EXERCICES ABOUT IMPRECISE PREDICTIVE INFERENCE ABOUT CATEGORICAL DATA

Bayes Theorem

□ Assumptions

- a prior $\boldsymbol{\theta} \sim \mathit{Diri}(\boldsymbol{\alpha})$ for the case $K = 2$
- data $\boldsymbol{a}$ with sampling distribution $\boldsymbol{a}|\boldsymbol{\theta} \sim \mathit{Mn}(n, \boldsymbol{\theta})$

□ 1) Show that

- $\theta_1|\boldsymbol{a} \sim \mathit{Beta}(\boldsymbol{a} + \boldsymbol{\alpha})$
- $a_1 \sim \mathit{BeBi}(n; \boldsymbol{\alpha})$

Hint: Use Bayes' theorem, and the equivalence between *Beta* and *Diri* for $K = 2$.

□ 2) Show,

- assuming future data $\boldsymbol{a'}$ sampled independently from the same population, *i.e.* $\boldsymbol{a'} \sim \mathit{Mn}(n'; \boldsymbol{\theta})$,
- that $a'_1|\boldsymbol{a} \sim \mathit{BeBi}(n'; \boldsymbol{a} + \boldsymbol{\alpha})$

Hint: Use Bayes' theorem a second time.

Expressions for the DiMn

☐ **Assumptions**: Consider a composition $\boldsymbol{a} = (a_1, \ldots, a_K)$, with $\sum_k a_k = n$ whose probability distribution is a Dirichlet-multinomial:

$$\boldsymbol{a} \sim \textit{DiMn}(n; \boldsymbol{\alpha})$$

☐ **1) Equivalent forms**

Show the equivalence between the three forms of the *DiMn* for $\boldsymbol{a}$, in terms of

- generalized binomial coefficients
- gamma functions
- ascending factorials

See: Mathematical functions & coefficients

☐ **2) Application:** Simplify the formula (defined for any integer n and any reals $0 < \alpha < s$)

$$\sum_{a=0}^{n} \binom{n}{a} \alpha^{[a]} (s-\alpha)^{[n-a]},$$

□ 3) Sequences and compositions

Consider the case $K = 2$ and an observed sequence of length $n = 4$, $S = (c_1, c_1, c_2, c_1)$, yielding the counts $a_1 = 3, a_2 = 1$.

- How many sequences yield the same composition in counts? Same question for any composition (a_1, a_2)?
- What is the probability $P(S)$ of sequence S?
- Express $P(S)$ as the ratio of two products. Can you find a graphical interpretation of that result?

Hint: Represent any sequence as a path on a plane with a_1 on the x-axis and a_2 on the y-axis.

Distribution DiMn
Particular cases

□ Assumptions

- Consider that the composition in counts, over K categories, $\boldsymbol{a}$ follows a *DiMn*$(n; \boldsymbol{\alpha})$

□ 1) Special case $\alpha = 1$:

- Show that, in this case, $\boldsymbol{a}$ has a uniform distribution over its domain $\mathcal{A}$.

- From previous result, deduce the number of possible compositions of size n over K categories, *i.e.* the cardinal of $\mathcal{A}$. Express this number as a binomial coefficient.

□ 2) Towards Haldane

- For the case $K = 2$ and $n = 2$, what are the possible compositions $\boldsymbol{a}$

- For each $\boldsymbol{a}$, give the expression of $P(\boldsymbol{a})$

- Calculate this distribution for $\alpha_1 = \alpha_2 = \frac{1}{2}$, for $\alpha_1 = \alpha_2 = \frac{1}{10}$

- What happens if $\alpha_1 = \alpha_2$ tends to 0?

DiMn: pooling and restriction

☐ **Assumptions**

- Consider $\boldsymbol{a} \sim \mathit{DiMn}(n; \boldsymbol{\alpha})$ for $K = 3$, *i.e.* $\boldsymbol{a} = a_1, a_2, a_3$ with fixed $\sum_k a_k = n$
- Let $a_{23} = a_2 + a_3$ be the count of the pooled category $c_{23} = (c_2 \; or \; c_3)$

☐ **1) Express** the overall distribution on $\boldsymbol{a}$, $P(\boldsymbol{a})$, as a function of the marginal $P(a_1, a_{23})$

☐ **2) What does this entail** for the following distributions?

