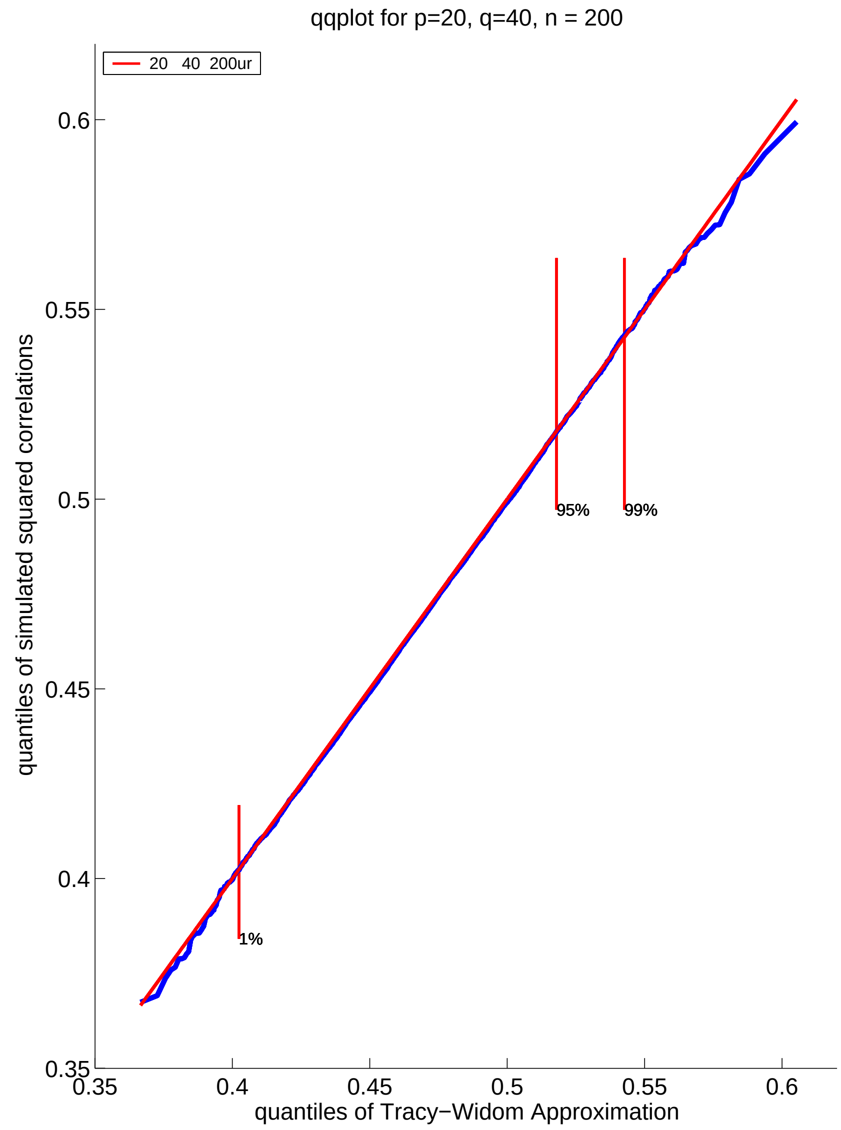


Finite (p, q, n) simulations of $u_1 = r_1^2$

$p = 20, q = 40,$
$n = 200$

$Y-$ axis: quantiles of simulated $u_1 = r_1^2$ $(10,000$ reps.)

$X-$ axis: quantiles of $\mu_+ + \sigma_+ W_1$



ctd.

TW- Percentile for W_1	p, q, n (μ, σ)	**20,40,200** (.494, .024)
-3.90	.01	.010
-3.18	.05	.052
-2.78	.10	.104
-1.91	.30	.311
-1.27	.50	.507
-0.59	.70	.706
0.45	**.90**	**.904**
0.98	**.95**	**.950**
2.02	**.99**	**.990**

E.g. $\mathbf{.904} = \hat{P}\left[\frac{u_1 - \mu_+}{\sigma_+} \leq 0.45 \mid (p, q, n) = (20, 40, 200)\right]$

ctd.

