New Mathematics from Biology:
Limit theorems for the number of matchings in random graphs

Grzegorz A. Rempala

Div of Biostatistics and MBI
The Ohio State University

Analysis Seminar, Math OSU
Feb 2014
Outline of Topics

1 Preliminaries
   - Random Matchings
   - Permanents

2 $U$-statistics
   - Canonical Decomposition
   - USIO Limits

3 $P$-statistics
   - Canonical Decomposition for $P$-stats
   - Limit Theorem

4 Limit Thm for MPM's
Let $G = (V_1, V_2; E)$ be a bipartite graph with $\#(V_1) = m$, $\#(V_2) = n$, and $E = \{(i, j) : i \in V_1, j \in V_2\}$. Assume that $m \leq n$.

With each edge $(i, j) \in E$ we associate a random variable (rv) $X_{i,j}$. Assume $X = [X_{ij}]$ are iid.

Any subgraph $\tilde{G} = (V_1, \tilde{V}_2; \tilde{E})$ of $G$ which represents bijection between $V_1$ and $\tilde{V}_2 \subset V_2$ is called a perfect matching (PM).

PM can be identified with the set $\tilde{V}_2 = \{j_1, \ldots, j_m\} \subset V_2$.

Let $h : \mathbb{R}^m \to \mathbb{R}$ be a symmetric measurable function (kernel). The space of such functions denote by $L_s(m)$.

With any PM $\{j_1, \ldots, j_m\}$ and corresponding random weights $(X_{i,j_l})$, $i, l = 1, \ldots, m$, we associate the value $h(X_{1,j_1}, \ldots, X_{m,j_m})$. 
PM's of interest in e.g., netwks connectivity models/ social nets etc

We would like to count different PM's in $G$

$M_{m \times n}$ space of $m \times n$ matrices with real entries. Define $\text{Per}_h : M_{m \times n} \rightarrow \mathbb{R}$, $m \times n$ permanent with kernel $h^{(m)}$ as

$$
\text{Per}_{h^{(m)}}(x) = \sum_{\{j_1, \ldots, j_m\} \in \{1, \ldots, n\}} \sum_{\sigma \in S_m} h^{(m)}(x_{\sigma(1)}, j_1, \ldots, x_{\sigma(m)}, j_m) \quad x \in M_{m \times n}
$$

A large sub-class of symmetric functionals on $X$

For product kernel this is just a “classical” permanent of $X$ (Cauchy 1812)

Averaged $\text{Per}_h(X)$ is a $P$-statistic, which emphasizes analogies with a $U$-statistic

Computing $\text{Per}_h(X)$ is NP-complete. Need approximations (e.g., weak limits)

Use ideas from $U$-statistics theory
Example: Number of monochromatic PM’s

- Let $P(X = k) = p_k$, $k = 0, 1, \ldots, N$, where $N \leq \infty$.
- Assume $X_{i,j} \overset{d}{=} X$.
- Let
  \[
  h_m(x_1, \ldots, x_m) = \sum_{k=1}^{N} \prod_{i=1}^{m} I(x_i = k) \in \{0, 1\}.
  \]
- $X_{i,j} = k > 0$ is the $k$-th color of the edge connecting the nodes $i$ and $j$.
- If $X_{i,j} = 0$ there is no edge (i.e., weight is zero) between $i$ and $j$.
- Then
  \[
  Per_{h_{m_n,n}}(X) = \mathcal{M}(m_n, n)
  \]
  counts monochromatic PM’s in a random bipartite graph $G$.
- Interested in the limiting distribution of $\mathcal{M}(m_n, n)$ as $m_n/n \to \lambda$.
For $1 \leq m \leq n$ define symmetrization operator $\sigma_m^n : L_s^m \rightarrow L_s^n$ by

$$[\sigma_m^n h](x_1, \ldots, x_n) = \sum_{1 \leq s_1 < \ldots < s_m \leq n} h(x_{s_1}, \ldots, x_{s_m}).$$

$(X_n)$ sequence of iid rv's

$$U_n(h) = \binom{n}{m}^{-1} [\sigma_m^n(h)](X_1, \ldots, X_n)$$

is called $U$-statistic with kernel $h$. Halmos (1946) Hoeffding (1948)

If $E|h(X_1, \ldots, X_m)| < \infty$ then $U_n(h)$ admits canonical decomposition, i.e. there exist symmetric functions $g_c : \mathbb{R}^c \rightarrow \mathbb{R}$ such that

$$E g_c(x_1, \ldots, x_{c-1}, X_c) = 0, \quad c = 1, 2, \ldots, m,$$

for any $x_1, \ldots, x_m$, and

$$U_n(h) - E U_n(h) = \sum_{c=1}^m \binom{m}{c} \binom{n}{c}^{-1} [\sigma_c^n g_c](X_1, \ldots, X_n).$$
The functions $g_c$'s are constructed as follows

- Let
  \[ h_c(x_1, \ldots, x_c) = E h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m) \]
  for $c = 1, 2, \ldots, m$, where $h_m = h$. Note that $E(h_c) = E(h)$ for any $c = 1, \ldots, m$.

