Probabilistic Modeling and Reasoning

Homework — 6

Nikolaos Liouliakis (AID25001)
Vasileios-Efraim Tsavalia (AID25006)

MSc in Artificial Intelligence and Data Analytics
University of Macedonia

Supervisor: Professor Dimitris Christou-Varsakelis

February 2025

Problem 1

The model selection was implemented in the *P1_coin_tossing.m* MATLAB script. To overcome numerical issues for large values of N_{Heads} and N_{Tails} the Log-Sum-Exp trick was used.

1. Direct Approach

The direct approach calculates the Bayes factor by summing terms involving high powers and small probabilities:

$$p_{\text{Data}|\text{Model}_i} = \sum \theta^{N_{\text{Heads}}}(1-\theta)^{N_{\text{Tails}}} \cdot \text{pdf}_i(\theta)$$

This approach works for the special case when $\theta = 0$ or $1-\theta = 0$ and $N_{\text{Heads}} = 0$ or $N_{\text{Tails}} = 0$ respectively because in MATLAB $0^0 = 1$, avoiding undefined behavior. However, it can still suffer from underflow when probabilities (θ or $1-\theta$) are raised to large powers (N_{Heads}, N_{Tails}).

2. Logarithmic Approach

The logarithmic approach mitigates underflow and precision issues by:

1. Computing the log of each term:

$$\log(\text{term}) = N_{\text{Heads}} \cdot \log(\theta) + N_{\text{Tails}} \cdot \log(1-\theta) + \log(\text{pdf}_i(\theta))$$

2. Using the `logsumexp` trick to sum terms in log-space:

$$\log\left(\sum \text{terms}\right) = c + \log\left(\sum \exp\left(\log(\text{terms}) - c\right)\right)$$

   where $c = \max(\log(\text{terms}))$ ensures numerical stability.
3. Finally, converting the log-ratio back to linear scale:

$$\text{Bayes Factor (accurate)} = \exp\left(\log(p_{\text{Data}|\text{Model}_1}) - \log(p_{\text{Data}|\text{Model}_2})\right)$$

This method avoids the pitfalls for the special cases mentioned above, as MATLAB handles `0 * ∞ = NaN` gracefully using `omitnan` during summation.

Problem 2

The given linear model is:

$$y_{t+1} = \sum_{k=1}^{K} w_k x_{t,k} = \mathbf{w}^\top \mathbf{x}_t,$$

where:

- $\mathbf{x}_t$ is the vector of features (factors) on day t,
- $\mathbf{w}$ is the vector of weights to estimate,
- $\mathbf{D} = \{(\mathbf{x}_t, y_{t+1})\}_{t=1}^{T-1}$ is the historical data,
- σ_t^2 represents the volatility for each day.

The returns y_t are assumed to follow a Gaussian distribution. The likelihood is:

$$p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}) = \prod_{t=2}^{T} \mathcal{N}\left(y_t|\mathbf{w}^\top \mathbf{x}_{t-1}, \sigma_t^2\right),$$

which expands to:

$$p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}) = \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(y_t - \mathbf{w}^\top \mathbf{x}_{t-1})^2}{2\sigma_t^2}\right)$$

a)

Log-likelihood

Taking the log of the likelihood:

$$\log p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}) = \sum_{t=2}^{T} \left[-\frac{1}{2}\log(2\pi\sigma_t^2) - \frac{(y_t - \mathbf{w}^\top \mathbf{x}_{t-1})^2}{2\sigma_t^2} \right].$$

Ignoring the term $-\frac{1}{2}\log(2\pi\sigma_t^2)$ (independent of $\mathbf{w}$), the negative log-likelihood (NLL) becomes:

$$\mathrm{NLL}(\mathbf{w}) = \frac{1}{2}\sum_{t=2}^{T} \frac{(y_t - \mathbf{w}^\top \mathbf{x}_{t-1})^2}{\sigma_t^2}.$$

