Charles Stein, covariance matrix estimation and some memories from one of his students

Wei-Liem Loh
National University of Singapore

21 June 2019

In memory of Charles and Margaret Stein
(1920 – 2016)

Baranchik, Alvin	1964
Chen, Louis H. Y.	1972
George, Edward	1981
Giri, Narayan	1962
Harris, Bernard	1958
Loh, Wei-Liem	1988
Morris, Carl	1966
Pong, James	1976
Portnoy, Stephen	1969
Srivastava, Muni	1964
Zaman, Asad	1978
Zidek, James	1967

- Stein's two-sample test for a linear hypothesis whose power is independent of the variance (1945).
  Charles showed that by using a two stage sequential procedure one could give a confidence interval for the mean of a normal distribution having a prescribed width $w > 0$ without knowing the variance.
- Hunt-Stein theorem, unpublished, (1946).
  A theorem stating conditions under which there exists a maximin invariant test in a problem of statistical hypothesis testing. It was eventually published in Erich Lehmann's well known text *Testing Statistical Hypotheses* (1959).
- The admissibility of Hotelling's T^2-test (1956).

- Stein's paradox: inadmissibility of the usual estimator for the mean of a multivariate normal distribution (1956).
- James-Stein shrinkage estimator of the multivariate normal normal mean when the dimension $d \geq 3$ (1961).
- Stein's method of probability approximations (1972).
- SURE: Stein's unbiased risk estimate (1973, 1981).
- plus many more, ...

- Proc. 4th Berkeley Symp. (1961):
Charles considered the problem of estimating the covariance matrix Σ given the sample covariance matrix S/n, $S \sim W_p(\Sigma, n)$, with respect to the convex loss function

$$L(\Sigma, \hat{\Sigma}) = \mathrm{tr}(\Sigma^{-1}\hat{\Sigma}) - \log|\Sigma^{-1}\hat{\Sigma}| - p.$$

- This loss function is called Stein's loss nowadays in honor of Charles.
- Charles constructed an estimator $\hat{\Sigma}$ which has a remarkably simple expression, is constant risk minimax and dominates any constant multiple of S (including the sample covariance matrix S/n).

- The Stein 1961 estimator is

$$\hat{\Sigma} = T' \begin{pmatrix} \frac{1}{n+p-1} & & & 0 \\ & \frac{1}{n+p-3} & & \\ & & \ddots & \\ 0 & & & \frac{1}{n+p-2p+1} \end{pmatrix} T$$

where $S = T'T$ with T an upper triangular matrix.

- Charles noted though that $\hat{\Sigma}$ is still inadmissible and has the unappealing feature of not being invariant under permutations of the variables.

- Since the 1950s, Charles observed that the eigenvalues of the sample covariance matrix S/n are much more spread out than the eigenvalues of the population covariance matrix Σ. Thus an approach to getting an improved estimator of Σ is to correct for the distortion of the sample eigenvalues.
- In his Rietz Lecture (1975), Charles developed a calculus on the eigenstructure of $S \sim W_p(\Sigma, n)$ which culminates in the Wishart identity: under suitable regularity conditions,

$$\mathbb{E}\mathrm{tr}\Sigma^{-1}g(S) = \mathbb{E}\mathrm{tr}[nS^{-1}g(S) + 2S\nabla\{g(S)S^{-1}\}]$$

  where $\nabla = (\nabla)_{p\times p}$, $\nabla_{ij} = (1/2)(1+\delta_{ij})\partial/\partial s_{ij}$ and g maps the set of $p \times p$ positive definite matrices to $\mathbb{R}^{p\times p}$.
- Independently of Charles, Leonard Haff (1978) also proved the Wishart identity using Stoke's theorem.

- Charles considered the class of orthogonally invariant estimators given by
$$\hat{\Sigma} = R\varphi(L)R'$$
where
(i) $L = diag(l_1, l_2, \ldots, l_p)$, $l_1 \geq l_2 \geq \ldots \geq l_p$, are the eigenvalues of S,
(ii) R is an orthogonal matrix such that $S = RLR'$,
(iii) $\varphi(L) = \mathrm{diag}(\varphi_1(L), \ldots, \varphi_p(L))$ with nonnegative elements.

