Exam PML (MO435)

Name: João Victor da Silva Guerra RA: 117410

Instructions:

- This is a take-home exam. You will have five days to submit your final answers. Hence, it is expected that you answer efficiently and effectively. Write a draft of your answers, and then edit them to deliver the best (concise) possible answers. **Only submit the final version of your answers.**
- You have 120 hours (5 days) to complete this exam. The deadline for the submission is Saturday 09/05 at 08:00.
- Answer the following questions by printing this exam, or in individual sheets of paper. In case you answer on individual sheets, you **must** identify each sheet with a sequence number and your RA. Simultaneously, you **must** identify each question with its number.
- Use clear writing to answer the questions. If your answers can't be read, they won't be graded.
- In case you answer on different sheets, use a similar amount of space as the one provided in this exam.
- This exam **must** be answered individually. It is **forbidden** to comment and discuss your exam and answers with other persons during the test period.
- It is **allowed** to check the book and other sources of information. But it is not OK to search for the question's answers directly.
- This exam has 2 questions for a total of 10 points. Check that your exam is complete.
- You should have received an invitation to Gradescope to your registered e-mail. If you didn't, then register using the code `9K72WB`. You must submit your exam exclusively through Gradescope assigned exam. To do so, scan or take a photo of your answer sheets, and follow the instructions on Gradescope to mark the pages that correspond to each question.

1. Bayesian Hierarchical Clustering (BHC) is similar to traditional agglomerative clustering techniques in that it is one pass, bottom-up method which initializes each point as its own cluster and iteratively merges pairs of clusters. The main difference is that uses statistical hypothesis testing to choose which cluster to merge.

   Assuming a dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ with $\mathcal{D}_i \subset \mathcal{D}$ as a set of data points at the leaves of the subtree T_i, the algorithm starts with N trivial trees, $\{T_i\}_{i=1}^{N}$, each containing a single data point $\mathcal{D}_i = \{x_i\}$. At each stage the algorithm merges two subtrees T_i and T_j into T_{ij} and its associated data is $\mathcal{D}_{ij} = \mathcal{D}_i \cup \mathcal{D}_j$.

   (a) At each merging step we consider two hypotheses: merge subtrees T_i and T_j together, as $M_{ij} = 1$, or not, $M_{ij} = 0$. Then, the probability of the data $\mathcal{D}_{ij}$ given the subtree T_{ij} is (2)

   $$p(\mathcal{D}_{ij} \mid T_{ij}) = p(\mathcal{D}_{ij} \mid M_{ij} = 1)p(M_{ij} = 1) + p(\mathcal{D}_{ij} \mid M_{ij} = 0)p(M_{ij} = 0). \quad (1)$$

   What should be the probability of the data given each hypothesis and why? In other words, define the probabilities $p(\mathcal{D}_{ij} \mid M_{ij} = 1)$ and $p(\mathcal{D}_{ij} \mid M_{ij} = 0)$ and explain why they have those definitions.

For the first hypothesis, where the subtrees T_i and T_j merge together ($M_{ij} = 1$), the data in D_{ij} is assumed to be generated independently and identically from the same probabilistic model ($p(x_i|\theta)$). However, we also have to consider the prior over the parameters θ of the model ($p(\theta|\lambda)$) with hyperparameters λ. As the model is latent, we need to marginalize with respect to all possible parameters θ. Hence, the probability of the data given the first hypothesis is:

$$p(D_{ij} \mid M_{ij} = 1) = \int p(D_i|\theta) \cdot p(\theta|\lambda)\, d\theta = \int \left[\prod_{x_i \in D_{ij}} p(x_i|\theta) \right] \cdot p(\theta|\lambda)\, d\theta$$

For the second hypothesis, where the subtrees T_i and T_j do not merge together ($M_{ij} = 0$), the data is assumed to be generated by each subtree independently (i.e. different models). Hence, the probability of the data given the second hypothesis is:

$$p(D_{ij} \mid M_{ij} = 0) = p(D_i|T_i) \cdot p(D_j|T_j)$$

where $p(D_i|T_i)$ and $p(D_j|T_j)$ are defined by (1) from the statment of the question.