- Let $\tilde{h}_c = h_c - E(h)$, $c = 1, \ldots, m$. Then
  \[ g_c(x_1, \ldots, x_c) = \sum_{i=1}^{c} (-1)^{c-i} \sum_{1 \leq j_1 < \ldots < j_i \leq c} \tilde{h}_i(x_{j_1}, \ldots, x_{j_i}). \quad (1) \]

- The number $r = \min\{c \geq 1 : h_c \neq 0\}$ is the level of non-degeneracy. Important for asymptotics.
• $r = 1$ we have the classical CLT (Hoeffding 1948)

$$\sqrt{n}[U_n(h) - E(U_n(h))] \xrightarrow{d} \mathcal{N}(0, m^2 E g_1^2(X_1))$$

• $r = 2$ Serfling (1980) $n\{U_n(h) - E(U_n(h))\}$ is asymptotically

$$\left(\begin{array}{c} m \\ 2 \end{array}\right) \sum_{k=1}^{\infty} \lambda_k (C_k - 1),$$

where $(C_k)$ are iid chi-square $\chi(1)$ rv's, and $(\lambda_k)$ are defined by

$$g_2(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x_1) \phi_k(x_2),$$

where $(\phi_k)$ is an orthonormal basis in $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$, $P_X$ being the distribution of $X_1$

• For any $r = c \geq 1$ Dynkin and Mandelbaum (1983) showed

$$\left(\begin{array}{c} m \\ c \end{array}\right)^{-1} n^{c/2}(U_n(h) - E(U_n(h))) \xrightarrow{d} J_c(g_c),$$

where $J_c$ is a multiple Wiener integral of the symmetric function $g_c$
Limit theorem for USIO (RW06)

Let \((X_i)\) be a sequence of iid rv's. Let \((U_n^{(m_n)})\) be a sequence of U-statistics defined for \((X_i)\) of the same level of degeneration \(c\) and with corresponding canonical sequences \(g_{m_n} = (g_{m_n,k})_{k \geq c}\). Assume that

(i) \(Eh_{m_n}^2 < \infty \) \(n = 1, 2, \ldots,\)

(ii) \(m_n/\sqrt{n} \to \lambda > 0\)

(iii) there exists a sequence of canonical fcns \(g = (g_k)_{k \geq c} \) \(k = c, c + 1, \ldots\) such that

\[
\sum_{k=c}^{\infty} \frac{\lambda^k}{k!} E \ g_k^2 < \infty \quad \text{and} \quad \sup_n \sum_{k=c}^{\infty} \frac{(m_n^2/n)^k}{k!} E \ g_k^2 < \infty
\]

as well as

\[
\sum_{k=c}^{m_n} \frac{(m_n^2/n)^k}{k!} E (g_{m_n,k} - g_k)^2 \to 0.
\]

Then

\[
Z_n = U_n^{(m_n)} - E \left( U_n^{(m_n)} \right) \xrightarrow{d} \sum_{k=c}^{\infty} \frac{\lambda^k/2}{k!} J_k(g_k).
\]
Canonical decomposition for $P$-stats

- Recall

$$Per_{h(m)}^{(m,n)}(x) = \sum_{\{j_1,\ldots,j_m\}\in\{1,\ldots,n\}} \sum_{\sigma\in S_m} h^{(m)}(x_{\sigma(1)},j_1,\ldots,x_{\sigma(m)},j_m), \quad x \in M_{m\times n}$$

- $\mathbb{X} \in M_{m\times n}$ of iid entries ($X_{11} \sim P_X$) and $E|h^{(m)}| < \infty$. $Per_{h(m)}^{(m,n)}$ may be decomposed, similarly to a $U$-statistic, as follows (R02).