- With respect to Stein's loss, the Stein unbiased risk estimate (SURE) is computed to be

$$\begin{aligned}
\mathrm{SURE} &= 2\sum_i \sum_{j\neq i} \varphi_i(L)/(l_i - l_j) + 2\sum_i \partial\varphi_i(L)/\partial l_i \\
&\quad +(n-p-1)\sum_i \varphi_i(L)/l_i - \sum_i \log\varphi_i(L) \\
&\quad +\text{terms without } \varphi_i\text{'s} \\
&= \sum_{i=1}^{p} \{n-p+1+2\sum_{j\neq i}\frac{l_i}{l_i-l_j}\}\psi_i - \log(\psi_i) \\
&\quad +2l_i\frac{\partial\psi_i(L)}{\partial l_i} + \text{terms without } \psi_i\text{'s},
\end{aligned}$$

  where $\psi_i = \varphi_i/l_i$.
- The above first appeared in Charles' Rietz lecture (1975).

- To minimize SURE with respect to the ψ_i's, Charles ignored the partial derivatives in SURE to obtain

$$\psi_i^s = \{(n-p+1) + 2l_i \sum_{j \neq i} 1/(l_i - l_j)\}^{-1}, i = 1, \ldots, p.$$

and hence

$$\varphi_i^s(L) = \frac{l_i}{n-p+1+2\sum_{j\neq i} l_i/(l_i - l_j)}, i = 1, \ldots, p.$$

- Noting that the constraint

$$\varphi_1^s(L) \geq \varphi_2^s(L) \geq \ldots \geq \varphi_p^s(L) \geq 0$$

may be violated, Charles proposed an isotonic regression to modify the φ_i^s's. This results in the Stein 1975 covariance matrix estimator:

$$\hat{\Sigma}^{ST} = R\varphi^{ST}(L)R',$$

where $\varphi^{ST} = \mathrm{diag}(\varphi_1^{ST}, \ldots, \varphi_p^{ST})$.

- The numerical performance of this estimator is excellent.
  (i) It reduces the risk drastically when the eigenvalues of Σ are close together.
  (ii) It is almost minimax. It exceeds the minimax risk only slightly when the population eigenvalues are extremely far apart.
  (iii) Stein's 1975 estimator has been called by many as the "gold standard".

- Charles' 1975 Rietz lectures were unpublished.
- In 1977, Charles gave a series of lectures in Leningrad. These lectures included his 1975 work on covariance matrix estimation. The lectures were to be translated and delivered in Russian.
- Charles told me when I was a student in the 1980s that the original English version of his lecture notes was lost and the Russian translated version was also misplaced.
- A few years back, I was really happy to find out that Charles' Leningrad lectures were published in 1986 in a Russian journal: Lectures on the theory of estimation of many parameters. *J. Math. Sci.* **34** (1986), 1373-1403.

  I am not sure if Charles was aware of this publication as he did not mention it to me when I was a student.

- The covariance matrix estimation problem that Charles was interested in is decision theoretic in nature. In particular, for sample size $n \geq p$ and p not too large; something like $p = 10, 50$.
- In the 1990s, due to advancement of the modern computer, a different kind of asymptotics emerged, namely both n and p tending to infinity where possibly $n < p$. This models the scenario of high dimension p and low sample size n.
- Random matrix theory becomes increasing important in statistics as a result. In particular, the Stieltjes transform and the Marčenko-Pastur equation.

- There has been a lot of work done on spectrum (eigenvalues) estimation since 2000 using random matrix theory.
- The following are three estimators that are very promising. These estimators are applicable even when $n < p$.
- Λ_p^{EK}, El-Karoui (2008)
- Λ_p^{LW}, Ledoit and Wolf (2012, 2015, 2018)
- Λ_p^{KV}, Kong and Valiant (2016)

- The following simulation study was done by Jun Wen, a former student of mine, to compare the performance of the three estimators with the Stein 1975 spectrum estimator $\hat{\Lambda}_p^{ST}$.
- Let Λ_p denote the population eigenvalues of Σ. Writing $\tilde{\Lambda}_p$ as a generic estimator for Λ_p, the loss function is

$$L(\tilde{\Lambda}_p, \Lambda_p) = \frac{1}{p} \sum_{i=1}^{p} |\tilde{\lambda}_i - \lambda_i|,$$

where $\tilde{\lambda}_i$ and λ_i are the ith smallest eigenvalues of $\tilde{\Lambda}_p$ and Λ_p respectively.