TW- Percentile for W_1	p, q, n (μ, σ)	**20,40,200** (.494, .024)	**5,10,50** (.478, .061)	**2,4,20** (.442, .117)	2 * SE
-3.90	.01	.010	.008	.000	(.002)
-3.18	.05	.052	.049	.009	(.004)
-2.78	.10	.104	.099	.046	(.006)
-1.91	.30	.311	.304	.267	(.009)
-1.27	.50	.507	.506	.498	(.010)
-0.59	.70	.706	.705	.711	(.009)
0.45	**.90**	**.904**	**.910**	**.911**	(.006)
0.98	**.95**	**.950**	**.955**	**.958**	(.004)
2.02	**.99**	**.990**	**.992**	**.995**	(.002)

E.g. $\mathbf{.904} = \hat{P}\left[\frac{u_1 - \mu_+}{\sigma_+} \leq 0.45 \mid (p, q, n) = (20, 40, 200)\right]$

Remarks

- $p^{-2/3}$ scale of variability for u_1
- 95th %tile $\doteq \mu_{p+} + \sigma_{p+}$, 99th %tile $\doteq \mu_{p+} + 2\sigma_{p+}$
- if $\mu_{p+} > .7$, **logit scale** $v_i = \log u_i/(1 - u_i)$ better:

$$\mu_{v+} = \log \frac{\mu_{p+}}{1 - \mu_{p+}}, \qquad \sigma_{v+} = v'(\mu_{p+})\sigma_{p+} = \frac{\sigma_{p+}}{\mu_{p+}(1 - \mu_{p+})}$$

- **Smallest** eigenvalue: with previous assumptions and $\gamma_0 < \phi_0$, $\sigma_{p-}^3 = \frac{1}{(2n-2)^2} \frac{\sin^4(\phi_p - \gamma_p)}{\sin \phi_p \ \sin \gamma_p}$ then

$$\frac{\mu_{p-} - u_p}{\sigma_{p-}} \xrightarrow{\mathcal{D}} W_1 \qquad (W_2)$$

- Corresponding limit distributions for $u_2 \geq \cdots \geq u_k$, $u_{p-k} \geq \cdots \geq u_{p-1}$, k **fixed**

Logit approximation

Percentile	TW	**50,50,150**	**5,5,15**	**2,2,6**	2 * SE
	(μ, σ)	(2.06, .127)	(1.93, .594)	(1.69, 1.11)	
-3.90	.01	.007	.002	.010	(.002)
-3.18	.05	.042	.023	.037	(.004)
-2.78	.10	.084	.062	.074	(.006)
-1.91	.30	.289	.262	.264	(.009)
-1.27	.50	.499	.495	.500	(.010)
-0.59	.70	.708	.725	.730	(.009)
0.45	**.90**	**.905**	**.919**	**.931**	(.006)
0.98	**.95**	**.953**	**.959**	**.966**	(.004)
2.02	**.99**	**.990**	**.991**	**.993**	(.002)

Testing Subsequent Correlations

Suppose: $\Sigma_{XY} = \begin{bmatrix} \rho_{\mathbf{1}}^{\mathbf{2}} & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \rho_{\mathbf{p}}^{\mathbf{2}} & 0 & \cdots & 0 \end{bmatrix}$ $\quad p \leq q, n-p$

If largest r correlations are large, test

$$\mathbf{H_r} : \rho_{r+1} = \rho_{r+2} = \ldots = \rho_p = 0?$$

Comparison Lemma (from SVD interlacing)

$$\mathcal{L}(u_{r+1}|p, q, n; S_{XY} \in \mathbf{H_r}) \overset{st}{<} \mathcal{L}(u_1|p, q-r, n; \mathbf{I})$$

$\Rightarrow$ conservative $P-$values for H_r via

$TW(p, q-r, n)$ approx'n to RHS

[Aside: $\mathcal{L}(u_1|p-r, q-r, n; I)$ may be better, but no bounds]

World Wheats Data ctd.

