Some nonstandard distributions and asymptotics for multivariate analysis

Akimichi Takemura

Univ. of Tokyo

August 10, 2010
In this talk I give a survey on some nonstandard techniques for multivariate distribution theory.

Based on joint works with Satoshi Kuriki, Yo Sheena, Hidehiko Kamiya and Tomonari Sei.

We cover the following topics.

1. Tube formula approximation to maximum type test statistics
2. Wishart matrix when population eigenvalues are infinitely dispersed
3. Star-shaped distributions
4. Multivariate distributions defined via optimal transport
Introduction

The four topics can be classified into two groups.

- Based on normality (including asymptotic normality):
  - Tube formula
  - Wishart matrix when population eigenvalues are infinitely dispersed

- Construction of non-normal distribution:
  - Star-shaped distributions
  - Distributions via optimal transport
1. Tube formula approximation to maximum type test statistic
Tube formula: some historical background

- Jacob Steiner already had Steiner’s formula (1840) for the volume of a tube of a convex set.
- Minkowski defined mixed volumes.
- Hotelling (1939) derived the tube formula for a one-dimensional curve and then H. Weyl immediately generalized it to a general dimension.
- Hotelling’s motivation was a nonlinear regression problem (explained later).
Euler characteristic method (independent development)

- Euler characteristic heuristic was initiated by R.J. Adler for approximating the distribution of the maximum of a random random field (Adler-Hasofer (1976), Adler’s book (1981)).
- This method has been vigorously developed by Adler and Keith Worsley\(^1\).
- Some important foundational work was done by Jonathan Taylor (2001 thesis).

\(^1\)Keith Worsley passed away in February of 2009, which is a great loss.
Two methods are equivalent

- Around 2000, I and Kuriki were sitting in a talk by Keith Worsley in ISM (Institute of Statistical Mathematics, Tokyo, Japan) and was surprised that he was doing the same computations as us.
- Takemura and Kuriki (2002) proved the equivalence of these two methods by using Morse theorem (for finite dimensional case).
- Tube method can be understood as finite dimensional specialization of Euler characteristic method.
- I should also mention that “abstract tube” by Naiman and Wynn is a discrete analog of tube formula.
Consider a nonlinear regression model:

$$Y_i = \mu + \beta f(x_i, \theta) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2)$$

For example $f(x_i, \theta) = \cos(\theta x_i), \quad i = 1, \ldots, n$. (Nonlinear in $\theta$)

Consider testing the null hypothesis

$$H : \beta = 0$$

Let $(\hat{\mu}, \hat{\beta}, \hat{\theta})$ be the maximum likelihood estimate.

Residuals:

$$e_i = Y_i - \hat{\mu} - \hat{\beta} f(x_i, \hat{\theta}), \quad i = 1, \ldots, n.$$
Back to Hotelling’s original problem

- The likelihood ratio test is written as
  \[
  \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} < c \Rightarrow \text{reject } H.
  \]
- The null distribution of the LRT is not standard (i.e. note a beta distribution) because of the singularity.
- Singularity: \( \theta \) can not be estimated under the null hypothesis \( H \).
- Let us consider the problem from a geometric viewpoint.

Recap of simple regression

- For simplicity omit \( \mu \) from the model (as in Hotelling’s original paper):
  \[
  Y_i = \beta f(x_i, \theta) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2)
  \]
For a given $\theta$, the LSE $\hat{\beta}(\theta)$ of $\beta$ is

$$
\hat{\beta}(\theta) = \frac{\sum_{i=1}^{n} Y_i f(x_i, \theta)}{\sum_{i=1}^{n} f(x_i, \theta)^2} = \frac{Y' f(\theta)}{\|f(\theta)\|^2},
$$

where $Y = (Y_1, \ldots, Y_n)'$, $f(\theta) = (f(x_1, \theta), \ldots, f(x_n, \theta))'$.

Let $\hat{Y}(\theta) = \hat{\beta}(\theta) f(\theta)$, $e(\theta) = Y - \hat{Y}(\theta)$. Then

$$
\|Y\|^2 = \|\hat{Y}(\theta)\|^2 + \|e(\theta)\|^2
$$

and minimizing $\|e(\theta)\|^2$ is equivalent to maximizing $\|\hat{Y}(\theta)\|^2$.

$$
\hat{Y}(\theta) = \frac{Y' f(\theta)}{\|f(\theta)\|} = Y' u(\theta), \quad u(\theta) = \frac{f(\theta)}{\|f(\theta)\|} \in S^{n-1}, \quad (1)
$$

where $S^{n-1} \subset \mathbb{R}^n$ is the unit sphere.
Back to Hotelling’s original problem

