Supplementary Materials for:

An Optimization-based Framework for the Transformation of Incomplete Biological Knowledge into a Probabilistic Structure and its Application to the Utilization of Gene/Protein Signaling Pathways in Discrete Phenotype Classification

Mohammad Shahrokh Esfahani and Edward R Dougherty

APPENDIX A

DIRICHLET DISTRIBUTION: DEFINITION AND PROPERTIES

The ratio \( \prod_{i=1}^{b} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_0)} \), called the multinomial Beta function, is denoted by \( B(\alpha) \) or \( B([\alpha_1, ..., \alpha_b]) \).

Five major properties of the Dirichlet distribution frequently used in this paper are listed below (for properties P1 – P3, refer to [29]-[30])

P1. If \( [p_1, p_2, ..., p_b] \sim D(\alpha_1, \alpha_2, ..., \alpha_b) \) and \( r_1, ..., r_l \) are integers such that \( 0 < r_1 < ... < r_l = b \), then

\[
(\sum_{i=1}^{r_1} p_i, \sum_{i=r_1+1}^{r_2} p_i, ..., \sum_{i=r_{l-1}+1}^{r_l} p_i) \sim D(\sum_{i=1}^{r_1} \alpha_i, \sum_{i=r_1+1}^{r_2} \alpha_i, ..., \sum_{i=r_{l-1}+1}^{r_l} \alpha_i).
\] (A.1)

P2. Under the assumption in P1, each \( p_i \) is (marginally) distributed as

\[
p_i \sim Beta(\alpha_i, \alpha_0 - \alpha_i).
\] (A.2)

P3. If \( [p_1, p_2, ..., p_b] \sim D(\alpha_1, \alpha_2, ..., \alpha_b) \), then for the first and the second moments,

\[
E[p_i] = \frac{\alpha_i}{\alpha_0}
\]

\[
E[p_i^2] = \frac{\alpha_i(\alpha_i + 1)}{\alpha_0(\alpha_0)}.
\]

P4. If the random vector \( p \) is distributed according to the Dirichlet distribution \( D(\alpha) \), where \( \alpha_0 = \sum_i \alpha_i \), then \( [32] \)

\[
E[\log p_k] = \psi(\alpha_k) - \psi(\alpha_0),
\]

where \( \psi \) is the digamma function.

Using the properties above, we prove two fundamental lemmas frequently used in our analysis.
Lemma 5. If \( [p_1, p_2, \ldots, p_b] \sim D(\alpha_1, \alpha_2, \ldots, \alpha_b) \), then for any Lebesgue-measurable function \( g : S_{b-1} \to \mathbb{R} \),

\[
E_p[p_i g(p)] = \frac{\alpha_i}{\sum_{k=1}^b \alpha_k} E_p'[g(p')],
\]

in which

\[
p' \sim D(\alpha'_1, \alpha'_2, \ldots, \alpha'_b); \alpha'_i = \alpha_i + 1, \alpha'_k = \alpha_k, k \neq i.
\]

Proof. Without loss of generality, we prove the property for \( i = 1 \). Expanding the expectation yields

\[
E_p[p_1 \log g(p)] = \int p_1 g(p) \frac{\Gamma(\sum_{k=1}^b \alpha_k)}{\prod_{k=1}^b \Gamma(\alpha_k)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_b^{\alpha_b-1} dp
\]

\[
= \frac{\Gamma(\alpha_1+1) \prod_{k=2}^b \Gamma(\alpha_k) \Gamma(\sum_{k=1}^b \alpha_k)}{\Gamma(\alpha_1 + \sum_{k=2}^b \alpha_k) \prod_{k=1}^b \Gamma(\alpha_k)} \int g(p) \frac{\Gamma(\alpha_1 + 1 + \sum_{k=2}^b \alpha_k)}{\Gamma(\alpha_1 + 1) \prod_{k=2}^b \Gamma(\alpha_k)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_b^{\alpha_b-1} dp
\]

\[
= \frac{\alpha_1 \prod_{i=2}^b \Gamma(\alpha_i) \Gamma(\sum_{k=1}^b \alpha_k)}{(\sum_{k=2}^b \alpha_k)(\sum_{k=1}^b \alpha_k) \prod_{k=1}^b \Gamma(\alpha_k)} E_p'[g(p')]; p' \sim D(\alpha_1 + 1, \alpha_2, \alpha_3),
\]

where in the last equality we used the fact that \( \Gamma(x + 1) = x\Gamma(x) \).

\[\square\]