$p = 4$ flour characteristics

$q = 5$ wheat characteristics

$n = 137$

$r_1^2 = .923$

$r_2^2 = .554$

$r_3^2 = .056$

$r_4^2 = .008$

r_1^2 significant (also by permutation test)

r_2^2? 99%–tile of $TW(p, q-1, n) = TW(4, 4, 137)$

$$\doteq \mu + 2\sigma \doteq 0.152 \ll 0.554 = r_2^2$$

$\rightarrow r_2^2$ significant (possible collinearity ...)

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

Joint distribution of latent roots, 1939

Fisher	Girshick	Hsu	Mood	Roy
Cambridge	*Columbia*	*London*	*Princeton*	*Calcutta*

$$f(x) = c \prod_{i=1}^{N} (1 - x_i)^{(\alpha-1)/2} (1 + x_i)^{(\beta-1)/2} \prod_{i<j} |x_i - x_j|$$

2 Wishart setting, but *notation change:*

$$x_i = 2u_i - 1, \qquad i = 1, \ldots, N = p; \quad \alpha = n - q - p; \beta = q - p$$

Random Matrix Theory

E.g. for largest eigenvalue: hard to marginalize to get at $P\{l_1 \le x\}$.
Key role: *determinants*, not independence:

$$\prod_{i<j}(l_i - l_j) = \det[l_i^{k-1}]_{1\le i,k\le p}$$

$$\prod_{i=1}^{p} I\{l_i \le x\} = \sum_{k=0}^{p}(-1)^k \binom{p}{k} \prod_{i=1}^{k} I\{l_i > x\}.$$

$\Rightarrow P\{l_1 \le x\}$ via *Fredholm* determinants.

Correlation kernel

For *complex* data $X_{kl} + iX'_{kl}$, joint density $f(x_1, \ldots, x_N)$

$$c \prod_1^N w(x_i) \prod_{i<j} (x_i - x_j)^2 = \frac{1}{N!} \det_{1 \le i,j \le N} [K_{N2}(x_i, x_j)]$$

with *correlation kernel* $K_{N2}(x,y) = \sum_{k=0}^{N-1} \phi_k(x)\phi_k(y)$

$\phi_k(x) = h_k^{-1/2} w^{1/2}(x) p_k(x)$ – orthonormalized polynomials, in *classical cases:*

$w(x)$	$p_k(x)$	Distribution
$e^{-x^2/2}$	Hermite $H_k(x)$	Gaussian
$e^{-x} x^\alpha$	Laguerre $L_k^\alpha(x)$	Wishart
$(1-x)^\alpha (1+x)^\beta$	Jacobi $P^{\alpha,\beta}(x)$	Multivariate Beta

Convergence of Kernels (i)

Airy kernel associated with T-W law F_2 (T-W, 1994)

$$K_A(s,t) = \frac{\mathsf{Ai}(s)\mathsf{Ai}'(t) - \mathsf{Ai}'(s)\mathsf{Ai}(t)}{s-t}$$

Approach: Uniform convergence of rescaled kernel

$$\sigma_N K_N(\mu_N + \sigma_N s, \mu_N + \sigma_N t) \to K_A(s,t) \qquad (*)$$

implies that of extreme eigenvalues (Soshnikov, 01)

$$\mathcal{L}(x_{(1)}, \dots, x_{(k)} | \mathbb{F}_N) \to \mathcal{L}(x_{(1)}, \dots, x_{(k)} | \mathbb{F}_2) \qquad \text{fixed } k$$

