Bayesian nonparametric methods for prediction and testing in ESTs analysis

Ramsés H. Mena

Joint work with Antonio Lijoi and Igor Prünster

IIMAS-UNAM, México

BNR 2007, Cambridge, UK.
August, 2007.

Analysis of ESTs

"Expressed sequence tags" (ESTs): Small pieces of DNA generated by partially sequencing randomly isolated gene transcripts that have been converted into cDNA.

- Each EST, putatively, represents a gene.
- ESTs analysis is an important tool for gene prediction, discovery and identification

Some problems of statistical interest:

- the coverage, defined as the proportion of unique genes in the cDNA library represented in the given sample of reads;
- the number of new unique genes to be observed in the additional sample;
- the discovery rate of new genes as a function of the future sample size.

- Typically a cDNA library contains a large and unknown number of differentially expressed genes: potentially infinite

- Each X_i denotes an EST $\approx$ CCCCCGTCTCTTTAAAAATATAT...
- Let $X^{(1,n)} = (X_1, \ldots, X_n)$ denote a base sample of ESTs, and K_n the number of "distinct" genes in $X^{(1,n)}$.
- Let $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ denote an additional sample
- $K_m^{(n)} := K_{n+m} - K_n$, the number of new species in $X^{(2,m)}$.
- ESTs info is typically reported in levels of expression: $\mathbf{N} = (N_1, \ldots, N_{K_n})$ where N_i denotes the frequency of gene i in the ESTs sample.

- Typically a cDNA library contains a large and unknown number of differentially expressed genes: potentially infinite

- Each X_i denotes an EST $\approx$ CCCCCGTCTCTTTAAAAATATAT...
- Let $X^{(1,n)} = (X_1, \ldots, X_n)$ denote a base sample of ESTs, and K_n the number of "distinct" genes in $X^{(1,n)}$.
- Let $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ denote an additional sample
- $K_m^{(n)} := K_{n+m} - K_n$, the number of new species in $X^{(2,m)}$.
- ESTs info is typically reported in levels of expression: $\mathbf{N} = (N_1, \ldots, N_{K_n})$ where N_i denotes the frequency of gene i in the ESTs sample.

- Typically a cDNA library contains a large and unknown number of differentially expressed genes: potentially infinite

- Each X_i denotes an EST $\approx$ CCCCCGTCTCTTTAAAAATATAT...
- Let $X^{(1,n)} = (X_1, \ldots, X_n)$ denote a base sample of ESTs, and K_n the number of "distinct" genes in $X^{(1,n)}$.
- Let $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ denote an additional sample
- $K_m^{(n)} := K_{n+m} - K_n$, the number of new species in $X^{(2,m)}$.
- ESTs info is typically reported in levels of expression: $\mathbf{N} = (N_1, \ldots, N_{K_n})$ where N_i denotes the frequency of gene i in the ESTs sample.

- Typically a cDNA library contains a large and unknown number of differentially expressed genes: potentially infinite

- Each X_i denotes an EST $\approx$ CCCCCGTCTCTTTAAAAATATAT...
- Let $X^{(1,n)} = (X_1, \ldots, X_n)$ denote a base sample of ESTs, and K_n the number of "distinct" genes in $X^{(1,n)}$.
- Let $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ denote an additional sample
- $K_m^{(n)} := K_{n+m} - K_n$, the number of new species in $X^{(2,m)}$.
- ESTs info is typically reported in levels of expression: $\mathbf{N} = (N_1, \ldots, N_{K_n})$ where N_i denotes the frequency of gene i in the ESTs sample.

- Typically a cDNA library contains a large and unknown number of differentially expressed genes: potentially infinite

- Each X_i denotes an EST $\approx$ CCCCCGTCTCTTTAAAAATATAT...
- Let $X^{(1,n)} = (X_1, \ldots, X_n)$ denote a base sample of ESTs, and K_n the number of "distinct" genes in $X^{(1,n)}$.
- Let $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ denote an additional sample
- $K_m^{(n)} := K_{n+m} - K_n$, the number of new species in $X^{(2,m)}$.
- ESTs info is typically reported in levels of expression: $\mathbf{N} = (N_1, \ldots, N_{K_n})$ where N_i denotes the frequency of gene i in the ESTs sample.

- Typically a cDNA library contains a large and unknown number of differentially expressed genes: potentially infinite

- Each X_i denotes an EST $\approx$ CCCCCGTCTCTTTAAAAATATAT...
- Let $X^{(1,n)} = (X_1, \ldots, X_n)$ denote a base sample of ESTs, and K_n the number of "distinct" genes in $X^{(1,n)}$.
- Let $X^{(2,m)} = (X_{n+1}, \ldots, X_{n+m})$ denote an additional sample
- $K_m^{(n)} := K_{n+m} - K_n$, the number of new species in $X^{(2,m)}$.
- ESTs info is typically reported in levels of expression: $\mathbf{N} = (N_1, \ldots, N_{K_n})$ where N_i denotes the frequency of gene i in the ESTs sample.

- Also typically clustered into $R = (R_1, R_2, \ldots)$, where R_i indicates the total number of genes with expression level i in the sample.
- Example: EST survey from cDNA library of 0-3 mm buds of tomato flowers (Quackenbush et al., 2000)
  - $n = 2586$ ESTs with $j = 1825$ unique genes.
  - $r_i = 1434, 253, 71, 33, 11, 6, 2, 3, 1, 2, 2, 1, 1, 1, 2, 1, 1$ with $i \in \{1, 2, \ldots, 14\} \cup \{16, 23, 27\} \Longrightarrow$ 1434 genes appear once, 253 genes appear twice, etc.
- Given the $r_i's$ the $n_i's$ are given by,

$$n_1 = \cdots = n_{1434} = 1$$
$$n_{1435} = \cdots = n_{1687} = 2$$
$$n_{1688} = \cdots = n_{1758} = 3$$
$$\ldots\ldots$$

- R is a sufficient statistic for the transcript abundance in the cDNA library

- Also typically clustered into $R = (R_1, R_2, \ldots)$, where R_i indicates the total number of genes with expression level i in the sample.
- Example: EST survey from cDNA library of 0-3 mm buds of tomato flowers (Quackenbush et al., 2000)
  - $n = 2586$ ESTs with $j = 1825$ unique genes.
  - $r_i = 1434, 253, 71, 33, 11, 6, 2, 3, 1, 2, 2, 1, 1, 1, 2, 1, 1$ with $i \in \{1, 2, \ldots, 14\} \cup \{16, 23, 27\} \Longrightarrow$ 1434 genes appear once, 253 genes appear twice, etc.