p	$\hat{\Lambda}_p^{ST}$	$\hat{\Lambda}_p^{LW}$	$\hat{\Lambda}_p^{EK}$	$\hat{\Lambda}_p^{KV}$
10	0.169	0.136	0.121	0.090
	(0.007)	(0.011)	(0.010)	(0.008)
20	0.120	0.098	0.082	0.056
	(0.005)	(0.008)	(0.007)	(0.005)
50	0.067	0.061	0.043	0.025
	(0.002)	(0.005)	(0.004)	(0.003)
500	0.0125	0.0153	0.0068	0.0040
	(0.0003)	(0.0016)	(0.0005)	(0.0007)
1000	0.0072	0.0081	0.0042	0.0022
	(0.0001)	(0.0011)	(0.0002)	(0.0004)

Table: Average loss when $\Lambda_p = \{1, \ldots, 1\}$, $n = 3p$ and $S \sim W_p(\Sigma, n)$

p	$\hat{\Lambda}_p^{ST}$	$\hat{\Lambda}_p^{LW}$	$\hat{\Lambda}_p^{EK}$	$\hat{\Lambda}_p^{KV}$
10	0.348	0.411	0.412	0.481
	(0.009)	(0.010)	(0.013)	(0.012)
20	0.306	0.326	0.305	0.362
	(0.005)	(0.007)	(0.009)	(0.009)
50	0.286	0.223	0.251	0.262
	(0.003)	(0.005)	(0.009)	(0.009)
500	0.3106	0.0952	0.2212	0.1648
	(0.0004)	(0.0038)	(0.0085)	(0.0081)
1000	0.3152	0.0850	0.2171	0.1787
	(0.0002)	(0.0031)	(0.0082)	(0.0089)

Table: Average loss when λ_i equals the $(i-0.5)/p$th theoretical quantile of $1+10\times\mathrm{Beta}(1,10), 1\leq i\leq p$, $n=3p$ and $S\sim W_p(\Sigma,n)$

p	$\hat{\Lambda}_p^{ST}$	$\hat{\Lambda}_p^{LW}$	$\hat{\Lambda}_p^{EK}$	$\hat{\Lambda}_p^{KV}$
10	1.18	1.27	1.30	1.66
	(0.03)	(0.04)	(0.03)	(0.04)
20	1.93	1.79	2.19	2.79
	(0.03)	(0.04)	(0.04)	(0.12)
50	4.47	3.07	5.06	6.14
	(0.03)	(0.06)	(0.10)	(0.50)
500	41.80	17.78	54.83	78.42
	(0.05)	(0.33)	(0.98)	(9.12)
1000	83.57	28.42	111.72	331.78
	(0.04)	(0.80)	(1.71)	(21.90)

Table: Average loss when $\Lambda_p = \{1, 2, \ldots, p\}$, $n = 3p$ and $S \sim W_p(\Sigma, n)$

- In particle physics, supersymmetry (SUSY) is a theory that proposes a relationship between two basic classes of elementary particles: bosons, which have an integer-valued spin, and fermions, which have a half-integer spin.
- There is no evidence at this time to show whether or not the theory of supersymmetry is physically correct.
- The mathematics of supersymmetry uses Grassmann anticommuting variables together with the usual mathematics (in particular complex analysis). This type of mathematics is called supermathematics (or superanalysis).

Supermathematics is well developed in the physics literature. Many books have been written on it. Some of these include

- Berezin, F. A. (1987). *Introduction to Superanalysis.* D. Reidel Publishing Company, Dordrecht.
- de Witt, B. S. (1992). *Supermanifolds*, 2nd edition. Cambridge Univ. Press, New York.
- Efetov, K. (1997). *Supersymmetry in Disorder and Chaos.* Cambridge Univ. Press, Cambridge.
- Khrennikov, A. (1997). *Superanalysis.* Kluwer Academic Pub., Dordrecht.
- Wegner, F. (2016). *Supermathematics and its Applications in Statistical Physics: Grassmann Variables and the Method of Supersymmetry*. Springer, Heidelberg.

- While the supersymmetry method is well known in physics, it appears that this method has little impact on statistics at this time.
- The supersymmetry method seems to have a dimension-reduction property for evaluating certain high dimensional integrals.
- As motivation, I shall now proceed directly to an example.

- Let S be a $p \times p$ sample covariance matrix such that nS has the Wishart distribution $W_p(\Sigma, n)$ where $n \geq p$ and Σ is a $p \times p$ population covariance matrix.
- The Stieltjes transform of S is defined to be

$$\begin{aligned} m(z) &= \frac{1}{p}\mathrm{tr}\{(S - z\mathbb{I}_p)^{-1}\} \\ &= \frac{1}{p}\sum_{i=1}^{p}\frac{1}{l_i - z}, \qquad \forall z \in \mathbb{C}\backslash[0,\infty), \end{aligned}$$

where $\mathbb{I}_p$ denotes the $p \times p$ identity matrix and $l_1 \geq l_2 \geq \ldots \geq l_p$ are the eigenvalues of S.