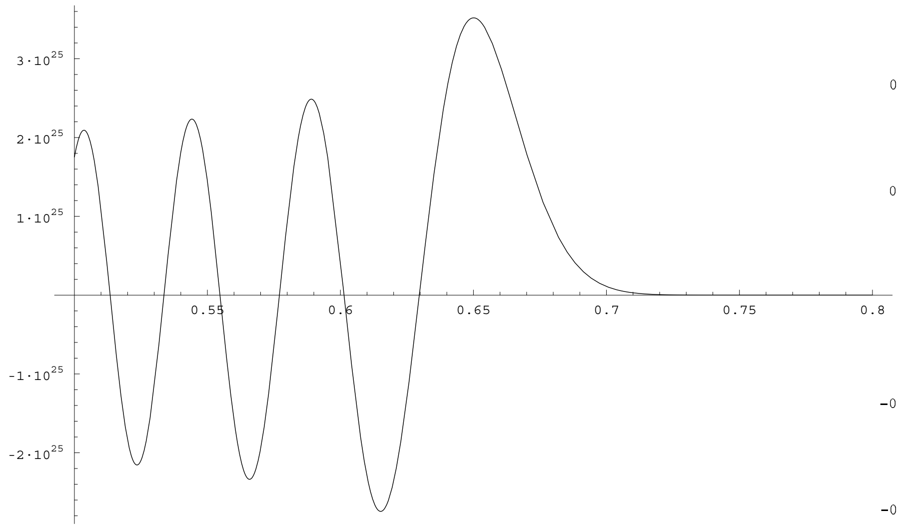
Key point: Choose μ_N, σ_N so that error in (*) drops from $O(N^{-1/3})$ to $O(N^{-2/3})$.

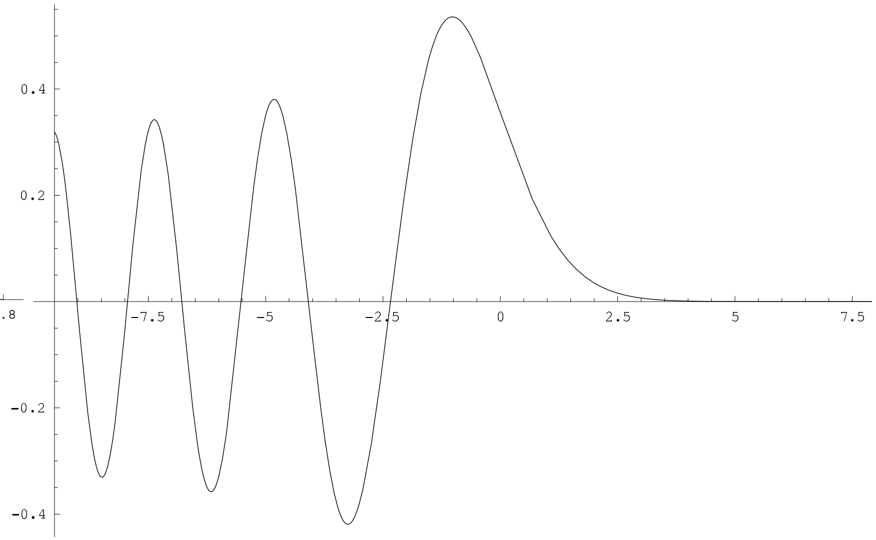
Convergence to Airy function

Example:

$$(\alpha, \beta, N) = (200, 100, 100) \leftrightarrow (p, q, n) = (100, 200, 500)$$

$(1-x)^{\alpha/2}(1+x)^{\beta/2}P_N^{\alpha,\beta}(x)$ Ai(s)





Convergence of Kernels (ii)

Jacobi polynomials $P_N^{\alpha,\beta}$ satisfy differential equation

$$W''(x) = \{\kappa^2 f(x) + g(x)\}W(x)$$

- treble asymptotics: $(\alpha, \beta, N) = (n-p-q, q-p, p)$ large
- Airy approximation at largest zero [Liouville-Green/ WKB]
- **Error Bounds** Olver, 74 constrain κ, f, g and yield error $O(N^{-2/3})$.