A.1 Proof of Lemma 1

Define \( C = \mathcal{X}\setminus(\bar{A} \cup \bar{B}) \). From property P1 we have

\[
\left( \sum_{i \in A} p_i, \sum_{i \in B} p_i, \sum_{i \in C} p_i \right) \sim D\left( \sum_{i \in A} \alpha_i, \sum_{i \in B} \alpha_i, \sum_{i \in C} \alpha_i \right).
\]

Hence, Lemma 5 indicates that

\[
E_p\left[ \frac{\sum_{i \in A} p_i}{\sum_{i \in A} p_i + \sum_{j \in B} p_j} \right] = \frac{\sum_{i \in A} p_i}{\sum_{i \in A} \alpha_i + \sum_{i \in B} \alpha_i + \sum_{i \in C} \alpha_i} E_p'[\frac{1}{\sum_{i \in A} p_i + \sum_{j \in B} p_j}],
\]

in which

\[
\left( \sum_{i \in A} p_i', \sum_{i \in B} p_i', \sum_{i \in C} p_i' \right) \sim D(1 + \sum_{i \in A} \alpha_i, \sum_{i \in B} \alpha_i, \sum_{i \in C} \alpha_i).
\]

Now, suppose \( (u, v, w) \sim D(\alpha_u, \alpha_v, \alpha_w) \), where \( u + v > 1 \). Then, since \( (u + v, w) \sim D(\alpha_u + \alpha_v, \alpha_w) \) (i.e. Beta-distributed), we may write \( E[\frac{1}{u+v}] = E[\frac{1}{z}] \), where \( (z, w) \sim D(\alpha_u + \alpha_v, \alpha_w) \). Finally we write

\[
E[\frac{1}{u+v}] = \frac{\Gamma(\alpha_u + \alpha_v + \alpha_w)}{\Gamma(\alpha_u + \alpha_v) \Gamma(\alpha_w)} \int \frac{1}{z} \frac{1}{\alpha_u + \alpha_v + \alpha_w - 1} dw = \frac{\Gamma(\alpha_u + \alpha_v + \alpha_w)}{\Gamma(\alpha_u + \alpha_v) \Gamma(\alpha_w)} \int \frac{1}{\alpha_u + \alpha_v + \alpha_w - 1} dw
\]

\[
= \frac{\Gamma(\alpha_u + \alpha_v + \alpha_w) \Gamma(\alpha_u + \alpha_w - 1) \Gamma(\alpha_w)}{\Gamma(\alpha_u + \alpha_v) \Gamma(\alpha_w)} = \frac{\alpha_u + \alpha_v + \alpha_w - 1}{\alpha_u + \alpha_v - 1}.
\]

3. \( S_{b-1} \) denotes the unit simplex in the two-dimensional Euclidean space.
According to equation (A.6), \( \alpha_u = 1 + \sum_{i \in A} \alpha_i \), \( \alpha_v = \sum_{i \in B} \alpha_i \), and \( \alpha_w = \sum_{i \in C} \alpha_i \). Combining equations (A.5) and (A.7), the proof for the first moment is finished.

For the variance, we find the second moment and then equation (18) would be the direct result of combining the first two moments. Writing the second moment, we have

\[
E_p\left[\sum_{i \in A} p_i / \left(\sum_{i \in A} p_i + \sum_{j \in B} p_j\right)^2\right] = \frac{\sum_{i \in A} \alpha_i (\sum_{i \in A} \alpha_i + 1)}{\left(\sum_{i \in A} \alpha_i \right) \left(\sum_{i \in A} \alpha_i + \sum_{j \in B} \alpha_j + 1\right)}
\]

(A.8)

where \( p'' \sim D(2 + \sum_{i \in A} \alpha_i, \sum_{i \in B} \alpha_i, \sum_{i \in C} \alpha_i) \). In equation (A.8), the equality comes from applying Lemma 5 twice. Then, similar to the approach leading to equation (A.7), we obtain

\[
E_p\left[\sum_{i \in A} p_i / \left(\sum_{i \in A} p_i + \sum_{j \in B} p_j\right)^2\right] = \frac{\sum_{i \in A} \alpha_i (\sum_{i \in A} \alpha_i + 1)}{\left(\sum_{i \in A} \alpha_i + \sum_{j \in B} \alpha_j \right) \left(\sum_{i \in A} \alpha_i + \sum_{j \in B} \alpha_j + 1\right)}
\]