Objective function for maximum likelihood

Maximizing the likelihood is equivalent to minimizing the NLL:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{t=2}^{T} \frac{(y_t - \mathbf{w}^\top \mathbf{x}_{t-1})^2}{\sigma_t^2}.$$

Gradient of the NLL

Expanding the quadratic term:

$$\mathrm{NLL}(\mathbf{w}) = \frac{1}{2}\sum_{t=2}^{T} \frac{1}{\sigma_t^2}\left[y_t^2 - 2y_t(\mathbf{w}^\top \mathbf{x}_{t-1}) + (\mathbf{w}^\top \mathbf{x}_{t-1})^2 \right].$$

$$\mathrm{NLL}(\mathbf{w}) = \frac{1}{2}\sum_{t=2}^{T} \frac{1}{\sigma_t^2}\left[y_t^2 - 2y_t\mathbf{x}_{t-1}^\top \mathbf{w} + \mathbf{w}^\top \mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top \mathbf{w} \right].$$

The gradient with respect to $\mathbf{w}$ is:

$$\frac{\partial \mathrm{NLL}(\mathbf{w})}{\partial \mathbf{w}} = -\sum_{t=2}^{T} \frac{y_t}{\sigma_t^2}\mathbf{x}_{t-1}^\top + \sum_{t=2}^{T} \frac{\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top}{\sigma_t^2}\mathbf{w}$$

Let

$$\mathbf{A_1} = \sum_{t=2}^{T} \frac{\mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top}{\sigma_t^2}$$

$$\mathbf{b_1} = \sum_{t=2}^{T} \frac{y_t}{\sigma_t^2}\mathbf{x}_{t-1}^\top$$

The solution/maximum likelihood estimate for $\mathbf{w}$ is:

$$\mathbf{w} = \mathbf{A_1}^{-1}\mathbf{b}$$

b)

Model Prior

The prior distribution over the weight vector $\boldsymbol{w}$ is defined as:

$$p(\boldsymbol{w} \mid M) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \boldsymbol{I}_{\|M\|_1}),$$

where M denotes the model as a binary vector, indexing the subset of factors being used and $\|M\|_1$ the l^1-norm. By representing M as a binary vector of length K, we encode which factors are included in the model. For example, with $K = 3$, we have $2^3 - 1 = 7$ models:

$$\{0,0,1\}, \{0,1,0\}, \{0,1,1\}, \{1,0,0\}, \{1,0,1\}, \{1,1,0\}, \{1,1,1\}.$$

Posterior Distribution

Using Bayes' rule and assuming a flat prior of M, the posterior distribution is:

$$p(M \mid \boldsymbol{D}) \propto p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M)$$

Marginal Likelihood for Model Selection

The likelihood for a given model M is:

$$p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M) = \int_{\mathbb{R}^{\|M\|_1}} p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, \boldsymbol{w}, M) \cdot p(\boldsymbol{w} \mid M) \, d^{\|M\|_1}\boldsymbol{w}$$

By substituting the expressions we get:

$$p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M) = \int \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(y_t - \mathbf{w}^\top \mathbf{x}_{t-1})^2}{2\sigma_t^2}\right) \cdot \frac{1}{\sqrt{(2\pi)^{\|M\|_1}}} \exp\left(-\frac{1}{2}\mathbf{w}^\top \mathbf{w}\right) d\boldsymbol{w}$$

$$= \frac{1}{\sqrt{(2\pi)^{\|M\|_1}}} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \int \exp\left(-\frac{(y_t - \mathbf{w}^\top \mathbf{x}_{t-1})^2}{2\sigma_t^2} - \frac{1}{2}\mathbf{w}^\top \mathbf{w}\right) d\boldsymbol{w}$$

$$= \frac{1}{\sqrt{(2\pi)^{\|M\|_1}}} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \int \exp\left(-\frac{1}{2}\mathbf{w}\left(\boldsymbol{I}_{\|M\|_1} + \sum_{t=2}^{T} \frac{\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^\top}{\sigma_t^2}\right)\mathbf{w}^\top + \sum_{t=2}^{T} \frac{y_t \boldsymbol{x}_{t-1}^\top}{\sigma_t^2}\mathbf{w} - \frac{1}{2}\sum_{t=2}^{T} \frac{y_t^2}{\sigma_t^2}\right) d\boldsymbol{w}$$