- Given the $r_i's$ the $n_i's$ are given by,

$$
\begin{aligned}
n_1 &= \cdots = n_{1434} = 1 \\
n_{1435} &= \cdots = n_{1687} = 2 \\
n_{1688} &= \cdots = n_{1758} = 3 \\
&\ldots\ldots
\end{aligned}
$$

- R is a sufficient statistic for the transcript abundance in the cDNA library

Gibbs–type priors ($\mathscr{G}_\sigma$): (Gnedin and Pitman (2005))

$\tilde{P}$ is a Gibbs type prior of order $\sigma \in [0,1)$, (denoted by $\mathscr{G}_\sigma$), iff the predictive distribution is given by

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(n)}\right] = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma)\, \delta_{X_j^*}(A), \tag{1}$$

where $X_1^*, \ldots, X_k^*$ are the k distinct genes in $X^{(n)}$ and $n_j > 0$ their frequencies s.t. $\sum_{j=1}^{k} n_j = n$. $P_0 := \mathbb{E}(\tilde{P})$. $\{V_{n,k}\}_{n \geq 1, 1 \leq k \leq n}$ is a set of weights s.t.

$$V_{n,k} = (n - k\sigma) V_{n+1,k} + V_{n+1,k+1}, \qquad V_{1,1} = 1.$$

PROBLEM: Find explicit values for $V_{n,k}$'s

Gibbs–type priors ($\mathscr{G}_\sigma$): (Gnedin and Pitman (2005))

$\tilde{P}$ is a Gibbs type prior of order $\sigma \in [0,1)$, (denoted by $\mathscr{G}_\sigma$), iff the predictive distribution is given by

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(n)}\right] = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma)\, \delta_{X_j^*}(A), \tag{1}$$

where $X_1^*, \ldots, X_k^*$ are the k distinct genes in $X^{(n)}$ and $n_j > 0$ their frequencies s.t. $\sum_{j=1}^{k} n_j = n$. $P_0 := \mathbb{E}(\tilde{P})$. $\{V_{n,k}\}_{n \geq 1, 1 \leq k \leq n}$ is a set of weights s.t.

$$V_{n,k} = (n - k\sigma) V_{n+1,k} + V_{n+1,k+1}, \qquad V_{1,1} = 1.$$

PROBLEM: Find explicit values for $V_{n,k}$'s

The Gibbs–structure allows to look at the predictive distributions as the result of two steps:

(1) X_{n+1} is a new gene with probability

$$V_{n+1,j+1}/V_{n,j},$$

whereas it equals one of the "old" $\{X_1^*, \ldots, X_j^*\}$ with probability

$$1 - V_{n+1,j+1}/V_{n,j} = (n - j\sigma)V_{n+1,j}/V_{n,j}$$

$\Longrightarrow$ This step depends on n and j but not on the expression levels $(n_1, \ldots, n_j)$.

(2)

(i) Given X_{n+1} is new, it is independently sampled from P_0.
(ii) Given X_{n+1} is a tie, it coincides with X_i^* with probability

$$(n_i - \sigma)/(n - k\sigma).$$

The Gibbs–structure allows to look at the predictive distributions as the result of two steps:

(1) X_{n+1} is a new gene with probability

$$V_{n+1,j+1}/V_{n,j},$$

whereas it equals one of the "old" $\{X_1^*, \ldots, X_j^*\}$ with probability

$$1 - V_{n+1,j+1}/V_{n,j} = (n - j\sigma)V_{n+1,j}/V_{n,j}$$

$\Longrightarrow$ This step depends on n and j but not on the expression levels $(n_1, \ldots, n_j)$.

(2)

(i) Given X_{n+1} is new, it is independently sampled from P_0.
(ii) Given X_{n+1} is a tie, it coincides with X_i^* with probability

$$(n_i - \sigma)/(n - k\sigma).$$

- $\mathscr{G}_\sigma$ selects (a.s.) discrete distributions
- $\mathscr{G}_\sigma$ induces an infinite Gibbs partition i.e. an exchangeable random partition with, for every $j, n \geq 1$, *Exchangeable Partition Probability Function* (EPPF) of the form

$$\Pi_j^{(n)}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i-1},$$

where $(a)_n = a(a+1)\ldots(a+n-1)$ is ascending factorial coefficient. Prob. of obtaining j distinct values with freq. $n_1, \ldots, n_j$, in a sample of size n of $\mathscr{G}_\sigma$

- The (prior) dist. for the number of distinct observations K_n, is given by

$$\mathbb{P}[K_n = k] = \frac{V_{n,k}}{\sigma^k} \mathscr{C}(n, k; \sigma), \tag{2}$$

where $\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$ is the generalised factorial coeff See Charalambides (2005).

- $\mathscr{G}_\sigma$ selects (a.s.) discrete distributions
- $\mathscr{G}_\sigma$ induces an infinite Gibbs partition i.e. an exchangeable random partition with, for every $j, n \geq 1$, *Exchangeable Partition Probability Function* (EPPF) of the form

$$\Pi_j^{(n)}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i - 1},$$

Prob. of obtaining j distinct values with freq. $n_1, \ldots, n_j$, in a sample of size n of $\mathscr{G}_\sigma$

- The (prior) dist. for the number of distinct observations K_n, is given by

$$\mathbb{P}[K_n = k] = \frac{V_{n,k}}{\sigma^k} \mathscr{C}(n, k; \sigma), \tag{2}$$

where $\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$ is the generalised factorial coeff See Charalambides (2005).