Real Case:

- quaternion determinants
- closed form formulas: Adler-Forrester-Nagao-van Moerbeke
- 'miraculous' cancellations $\rightarrow O(N^{-2/3})$

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues
- Ex. 2: CCA etc - eigenvalues
  - Canonical Correlation Analysis - description
  - A basic setting - 2 independent Wisharts
  - Asymptotics - empirical spectrum
  - Asymptotics - largest root approximation
    - its accuracy & application
    - some comments on derivation
    - loose analogy with $t-$approximation
- Ex. 3: sparse PCA - eigenvectors
- Some related problem areas

Loose Analogy

	$t-$**statistic** $\sqrt{n}\bar{x}/s$	**largest root** u_1 **of** A, B
Model:	$X_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$	$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N(0, \Sigma)$
	$H_0 : \mu = 0$	$H_0 : \Sigma_{XY} = 0$
Exact Law:	$t \sim t_{n-1}$	$u_1 \sim JOE_p(n-q-p, q-p)$
Approx Law:	$\Phi(x) = \int_{-\infty}^{x} \phi(s) ds$	$F_1(x) = \exp\{-\frac{1}{2} \int_x^\infty q(s) + (x-s)^2 q(s) ds\}$

Loose Analogy

	$t-$**statistic** $\sqrt{n}\bar{x}/s$	**largest root** u_1 **of** A, B
Model:	$X_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$	$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N(0, \Sigma)$
	$H_0 : \mu = 0$	$H_0 : \Sigma_{XY} = 0$
Exact Law:	$t \sim t_{n-1}$	$u_1 \sim JOE_p(n - q - p, q - p)$
Approx Law:	$\Phi(x) = \int_{-\infty}^{x} \phi(s)ds$	$F_1(x) = \exp\{-\frac{1}{2}\int_x^\infty q(s) + (x - s)^2 q(s)ds\}$

Second-Order Accuracy

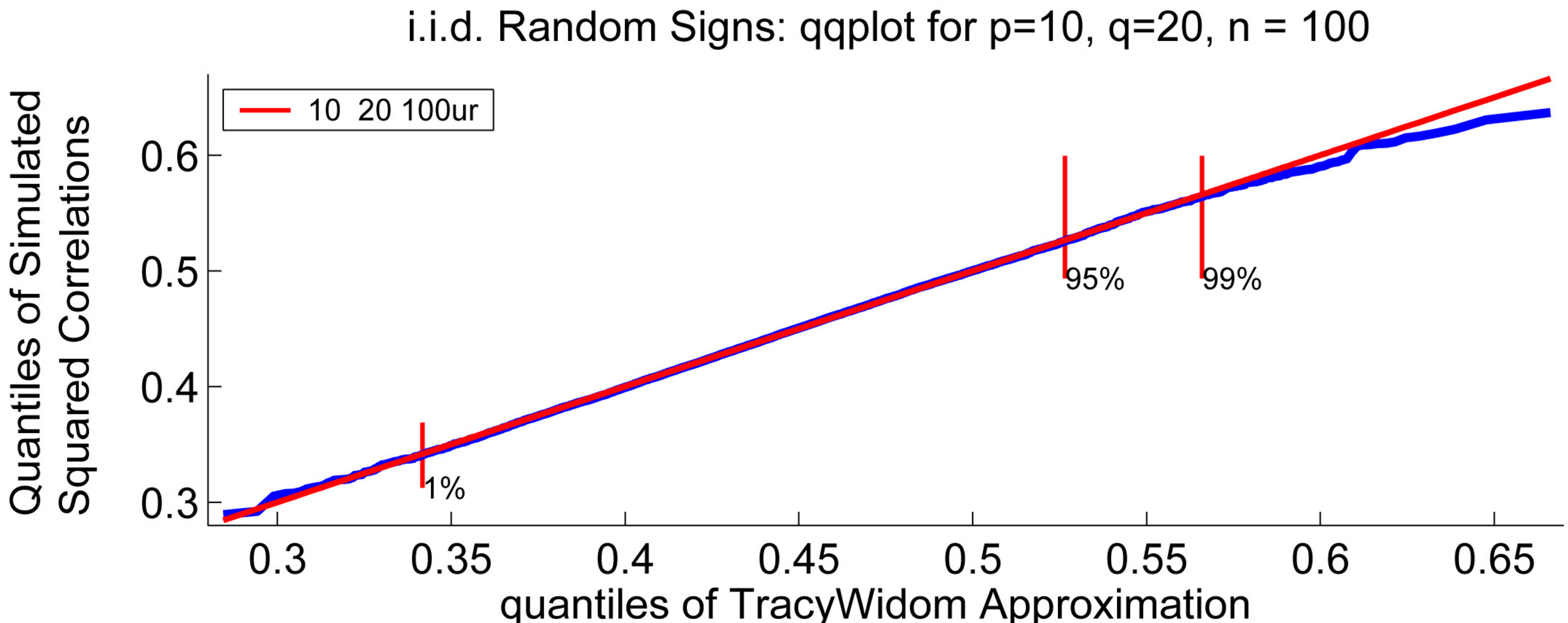
$$P\{t_{n-1} \leq x\} = \Phi(x) + O\Big[\Big(\frac{1}{\sqrt{n}}\Big)^2\Big]$$