- To calculate $\mathbb{E}m(z)$, classical multivariate normal theory, cf. Alan T. James (*Ann. Math. Statist.*, 1964) gives the expression

$$\begin{aligned}\mathbb{E}m(z) &= \int_{l_1 > \ldots > l_p} m(z)(\frac{n}{2})^{np/2} \frac{\pi^{p^2/2}|\Sigma|^{-n/2}}{\Gamma_p(n/2)\Gamma_p(p/2)} \\ &\quad \times \prod_{i=1}^{p} l_i^{(n-p-1)/2} \prod_{i<j}(l_i - l_j) \\ &\quad \times \int_{O(p)} e^{\mathrm{tr}(-n\Sigma^{-1}HLH'/2)}(dH)(dL). \qquad (1)\end{aligned}$$

Here $L = \mathrm{diag}(l_1, \ldots, l_p)$, $O(p)$ denotes the group of $p \times p$ orthogonal matrices and (dH) is its Haar measure.
$\Gamma_p(.)$ is the multivariate Gamma function given by

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma(a - \frac{i-1}{2}), \forall \mathrm{Re}(a) > \frac{p-1}{2}.$$

- No simple expression is known for the integral

$$\int_{O(p)} e^{\mathrm{tr}(-n\Sigma^{-1}HLH'/2)}(dH).$$

  The above integral is sometimes called the Harish-Chandra-Itzykson-Zuber type integral.
- Harish-Chandra (*Amer. J. Math.*, 1958), obtained relatively simple closed form expressions of the above integral over a number of groups, such as the unitary group, but not the orthogonal group.

- It is also well known that, cf. R. J. Muirhead (1982), *Aspects of Multivariate Statistical Theory*,

$$\int_{O(p)} e^{\mathrm{tr}(-n\Sigma^{-1}HLH'/2)}(dH) = {}_0F_0(-nL/2, \Sigma^{-1})$$
$$= \sum_{k=0}^{\infty}\sum_{\kappa} \frac{C_\kappa(-nL/2)C_\kappa(\Sigma^{-1})}{k!C_\kappa(\mathbb{I}_p)},$$

 where $\sum_\kappa$ denotes summation over all partitions $\kappa = (k_1, \ldots, k_p)$, $k_1 \geq \ldots \geq k_p \geq 0$, of k, $C_\kappa(X)$ is the zonal polynomial of X corresponding to κ and ${}_0F_0(-nL/2, \Sigma^{-1})$ is a hypergeometric function with matrix arguments.
- Unfortunately, a resummation of the above series is a well known difficult problem and, as far as we know, no good algorithm currently exists.

Suppose $n \geq 5$ and $p \geq 1$. Then using the supersymmetry method, it can be proved that

$$
\begin{aligned}
\mathbb{E}m(z) \;=\; & -\frac{1}{8p}\int_0^\infty\int_0^\infty \frac{|r_a - r_b|(r_a r_b)^{(n-3)/2}}{\prod_{i=1}^p \sqrt{(nz - r_a\lambda_i)(nz - r_b\lambda_i)}} e^{-(r_a+r_b)/2} \\
& \times \sum_{k=0}^{n\wedge(p-1)} \frac{(-1)^k}{(n-k)!}(p-k)n^{p-k}z^{p-k-1}\Big\{ n(n-1)e_k(\Lambda) \\
& \quad +(n-1)\sum_{i=1}^{p}\lambda_i^2 e_{k-1}(\Lambda^{[-i]})\Big(\frac{r_a}{nz - r_a\lambda_i} + \frac{r_b}{nz - r_b\lambda_i}\Big) \\
& \quad + \sum_{j,l:1\leq j\neq l\leq p} \frac{r_a r_b \lambda_j^2\lambda_l^2 e_{k-2}(\Lambda^{[-j,-l]})}{(nz - r_a\lambda_j)(nz - r_b\lambda_l)}\Big\} dr_a dr_b, \qquad (2)
\end{aligned}
$$

where $\Lambda = \{\lambda_1, \ldots, \lambda_p\}$ are the eigenvalues of Σ where
$\lambda_1 \geq \ldots \geq \lambda_p \geq 0$.
$\Lambda^{[-i]} = \{\lambda_1, \ldots, \lambda_p\}\backslash\{\lambda_i\}$ and $\Lambda^{[-j,-l]} = \{\lambda_1, \ldots, \lambda_p\}\backslash\{\lambda_j, \lambda_l\}$.
$e_k(.)$'s are elementary symmetric polynomials.