- $\mathscr{G}_\sigma$ selects (a.s.) discrete distributions
- $\mathscr{G}_\sigma$ induces an infinite Gibbs partition i.e. an exchangeable random partition with, for every $j, n \geq 1$, *Exchangeable Partition Probability Function* (EPPF) of the form

$$\Pi_j^{(n)}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i-1},$$

Prob. of obtaining j distinct values with freq. $n_1, \ldots, n_j$, in a sample of size n of $\mathscr{G}_\sigma$

- The (prior) dist. for the number of distinct observations K_n, is given by

$$\mathbb{P}[K_n = k] = \frac{V_{n,k}}{\sigma^k} \mathscr{C}(n, k; \sigma), \tag{2}$$

where $\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$ is the generalised factorial coeff See Charalambides (2005).

Dirichlet process (Ferguson, 1973)

The Dirichlet process $\mathscr{D}_{(\theta,P_0)}$ can be seen as $\mathscr{G}_{\sigma=0}$ and $V_{n,k} = \frac{\theta^k}{(\theta)_n}$

- The resulting EPPF (Ewens (1972) and Antoniak (1974))

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\theta^k}{(\theta)_n}\prod_{j=1}^{k}(n_j-1)! \tag{3}$$

- The (prior) dist. for K_n is given by

$$\mathbb{P}[K_n = k] = \frac{\theta^k}{(\theta)_n}|s(n,k)| \tag{4}$$

where $s(n,k)$ for $n \geq k \geq 1$ denote the Stirling numbers of the first type.

Dirichlet process (Ferguson, 1973)

The Dirichlet process $\mathscr{D}_{(\theta,P_0)}$ can be seen as $\mathscr{G}_{\sigma=0}$ and $V_{n,k} = \frac{\theta^k}{(\theta)_n}$

- The resulting EPPF (Ewens (1972) and Antoniak (1974))

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\theta^k}{(\theta)_n}\prod_{j=1}^{k}(n_j-1)! \tag{3}$$

- The (prior) dist. for K_n is given by

$$\mathbb{P}[K_n = k] = \frac{\theta^k}{(\theta)_n}\,|s(n,k)| \tag{4}$$

where $s(n,k)$ for $n \geq k \geq 1$ denote the Stirling numbers of the first type.

Two parameter Poisson-Dirichlet (Pitman, 1995)

The two parameter Poisson-Dirichlet process with two parameters $\mathscr{PD}_{(\theta,\sigma,P_0)}$ can be seen as $\mathscr{G}_\sigma$ with $V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}$.

- Predictive distribution

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(1,n)}\right] = \frac{\theta + k\sigma}{\theta + n} P_0(A) + \frac{1}{\theta + n} \sum_{j=1}^{k} (n_j - \sigma)\delta_{X_j}(A).$$

- Prior distribution for K_n

$$\mathbb{P}[K_n = k] = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma^k (\theta + 1)_{n-1}} \mathscr{C}(n, k; \sigma).$$

where

$$\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$$

Two parameter Poisson-Dirichlet (Pitman, 1995)

The two parameter Poisson-Dirichlet process with two parameters $\mathscr{PD}_{(\theta,\sigma,P_0)}$ can be seen as $\mathscr{G}_\sigma$ with $V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}$.

- Predictive distribution

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(1,n)}\right] = \frac{\theta + k\sigma}{\theta + n} P_0(A) + \frac{1}{\theta + n} \sum_{j=1}^{k} (n_j - \sigma)\delta_{X_j}(A).$$

- Prior distribution for K_n

$$\mathbb{P}[K_n = k] = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma^k (\theta + 1)_{n-1}} \mathscr{C}(n, k; \sigma).$$

where

$$\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$$

Prediction of new species

- Let $(X_n)_{n\geq 1}$ be an exchangeable sequence modulated by $\mathscr{G}_\sigma$
- (a) Given a base sample $X^{(1,n)}$, how many new species are expected in the additional sample $X^{(2,m)}$?

"Posterior" distribution on the number of new genes

If $X^{(1,n)}$ contains j distinct genes, then

$$\mathbb{P}\left[K_m^{(n)} = k \middle| X^{(1,n)}\right] = \mathbb{P}\left[K_m^{(n)} = k \middle| K_n = j\right]$$

$$= \frac{V_{n+m,j+k}}{V_{n,j}} \frac{(-1)^{m-k}}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma), \qquad (5)$$

where $\mathscr{C}(m,k;\sigma,-n+j\sigma) = \frac{(-1)^n}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (-\sigma j - \gamma)_n$ denotes the Non-central generalised factorial coefficient. K_n is enough to predict the number of new genes.

Prediction of new species

- Let $(X_n)_{n\geq 1}$ be an exchangeable sequence modulated by $\mathscr{G}_\sigma$
- (a) Given a base sample $X^{(1,n)}$, how many new species are expected in the additional sample $X^{(2,m)}$?

"Posterior" distribution on the number of new genes

If $X^{(1,n)}$ contains j distinct genes, then

$$\mathbb{P}\left[K_m^{(n)} = k \middle| X^{(1,n)}\right] = \mathbb{P}\left[K_m^{(n)} = k \middle| K_n = j\right]$$

$$= \frac{V_{n+m,j+k}}{V_{n,j}} \frac{(-1)^{m-k}}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma), \qquad (5)$$

where $\mathscr{C}(m,k;\sigma,-n+j\sigma) = \frac{(-1)^n}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (-\sigma j - \gamma)_n$ *denotes the Non-central generalised factorial coefficient.* K_n is enough to predict the number of new genes.

Prediction of new species

- Let $(X_n)_{n\geq 1}$ be an exchangeable sequence modulated by $\mathscr{G}_\sigma$
- (a) Given a base sample $X^{(1,n)}$, how many new species are expected in the additional sample $X^{(2,m)}$?