Which can be written as:

$$= \frac{1}{\sqrt{(2\pi)^{\|M\|_1}}} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \int \exp\left(-\frac{1}{2}\mathbf{w}\boldsymbol{A}\mathbf{w}^\top + \boldsymbol{b}^\top \mathbf{w} + c\right) d\boldsymbol{w}$$

where:

$$\boldsymbol{A} = \boldsymbol{I}_{\|M\|_1} + \sum_{t=2}^{T} \frac{\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^\top}{\sigma_t^2},$$

$$\boldsymbol{b} = \sum_{t=2}^{T} \frac{y_t \boldsymbol{x}_{t-1}}{\sigma_t^2},$$

$$c = -\frac{1}{2}\sum_{t=2}^{T} \frac{y_t^2}{\sigma_t^2}$$

This integral can be solved analytically using:

$$\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}x^\intercal Ax + b^\intercal x + c\right) d^n x = \sqrt{\det(2\pi A^{-1})} e^{\frac{1}{2}b^\intercal A^{-1}b+c}$$

Therefore, it simplifies to:

$$p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M) = \frac{1}{\sqrt{(2\pi)^{\|M\|_1}}} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \sqrt{\det(2\pi A^{-1})} e^{\frac{1}{2}b^\intercal A^{-1}b+c}$$

Taking the logarithm and then multiplying by two, we get:

$$2\log(p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M)) = -\|M\|_1 \log(2\pi) - \sum_{t=2}^{T} \log(2\pi\sigma_t^2) + \log\det(2\pi \boldsymbol{A}^{-1}) + \boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{b} + \frac{c}{2}$$

$$= -\|M\|_1 \log(2\pi) - \sum_{t=2}^{T} \log(2\pi\sigma_t^2) + \log\left((2\pi)^{\|M\|_1} \det(\boldsymbol{A}^{-1})\right) + \boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{b} - \sum_{t=2}^{T} \frac{y_t^2}{\sigma_t^2}$$

$$= -\sum_{t=2}^{T} \log(2\pi\sigma_t^2) + \log\det(\boldsymbol{A}^{-1}) + \boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{b} - \sum_{t=2}^{T} \frac{y_t^2}{\sigma_t^2}$$

Bayesian Model Selection

Finally, we compute the likelihood for each model M and select the model with the highest likelihood:

$$M^* = \arg\max_M p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M)$$

$$= \arg\max_M (2\log(p(\boldsymbol{y}_{2:T} \mid \boldsymbol{x}_{1:T-1}, M)))$$

$$= \arg\max_M \left(-\sum_{t=2}^{T} \log(2\pi\sigma_t^2) + \log\det(\boldsymbol{A}^{-1}) + \boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{b} - \sum_{t=2}^{T} \frac{y_t^2}{\sigma_t^2}\right)$$

c)

To find the model that best fits the data in the *dodder.txt* the Matlab script named *P2˙Model˙selection.m* was used, which gave the results:

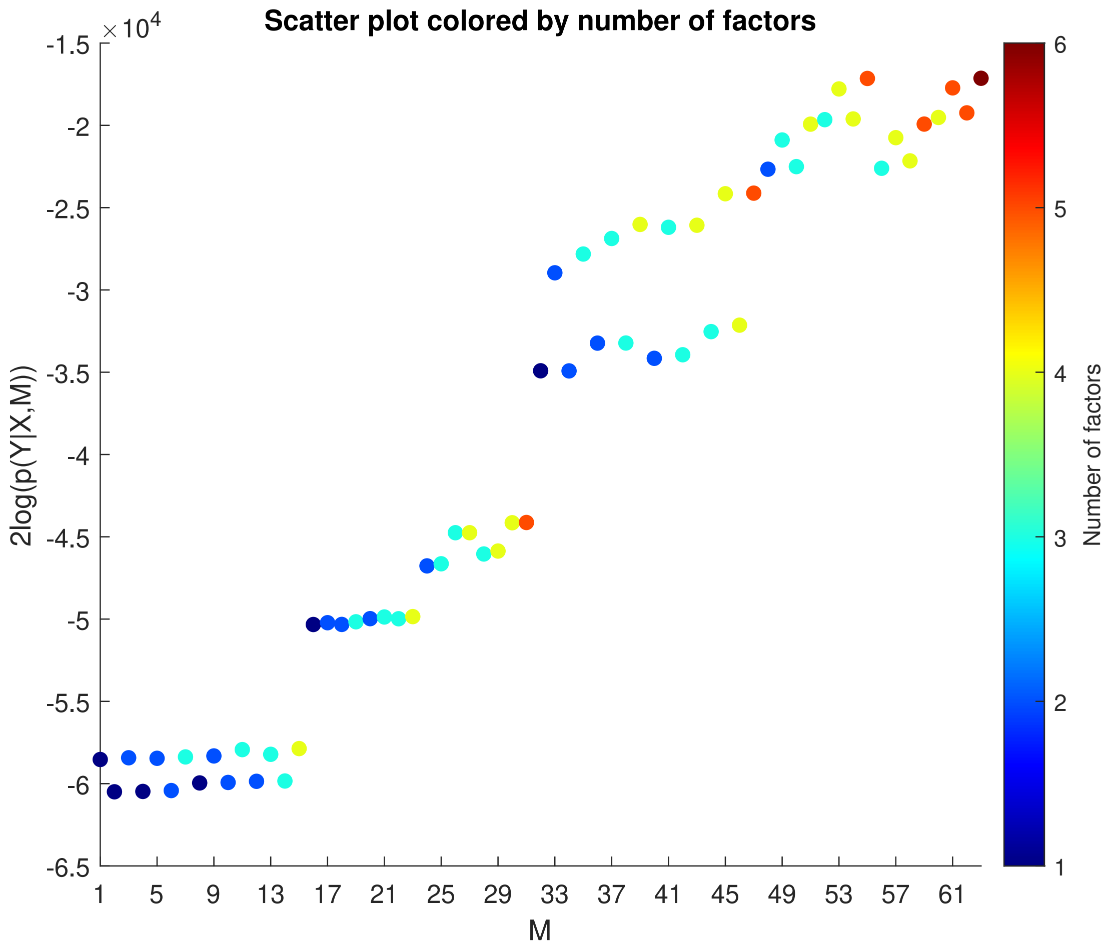

Figure 1: Log likelihood for every possible value for M

Given those values, we conclude that $M^* = 63$, which means we should keep all 6 the factors. Probably there is some numerical mistake in calculations due to the limited perdition of floating numbers and the actual maximum value for the likelihood is for $M = 55$ which has only 5 factors.

Problem 3

Step 1: Definitions and Notation

Let the observed classifications from the table be represented as counts:

$$\mu^a = (\mu_1^a, \mu_2^a, \mu_3^a), \quad \mu^b = (\mu_1^b, \mu_2^b, \mu_3^b), \quad \mu^c = (\mu_1^c, \mu_2^c, \mu_3^c)$$

where μ^a, μ^b, μ^c are the counts of categories $1, 2, 3$ for Persons 1, 2, and 3, respectively.