- For any finite set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$, $m \in \mathbb{Z}^+$, we define the elementary symmetric polynomials in Θ to be

$$
\begin{aligned}
e_0(\Theta) &= 1, \\
e_k(\Theta) &= \sum_{1 \leq j_1 < j_2 < \ldots < j_k \leq m} \theta_{j_1} \ldots \theta_{j_k}, \qquad \forall k = 1, \ldots, m, \\
e_k(\Theta) &= 0, \qquad \forall k \in \mathbb{Z} \backslash \{0, \ldots, m\},
\end{aligned}
$$

and $e_0(\emptyset) = 1, e_k(\emptyset) = 0$ for $k \neq 0$ where $\emptyset$ denotes the empty set.

- The expressions of $\mathbb{E}m(z)$ given by equations (1) and (2) are rather different.
- An important difference is that the dimension of the integral in (1) is of order $O(p^2)$ while the dimension of the integral in (2) is just 2.
- This indicates that the supersymmetry method possesses a dimension reduction property.

- Let $q \in \mathbb{Z}^+$. Anticommuting variables χ_i, $i = 1, \ldots, q$, are defined formally as mathematical objects obeying the following anticommutative rules:

$$\chi_i \chi_j = -\chi_j \chi_i, \qquad \forall 1 \le i, j \le q.$$

 Complex numbers commute with anticommuting variables, i.e. $z\chi_i = \chi_i z$ for all $z \in \mathbb{C}$ and $i = 1, \ldots, q$. Consequently, $\chi_i^2 = 0$ for all i and any function of $\chi_1, \ldots, \chi_q$ is a finite polynomial of the form

$$\sum_{\theta_1, \ldots, \theta_q \in \{0,1\}} z_{\theta_1, \ldots, \theta_q} \chi_1^{\theta_1} \ldots \chi_q^{\theta_q},$$

 where $z_{\theta_1, \ldots, \theta_q} \in \mathbb{C}$ are constants.

- Like complex numbers, let χ_i^* be the complex conjugate of χ_i, $i = 1, \dots, q$. χ_i^* anticommutes with each other and with $\chi_1, \dots, \chi_q$. Also we have $(\chi_1 \dots \chi_q)^* = \chi_1^* \dots \chi_q^*$ and $(\chi_i^*)^* = -\chi_i$.
- Let k, k_1, k_2 be positive integers such that $k = k_1 + k_2$. A $k \times k$ supermatrix $\mathcal{F}$ is defined via block construction

$$\mathcal{F} = \begin{pmatrix} A & \mathscr{B} \\ \mathscr{C} & D \end{pmatrix},$$

where A, D are $k_1 \times k_1, k_2 \times k_2$ matrices, respectively, whose entries are commuting elements, and $\mathscr{B}, \mathscr{C}$ are $k_1 \times k_2, k_2 \times k_1$ matrices, respectively, whose entries are anticommuting elements.

- The $k_1 \times k_1$ matrix A is called the bosonic part of supermatrix $\mathcal{F}$.
- The $k_2 \times k_2$ matrix D is called the fermionic part of $\mathcal{F}$.
- The supertrace of $\mathcal{F}$ is defined as

$$\mathrm{str}(\mathcal{F}) = \mathrm{tr}(A) - \mathrm{tr}(D).$$

- The superdeterminant of $\mathcal{F}$ is defined as $\mathrm{sdet}(\mathcal{F}) = |A - \mathscr{B} D^{-1} \mathscr{C}||D|^{-1}$ if D^{-1} exists.
- Also $\mathrm{sdet}^{-1}(\mathcal{F}) = |A|^{-1}|D - \mathscr{C} A^{-1} \mathscr{B}|$ if A^{-1} exists.

- Integration over anticommuting variables is defined formally as: for any anticommuting variable χ, $\int d\chi = \int d\chi^* = 0$ and $\int \chi d\chi = \int \chi^* d\chi^* = c$, a constant. The actual value of c can be arbitrary. Here we take $c = 1/\sqrt{2\pi}$.
- Multiple integrals are carried out iteratively. In particular, we have

$$\int \int \exp(z\chi^*\chi) d\chi d\chi^* = \int \int (1 + z\chi^*\chi) d\chi d\chi^* = z, \quad (3)$$

for any constant $z \in \mathbb{C}$ since $(\chi^*\chi)^k = 0$ for all $k \geq 2$.
- Interestingly, equation (3) lies at the heart of the way the supersymmetry method is used.