"Posterior" distribution on the number of new genes

If $X^{(1,n)}$ contains j distinct genes, then

$$\mathbb{P}\left[K_m^{(n)} = k \middle| X^{(1,n)}\right] = \mathbb{P}\left[K_m^{(n)} = k \middle| K_n = j\right]$$
$$= \frac{V_{n+m,j+k}}{V_{n,j}} \frac{(-1)^{m-k}}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma), \qquad (5)$$

where $\mathscr{C}(m,k;\sigma,-n+j\sigma) = \frac{(-1)^n}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (-\sigma j - \gamma)_n$ *denotes the Non-central generalised factorial coefficient.* K_n is enough to predict the number of new genes.

Prediction of new species

- Let $(X_n)_{n\geq 1}$ be an exchangeable sequence modulated by $\mathscr{G}_\sigma$
- (a) Given a base sample $X^{(1,n)}$, how many new species are expected in the additional sample $X^{(2,m)}$?

"Posterior" distribution on the number of new genes

If $X^{(1,n)}$ contains j distinct genes, then

$$\mathbb{P}\left[K_m^{(n)} = k \middle| X^{(1,n)}\right] = \mathbb{P}\left[K_m^{(n)} = k \middle| K_n = j\right]$$
$$= \frac{V_{n+m,j+k}}{V_{n,j}} \frac{(-1)^{m-k}}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma), \qquad (5)$$

where $\mathscr{C}(m,k;\sigma,-n+j\sigma) = \frac{(-1)^n}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (-\sigma j - \gamma)_n$ *denotes the Non-central generalised factorial coefficient.* K_n is enough to predict the number of new genes.

Predictive distributions of interest

- Case: Dirichlet process $\mathscr{D}_{(\theta,P_0)}$

$$\mathbb{P}\left[K_m^{(n)} = k \,\middle|\, X_j^{(1,n)}\right] = \frac{\theta^k(\theta)_n}{(\theta)_{n+m}} \sum_{l=k}^{m} \binom{m}{l} |s(l,k)|\,(n)_{m-l}.$$

  - The probability of discovering k new genes doesn't depend on j observed genes $\Rightarrow$ undesirable property!
  - Characteristic property of $\mathscr{D}_{(\theta,P_0)}$ among the class of Gibbs priors $\mathscr{G}_\sigma$

- Caso: Two parameter Poisson-Dirichlet process $\mathscr{PD}_{(\theta,\sigma,P_0)}$

$$\mathbb{P}[K_m^{(n)} = k \mid X_j^{(1,n)}] \tag{6}$$

$$= \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{(-1)^{m-k}\prod_{i=j}^{j+k-1}(\theta+i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma)$$

Predictive distributions of interest

- Case: Dirichlet process $\mathscr{D}_{(\theta,P_0)}$

$$\mathbb{P}\left[K_m^{(n)} = k \,\middle|\, X_j^{(1,n)}\right] = \frac{\theta^k (\theta)_n}{(\theta)_{n+m}} \sum_{l=k}^{m} \binom{m}{l} |s(l,k)| \, (n)_{m-l}.$$

  - The probability of discovering k new genes doesn't depend on j observed genes $\Rightarrow$ undesirable property!
  - Characteristic property of $\mathscr{D}_{(\theta,P_0)}$ among the class of Gibbs priors $\mathscr{G}_\sigma$

- Caso: Two parameter Poisson-Dirichlet process $\mathscr{PD}_{(\theta,\sigma,P_0)}$

$$\mathbb{P}[K_m^{(n)} = k \mid X_j^{(1,n)}] \tag{6}$$

$$= \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{(-1)^{m-k} \prod_{i=j}^{j+k-1} (\theta + i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma)$$

Predictive distributions of interest

- Case: Dirichlet process $\mathscr{D}_{(\theta,P_0)}$

$$\mathbb{P}\left[K_m^{(n)} = k \,\middle|\, X_j^{(1,n)}\right] = \frac{\theta^k (\theta)_n}{(\theta)_{n+m}} \sum_{l=k}^{m} \binom{m}{l} |s(l,k)| \, (n)_{m-l}.$$

  - The probability of discovering k new genes doesn't depend on j observed genes $\Rightarrow$ undesirable property!
  - Characteristic property of $\mathscr{D}_{(\theta,P_0)}$ among the class of Gibbs priors $\mathscr{G}_\sigma$

- Caso: Two parameter Poisson-Dirichlet process $\mathscr{PD}_{(\theta,\sigma,P_0)}$

$$\mathbb{P}[K_m^{(n)} = k \mid X_j^{(1,n)}] \tag{6}$$

$$= \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{(-1)^{m-k} \prod_{i=j}^{j+k-1} (\theta + i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma)$$

Predictive distributions of interest

- Case: Dirichlet process $\mathscr{D}_{(\theta,P_0)}$

$$\mathbb{P}\left[K_m^{(n)} = k \,\middle|\, X_j^{(1,n)}\right] = \frac{\theta^k(\theta)_n}{(\theta)_{n+m}} \sum_{l=k}^{m} \binom{m}{l} |s(l,k)|\,(n)_{m-l}.$$

  - The probability of discovering k new genes doesn't depend on j observed genes $\Rightarrow$ undesirable property!
  - Characteristic property of $\mathscr{D}_{(\theta,P_0)}$ among the class of Gibbs priors $\mathscr{G}_\sigma$

- Caso: Two parameter Poisson-Dirichlet process $\mathscr{PD}_{(\theta,\sigma,P_0)}$

$$\mathbb{P}[K_m^{(n)} = k \mid X_j^{(1,n)}] \tag{6}$$

$$= \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{(-1)^{m-k}\prod_{i=j}^{j+k-1}(\theta+i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma)$$

(b) Which is the probability of observing a new gene in the $(n+m+1)$-th sampling, independently of what happened in the additional sample, i.e. without observing $X^{(2,m)}$?