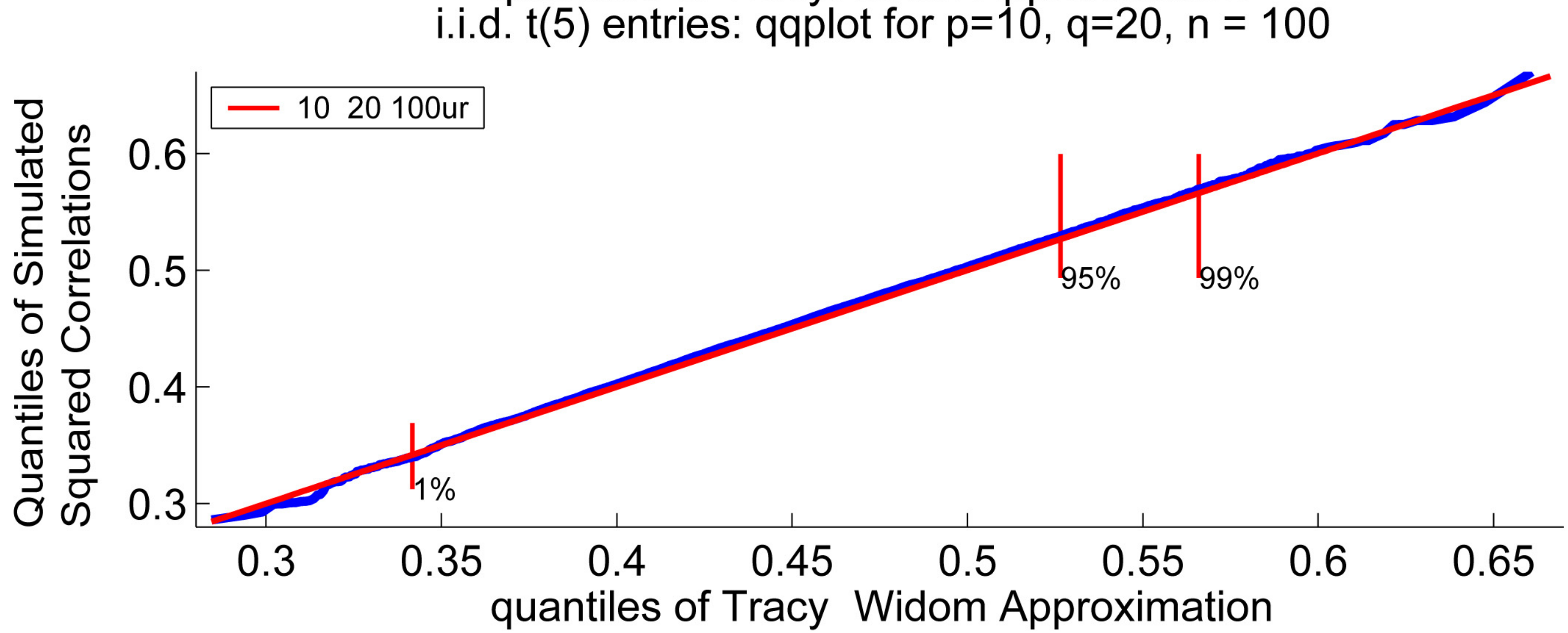
Correlation functions

$$\sigma_N K_N(\mu_N + \sigma_N s, \mu_N + \sigma_N t) \to K_A(s, t) + O(N^{-2/3})$$