Let $X_1, \ldots, X_n$ be an i.i.d. sequence of random vectors from $N_p(0, \Sigma)$. Define the $p \times n$ data matrix $Y = (X_1, \ldots, X_n)$. Then $S = YY'/n$.

Following Recher, et al. (*Phys. Rev. Lett.*, 2010), we begin with the generating function

$$Z(\mathfrak{z}_0, \mathfrak{z}_1) = \mathbb{E}\frac{|\mathfrak{z}_1 \mathbb{I}_p - YY'|}{|\mathfrak{z}_0 \mathbb{I}_p - YY'|}, \qquad \forall \mathfrak{z}_0, \mathfrak{z}_1 \in \mathbb{C}, \mathrm{Im}(\mathfrak{z}_0) > 0.$$

By invariance, we shall without loss of generality assume that $\Sigma = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ where $\lambda_1 \geq \ldots \geq \lambda_p$. Then

$$Z_{(\mathfrak{z}_0, \mathfrak{z}_1)} = \mathbb{E}\prod_{i=1}^{p} \frac{\mathfrak{z}_1 - nl_i}{\mathfrak{z}_0 - nl_i},$$

where $l_1, \ldots, l_p$ are the eigenvalues of S, the sample covariance matrix.

Now differentiating with respect to $\mathfrak{z}_1$ and then setting $\mathfrak{z}_1 = \mathfrak{z}_0 = z$, we obtain

$$\begin{aligned}\frac{\partial}{\partial \mathfrak{z}_1} Z(z, \mathfrak{z}_1)|_{\mathfrak{z}_1 = z} &= \mathbb{E}\sum_{j=1}^{p} \frac{1}{z - nl_j} \\ &= \mathbb{E}\frac{1}{n}\sum_{j=1}^{p}(\frac{z}{n} - l_j)^{-1} \quad = \quad -\frac{p}{n}\mathbb{E}m_p(\frac{z}{n}).\end{aligned}$$

- Thus to compute $\mathbb{E}m(z)$, it suffices to compute the generating function $Z(\mathfrak{z}_0, \mathfrak{z}_1)$.

From the joint distribution of the elements of Y, we have

$$\begin{aligned}& Z(\mathfrak{z}_0, \mathfrak{z}_1) \\ = \quad & \frac{1}{(2\pi)^{pn/2}|\Lambda|^{n/2}} \int_{\mathbb{R}^{np}} \frac{|\mathfrak{z}_1 I_p - YY'|}{|\mathfrak{z}_0 I_p - YY'|} e^{-\mathrm{tr}(Y'\Lambda^{-1}Y)} d[Y], \quad (4)\end{aligned}$$

where $Y = (y_{ij})_{1\le i\le p, 1\le j\le n}$ and $d[Y] = \prod_{i=1}^{p}\prod_{j=1}^{n} dy_{ij}$.

- In what follows, the motivation is to embed $\mathbb{R}^{np}$ in a larger superspace (with Grassmann anticommuting variables) in order to facilitate the integration for $Z(\mathfrak{z}_0, \mathfrak{z}_1)$.
- A crude analogy is the following: In many situations, in order to evaluate an integral on the real line $\mathbb{R}$, a common strategy is to embed $\mathbb{R}$ in the complex plane $\mathbb{C}$ and then apply contour integration techniques.
- For $j = 1, \ldots, p$, let $u_j, v_j \in \mathbb{R}$ and ζ, ζ^* be anticommuting variables. Define the $4 \times p$ matrix A by

$$A = \begin{pmatrix} u_1 & u_2 & \ldots & u_p \\ v_1 & v_2 & \ldots & v_p \\ \zeta_1 & \zeta_2 & \ldots & \zeta_p \\ -\zeta_1^* & -\zeta_2^* & \ldots & -\zeta_p^* \end{pmatrix}.$$

- The denominator of the ratio of determinants in equation (4) can be expressed as a Gaussian integral over a vector consisting of ordinary variables.
  Writing $u = (u_1, \dots, u_p)'$ and $v = (v_1, \dots, v_p)'$, we have

$$\begin{aligned} & \int \{\prod_{j=1}^{p} du_j v_j\} e^{-\mathbf{i}u'(\mathfrak{z}_0 \mathbb{I}_p - YY')u/2} e^{-\mathbf{i}v'(\mathfrak{z}_0 \mathbb{I}_p - YY')v/2} \\ = & \frac{\mathbf{i}^{-p}(2\pi)^p}{|\mathfrak{z}_0 \mathbb{I}_p - YY'|}. \end{aligned} \tag{5}$$