- The rejection region of LRT is written as
  \[
  \min_{\theta} \frac{\|e(\theta)\|^2}{\|Y\|^2} < c
  \]

- This is equivalent to
  \[
  c' < \max_{\theta} \frac{(Y'u(\theta))^2}{\|Y\|^2} = \max_{\theta} (\tilde{Y}'u(\theta))^2, \quad \tilde{Y} = Y/\|Y\| \in S^{n-1}.
  \]

- Write \( M = \{u(\theta) \mid \theta \in \Theta\} \), where \( \Theta \) is the range of \( \theta \). Then \( M \subset S^{n-1} \). If \( \theta \) is a scalar, then \( M \) is a curve in \( S^{n-1} \).

- Maximization in (1) is the projection of \( \tilde{Y} \) onto \( M \).
Projection onto $M$

Figure: Projection onto $M$
Let $z = (z_1, \ldots, z_n)' \sim N_n(0, I_n)$.

Let $M \subset S^{n-1}$ be a $C^2$-submanifold of dimension $d = \dim M$ with piecewise smooth boundaries.

Let

$$Z(u) = u'z = \sum_{i=1}^{n} u_i z_i, \quad u = (u_1, \ldots, u_n)' \in M.$$ 

Also consider a standardized random field

$$Y(u) = u'z/\|z\|, \quad u \in M, \quad \|z\| = \sqrt{z'z}.$$
We want to evaluate the distributions of maxima, corresponding to maximum type test statistics:

\[ T = \max_{u \in M} Z(u), \quad U = \max_{u \in M} Y(u). \]

The tube method gives an approximation of the tail probabilities

\[ P(T \geq x), \quad x \uparrow \infty, \quad \text{and} \quad P(U \geq x), \quad x \uparrow 1. \]
Spherical tube and its volume

Evaluation of the distribution reduces to the evaluation of the volume of a spherical tube around $M$.

**Figure:** Spherical tube around $M$
Spherical tube and its volume

- Let
  \[ M_\theta = \left\{ v \in S^{n-1} \mid \min_{u \in M} \cos^{-1}(u'v) \leq \theta \right\} \]
  denote the tube around \( M \) with radius \( \theta \).
- Let \( \text{Vol}(M_\theta) \) denote the \((n - 1)\)-dimensional spherical volume of \( M_\theta \).
- By definition
  \[ P\left( \max_{u \in M} Y(u) \geq \cos \theta \right) = \frac{\text{Vol}(M_\theta)}{\Omega_n}, \]
  where
  \[ \Omega_n = \text{Vol}(S^{n-1}) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \]
  and \( \bar{B}_{a,b}(\cdot) \) denotes the upper probability of beta distribution with parameter \((a, b)\).
Tube formula for the volume of a spherical tube

- **Tube formula:** For $\theta$ smaller than the critical radius,

$$\text{Vol}(M_\theta) = \Omega_n \left\{ w_{d+1} B_{\frac{d+1}{2}, \frac{n-d-1}{2}} (\cos^2 \theta) + w_d B_{\frac{d}{2}, \frac{n-d}{2}} (\cos^2 \theta) \\
+ \cdots + w_1 B_{\frac{1}{2}, \frac{n-1}{2}} (\cos^2 \theta) \right\},$$

where $w_1, \ldots, w_{d+1}$ are geometric invariants of $M$, which can be evaluated by differential geometric methods.

- In particular $w_{d+1} = \text{Vol}(M)/\Omega_n$, $w_d = \text{Vol}(\partial M)/\Omega_n$.

- The Euler characteristic of $\chi(M)$ of $M$:

$$\chi(M) = 2 \sum_{e=0}^{m} w_{m+1-e} = \begin{cases} 
2(w_1 + w_3 + \cdots + w_{m+1}) & (m: \text{even}) \\
2(w_1 + w_3 + \cdots + w_m) & (m: \text{odd}),
\end{cases}$$
Critical radius

- Critical radius of the tube $M_\theta$ is the supremum of $\theta$, such that $M_\theta$ does not have a self-intersection.