- Assume we observe $X^{(1,n)}$ with j genes and $X^{(2,m)}$ with k new genes: the discovery probability is given by

$$\mathbb{P}\left[K_1^{(n+m)} = 1 \middle| X^{(1,n)}, X^{(2,m)}\right] = \mathbb{P}\left[K_1^{(n+m)} = 1 \middle| K_n = j, K_m^{(n)} = k\right].$$

- Without observing $X^{(2,m)}$, we estimate the random probability

$$D_m^{(n:j)} := \mathbb{P}\left[K_1^{(n+m)} = 1 \middle| K_n = j, K_m^{(n)}\right]$$

The randomness in $D_m^{(n:j)}$ is due to the randomness in $K_m^{(n)}$.

(b) Which is the probability of observing a new gene in the $(n+m+1)$-th sampling, independently of what happened in the additional sample, i.e. without observing $X^{(2,m)}$?

- Assume we observe $X^{(1,n)}$ with j genes and $X^{(2,m)}$ with k new genes: the discovery probability is given by

$$\mathbb{P}\left[K_1^{(n+m)} = 1 \middle| X^{(1,n)}, X^{(2,m)}\right] = \mathbb{P}\left[K_1^{(n+m)} = 1 \middle| K_n = j, K_m^{(n)} = k\right].$$

- Without observing $X^{(2,m)}$, we estimate the random probability

$$D_m^{(n:j)} := \mathbb{P}\left[K_1^{(n+m)} = 1 \middle| K_n = j, K_m^{(n)}\right]$$

The randomness in $D_m^{(n:j)}$ is due to the randomness in $K_m^{(n)}$.

- If $m = 0$, then the probability of observing a new gene at the $(n + m + 1)$–th draw is simply the one–step predictive distribution

$$\hat{D}^{(n)} := \mathbb{P}\left[K_1^{(n)} = 1 \middle| K_n = j\right] = \frac{V_{n+1,j}}{V_{n,j}}.$$

Correspondingly, the sample coverage, defined as the proportion of unique genes present in the sample, is given by

$$\hat{C}^{(n)} := 1 - \hat{D}^{(n)} = 1 - \frac{V_{n+1,j}}{V_{n,j}}$$

$\Longrightarrow$ Bayesian nonparametric analog of the Turing estimator

The Bayes estimator, under quadratic loss, of the probability of observing a new gene in the $(n+m+1)$–th draw, conditional on $X^{(1,n)} = j$, is given by

$$\hat{D}_m^{(n:j)} = \sum_{k=0}^{m} \frac{V_{n+m+1,j+k+1}}{V_{n,j}} \frac{(-1)^{m-k}}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma).$$

$\Longrightarrow$ Bayesian version of the Good–Toulmin estimator.

- Dirichlet process: $\mathscr{D}(\theta, P_0)$:

$$\hat{D}_m^{(n:j)} = \frac{\theta}{(\theta+n)_{m+1}} \sum_{k=0}^{m} \theta^k \sum_{l=k}^{m} \binom{m}{l} |s(l,k)| \, (n)_{m-l} \qquad (7)$$

- Two parameter Poisson-Dirichlet process: $\mathscr{PD}(\theta, \sigma, P_0)$:

$$\hat{D}_m^{(n:j)} = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{k=0}^{m} \frac{(-1)^{m-k} \prod_{i=j}^{j+k} (\theta + i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma). \qquad (8)$$

Prior specification

We adopt the two parameter PD process as prior and elicit (θ, σ) in the following ways:

- Empirical Bayes specification: σ and θ are fixed so to maximize the EPPF corresponding to the observed sample $(j, n_1, \ldots, n_j)$, i.e.

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{i=1}^{j}(1-\sigma)_{n_i-1},$$

- Bayesian specification: a prior q is placed on σ and θ such that

$$\mathbb{E}[K_n] = \int_0^{\infty}\int_0^1 \left(\frac{(\theta+\sigma)_n}{\sigma(\theta+1)_{n-1}} - \frac{\theta}{\sigma}\right) q(\mathrm{d}\sigma, \mathrm{d}\theta) = j$$

Prior specification

We adopt the two parameter PD process as prior and elicit (θ, σ) in the following ways:

- Empirical Bayes specification: σ and θ are fixed so to maximize the EPPF corresponding to the observed sample $(j, n_1, \ldots, n_j)$, i.e.

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{i=1}^{j}(1-\sigma)_{n_i-1},$$

- Bayesian specification: a prior q is placed on σ and θ such that

$$\mathbb{E}[K_n] = \int_0^{\infty}\int_0^{1}\left(\frac{(\theta+\sigma)_n}{\sigma(\theta+1)_{n-1}} - \frac{\theta}{\sigma}\right) q(\mathrm{d}\sigma, \mathrm{d}\theta) = j$$

Prior specification

We adopt the two parameter PD process as prior and elicit (θ, σ) in the following ways:

- Empirical Bayes specification: σ and θ are fixed so to maximize the EPPF corresponding to the observed sample $(j, n_1, \ldots, n_j)$, i.e.

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{i=1}^{j}(1-\sigma)_{n_i-1},$$

- Bayesian specification: a prior q is placed on σ and θ such that

$$\mathbb{E}[K_n] = \int_0^{\infty}\int_0^1 \left(\frac{(\theta+\sigma)_n}{\sigma(\theta+1)_{n-1}} - \frac{\theta}{\sigma}\right) q(\mathrm{d}\sigma, \mathrm{d}\theta) = j$$