- Using (3), the determinant in the numerator of (4) can be expressed as a Gaussian integral over a vector consisting of Grassmann anticommuting variables.
  Writing $\zeta = (\zeta_1, \ldots, \zeta_p)'$, $\zeta^* = (\zeta_1^*, \ldots, \zeta_p^*)'$ and $\zeta^\dagger$ be the conjugate transpose of ζ, we have

$$\int \prod_{j=1}^{p} d\zeta_j^* \zeta_j \exp\Big\{ -\frac{\mathbf{i}}{2}\{\zeta^\dagger(\mathfrak{z}_1\mathbb{I}_p - YY')\zeta - \zeta'(\mathfrak{z}_1\mathbb{I}_p - YY')\zeta^*\}\Big\}$$
$$= \mathbf{i}^p(2\pi)^{-p}|\mathfrak{z}_1\mathbb{I}_p - YY'|. \tag{6}$$

- Writing $D = \mathrm{diag}(\mathfrak{z}_0, \mathfrak{z}_0, \mathfrak{z}_1, \mathfrak{z}_1)$ and $A^\dagger$ to be the conjugate transpose of A, it follows from (5) and (6) that

$$\begin{aligned} & \frac{|\mathfrak{z}_1 I_p - YY'|}{|\mathfrak{z}_0 I_p - YY'|} \\ = \quad & \int d[A] \exp\left(\frac{\mathbf{i}}{2}\mathrm{str}(DA^\dagger A - A^\dagger YY'A)\right), \end{aligned} \tag{7}$$

 where $d[A] = (2\pi)^{-p} \prod_{j=1}^{p} du_j dv_j d\zeta_j^* d\zeta_j$.
- Since the imaginary part of $\mathfrak{z}_0$ is greater than zero, the convergence of the Gaussian integral is ensured.

- Plugging (7) into (4), we obtain

$$
\begin{aligned}
& Z(\mathfrak{z}_0, \mathfrak{z}_1) \\
= \;& \frac{1}{(2\pi)^{np/2}|\Lambda|^{n/2}} \int d[A] \exp\left(\frac{\mathbf{i}}{2}\mathrm{str}(DA^\dagger A)\right) \\
& \times \int d[Y] \exp\left(-\frac{1}{2}\mathrm{tr}\{Y'(\Lambda^{-1} + \mathbf{i}AA^\dagger)Y\}\right) \\
= \;& \int d[A] \exp\left(\frac{\mathbf{i}}{2}\mathrm{str}(DA^\dagger A)\right)|\mathbb{I}_p + \mathbf{i}AA^\dagger\Lambda|^{-n/2}. \qquad (8)
\end{aligned}
$$

The last equality is obtained by evaluating the Gaussian integral over Y.

- Y has been integrated out. The next step is to integrate out the $4 \times p$ matrix A whose last 2 rows comsist of Grassmann variables.

- Using the fact that $\mathrm{tr}(A^\dagger A\Lambda) = \mathrm{str}(A\Lambda A^\dagger)$, we obtain

$$|\mathbb{I}_p + \mathbf{i}A^\dagger A\Lambda| = \mathrm{sdet}(\mathbb{I}_4 + \mathbf{i}A\Lambda A^\dagger).$$

- This equality reduces the matrix dimensionality from $p \times p$ to a 4×4 supermatrix.
- It follows from (8) that

$$Z(\mathfrak{z}_0, \mathfrak{z}_1) = \int d[A] e^{\mathbf{i}\mathrm{str}(DA^\dagger A)/2} \mathrm{sdet}^{-n/2}(\mathbb{I}_4 + \mathbf{i}A\Lambda A^\dagger), \quad (9)$$

where $d[A] = (2\pi)^{-p} \prod_{j=1}^{p} du_j dv_j d\zeta_j^* d\zeta_j$.
- For brevity, the domain of the above integral is omitted and is taken to be over the region in which the variables are defined.