**Figure:** Tubes with a radius equal to the critical radius.
Tail probability for $T$ (non-standardized maximum)

- For $T = \max_{u \in M} Z(u)$ we need integration of the tube formula in $\|z\|$.
- By integration on $\|z\|$ we have

$$
P\left(\max_{u \in M} Z(u) \geq x\right) = w_{d+1} \bar{G}_{d+1}(x^2) + w_d \bar{G}_d(x^2) + \cdots$$

$$+ w_1 \bar{G}_1(x^2) + O(\bar{G}_n(x^2(1 + \tan^2 \theta_c))),$$

where $\bar{G}_a(\cdot)$ is the upper probability of $\chi^2$ distribution with $a$ degrees of freedom and $\theta_c$ is the critical radius.
Application to a multilinear form

Consider a model for interaction in the two-way layout (Johnson and Graybill (1972)):

\[ x_{ij} = \alpha_i + \beta_j + \phi u_i v_j + \epsilon_{ij}, \]

\[ \epsilon_{ij} \sim N(0, \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J. \]

This is a “rank one” interaction model.

Consider testing the null hypothesis of no interaction \( H : \phi = 0 \) against this model.

The LR statistic is given by the largest singular value of (doubly centered) data matrix

\[ (z_{ij})_{I \times J}, \quad \text{with} \quad z_{ij} = x_{ij} - \bar{x}_i. - \bar{x}_j + \bar{x}_{..}. \]
Application to a multilinear form

Consider an extension to the 3-way layout (Boik and Marasinghe, Kawasaki and Miyakawa):

\[ x_{ijk} = (\alpha \beta)_{ij} + (\beta \gamma)_{jk} + (\gamma \alpha)_{ki} + \phi u_i v_j w_k + \varepsilon_{ijk}, \]
\[ \varepsilon_{ijk} \sim N(0, \sigma^2), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K. \]

Let

\[ z_{ijk} = x_{ijk} - \bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.k} + \bar{x}_{i..} + \bar{x}_{.j} + \bar{x}_{..k} - \bar{x}_{...}. \]

The LR statistic for testing \( H_0 : \phi = 0 \) is given by the “largest singular value of 3-way array”:

\[ T = \max_{ijk} (u_i v_j w_k z_{ijk}) \]

subject to

\[ \sum u_i^2 = \sum v_j^2 = \sum w_k^2 = 1. \]
Application to a multilinear form

- The null distribution of the LR statistic $T$ has the canonical form
  $T = \max_{u \in M} u' z$, where
  $z \in \mathbb{R}^{(I-1)(J-1)(K-1)} \sim N(0, I_{(I-1)(J-1)(K-1)})$

  and $M = S^{I-2} \otimes S^{J-2} \otimes S^{K-2} \subset S^{(I-1)(J-1)(K-1)-1}$.

- For example, when $I = J = K = 3$,
  
  $P(T \geq c) \sim \pi \tilde{G}_4(c^2) - \frac{3\pi}{2} \tilde{G}_2(c^2)$

  as $c \to \infty$;

  $P(U \geq c) \sim \pi \tilde{B}_{4,4}(c^2) - \frac{3\pi}{2} \tilde{B}_{2,6}(c^2)$

  for $c \geq 2/\sqrt{7}$. 
Application to a multilinear form

Figure: Tail probability of $T (I = J = K = 3)$
Other uses of tube formula by Kuriki and Takemura include

- Test of multivariate normality based on maximized higher order cumulants
- Anderson-Stephens statistic for testing uniformity on the sphere
- Maximum covariance difference test for equality of two covariance matrices
- Distribution of projection pursuit index
2. Asymptotic distribution of Wishart matrix when the population eigenvalues are infinitely dispersed
Various asymptotics for Wishart distribution

- Let $W = (w_{ij})$ be distributed according to Wishart distribution $W_p(n, \Sigma)$, where $p$ is the dimension, $n$ is the degrees of freedom and $\Sigma$ is the covariance matrix.
- We are interested in the joint distribution of the eigenvalues and the eigenvectors of $W$.
- Exact distribution in terms of hypergeometric function of matrix arguments is still difficult.
- Various approximations
  - Classical (large $n$, fixed $p$): Anderson (1963) and many subsequent authors.
  - Random matrix theory (large $p$) is now a large research field.

Here we propose yet another type of asymptotics.
Wishart matrix when population eigenvalues are dispersed

Infinitely dispersed population eigenvalues

- Denote the spectral decompositions of $W$ and $\Sigma$ by
  
  $$W = GLG', \quad \Sigma = \Gamma \Lambda \Gamma'.$$

- $G$ and $\Gamma$: $p \times p$ orthogonal matrices.

- $L = \text{diag}(l_1, \ldots, l_p)$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ are diagonal with the eigenvalues $l_1 \geq \ldots \geq l_p > 0$, $\lambda_1 \geq \ldots \geq \lambda_p > 0$ of $W$ and $\Sigma$.