(b) Assume that each data point has a single binary feature. What is an optimal choice for the likelihood distribution of the data $p(x_i \mid \theta_i)$ and the parameter's prior $p(\theta_i \mid \beta)$? And using them, what is the probability $p(\mathcal{D}_{ij} \mid M_{ij} = 1)$? (Show your work.) (3)

As stated, we assume a one-dimensional feature $x \in [0,1]$. Hence, the optimal choice for the likelihood distribution of the data ($p(x_i|\theta)$) is a Bernoulli distribution and for the parameter's prior ($p(\theta|\alpha,\beta)$), we choose the conjugate prior of the Bernoulli distribution, the beta distribution. Then, the integral in the probability $p(D_{ij}|M_{ij}=1)$ is tractable (i.e. has a closed form).

$$p(x_i|\theta) = \text{Ber}(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} \quad (1)$$

$$p(\theta|\alpha,\beta) = \text{Beta}(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}; \text{ for } 0 \leqslant \theta \leqslant 1 \quad (2)$$

The probability $p(D_{ij}|M_{ij}=1)$ has the following form:

$$p(D_{ij}|M_{ij}=1) = \int \left[\prod_{x_i \in D_{ij}} p(x_i|\theta)\right] \cdot p(\theta|\lambda) \cdot d\theta \quad (3)$$

Replacing (1) and (2) in (3):

$$p(D_{ij}|M_{ij}=1) = \int_0^1 \left[\prod_{x_i \in D_{ij}} \theta^{x_i}(1-\theta)^{1-x_i}\right] \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \, d\theta$$

$$= \int_0^1 \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \, d\theta$$

For simplicity, we denote $n_i = \sum x_i$ (i.e. the number of ones in D_{ij}) and n is the total number of data points in D_{ij}. Hence,

$$p(D_{ij}|M_{ij}=1) = \frac{1}{B(\alpha,\beta)} \cdot \int_0^1 \theta^{n_i+\alpha-1}(1-\theta)^{n-n_i+\beta-1} \, d\theta \quad (4)$$

As we know, the beta function has the following form:

$$B(x,y) = \int_0^1 \theta^{x-1}(1-\theta)^{y-1} \, d\theta \quad (5)$$

From the integral of (4), we recognize (5) for $x = n_i + \alpha$ and $y = n - n_i + \beta$. Hence, the integral of (4) is equal $B(n_i+\alpha, n-n_i+\beta)$ and the probability is:

$$p(D_{ij}|M_{ij}=1) = \frac{B(n_i+\alpha, n-n_i+\beta)}{B(\alpha,\beta)}$$

⊛ Note: The hyperparameters β were renamed to $\lambda = (\alpha,\beta)$ to agree with item (a).

(c) One important step on BHC is learning the model's hyperparameters β used in the prior $p(\theta \mid \beta)$. (2)
Derive the gradient $\frac{\partial p(\mathcal{D}_{ij} \mid T_{ij})}{\partial \beta}$ and explain how you will use it to learn β. (Show your work.)

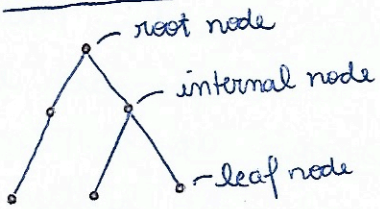
First, let us define some concepts and useful tools:

$$\frac{\partial B(x,y)}{\partial x} = B(x,y) \cdot \left[\Psi_0(x) - \Psi_0(x+y) \right] \quad (A)$$

$$\frac{\partial B(x,y)}{\partial y} = B(x,y) \cdot \left[\Psi_0(y) - \Psi_0(x+y) \right] \quad (B)$$

digamma function $\equiv \Psi_0(x) = \frac{d}{dx} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$

Decision tree:



The probability of the data D_{ij} given the subtree T_{ij} is:

$$p(D_{ij} \mid T_{ij}) = p(D_{ij} \mid M_{ij}=1) \cdot p(M_{ij}=1) + p(D_{ij} \mid M_{ij}=0) \cdot p(M_{ij}=0) \quad (1)$$

For simplicity, we denote that the prior is $p(M_{ij}=1) = \pi_{ij}$. There is a proper procedure to define π_{ij}.

Then, implicitly derive (1) with respect to $\lambda = (\alpha, \beta)$:

$$\frac{\partial p(D_{ij} \mid T_{ij})}{\partial \lambda} = \left(\frac{\partial p(D_{ij} \mid T_{ij})}{\partial \alpha}, \frac{\partial p(D_{ij} \mid T_{ij})}{\partial \beta} \right)$$

$$\frac{\partial p(D_{ij} \mid T_{ij})}{\partial \lambda_\ell} = \frac{\partial p(D_{ij} \mid M_{ij}=1)}{\partial \lambda_\ell} \cdot \pi_{ij} + \frac{\partial p(D_i \mid T_i)}{\partial \lambda_\ell} \cdot p(D_j \mid T_j) \cdot (1-\pi_{ij}) + p(D_i \mid T_i) \cdot \frac{\partial p(D_j \mid T_j)}{\partial \lambda_\ell} \cdot (1-\pi_{ij}) \quad (2)$$