A.2 Proof of Lemma 2

First define \( B_{k+y}^{0,0} \), \( k \in \{1, \ldots, 2^M\} \) and \( y \in \{0, 1\} \), as in the Lemma. Moreover, we use \( Z_{B_{k+y}^{0,0}} \) to denote \( \sum_{i \in B_{k+y}^{0,0}} u_i \). Then, based on the assumptions, we have

\[
(Z_{B_{k+y}^{0,0}}, Z_{B_{k+y}^{0,1}}, 1 - Z_{B_{k+y}^{0,0}} - Z_{B_{k+y}^{0,1}}) \sim Mult\left(\sum_{i \in B_{k+y}^{0,0}} p_i, \sum_{i \in B_{k+y}^{0,1}} p_i, 1 \right)
\]

Now, we expand the expected conditional entropy as follows:

\[
E[H|Z_{A_0}, \ldots, Z_{A_M}] = \sum_{k=1}^{2^M} \sum_{y=0}^1 E[\Pr(Z_{A_0} = y, f_{dec}(Z_{A_1}, \ldots, Z_{A_M}) = k) \log \Pr(Z_{A_0} = y, f_{dec}(Z_{A_1}, \ldots, Z_{A_M}) = k)]
\]

\[-E[\Pr(Z_{A_0} = y, f_{dec}(Z_{A_1}, \ldots, Z_{A_M}) = k) \log \Pr(f_{dec}(Z_{A_1}, \ldots, Z_{A_M}) = k)],
\]

(A.9)

where \( f_{dec}(Z_{A_1}, \ldots, Z_{A_M}) \) is a function mapping the binary-valued vector \( (Z_{A_1}, \ldots, Z_{A_M}) \) to its corresponding decimal number. Then one may write

\[
(Z_{A_0} = y, f_{dec}(Z_{A_1}, \ldots, Z_{A_M}) = k) \equiv p Z_{B_{k+y}^{0,0}}.
\]

Hence, equation (A.9) can be rewritten as

\[
E[H|Z_{A_0}, \ldots, Z_{A_M}] = \sum_{k=1}^{2^M} \sum_{y=0}^1 E[\log(\sum_{i \in B_{k+y}^{0,0}} p_i)]
\]

\[-E[\log(\sum_{i \in B_{k+y}^{0,0}} p_i) + \sum_{i \in B_{k+y}^{0,1}} p_i],
\]

(A.10)

where from property P1, \( (\sum_{i \in B_{k+y}^{0,0}} p_i, \sum_{i \in B_{k+y}^{0,1}} p_i, 1 - \sum_{i \in B_{k+y}^{0,0}} p_i - \sum_{i \in B_{k+y}^{0,1}} p_i) \sim D(\sum_{i \in B_{k+y}^{0,0}} \alpha_i, \sum_{i \in B_{k+y}^{0,1}} \alpha_i, \alpha_0 - \sum_{i \in B_{k+y}^{0,0}} \alpha_i - \sum_{i \in B_{k+y}^{0,1}} \alpha_i) \). That being said, applying Lemma 1 to equation (A.10), knowing property P4, the result can be readily derived.
APPENDIX B

MAXIMUM ENTROPY AND MAXIMAL DATA INFORMATION

B.1 Maximum Entropy Method

B.1.1 General Methodology

The MaxEnt prior distribution is the solution to the following problem:

\[
\max_{P: X \sim P} H_\theta[\theta] = -E_\theta[\ln \pi(\theta)] \tag{B.1}
\]

Subject to: \(E_\theta[f_i(\theta)] = \beta_i; i = 1, ..., m,\)

where \(f_i(.)\) are some measurable functions and \(\beta_i\) are known due to prior knowledge. The solution to the optimization problem in equation (B.1) is given by

\[
\pi(\theta) = Z^{-1} \exp\{\sum_{i=1}^m \gamma_i f_i(\theta)\},
\]

where \(Z\) and \(\gamma_i\)'s are a normalizing factor and the Lagrange multipliers, respectively, computed using the constraints in equation (B.1).

B.1.2 Multinomial with Dirichlet prior

Considering the Dirichlet prior, when there is no prior information in the form of expectations available, the MaxEnt prior, provided that the parameter \(\alpha_0\) is known, is the solution to the following optimization problem:

\[
\max_{\alpha \in S_{\alpha_0}^b} \sum_{k=1}^b \log \Gamma(\alpha_k) - (\alpha_k - 1)\psi(\alpha_k). \tag{B.2}
\]

Introducing the Lagrange multiplier \(\lambda\), we may write

\[
\max_{\alpha \in S_{\alpha_0}^b} \sum_{k=1}^b \log \Gamma(\alpha_k) - (\alpha_k - 1)\psi(\alpha_k) + \lambda(\alpha_0 - \sum_{k=1}^b \alpha_k), \tag{B.3}
\]

whose derivative with respect to any individual \(\alpha_k\) is given by

\[-(\alpha_k - 1)\psi'(\alpha_k) + \lambda = 0. \tag{B.4}\]

Knowing that the function \((\alpha_k - 1)\psi'(\alpha_k)\) is monotonic, the maximum is attained when \(\alpha_i = \alpha_j; \forall i, j\).