Using the provided table, we get:

$$\mu^a = (13, 3, 4), \quad \mu^b = (4, 9, 7), \quad \mu^c = (8, 8, 4)$$

The combined counts are:

$$\mu_{\text{total}} = \mu^a + \mu^b + \mu^c = (25, 20, 15)$$

The Dirichlet normalization constant is:

$$Z(u) = \frac{\prod_q \Gamma(u_q)}{\Gamma\left(\sum_q u_q\right)}$$

Assuming a uniform prior on counts ($u_q = 1$), this simplifies to:

$$Z(u) = \frac{\prod_q \Gamma(1)}{\Gamma(3)} = \frac{1}{2}$$

Step 2: Likelihoods of the Models

1. Likelihood for H_{indep}:

$$p(o_a, o_b, o_c \mid H_{\text{indep}}) = p(H_{\text{indep}}) \cdot \frac{Z(u+\mu^a)}{Z(u)} \cdot \frac{Z(u+\mu^b)}{Z(u)} \cdot \frac{Z(u+\mu^c)}{Z(u)}$$

2. Likelihood for H_{same}:

$$p(o_a, o_b, o_c \mid H_{\text{same}}) = p(H_{\text{same}}) \cdot \frac{Z(u+\mu_{\text{total}})}{Z(u)}$$

Step 3: Bayes Factor

Assuming no prior preference amongst hypotheses, the Bayes factor is:

$$\text{Bayes Factor} = \frac{p(o_a, o_b, o_c \mid H_{\text{indep}})}{p(o_a, o_b, o_c \mid H_{\text{same}})} = \frac{\frac{Z(u+\mu^a)}{Z(u)} \cdot \frac{Z(u+\mu^b)}{Z(u)} \cdot \frac{Z(u+\mu^c)}{Z(u)}}{\frac{Z(u+\mu_{\text{total}})}{Z(u)}}$$

Simplify:

$$\text{Bayes Factor} = \frac{Z(u+\mu^a) \cdot Z(u+\mu^b) \cdot Z(u+\mu^c)}{Z(u+\mu_{\text{total}}) \cdot Z(u) \cdot Z(u)}$$

Step 4: Substitution of Counts and Evaluation

For the evaluation, the MATLAB script *P3_hypotheses_test.m* was used, which gives the result of:

$$\text{Bayes Factor} = 2.7586$$

Problem 4: Mixture Model

a) Posterior Probability

The posterior probability that a sample x was generated from $f_1(x)$ is computed using Bayes' rule:

$$P(f_1|x) = \frac{P(f_1) \cdot P(x|f_1)}{P(x)} = \frac{\theta f_1(x)}{\theta f_1(x) + (1-\theta) f_2(x)}$$

b) Log-Likelihood Using Indicator Function

Let $\mathbb{I}(c_i = 1)$ and $\mathbb{I}(c_i = 0)$ denote indicator functions, where:

$$\mathbb{I}(c_i = 1) = \begin{cases} 1, & \text{if } c_i = 1, \\ 0, & \text{otherwise} \end{cases} \quad \mathbb{I}(c_i = 0) = 1 - \mathbb{I}(c_i = 1)$$

The likelihood of the dataset $\{c_1, x_1, \dots, c_n, x_n\}$ (given i.i.d. coin flips) is:

$$P(c_1, x_1, \dots, c_n, x_n|\theta) = \prod_{i=1}^{n} P(c_i, x_i|\theta)$$

$$P(c_1, x_1, \dots, c_n, x_n|\theta) = \prod_{i=1}^{n} [\theta f_1(x_i)]^{\mathbb{I}(c_i=1)}[(1-\theta)f_2(x_i)]^{\mathbb{I}(c_i=0)}$$

Taking the natural logarithm, the log-likelihood becomes:

$$\log P(c_1, x_1, \dots, c_n, x_n|\theta) = \sum_{i=1}^{n} [\mathbb{I}(c_i = 1)\log(\theta f_1(x_i)) + \mathbb{I}(c_i = 0)\log((1-\theta)f_2(x_i))]$$