- The population eigenvalues become infinitely dispersed if
  
  $$\rho = \rho(\Sigma) = \max(\frac{\lambda_2}{\lambda_1}, \frac{\lambda_3}{\lambda_2}, \ldots, \frac{\lambda_p}{\lambda_{p-1}}) \rightarrow 0.$$
Correct normalizations for infinite dispersion

Let

\[ \tilde{G} = (\tilde{g}_{ij}) = \Gamma' G. \]

Columns of \( \tilde{G} \) are eigenvectors in a standardized coordinate system. \( \tilde{G} \) is close to the identity matrix \( I_p \).

Define

\[ f_i = \frac{l_i}{\lambda_i}, \quad 1 \leq i \leq p, \]

\[ q_{ij} = \tilde{g}_{ij} \frac{l_j^{\frac{1}{2}} \lambda_i^{-\frac{1}{2}}}{l_j^{\frac{1}{2}} \lambda_i^{-\frac{1}{2}}} = \tilde{g}_{ij} f_j^{\frac{1}{2}} \lambda_j^{\frac{1}{2}} \lambda_i^{-\frac{1}{2}}, \quad 1 \leq j < i \leq p. \]

(lower triangular part for \( \tilde{G} \)).
Asymptotic distribution under infinite dispersion

Theorem 1 (Takemura-Sheena (2005))

As \( \rho = \max(\lambda_2/\lambda_1, \ldots, \lambda_p/\lambda_{p-1}) \to 0 \),

\[
f_i \xrightarrow{d} \chi^2_{n-i+1}, \quad 1 \leq i \leq p,
\]

\[
q_{ij} \xrightarrow{d} N(0, 1), \quad 1 \leq j < i \leq p,
\]

and \( f_i \) \((1 \leq i \leq p)\), \( q_{ij} \) \((1 \leq j < i \leq p)\) are asymptotically mutually independently distributed.

Note the similarity of this result to the triangular decomposition of \( W \) when \( \Sigma = I_p \).
It just comes from the Wishart density.

Let $\Sigma = \Lambda$ be diagonal without loss of generality. Then $\tilde{g}_{ij} = g_{ij}$.

The joint density of $l_1, \ldots, l_p$ and $g_{ij}, i > j$, is

$$c\frac{|L|^{(n-p-1)/2}}{\Lambda^{n/2}} \exp\left( -\frac{1}{2} \text{tr} GLG' \Lambda^{-1} \right) \prod_{i<j} (l_i - l_j) \times J(G),$$

where $J(G)$ is the Jacobian from lower triangular part of $G$ to the uniform measure on the orthogonal group. This Jacobian actually does not matter.
Why should the result hold?

- The Jacobian of the transformation

\[
f_i = \frac{l_i}{\lambda_i}, \quad q_{ij} = g_{ij} \sqrt{f_j \lambda_j / \lambda_i}
\]

is given as

\[
\left| \frac{\partial (l_i, g_{ij})}{\partial (f_i, q_{ij})} \right| = \prod_{i=1}^{p} \lambda_i^{i-(p-1)/2} \prod_{i=1}^{p} f_i^{-(p-i)/2}.
\]

- Hence the joint density of \((f_i, q_{ij})\) is written as

\[
c \prod_{i=1}^{p} f_i^{(n-i-1)/2} \exp\left(-\frac{1}{2} \text{tr} GG' \Lambda^{-1}\right) \prod_{i<j} \left(1 - \frac{\lambda_j f_j}{\lambda_i f_i}\right) J(G)
\]
Why should the result hold?

- As $\rho \to 0$
  \[
  \prod_{i<j} (1 - \frac{\lambda_j f_j}{\lambda_i f_i}) J(G) \to J(I_p).
  \]

- The essential part is $\exp(-\frac{1}{2} \text{tr} GLG' \Lambda^{-1})$:
  \[
  \text{tr} GLG' \Lambda^{-1} = \sum_{i,j} g_{ij}^2 l_j / \lambda_j = \sum_i g_{ii}^2 l_i / \lambda_i + \sum_{i>j} q_{ij}^2 + \sum_{i<j} g_{ij}^2 l_j / \lambda_j
  \]
  \[
  \to \sum_i f_i + \sum_{i>j} q_{ij}^2
  \]

- Hence the joint density converges to
  \[
  c' \prod_{i=1}^p f_i^{-(n-i-1)/2} \exp\left(-\frac{1}{2} \sum_i f_i - \frac{1}{2} \sum_{i>j} q_{ij}^2\right).
  \]
Some generalizations

- The result can be generalized to blockwise dispersion of population eigenvalues, e.g., for some \( q \)

\[
\lambda_1 > \cdots > \lambda_q \quad \Rightarrow \quad \lambda_{q+1} > \cdots > \lambda_p.
\]

Then blockwise asymptotic independence result holds.