Hence, the Dirichlet prior shape is given by

\[\alpha = \alpha_0 1_b/b.\]
B.2 Maximal data information prior

B.2.1 General methodology

The simple MDIP is the solution to the following optimization problem:

$$\max \ H[\theta] - E_\theta[H[f(x|\theta)]]$$

Subject to: $E_\theta[g_i(\theta)] = \beta_i; i = 1, \ldots, m$ (B.6)

whose solution is given by

$$\pi(\theta) \propto \exp\{-H[f(x|\theta)] + \sum_{i=1}^{m} \gamma_i g_i(\theta)\}.$$  

B.2.2 Multinomial with Dirichlet prior

For example, for the multinomial model (discrete setting), the MDIP prior is given by

$$\pi(\theta) \propto (1 - b - 1 \sum_{i=1}^{b} \theta_i^{b-1}) \prod_{i=1}^{b} \theta_i^{\beta_i} \exp\{\sum_{i=1}^{m} \gamma_i g_i(\theta)\},$$

where a normalization factor is needed to result in a proper MDIP. This factor is computed for binomial and trinomial cases in [27].

APPENDIX C

PROOF OF LEMMA 3

Since the digamma function $\psi(x)$ is concave in $x$, the result for the REML objective function holds because the only non-linear part of it is the negative of the digamma function, making it convex provided that the feasible region is convex.

For the RMEP function, ignoring the linear function, we need to show that the term $- \sum_{k=1}^{b} \left[ \ln \Gamma(\alpha_i) - (\alpha_i - 1)\psi(\alpha_i) \right]$ is convex in $\alpha$. Since this term is the summation of a single function on elements of the vector, $\alpha_i$, it is sufficient to show that each individual summand is convex in its element, i.e. that $h(x) := (x - 1)\psi(x) - \ln \Gamma(x)$ is convex in $x$. It is enough to show that the second derivative of $h(x)$ is positive in its domain. Thus, we show that the derivative is increasing. For the derivative, one may write

$$f'(x) = (x - 1)\psi'(x) = x\psi'(x) - \psi'(x).$$

We split the problem into two parts: $x \in (0, 1]$ and $x \in (1, \infty)$. For $x \in (0, 1]$, $f''(x) = \psi'(x) + (x - 1)\psi''(x)$. Since $\psi(x)$ is increasing and concave, and $x - 1$ is negative, the second derivative is positive for $x \in (0, 1]$. Now consider $x \in (1, \infty)$. Expanding $\psi'(x)$, yields $\psi'(x) = \frac{1}{x} + \frac{1}{2x^2} + o(\frac{1}{x^3})$. Hence $x\psi'(x) = 1 + \frac{1}{2x} + \frac{1}{6x^2} +$
o(1/x^3). Then, the whole term can be rewritten as
\[
f'(x) = 1 - \frac{1}{2x} - \frac{1}{3x^2} - \frac{1}{6x^3} - \frac{1}{30x^4} + \frac{1}{30x^5} + o(\frac{1}{x^6}).
\]
Because \(x > 1\), the dominant term is \(-\frac{1}{2x}\), which is increasing in \(x \in (1, \infty)\). Finally, since each summand is convex, the linear combination is convex too.

Now consider the RMDIP function
\[
-\sum_{k=1}^{b} \left[ \ln \Gamma(\alpha_i) - (\alpha_i - 1)\psi(\alpha_i) + \alpha_i \frac{\psi(\alpha_i+1)}{\alpha_0} \right],
\]
where similar to above we only consider one term: \((x - 1)\psi(x) - x \frac{\psi(x+1)}{\alpha_0} - \ln \Gamma(x)\). Since \(\psi(x+1) = \psi(x) + \frac{1}{x}\), by defining \(\beta = \frac{\alpha_0 - 1}{\alpha_0}\) we can rewrite the function as
\[
(x - 1)\psi(x) - \frac{1}{\alpha_0} (1 + x\psi(x)) - \ln \Gamma(x) = (\beta x - 1)\psi(x) - \ln \Gamma(x),
\]
for which the derivative is given by
\[
\psi(x)(\beta x - 1) + \psi(x)(\beta - 1).
\]
Since for \(\alpha_0 > 1\), \(\beta \in (0, 1)\), the second term above, \(\psi(x)(\beta - 1)\), is convex. Similar to the RMEP function, we can show that the first term is also convex, again due to having \(\beta < 1\). The summation of two convex functions leads to another convex function, thereby finishing the proof.