- Empirical Bayes $\Rightarrow (\hat{\sigma}, \hat{\theta}) = (0.612, 741)$

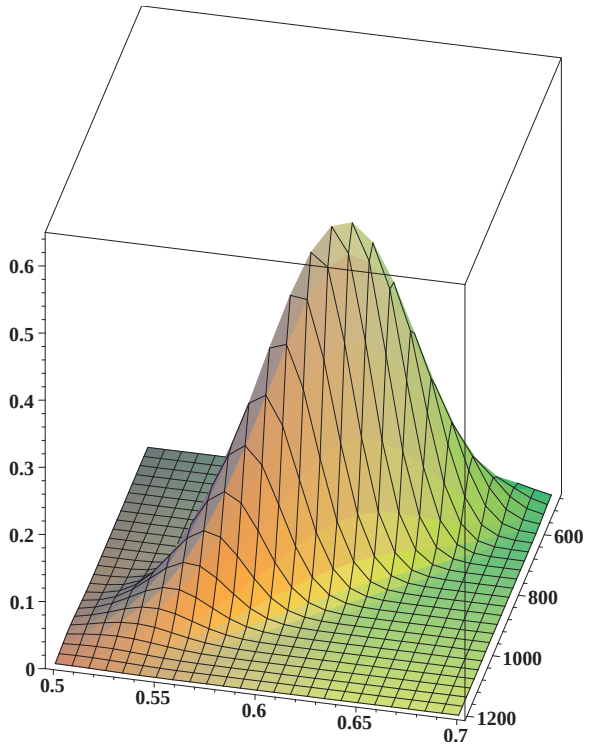
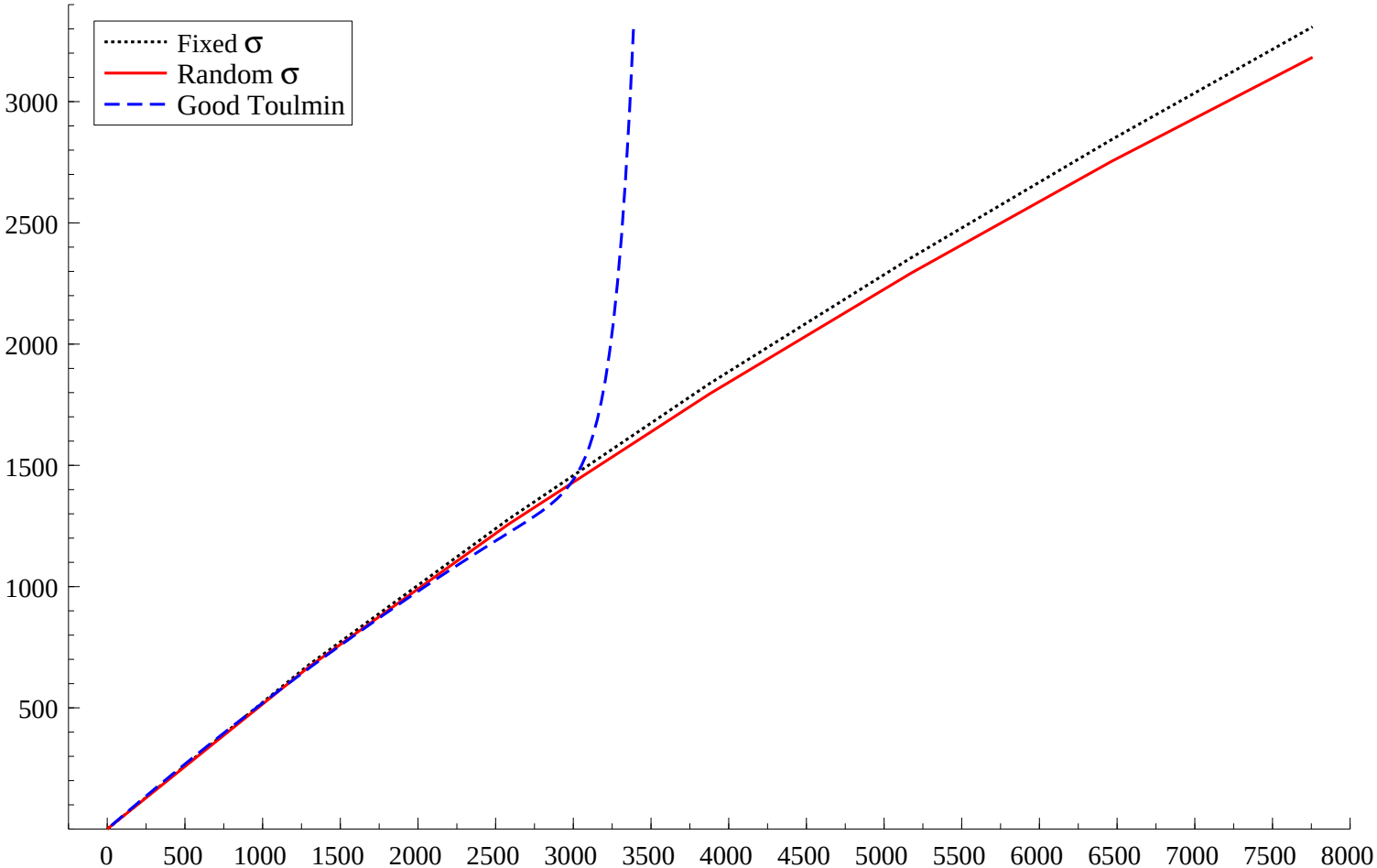
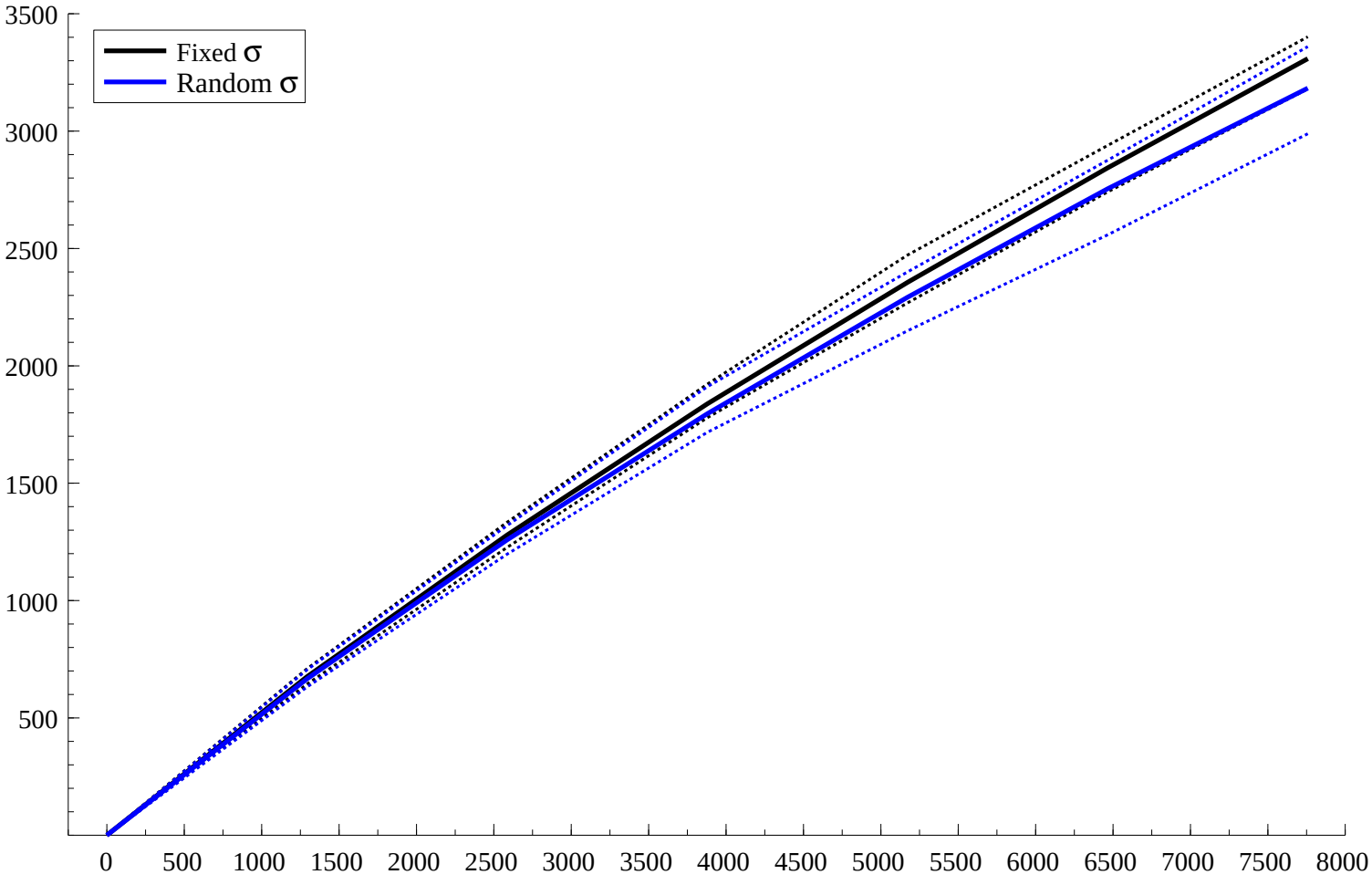