- We can also derive an asymptotic expansion in terms of

\[
\frac{\lambda_2}{\lambda_1}, \ldots, \frac{\lambda_p}{\lambda_{p-1}}.
\]
Our original motivation for the result was decision theoretic investigation of risk functions of estimators of the population covariance matrix at the boundary of the parameter space.

We proved results on “tail minimaxity” and non-minimaxity of some estimators.

We here present an application of blockwise result to a hypothesis testing problem.

Consider the null hypothesis on the $m$th ($m = 1, \ldots, p$) population eigenvalue

$$H_0^{(m)}: \lambda_m \geq \lambda_m^*$$

against the alternative $H_1^{(m)}: \lambda_m < \lambda_m^*$. 
Some uses of infinite dispersion

**Theorem 2 (Sheena-Takemura (2007))**

For testing hypothesis $H_0^{(m)}$ against $H_1^{(m)}$, a test with significance level $\gamma$ is given with the rejection region

$$l_m \leq l_m^*(\gamma),$$

where $l_m^*(\gamma)$ is the lower 100$\gamma$% point of the smallest eigenvalue of $W_m(n, \lambda^*_m I_m)$.

This theorem follows from the fact that the limiting case $\lambda_{m+1}/\lambda_m \to 0$, $\lambda_1 = \cdots = \lambda_m$, $\lambda_{m+1} = \cdots = \lambda_p$, gives the least favorable distribution for the testing problem.
Some thoughts on different asymptotics

- Are various asymptotics mutually exclusive?
- The mixed asymptotics may be interesting
- Prof. Fujikoshi and his collaborators are working on asymptotic expansions where both $n$ and $p$ are large. They report good numerical results.
- It may be feasible to mix infinite dispersion and large $n$. 
3. A generalization of elliptically contoured distributions to star-shaped distributions
Contours of the density of elliptically contoured distributions are proportional ellipsoids.

How about densities with square contours?

How about something like the following?

Figure: From elliptical contours to general contours
Star-shaped distributions

From elliptical contours to star-shaped contoured

- Suppose that the contours of the density are star-shaped and proportional with respect to the origin (concentric around the origin).
- Every contour is obtained by a magnification of one specific contour, obtained by multiplication by a positive constant.
- Then the independence of “length” and “direction” holds!
  - “Length”: on which contour the observation falls
  - “Direction”: the relative position on the contour

- Example in $\mathbb{R}^2$

$$g = g(x, y) = \max\{|x|, |y|\}, \quad \theta = \theta(x, y) = \text{“angle”}$$
From elliptical contours to star-shaped contours

Figure: Square contours

- Under this kind of density, $g$ and $\theta$ are independent.
- Distribution of $\theta$ only depends on the shape (i.e. square) of the contour.
General theory

The results on the previous page can be proved in general by the theory of invariant measures.

- Let a group $\mathcal{G}$ act on the sample space $\mathcal{X}$ from the left:

$$ (g, x) \mapsto gx : \mathcal{G} \times \mathcal{X} \to \mathcal{X}. $$

- $\mathcal{G}x = \{gx : g \in \mathcal{G}\}$: the orbit containing $x \in \mathcal{X}$

- $\mathcal{X}/\mathcal{G} = \{\mathcal{G}x : x \in \mathcal{X}\}$: the orbit space.

- In the previous example, $\mathcal{G} = \mathbb{R}^*_+$ is the multiplicative group of positive reals and the orbits are rays emanating from the origin.
General theory

- Let $\mathcal{G}_x = \{ g \in \mathcal{G} \mid gx = x \}$ denote the isotropy subgroup (stabilizer) of $x$.
- When $\mathcal{G}_x = \{ e \}$ for all $x \in \mathcal{X}$, the action is said to be free, where $e$ denotes the identity element of $\mathcal{G}$.
- For discussing star-shaped distributions, we only need to consider the case of free action.
- For simplicity we assume free action in this talk. However the whole theory can be generalized to the case of non-free action.
General theory

- A cross section $\mathcal{Z} \subset \mathcal{X}$: a set $\mathcal{Z}$ intersects each orbit $Gx$, $x \in \mathcal{X}$, exactly once.
- **Orbital decomposition**: each $x \in \mathcal{X}$ is uniquely written as $x = gz = g(x)z(x)$, $g \in G$, $z \in \mathcal{Z}$.