Claim: $P\{u_1 \leq \mu_N + \sigma_N x\} = F_1(x) + O(N^{-2/3})$

Non Gaussian Data





Agenda

- Orientation
- Ex. 1: PCA - eigenvalues [v. brief]
- Ex. 2: CCA etc - eigenvalues [main]
- Ex. 3: sparse PCA - eigenvectors [brief]
- Some related problem areas [mention]

Ex. 3: PCA - Estimating Eigenvectors

High-p signal processing areas, e.g.

hyperspectral data, face recognition, ECGs

routinely use dimensionality reduction techniques

variable/feature selection, PCA, transform domains

Combined (in some order) to aim for sparse representation, e.g.:

- transform to wavelet basis (dimension n)
- select high variance variables (dimension $k \ll n$)
- PCA on reduced subset ($O(k^3)$ vs. $O(n^3)$)

Clear speed benefits; **here:** helps with consistency issues

Orthogonal Factor models in PCA

$$x_i = \mu + v_i \rho + \sigma z_i \qquad i = 1, \ldots, n$$

- $\rho \in \mathbb{R}^p$, single component to be estimated
- $v_i \stackrel{i.i.d.}{\sim} N(0,1)$ random effects
- $z_i \stackrel{i.i.d.}{\sim} N_p(0,I)$ Gaussian noise (e.g. $\sigma = 1$)

Orthogonal Factor models in PCA

$$x_i = \mu + v_i \rho + \sigma z_i \qquad i = 1, \ldots, n$$

- $\rho \in \mathbb{R}^p$, single component to be estimated
- $v_i \overset{i.i.d.}{\sim} N(0,1)$ random effects
- $z_i \overset{i.i.d.}{\sim} N_p(0, I)$ Gaussian noise (e.g. $\sigma = 1$)

A Multicomponent Model

$$x_i = \sum_{j=1}^{m} v_i^j \rho^j + \sigma z_i, \qquad i = 1, \ldots, n$$

ρ^j unknown, orthogonal, $\|\rho^1\| \geq \cdots \geq \|\rho^m\|$; for asymptotics

$$(\|\rho^1(n)\|, \ldots, \|\rho^j(n)\|, \ldots) \xrightarrow{\ell_1} (\varrho_1, \ldots, \varrho_j, \ldots).$$

Inconsistency

In either single (or multi-) component model,

Theorem (Lu) If $p/n \to c > 0$, then

$$\liminf_{n\to\infty} \; E \sin \angle(\hat{\rho}, \rho) > 0.$$

- Noise does not average out in PCA if too many dimensions p relative to n.
- Suggests: reduce p to $k \ll p$ before starting PCA

Sparsity and PCA

In basis $\{e_\nu(t)\}$, a population p.c. $\{\rho\}$ has coefficents $\{\rho_\nu\}$:

$$\rho(t) = \sum_{\nu=1}^{p} \rho_\nu e_\nu(t).$$

Sparsity and weak ℓ_p Say $\rho \in w\ell_p(C)$ if

$$|\rho_{(\nu)}| \leq C\nu^{-1/p}, \qquad \nu = 1, 2, \ldots$$

- p small $\Rightarrow$ rapid decay of ordered coefficients
- **choose basis** to exploit sparsity

Consistency of Sparse PCA

- Single component model. Suppose (i) $p/n \to c > 0$,
  (ii) $\|\rho(n)\| \to \varrho > 0$.
- Assume *Sparsity*: $\rho(n) \in w\ell_p(C)$ uniformly in n
- Subset selection rule:

$$\hat{I} = \{\nu : \hat{\sigma}_\nu^2 > \sigma^2(1 + c\sqrt{2\log p}\sqrt{2/n})\}$$

- Let $\hat{\rho}$ denote sparse PCA estimate based on $\hat{I}$.