- $P(a_1, a_{23})$
- $P(a_2, a_3 | a_{23})$

☐ **3) Recursion**: The preceding example can be viewed as (i) defining a tree underlying the set of categories C, $T = \{c_1, c_{23} = \{c_2, c_3\}\}$, and (ii) "cutting" tree T at node c_{23}. What would be obtained for $K = 5$ categories underlied by tree $T = \{c_{1234} = \{c_1, c_{234} = \{c_2, c_3, c_4\}\}, c_5\}$

Bayesian prediction

□ Assume the following prior and posterior predictive distributions

- K is fixed
- $\boldsymbol{a} \sim$ *DiMn*$(n; \boldsymbol{\alpha})$
- $\boldsymbol{a'} \sim$ *DiMn*$(n'; \boldsymbol{a} + \boldsymbol{\alpha})$

□ Answer the following questions

- First, consider the prior prediction for $n = 1$. What is the probability that $a_k = 1$?
- Now, consider the posterior prediction for $n' = 1$. What is the probability that $a'_k = 1$?
- Same questions, with assuming also that the prior is a symmetric Dirichlet, *i.e.* $\alpha_k = \alpha$
- Now, consider the "bag of marbles" data, with observed data: 1 red, 2 green, 2 light blue, 1 dark blue. Under the same assumptions, what is the probability that $a'_{blue} = 1$ for $n' = 1$?
- Is there a problem?

Imprecision and s

□ Assumptions

- Prior uncertainty is modelled by an IDMM(s)
- Denote by B_j the event that next observation will be from category c_j (possibly not elementary)

□ Questions

- Find the prior lower and upper probabilities, $\underline{P}(B_j)$ and $\overline{P}(B_j)$.
- After observing data $\boldsymbol{a}$, find the posterior lower and upper probabilities, $\underline{P}(B_j|\boldsymbol{a})$ and $\overline{P}(B_j|\boldsymbol{a})$.
- Define the imprecision about an event by $\Delta(\cdot) = \overline{P}(\cdot) - \underline{P}(\cdot)$. What are $\Delta(B_j)$ and $\Delta(B_j|\boldsymbol{a})$?
- Compute the ratio of these two imprecisions. When is it equal to 2, to 10?
- Apply the preceding results to the "bag of marbles" example, with B_j being the event that the next observation is blue.

Confirming a universal law

□ Assumptions

- There are K basic categories
- Amongst n observations, all were found to belong to c_1, *i.e.* $a_1 = n$
- You envisage to collect n' more data, and you consider the hypothesis H_0 that these future data might all be of type c_1 again, *i.e.* that $a'_1 = n'$.

□ 1) Bayesian answers

- Under a standard Bayesian model, with prior *Diri*$(\boldsymbol{\alpha})$, what is the expression $P = P_{\boldsymbol{\alpha}}(H_0|\boldsymbol{a})$?
- What is the value of P under Haldane's model, *i.e.* $\boldsymbol{\alpha} = 0$?
- What is the value of P under Bayes-Laplace's model, *i.e.* $\boldsymbol{\alpha} = 1$, assuming $K = 2$, and then $K = 3$?

- Under Bayes-Laplace's model, find the expressions of P for the special cases, $n' = 1$, $n' = n$ and $n' \to \infty$, assuming either $K = 2$ or $K = 3$.

☐ **2) IDMM answers**

- Under the prior IDMM(s), find the lower and upper probabilities of the same event: $\underline{P} = \underline{P}(H_0|\boldsymbol{a})$ and $\overline{P} = \overline{P}(H_0|\boldsymbol{a})$.

- What are these L&U probabilities for an IDMM with $s = 1$, $s = 2$, and as $s \to 0$ or $s \to \infty$?

- Under the IDMM with $s = 1$, find the expressions of $\underline{P}$ and $\overline{P}$ for the special cases, $n' = 1$, $n' = n$ and $n' \to \infty$.

- Do we need to make assumptions about K?

- Compare these results with those of part 1.

☐ **3) Iguana example**:

Bernardo & Smith (1994) consider the example of $n = 90$ iguanas all found with the same skin pattern on an island where the overall number of iguanas is estimated to be $n^* = n + n' = 100,000$. Find the preceding Bayesian and IDMM($s = 1$) answers for that example.