- $g(x)$: equivariant part of $x$, “length”
- $z(x)$: invariant part of $x$, “direction”
A relatively invariant dominating measure $\lambda$ on $\mathcal{X}$ with multiplier $\chi$:

$$\lambda(d(gx)) = \chi(g)\lambda(dx), \; g \in \mathcal{G}.$$  

($\chi(g)$ can be thought of as the Jacobian)

- We call a density $f(x)$ w.r.t. $\lambda$ cross-sectionally contoured if it is of the form

  $$f(x) = f_G(g(x)).$$

- Namely, $f(x)$ only depends on the “length” $g(x)$.
- The contours of $f$ are $\{x \mid g(x) = c\}$. 

General theory
Theorem 3 (Kamiya-Takemura-Kuriki (2008))

Suppose that $x$ is distributed according to a cross-sectionally contoured distribution $f_g(g(x))\lambda(dx)$. Under some regularity conditions on the topology of $\mathcal{X}$ and the orbits, we have:

1. $g = g(x)$ and $z = z(x)$ are independently distributed.
2. The distribution of $z$ only depends on the cross section $\mathcal{Z}$. 
Back to the star-shaped distribution

- \( G = \mathbb{R}_+^*, \ X = \mathbb{R}^p - \{0\}. \)
- \( G \) freely acts on \( X \) as:
  \[
  (g, (x_1, \ldots, x_p)) \mapsto (gx_1, \ldots, gx_p).
  \]
- The orbits under this action are rays emanating from the origin.
- A cross section \( \mathcal{Z} \) is the boundary of a star-shaped set.
- The Lebesgue measure \( dx \) is relatively invariant with multiplier \( \chi(g) = g^p. \)
- We call a distribution with the density of the form \( f(x) = f_G(g(x)) \) a **star-shaped distribution**.
- From Theorem 3, under the star-shaped distribution \( g(x) \) and \( z(x) \) are independent and the distribution of \( z(x) \) depends only on \( \mathcal{Z}. \)
Application to random matrices

- The general theory can be applied to distributions other than star-shaped distributions.
- We consider a pair of Wishart matrices.
- Let \( W_1 = (w_{1,ij}) \) and \( W_2 = (w_{2,ij}) \) be two \( p \times p \) positive definite matrices.
- The sample space \( \mathcal{X} \) is \( \{ W = (W_1, W_2) : W_1, W_2 \in PD(p) \} \).
- As a dominating measure we consider

\[
\lambda(dW) = (\det W_1)^a (\det W_2)^b dW_1 dW_2, \quad (2)
\]

where \( a, b > (p - 1)/2 \) and
\[
dW_1 = \prod_{1 \leq i \leq j \leq p} dw_{1,ij}, \quad dW_2 = \prod_{1 \leq i \leq j \leq p} dw_{2,ij}.
\]
We consider the action of the lower triangular group.

Let $LT(p)$ denote the group consisting of $p \times p$ lower triangular matrices with positive diagonal elements.

$G = LT(p)$ freely acts on $X$ by

$$( T, (W_1, W_2)) \mapsto (TW_1 T', TW_2 T'), \quad T \in LT(p).$$

We consider the Cholesky decomposition of $W_1 + W_2 = TT'$. Then the following beta-type cross section is standard:

$$Z' = \{(U, I_p - U) : 0 < U < I_p\} \subset PD(p) \times PD(p),$$
The orbital decomposition of $W = (W_1, W_2)$ w.r.t. $Z'$

$$(W_1, W_2) = (TUT', T(I_p - U)T'), \quad T = T(W), \quad U = U(W).$$ (3)

- $W_1 + W_2$ and hence $T(W)$ is considered as “length”.
- $U(W) = T^{-1}W_1(T^{-1})'$ is considered as “direction”.
Application to random matrices

- We now consider a general cross section.
- Let $S(U)$ be a function from $\{U : 0 < U < I_p\}$ to $LT(p)$.
- Define a cross section
  \[ Z = \{ (S(U)US(U)', S(U)(I_p - U)S(U)' ) , 0 < U < I_p \} \]
- Corresponding to $Z$ define
  \[ g(W) = T(W)S(U(W))^{-1}. \] (4)
Application to random matrices

Consider a density

\[ f(W) = f_g(g(W)) \] (5)

with respect to \( \lambda(dW) \) in (2),

Let \( W \) have the density (5). By Theorem 3, \( g(W) \) and \( U(W) \) are independently distributed.