$$Per_{h(m)}^{(m,n)}(\mathbb{X}) = E\left(Per_{h(m)}^{(m,n)}(\mathbb{X})\right) + m! \binom{n}{m} \sum_{k=1}^{m} \frac{(n-k)!}{n!} W_{g_{m,k}}^{(m,n)},$$

where

$$W_{g_{m,k}}^{(m,n)} = \sum_{1 \leq i_1 < \ldots < i_k \leq m} \sum_{1 \leq j_1 < \ldots < j_k \leq n} \sum_{\sigma\in S_m} g_{m,k} \left(X_{i_\sigma(1)},j_1,\ldots,X_{i_\sigma(k)},j_k\right)$$

and (as for $U$-stats before)

$$g_{m,k}(w_1,\ldots,w_k) = \int_{\mathbb{R}^m} h^{(m)}(z_1,\ldots,z_m) \left(\prod_{r=1}^{k} (\delta_{w_r}(dz_r) - P_X(dz_r))\right) \left(\prod_{s=k+1}^{m} P_X(dz_s)\right)$$
Canonical decomposition for $P$-stats (cont)

- Let $X[k|k]$ be a $k \times k$ sub-matrix of $X$. Note that the previous formula may be written as

$$
\left\{ \binom{n}{m_n} m_n! \right\}^{-1} \left\{ Per_{h(m_n)}^{(m_n,n)}(X) - E(Per_{h(m_n)}^{(m_n,n)}(X)) \right\}
$$

$$
= \sum_{k=1}^{m} \frac{(n-k)!}{n!} W_{g_{m,k}}^{(m,n)} = \sum_{k=1}^{m} \frac{(n-k)!}{n!} \sigma_{(m,n)}^{(k,k)} \left( Per_{g_{m,k}}^{(k,k)}(X[k|k]) \right)
$$

- If $E \left[ \left( h^{(m_n)} \right)^2 \right] < \infty$, then

  (i) $Cov(W_{g_{m,k}}^{(m,n)}, W_{g_{m,l}}^{(m,n)}) = 0 \quad l \neq k$

  (ii) $Var(W_{g_{m,k}}^{(m,n)}) = k! \binom{n}{k} \binom{m}{k} E g_{m,k}^2$

- Similarly as for a $U$-statistic we say that $c$ is a non-degeneracy level of a $P$-statistic if $g_{k}^{(m)} \equiv 0$ for $k < c$
Limit theorem for PSIO (RW06)

Consider a sequence \( \left( P_{\text{Per}}^{(m_n,n)}(X) \right)_n \) where \( X \) is embedded in an infinite matrix \( X_\infty \) of iid entries. Assume \( c \) to be the common level of degeneracy of the associated canonical seqs \( g_{m_n} = (g_{m_n,k}) \) for \( k = c, c+1, \ldots, m_n \) as well as

(i) \( E \left[ (h^{(m_n)})^2 \right] < \infty \) for \( n = 1, 2 \ldots \)

(ii) \( m_n/n \to \lambda > 0 \)

(iii) there exists a sequence of canonical fcns \( g = (g_k) \) \( k = c, c+1, \ldots \) such that

\[
\sum_{k=c}^{\infty} \frac{\lambda^k}{k!} E g_k^2 < \infty \quad \text{and} \quad \sup_n \sum_{k=c}^{\infty} \frac{(m_n/n)^k}{k!} E g_k^2 < \infty
\]

as well as

\[
\sum_{k=c}^{m_n} \frac{(m_n/n)^k}{k!} E (g_{m_n,k} - g_k)^2 \to 0.
\]

Then

\[
\frac{1}{\binom{m_n}{n} m_n!} \left[ P_{\text{Per}}^{(m_n,n)}(X) - E \left( P_{\text{Per}}^{(m_n,n)}(X) \right) \right] \xrightarrow{d} \sum_{k=c}^{\infty} \frac{\lambda^{k/2}}{k!} J_k(g_k)
\]
Note that conditions (i)–(ii) are easily verifiable

Need some simple criteria to help verify (iii)

For instance

(a) \( g_{m\cdot n, k} \to g_k \) point-wise
for any \( k = c, c + 1, \ldots \)

(b) \( \exists A, B \ E(g_k^2) \leq AB^k \)
for any \( k = c, c + 1, \ldots \)

(c) \( \exists C, D \ E(g_{m\cdot n, k}^2) \leq CD^k \)
for any \( k = c, c + 1, \ldots, m_n \)
and \( n = 1, 2 \ldots \)

Then (a)–(c) \( \Rightarrow \) (iii)

or (a)–(c) with \( g_k, g_{m\cdot n, k} \) replaced by \( h_k, h_{m\cdot n, k} \) \( \Rightarrow \) (iii).
Example: Number of monochromatic PM’s

- Let $P(X = k) = p_k$, $k = 0, 1, \ldots, N$, where $N \leq \infty$.
- Assume $X_{i,j} \overset{d}{=} X$.
- Let
  \[
  h_m(x_1, \ldots, x_m) = \sum_{k=1}^{N} \prod_{i=1}^{m} I(x_i = k).
  \]
- $X_{i,j} = k > 0$ is the $k$-th color of the edge connecting the nodes $i$ and $j$.
- if $X_{i,j} = 0$ there is no edge (i.e., weight is zero) between $i$ and $j$.
- Then
  \[
  \text{Per}_{h_{m,n}}(X) = \mathcal{M}(m_n, n)
  \]
  counts monochromatic PM’s in a random bipartite graph $G$.