- The next step is also called generalized Hubbard-Stratonovich transformation in the physics literature.
- We shall now express the integral in (9) in terms of Gaussian integrals to facilitate the integration.
- Motivated by the delta function representation:

$$\delta(\sigma - A^{\dagger}\Lambda A) \sim \int d[\rho] e^{-\mathbf{i}\mathrm{str}\{\rho(\sigma - A^{\dagger}\Lambda A)\}},$$

it can be shown that

$$\mathrm{sdet}^{-n/2}(I_4 + \mathbf{i}A\Lambda A^{\dagger}) = \lim_{\varepsilon \to 0^+} \int d[\rho] \ell_{\varepsilon}(\rho) e^{-\mathbf{i}\mathrm{str}(A\Lambda A^{\dagger}\rho)/2},$$

where

$$\ell_{\varepsilon}(\rho) = \int d[\sigma] \mathrm{sdet}^{-n/2}(I_4 + \mathbf{i}\sigma) e^{\{\mathbf{i}\mathrm{str}(\sigma\rho) - 4\varepsilon\sigma_1^2\}/2}. \qquad (10)$$

ρ is a 4×4 supermatrix defined by

$$\rho = \begin{pmatrix} \rho_0 & \omega \\ \omega^{\dagger} & \mathbf{i}\rho_1 \mathbb{I}_2 \end{pmatrix}$$

ω is a 2×2 matrix with anticommuting elements given by

$$\omega = \begin{pmatrix} \psi & \psi^* \\ \phi & \phi^* \end{pmatrix}.$$

The differentials are defined as

$$d[\rho] \quad = \quad (4\pi)^{-2} d\rho_{aa} d\rho_{ab} d\rho_{bb} d\rho_1 d\psi d\psi^* d\phi d\phi^*.$$

The 4×4 supermatrix σ is similarly defined as ρ.

- Finally we obtain

$$
\begin{aligned}
Z(\mathfrak{z}_0, \mathfrak{z}_1) &= \int d[A] \lim_{\varepsilon \to 0^+} \int d[\rho] \ell_\varepsilon(\rho) e^{-\mathbf{i}\,\mathrm{str}(A\Lambda A^\dagger \rho)/2} e^{\mathbf{i}\,\mathrm{str}(DAA^\dagger)/2} \\
&= \lim_{\varepsilon \to 0^+} \int d[\rho] \ell_\varepsilon(\rho) \int d[A] e^{\mathbf{i}\,\mathrm{str}(DAA^\dagger - A\Lambda A^\dagger \rho)/2} \\
&= \lim_{\varepsilon \to 0^+} \int d[\rho] \ell_\varepsilon(\rho) \prod_{j=1}^{p} \mathrm{sdet}^{-1/2}(D - \rho \lambda_j).
\end{aligned}
$$

- Since $d[\rho] = (4\pi)^{-2} d\rho_{aa} d\rho_{ab} d\rho_{bb} d\rho_1 d\psi d\psi^* d\phi d\phi^*$, the domain of the above integral does not depend on p.
- What remains to be done is to integrate the last integral explicitly to obtain the an exact expression of $\mathbb{E}m(z)$ in terms of a double integral, i.e. (2). This can be done by "brute force", if necessary.

In conclusion, we list some other results on random matrix theory proved via the supersymmetry method.

- Recher, Kieburg, Guhr and Zirnbauer (*Phys. Rev. Lett.*, 2010) expressed the expectation of the density of the empirical spectral distribution of a real Wishart matrix in term of a double integral.
- Wirtz and Guhr (*Phys. Rev. Lett.*, 2013) obtained a relatively simple exact expression for the expected distribution of the smallest eigenvalue of a real correlated Wishart matrix.
- The real Wishart distribution, though most relevant to statistics, is one of the more technically difficult class of matrices to work with.
- Similar, though simpler, results has been obtained in the literature for the complex Wishart distribution.

- Berezin, F. A. (1987). *Introduction to Superanalysis.* D. Reidel Publishing Company, Dordrecht.
- Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.
- Recher, C., Kieburg, M. and Guhr, T. (2010). Eigenvalue densities of real and complex Wishart correlation matrices. *Phys. Rev. Lett.* **105** 244101.
- Recher, C., Kieburg, M., Guhr, T. and Zirnbauer, M. R. (2012). Supersymmetry approach to Wishart correlation matrices: exact results. *J. Statist. Phys.* **148** 981-998.
- Wen, J. (2017). *On High-Dimensional Covariance Matrices Estimation*. PhD thesis, National Univ. of Singapore.
- Wirtz, T. and Guhr, T. (2013). Distribution of the smallest eigenvalue in the correlated Wishart model. *Phys. Rev. Lett.* **111** 094101.