Note that Theorem 3 states the independence of \( g(W) \) and \( S(U(W))U(W)S(U(W))' \). However since \( S(U(W))U(W)S(U(W))' \) and \( U(W) \) are in one-to-one correspondence (\( \mathcal{Z} \) and \( \mathcal{Z}' \) are in one-to-one correspondence), \( g(W) \) and \( U(W) \) are independently distributed.
4. Multivariate distributions defined via optimal transport
Multivariate distributions defined via optimal transport

- Transformation of a variable is convenient for univariate statistical analysis
  → ex. Box-Cox transformation
- In the univariate case, by a monotone transformation of the form $F^{-1}(G(Y))$, $Y \sim G$, we can transform any cumulative distribution function $G$ to any other cumulative distribution function $F$.
- A multivariate generalization?
  → The optimal transport theory is one answer.
- **HINT**: A monotone function is a derivative of a convex function in $\mathbb{R}^1$. 
Figure: The slope is strictly increasing
Let \( Y = (Y_1, \ldots, Y_p)' \sim N_p(0, I). \)

Let \( \psi(x) = \psi(x_1, \ldots, x_p) \) be a smooth and strictly convex function. We call \( \psi \) a “potential function”.

Assume that the gradient map

\[
    x \mapsto \nabla \psi(x) = \left( \frac{\partial \psi}{\partial x_1}, \ldots, \frac{\partial \psi}{\partial x_p} \right)'
\]

is onto \( \mathbb{R}^p \). (“co-finiteness” of \( \psi \).) \( \nabla \psi(x) \) is injective because \( \psi \) is strictly convex.

- For \( x, y \in \mathbb{R}^p, x \neq y \), let \( \tilde{\psi}(t) = \psi(x + t(y - x)) \), which is smooth and strictly convex in \( t \in \mathbb{R} \). Then \( \tilde{\psi}'(t) \) is strictly increasing.
- \( \tilde{\psi}'(0) = (y - x)' \nabla \psi(x), \tilde{\psi}'(1) = (y - x)' \nabla \psi(y) \).
- If \( \nabla \psi(x) = \nabla \psi(y) \), then \( \tilde{\psi}'(0) = \tilde{\psi}'(1) \) (a contradiction).

Therefore under co-finiteness, \( \nabla \psi : \mathbb{R}^p \to \mathbb{R}^p \) is bijective and smooth.
Multivariate distributions via optimal transport

- Then we can define a random vector $X$ on $\mathbb{R}^p$ by the solution of
  \[ \nabla \psi(X) = Y \] (6)

- Let
  \[ \nabla \nabla' \psi(x) = \left( \frac{\partial^2 \psi(x)}{\partial x_i \partial x_j} \right)_{i,j=1,\ldots,p} \]
  denote the Hessian matrix of $\psi$.

- Then the density function of $x$ is explicitly written as
  \[ p(x) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \|\nabla \psi(x)\|^2\right) \det(\nabla \nabla' \psi(x)). \]
  (No need to worry about the normalizing constant.)
A trivial example: \( \psi(x) = a'x + \frac{1}{2}x'Bx \) \( (B \text{ is positive definite}) \)

\[ \nabla \psi(x) = a + Bx \]

Hence a positive definite quadratic form corresponds to an affine transformation.

The basic fact is that any distribution can be obtained from \( N_p(0, I) \) by this kind of transformation.
Basic theorem

Theorem 4 (McCann (1995))

Let $P$ and $Q$ be two (arbitrary) probability measures on $\mathbb{R}^p$ absolutely continuous w.r.t. the Lebesgue measure. Let $Y$ be a random vector with the distribution $Q$. Then there exists a convex function $\psi$ on $\mathbb{R}^p$ such that the solution $X$ of $\nabla \psi(X) = Y$ has the distribution $P$. The function $\psi$ is unique up to an arbitrary constant.
Basic theorem

- This theorem is a consequence of the theory of optimal transport.
- Consider the problem minimizing the cost
  \[
  \int_{\mathbb{R}^p} \|x - T(x)\|^2 P(dx)
  \]
  subject to a one-to-one map \( T \) that pushes \( P \) forward to \( Q \):
  \[
  X \sim P \Rightarrow T(X) \sim Q.
  \]
- It can be shown that the optimal map \( T \) is the gradient map \( \nabla \psi \) in Theorem 4.
Let $p = 3$ and consider the potential function

$$\psi(x_1, x_2, x_3) = \frac{1}{2}x'x + \theta \sum_{\lambda=1}^{4} \arctan(e'_\lambda x), \quad |\theta| < \frac{2}{3\sqrt{3}},$$

where $e_1 = (1, 1, 1)'$, $e_2 = (-1, 1, -1)'$, $e_3 = (-1, -1, 1)'$, $e_4 = (1, -1, -1)'$.