In the sequel write $h_m, \text{Per}(h_m)$ etc.
If \( h_m(x_1, \ldots, x_m) = \sum_{k=1}^{N} \prod_{i=1}^{m} I(x_i = k) \)
then \( E(h_m) = \sum_{k=1}^{N} p_k^m \) and,
\[
g_{m,k}(x_1, \ldots, x_k) = \sum_{i=1}^{N} p_i^m \prod_{l=1}^{k} \left( \frac{1}{p_i} I(x_l = i) - 1 \right).
\]

Consider first the 'finite-uniform' case, i.e., \( N < \infty \) and \( p_i = p, \ i = 1, \ldots, N \).
Then \( E(h_m) = Np^m \). Replace \( h_m \) by \( \tilde{h}_m = h_m/(Np^m) \)
\[
g_{m,k}(x_1, \ldots, x_k) = g_k(x_1, \ldots, x_k) = \frac{1}{N} \sum_{i=1}^{N} \prod_{l=1}^{k} \left( \frac{1}{p} I(x_l = i) - 1 \right)
= \frac{1}{N} \sum_{i=1}^{N} \phi_i^{\otimes k}(x_1, \ldots, x_k),
\]
where \( \phi_i(x) = \frac{1}{p} I(x = i) - 1, \ i = 1, \ldots, N \). By \( P \)-stats theorem
\[
\frac{\mathcal{M}(m,n)}{Np^m \binom{n}{m} m!} \xrightarrow{d} 1 + \frac{1}{N} \sum_{k=1}^{\infty} \frac{\lambda^{k/2}}{k!} \sum_{r=1}^{N} H_k(J_1(\phi_r)) = \frac{1}{N} \exp(-\lambda/2) \sum_{r=1}^{N} \exp \left( \sqrt{\lambda Z_r} \right),
\]
where \((Z_1, \ldots, Z_N)\) is an \( N \)-variate centered normal random vector with
covariances: \( E(Z_r Z_s) = -1 \) for \( r \neq s \) and \( EZ_r^2 = (1 - p)p^{-1} \),
\( r, s \in \{1, \ldots, N\} \).
In the general case, i.e. \( N \leq \infty \), let \( p = \max_{1 \leq i \leq N} p_i \) and let \( j_1, \ldots, j_K \) be such that \( p_{j_s} = p; \ s = 1, \ldots, K \). Assume \( K \geq 2 \).

Defining \( \tilde{h}_m = h_m/(Kp^m) \) (note \( E\tilde{h}_m \rightarrow 1 \)) gives corresp canonical fcns

\[
 g_{m,k}(x_1, \ldots, x_k) = \frac{1}{K} \sum_{i=1}^{N} \left( \frac{p_i}{p} \right)^m \prod_{l=1}^{k} \left( \frac{1}{p_i} I(x_l = i) - 1 \right) \quad k = 1, \ldots, m.
\]

Define \( \psi_s = \phi_{j_s}, \ s = 1, \ldots, K \) and note that for (iii) we may take

\[
 g_k(x_1, \ldots, x_k) = \frac{1}{K} \sum_{s=1}^{K} \prod_{l=1}^{k} \left( \frac{1}{p} I(x_l = j_s) - 1 \right) = \frac{1}{K} \sum_{s=1}^{K} \psi_s^\otimes_k(x_1, \ldots, x_k)
\]

for \( k = 1, 2, \ldots \).

Consequently,

\[
 \frac{M(m,n)}{Kp^m\binom{n}{m}m!} \overset{d}{\rightarrow} 1 + \frac{1}{K} \sum_{k=1}^{\infty} \frac{\lambda^{k/2}}{k!} \sum_{s=1}^{K} H_k(J_1(\psi_s)) = \frac{1}{K} \exp(-\lambda/2) \sum_{s=1}^{K} \exp\left(\sqrt{\lambda}Z_s\right)
\]

where \((Z_1, \ldots, Z_K)\) is a \( K \)-dimensional centered normal random vector with covariances \( E(Z_rZ_s) = -1 \) for \( r \neq s \) and \( E(Z_s^2) = (1 - p)p^{-1} \), \( r, s \in \{1, \ldots, K\} \). (Note: \( p \leq 1/2 \))
In Further Reading

Rempala, G. A. and Wesolowski J.
Symmetric Functionals on Random Matrices and Random Matchings Problems.
Springer-Verlag, New York 2008