Figure: Distribution of K_n and the n_i's (PD) as (θ, σ) vary .

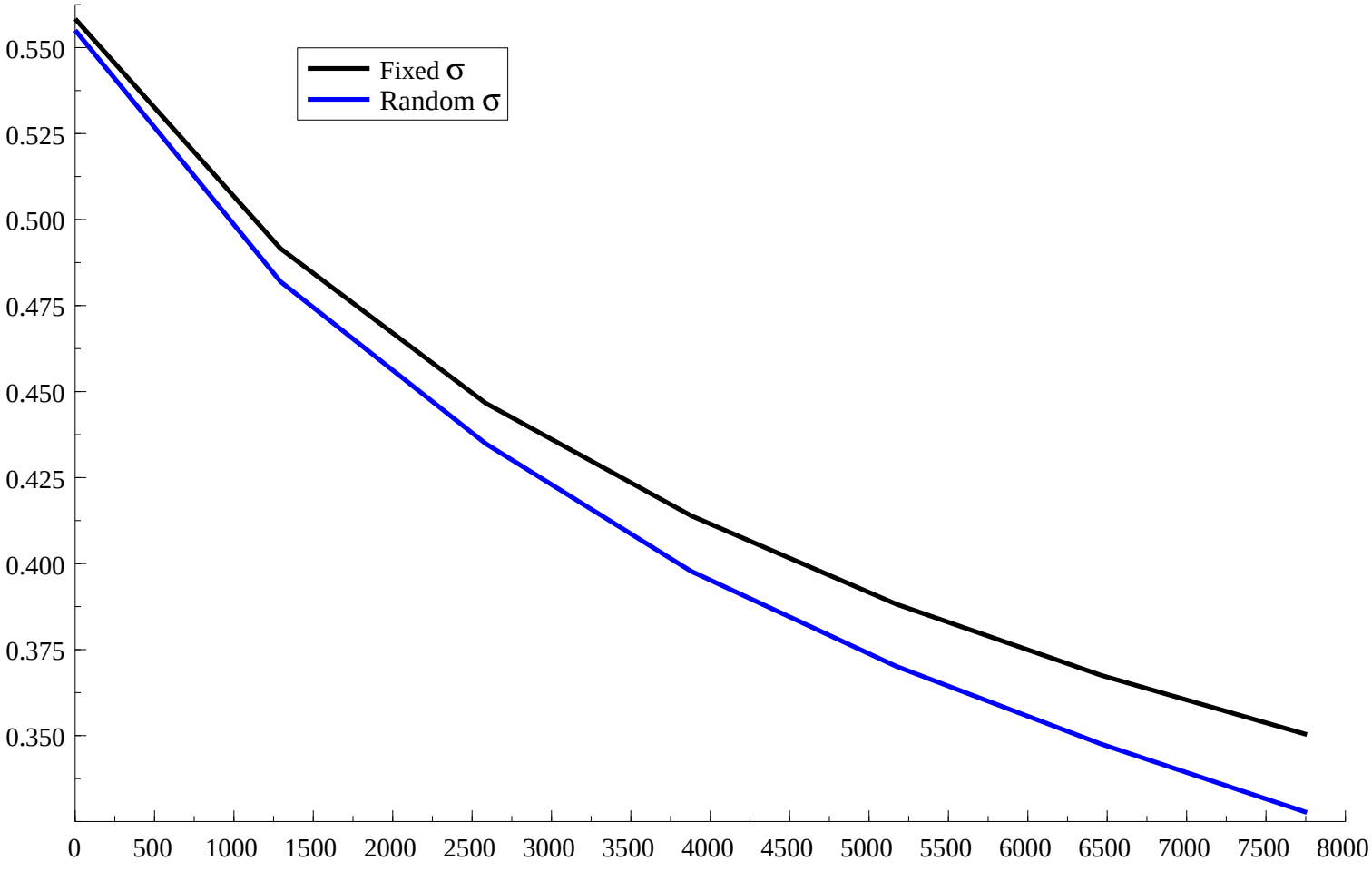
Expected number of new genes (PD) and Good-Toulmin