Note that $\psi$ is convex for sufficiently small $|\theta|$.

Then the solution of $\nabla \psi(X) = Y$ has the 3-dimensional interaction as long as $\theta \neq 0$. 
An example of 3-dimensional interaction

(a) $p(x_1, x_2 | x_3 = 1)$.

(b) $p(x_1, x_2 | x_3 = -1)$.

Figure: Conditional densities of 3-dim. interaction model ($\theta = 0.15$).
An example of data analysis

- We applied this model with a general quadratic form to detect three-dimensional interaction of a real data.
- The data consists of scores of high-jump ($X_1$), 400m ($X_2$), and 110m hurdle ($X_3$) of 50 decathlon players (Miyakawa 1997).
- Preprocess: We first normalized the data such that the sample mean and variance of each variable are 0 and 1, respectively.
- Then the estimated potential function $\psi(x_1, x_2, x_3)$ was

$$\frac{1}{2}x' \begin{pmatrix} 1.054 & -0.094 & 0.137 \\ -0.094 & 1.138 & -0.307 \\ 0.137 & -0.307 & 1.148 \end{pmatrix} x + 0.186 \sum_{\lambda=1}^{4} \arctan(e_{\lambda}'x).$$

AIC of this model was 10.4 less than the Gaussian model.
An example of data analysis

- Indeed two empirical conditional correlations are quite different:
  \( \text{cor}(X_2, X_3 \mid X_1 > 0) = 0.688 \) and \( \text{cor}(X_2, X_3 \mid X_1 < 0) = -0.049 \).

Figure: 110mH v.s. 400m.

(a) all
Cor. = 0.465

(b) \( X_1 > 0 \) "good high-jumper"
Cor. = 0.688

(c) \( X_1 < 0 \) "bad high-jumper"
Cor. = -0.049
Exact sampling

- Sampling from $N_p(0, I)$ is easy.
- Once we determine the gradient map $\nabla \psi$, we can sample from $p(x)$ exactly by solving $\nabla \psi(X) = Y$, where $Y \sim N_p(0, I)$.
- The equation is equivalent to the following convex optimization problem:
  $$X = \arg\min_{x \in \mathbb{R}^p} \{\psi(x) - Y'x\}.$$ 
- We can efficiently solve this optimization problem by Newton’s method.
A family of distributions based on gradient maps

- Let $\psi_1(x), \ldots, \psi_m(x)$ be smooth convex functions such that $\nabla \psi_i$ is onto $\mathbb{R}^p$, $i = 1, \ldots, m$.
- Let $\Theta$ be a convex subset of $\mathbb{R}^m$.
- Consider
  \[
  \psi(x) = \psi(x, \theta) = \theta_1 \psi_1(x) + \cdots + \theta_m \psi_m(x), \quad x \in \mathbb{R}^p, \quad \theta \in \Theta. \tag{7}
  \]
- This defines a family of distributions, which we call a “g-model”.
- This family has the following nice property.

**Theorem 5 (Sei (2010))**

*The log-likelihood function of g-model is concave.*
Transformation from a uniform distribution

- We can also take the uniform distribution on the unit cube $[0, 1]^p$ as the standard distribution: $Y \sim U[0, 1]^p$.
- If we appropriately specify $\psi$, we can obtain a class of multivariate distributions on the cube.
- Example in $p = 2$.
  - Let
    \[
    \psi(x_1, x_2 \mid \theta) = \frac{1}{2} (x_1^2 + x_2^2) - \frac{\theta}{\pi^2} \cos(\pi x_1) \cos(\pi x_2)
    \]
  - Then the density of $x_1, x_2$ is given as
    \[
    p(x_1, x_2 \mid \theta) = \det \begin{pmatrix}
    1 + \theta \cos(\pi x_1) \cos(\pi x_2) & -\theta \sin(\pi x_1) \sin(\pi x_2) \\
    -\theta \sin(\pi x_1) \sin(\pi x_2) & 1 + \theta \cos(\pi x_1) \cos(\pi x_2)
    \end{pmatrix}
    = 1 + 2\theta \cos(\pi x_1) \cos(\pi x_2) + \frac{\theta^2}{2} \left( \cos(2\pi x_1) + \cos(2\pi x_2) \right).
    \]
Transformation from a uniform distribution

(a) $\theta = 0.5$.  
(b) $\theta = -0.5$.  

Figure: The probability density $p(x_1, x_2|\theta)$ for $\theta = \pm 0.5$. 
Multivariate distributions defined via optimal transport