**APPENDIX D**

**OPTIMAL PRECISION FACTOR**

First we calculate the MSE of the histogram estimator:

\[
\text{MSE}_{\text{hist}} = \sum_{k=1}^{b} E\left[ \frac{u_k^n}{n} - p_k^{true} \right] = \sum_{k=1}^{b} E\left[ \frac{u_k^n}{n^2} - \frac{2u_k n}{n} p_k^{true} \right] + p_k^{true^2} = \sum_{k=1}^{b} \left[ \frac{1}{n^2} \times (np_k^{true}(1 - p_k^{true}) + n^2 p_k^{true^2}) - 2/n (np_k^{true} p_k^{true}) + p_k^{true^2} \right] = \sum_{k=1}^{b} \frac{p_k^{true} - p_k^{true^2}}{n} = \frac{1 - \sum_{k=1}^{b} p_k^{true^2}}{n} = \frac{1 - \|p^{true}\|^2}{n}.
\]

Now, we continue this part, by deriving the MSE of the Bayesian estimate via (35):
\[ \text{MSE}_{\text{Bayesian}} = \sum_{k=1}^{b} E[m_k^* - p_k^{\text{true}}]^2 = \sum_{k=1}^{b} E[m_k^{*2} - 2m_k^* p_k^{\text{true}}] + p_k^{\text{true}}^2 \]  
\text{(D.1)}

Manipulating the terms after decomposition, we would have

\[ E[m_k^{*2}] = E[u_k^2 + 2u_k\alpha_k + \alpha_k^2] \]
\[ = \frac{1}{(n + \alpha_0)^2}[np_k^{\text{true}}(1 - p_k^{\text{true}}) + n^2p_k^{\text{true}}^2 + 2np_k^{\text{true}}\alpha_k + \alpha_k^2], \]
\[ E[m_k^* p_k^{\text{true}}] = \frac{1}{n + \alpha_0}[np_k^{\text{true}} + \alpha_k p_k^{\text{true}}] \]

After some simplifications, equation (D.1) can be rewritten as follows

\[ \text{MSE}_{\text{Bayesian}} = \sum_{k=1}^{b} \left( np_k^{\text{true}}(1 - p_k^{\text{true}}) + (\alpha_k - \alpha_0 p_k^{\text{true}})^2 \right) \]
\[ = \frac{n(1 - \|p_k^{\text{true}}\|^2) + \alpha_0^2\|m^* - p^{\text{true}}\|^2}{(n + \alpha_0)^2} \]
\[ = \frac{n^2\text{MSE}_{\text{hist}} + \alpha_0^2\|m^* - p^{\text{true}}\|^2}{(n + \alpha_0)^2} \]
\[ = \frac{\text{MSE}_{\text{hist}} + \alpha_0^2/n^2\|m^* - p^{\text{true}}\|^2}{(1 + \alpha_0/n)^2} \]
\[ = \frac{\text{MSE}_{\text{hist}} + \beta^2\|m^* - p^{\text{true}}\|^2}{(1 + \beta)^2} \]
\[ = \frac{\text{MSE}_{\text{hist}}}{(1 + \beta)^2} + \frac{\|m^* - p^{\text{true}}\|^2}{(1 + 1/\beta)^2}, \]

where we used \( \beta \) to denote the ratio \( \alpha_0/n \). Finally, the first derivative of the \( \text{MSE}_{\text{Bayesian}} \) would be written as follows

\[ 2\beta\|m^* - p^{\text{true}}\|^2 - \frac{\text{MSE}_{\text{hist}}}{(1 + \beta)^2} \]
whose only zero is attained at \( \beta = \frac{\text{MSE}_{\text{hist}}}{\|m^* - p^{\text{true}}\|^2} \). It can be readily seen that this point corresponds to the minimum of \( \text{MSE}_{\text{Bayesian}} \), finishing the proof.

**APPENDIX E**

**RESULTS: p53 Pathways**

In this part, some of the results associated with the DNA-damage pathways are shown.

**REFERENCES**

Fig. E.1. Expected difference between the true model and estimated probability masses estimated by histogram and Bayesian approach via constructed priors associated with p53 pathways. In the REML case, \( n = 5 \) points from each class are held out for prior construction. The first and second rows show the results for class 0 and 1, respectively.