Expected number of new genes



Discovery probability



ESTs clustering errors

- Denote by $C_i = j$ the gene membership variable, e.g. $X_i = X_j^*$ iff $C_i = j$
- The ESTs sequencing is exposed to some error, namely the tag assigned to gene X_j might not fully identify the gene structure. Essentially this result in bad clustering

Error I When two sequences should identify the same gene but in the clustering they don't!,e.g. $X_i = X_j^* \Rightarrow C_i \neq j$

Error II When two sequences are different but they are clustered into the same group,e.g $X_i \neq X_j^* \Rightarrow C_i = j$

- Error I 10 times higher for sequences from the 5' than sequences form the 3' (30% vs 3%)
- Error II $< 1.5\%$ for nucleotides sequenced from both ends of the molecule, 5' or 3'

- Error I $\Rightarrow$ multiple unigenes can represent the same gene, hence more unique genes than actually present are recorded
- Wang et al. 2004 proposed a method for correcting the insufficient overlap error (ISO error).

- Is the clustering structure affected by the ISO correction or does it only reduce "proportionally" the estimates?
  - Interesting to decide whether the libraries with and without ISO correction have the same clustering structurure
  - For deciding whether is worth to merge/split libraries
  - etc...

Gibbs prior ($\mathscr{G}_\sigma$)

- Within a Gibbs prior model, the clustering structure is dictated by σ, whereas $V_{n,k}$ (or simply θ in the two parameter PD case) affects only the rate at which new clusters are produced.
- Consider the two parameter PD model with Bayesian specification:
  - We want to test

$$H_0 : (\sigma_1, \theta_1) = (\sigma_2, \theta_2) \qquad \text{vs} \qquad H_1 : (\sigma_1, \theta_1) \neq (\sigma_2, \theta_2)$$

by computing the Bayes factor

$$\mathrm{BF}_{01} = \frac{\int_0^1 \int_0^\infty \mathrm{EPPF}^{(1)}_{k,n}(\mathbf{n}; \sigma, \theta)\, \mathrm{EPPF}^{(2)}_{k,n}(\mathbf{n}; \sigma, \theta)\, \pi_\sigma(\mathrm{d}\sigma)\pi_\theta(\mathrm{d}\theta)}{\prod_{j=1}^2 \int_0^1 \int_0^\infty \mathrm{EPPF}^{(j)}_{k,n}(\mathbf{n}; \sigma, \theta)\, \pi_{j,\sigma}(\mathrm{d}\sigma)\, \pi_{j,\theta}(\mathrm{d}\theta)} \tag{9}$$

ISO correction to arabidopsis thaliana cDNA libraries

- ISO correction applied to two different libraries ABGR and root (Wang *et. al* 2004)

Library	n	j
ABGR before ISO	5811	3116
ABGR after ISO	5812	2883
Root before ISO	5880	3368
Root after ISO	5891	3126

$-2 \ln \mathrm{BF}_{ABGR} = -11.162$ $\quad$ $-2 \ln \mathrm{BF}_{Root} = -11.105$

additional sample	$\hat{K}_m^{(n)}$ for ABGR		$\hat{K}_m^{(n)}$ for Root	
	before ISO	after ISO	before ISO	after ISO
1000	383	334	418	363
2000	1765	1508	1935	1652
10000	3267	2758	3591	3028
30000	5156	6640	8985	7306

Merging libraries

- Do different libraries from the same organism exhibit the same clustering structure? Is there some gain in splitting them?

Library	n	j
ABGR	5812	2883
Root	5891	3126
ABGR+Root	11529	5243
Silique	12330	5093
Flower bud	5503	2546
Silique+Flower bud	17784	6595

- ABGR+Root:
  - $-2\ln BF = -37.564$ against ABGR;
  - $-2\ln BF = -10.857$ against Root.
- Silique+Flower bud:
  - $-2\ln BF = 1.477$ against Silique;
  - $-2\ln BF = -159.899$ against Flower bud.

The reason is that they share too few genes!

Merging libraries

- Do different libraries from the same organism exhibit the same clustering structure? Is there some gain in splitting them?

Library	n	j
ABGR	5812	2883
Root	5891	3126
ABGR+Root	11529	5243
Silique	12330	5093
Flower bud	5503	2546
Silique+Flower bud	17784	6595

- ABGR+Root:
  - $-2 \ln \mathrm{BF} = -37.564$ against ABGR;
  - $-2 \ln \mathrm{BF} = -10.857$ against Root.
- Silique+Flower bud:
  - $-2 \ln \mathrm{BF} = 1.477$ against Silique;
  - $-2 \ln \mathrm{BF} = -159.899$ against Flower bud.

The reason is that they share too few genes!

Merging libraries

- Do different libraries from the same organism exhibit the same clustering structure? Is there some gain in splitting them?

Library	n	j
ABGR	5812	2883
Root	5891	3126
ABGR+Root	11529	5243
Silique	12330	5093
Flower bud	5503	2546
Silique+Flower bud	17784	6595

- ABGR+Root:
  - $-2 \ln \mathrm{BF} = -37.564$ against ABGR;
  - $-2 \ln \mathrm{BF} = -10.857$ against Root.
- Silique+Flower bud:
  - $-2 \ln \mathrm{BF} = 1.477$ against Silique;
  - $-2 \ln \mathrm{BF} = -159.899$ against Flower bud.

The reason is that they share too few genes!

Thanks!