Theorem $\angle(\hat{\rho}, \rho) \stackrel{a.s.}{\longrightarrow} 0.$

Later work (D. Paul): rates of convergence, lower bounds.

Example 3: Summary for Sparse PCA

- initial dimension reduction before PCA
  - otherwise, inconsistency!

- use basis with sparse representation
  - so that little is lost in initial dimension reduction

- Background role for large random matrices
  - Small perturbations of symmetric matrices
  - a.s. bounds for extreme eigenvalues of large matrices

Agenda

- Orientation
- Ex. 1: PCA - eigenvalues [v. brief]
- Ex. 2: CCA etc - eigenvalues [main]
- Ex. 3: sparse PCA - eigenvectors [brief]
- Some related problem areas [mention]

Some Related Problem Areas

- Extreme Sample Eigenvalues in large n, p setting.
  - Limiting distributions and approximations under "alternative" hypotheses
  - First order (strong law behavior) under dependence
  - Validity of bootstrap confidence intervals

Some Related Problem Areas

- Extreme Sample Eigenvalues in large n, p setting.
  - Limiting distributions and approximations under "alternative" hypotheses
  - First order (strong law behavior) under dependence
  - Validity of bootstrap confidence intervals
- Sample Eigenvectors (associated with extreme eigenvalues)
  - Consistency and asymptotic distribution as p grows
  - Effect of regularization (as in functional data)
  - possibilities for "sparse" versions of PCA [Lu]

Some Related Problem Areas

- Extreme Sample Eigenvalues in large n, p setting.
  - Limiting distributions and approximations under "alternative" hypotheses
  - First order (strong law behavior) under dependence
  - Validity of bootstrap confidence intervals
- Sample Eigenvectors (associated with extreme eigenvalues)
  - Consistency and asymptotic distribution as p grows
  - Effect of regularization (as in functional data)
  - possibilities for "sparse" versions of PCA [Lu]
- Empirical distributions of eigenvalues
  - (further) statistical uses of Marcenko-Pastur law
  - statistical potential of central limit theorems for linear statistics of eigenvalues $\sum h(l_i)$.

Some Related Problem Areas, Ctd.

- Estimation of large covariance matrices
  - Sparsity models for non-unit variances (subspaces of elevated variance)
  - Prior distributions on covariance matrices
  - Frequentist properties of Bayes estimates

Some Related Problem Areas, Ctd.

- Estimation of large covariance matrices
  - Sparsity models for non-unit variances (subspaces of elevated variance)
  - Prior distributions on covariance matrices
  - Frequentist properties of Bayes estimates
- Classification & clustering
  - kernel PCA and ICA in statistical learning theory
  - models for large covariance structures

Some Related Problem Areas, Ctd.

- Estimation of large covariance matrices
  - Sparsity models for non-unit variances (subspaces of elevated variance)
  - Prior distributions on covariance matrices
  - Frequentist properties of Bayes estimates
- Classification & clustering
  - kernel PCA and ICA in statistical learning theory
  - models for large covariance structures
- Issues from application domains
  - meteorology/climate, signal processing (multiple input, multiple output (MIMO)), face recognition, document retrieval, hyperspectral imagery

Publicity

Program, Fall Semester 2005-06, at
SAMSI [Statistics and Applied MathS Institute], North Carolina

"High dimensional multivariate statistics and random matrices"

- statistical focus: spectral properties – eigenvalues, eigenvectors
- RMT focus: RMT applications with (potential) relevance to statistics
- connections with selected areas of application

Info: contact IMJ or Craig Tracy (UC Davis)

THANK YOU!