Expanding the terms:

$$\log P(c_1, x_1, \dots, c_n, x_n|\theta) = \sum_{i=1}^{n} [\mathbb{I}(c_i = 1)(\log\theta + \log f_1(x_i)) + \mathbb{I}(c_i = 0)(\log(1-\theta) + \log f_2(x_i))]$$

c) Maximum Likelihood Estimator (MLE) for θ

To find the MLE, maximize the log-likelihood with respect to θ:

$$\log P(c_1, x_1, \dots, c_n, x_n|\theta) = \sum_{i=1}^{n} [\mathbb{I}(c_i = 1)\log\theta + \mathbb{I}(c_i = 0)\log(1-\theta)] + (\text{terms independent of } \theta)$$

Let:

$$N_1 = \sum_{i=1}^{n} \mathbb{I}(c_i = 1), \quad N_2 = \sum_{i=1}^{n} \mathbb{I}(c_i = 0)$$

Therefore, the log-likelihood can be rewritten as:

$$\log P(c_1, x_1, \dots, c_n, x_n|\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

Take the derivative with respect to θ:

$$\frac{\partial}{\partial\theta}[N_1 \log\theta + N_2 \log(1-\theta)] = \frac{N_1}{\theta} - \frac{N_2}{1-\theta}$$

Set the derivative to zero:

$$\frac{N_1}{\theta} = \frac{N_2}{1-\theta}$$

Solve for θ:

$$\theta = \frac{N_1}{N_1 + N_2} = \frac{N_1}{n}$$

Thus, the MLE is:

$$\theta_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(c_i = 1)$$

Problem 5: EM Algorithm

E-Step

Compute the probability that x_i was generated from $f_1(x)$ or $f_2(x)$, given the current estimate of θ:

$$q_{i,1} = P(c_i = 1|x_i, \theta) = \frac{\theta f_1(x_i)}{\theta f_1(x_i) + (1-\theta) f_2(x_i)}$$

$$q_{i,0} = P(c_i = 0|x_i, \theta) = \frac{(1-\theta) f_2(x_i)}{\theta f_1(x_i) + (1-\theta) f_2(x_i)}$$

M-Step

We write the Energy term:

$$\sum_{i=1}^{n} \langle \log P(c_i, x_i|\theta) \rangle_{q(c_i|x_i,\theta)}$$

$$\sum_{i=1}^{n} \langle \log P(x_i|c_i, \theta) P(c_i \mid \theta) \rangle_{q(c_i|x_i,\theta)}$$

$$\sum_{i=1}^{n} \langle \log P(x_i|c_i, \theta) + \log P(c_i \mid \theta) \rangle_{q(c_i|x_i,\theta)}$$

$$\sum_{i=1}^{n} q_{i,1}\big(\log P(x_i|c_i = 1, \theta) + \log P(c_i = 1 \mid \theta)\big) + q_{i,0}\big(\log P(x_i|c_i = 0, \theta) + \log P(c_i = 0 \mid \theta)\big)$$

$$\sum_{i=1}^{n} q_{i,1}\big(\log f_1(x_i) + \log \theta\big) + q_{i,0}\big(\log f_2(x_i) + \log(1-\theta)\big)$$

Let:

$$Q_1 = \sum_{i=1}^{n} q_{i,1}, \quad Q_2 = \sum_{i=1}^{n} q_{i,2}, \quad Q_1 + Q_2 = n$$

Therefore, we can write it as:

$$Q_1\big(\log f_1(x_i) + \log \theta\big) + Q_0\big(\log f_2(x_i) + \log(1-\theta)\big)$$

Keeping only the terms dependent on θ we have:

$$Q_1 \log \theta + Q_2 \log(1-\theta)$$

Take the derivative with respect to θ:

$$\frac{\partial}{\partial \theta}[Q_1 \log \theta + Q_2 \log(1-\theta)] = \frac{Q_1}{\theta} - \frac{Q_2}{1-\theta}$$

Set the derivative to zero:

$$\frac{Q_1}{\theta} = \frac{Q_2}{1-\theta}$$

Solve for θ:

$$\theta = \frac{Q_1}{Q_1 + Q_2} = \frac{Q_1}{n}$$