Considering the base case (leaf node), the second term of (1) is equal to zero since there are no subtree, and the prior of the first term is one. For a leaf node k, we have from (1) as follows:

$$\frac{\partial p(D_k \mid T_k)}{\partial \alpha} = \frac{\partial p(D_k \mid M_k=1)}{\partial \alpha} \cdot \cancel{p(M_k=1)}^{1} \quad (3)$$

As we know from item (b), the $p(D_k \mid M_k=1)$ has the following form:

$$p(D_k \mid M_k=1) = \frac{B(\alpha + k_1, \beta + k - k_1)}{B(\alpha, \beta)} \quad (4)$$

, where k_1 is the number of ones in node k and k is the total number of data points in node k.

Replacing (4) in (3):

$$\frac{\partial p(D_k \mid T_k)}{\partial \alpha} = \frac{\partial}{\partial \alpha} \left(\frac{B(\alpha+k_1, \beta+k-k_1)}{B(\alpha,\beta)} \right) = \frac{\frac{\partial B(\alpha+k_1, \beta+k-k_1)}{\partial \alpha} B(\alpha,\beta) - B(\alpha+k_1, \beta+k-k_1) \cdot \frac{\partial B(\alpha,\beta)}{\partial \alpha}}{B(\alpha,\beta)^2}$$

$$= \frac{B(\alpha+k_1, \beta+k-k_1) \left[\Psi_0(\alpha+k_1) - \Psi_0(\alpha+\beta+k) \right] \cdot \cancel{B(\alpha,\beta)} - B(\alpha+k_1, \beta+k-k_1) \cdot \cancel{B(\alpha,\beta)} \left[\Psi_0(\alpha) - \Psi_0(\alpha+\beta) \right]}{B(\alpha,\beta)^{\cancel{2}}}$$

$$= \frac{B(\alpha+k_1, \beta+k-k_1)}{B(\alpha,\beta)} \left[\Psi_0(\alpha+k_1) + \Psi_0(\alpha+\beta) - \Psi_0(\alpha+\beta+k) - \Psi_0(\alpha) \right] \quad (5)$$

Similarly,

$$\frac{\partial p(D_k|T_k)}{\partial \beta} = \frac{B(\alpha+k_1, \beta+k-k_1)}{B(\alpha,\beta)} \cdot \left[\psi_0(\beta+k-k_1) + \psi_0(\alpha+\beta) - \psi_0(\alpha+\beta+k) - \psi_0(\beta) \right] \quad (6)$$

With (5) and (6), we have expressions for $\frac{\partial p(D_{ij}|M_{ij}=1)}{\partial \lambda_e}$, $\frac{\partial p(D_i|T_i)}{\partial \lambda_e}$ and $\frac{\partial p(D_j|T_j)}{\partial \lambda_e}$. Further, with (4), we have expressions for $p(D_i|T_i)$ and $p(D_j|T_j)$. Last, we also need to compute $\pi_k = p(M_k=1)$, for each node k with children i and j. This can be compute as follows: initialize $d_i = \gamma$ and $\pi_i = 1$ for each leaf node i; then as the tree is built, for each node k, compute $d_k = \gamma \cdot \Gamma(n_k) + d_i d_j$, and $\pi_k = \frac{\gamma \Gamma(n_k)}{d_k}$, where i and j are k's left and right children. Hence, we can efficiently compute (2) recursively from the leaf nodes to the root node as the tree is built.

Now, we can optimize the hyperparameters as we possess the gradient of them. Then, we can apply an Expectation maximization (EM) algorithm. In the E-step, we calculate the best tree structure given the current hyperparameters $\lambda_t = (\alpha_t, \beta_t)$. After that, in the M-step, we find the best hyperparameters for the current best tree, using a gradient-descent approach (e.g. $\lambda_{t+1} = \lambda_t - \eta_t \cdot g(\lambda_t)$, where η_t is the learning rate and $g(\lambda_t)$ is the gradient of the hyperparameters at the time t). This procedure is repeated until convergence. However, the EM algorithm finds a local optima (not necessarily global optima). Then, the results are sensitive to the initial values of the hyperparameters and, it can be overcomed by multiple random restarts.

⊕ Note: The hyperparameters β were renamed to $\lambda = (\alpha, \beta)$ to agree with item (a) and item (b).

2. Consider a two dimensional Gaussian distribution $p(x)$ centered at $(0,0)$. Let the variances in the principal directions be σ_x^2 and σ_y^2.

   (a) What is the optimal variance if this distribution is approximated by a spherical Gaussian (that is its variance is the same in both principal directions) with variance σ_q^2, optimized by the reverse KL? (Show your work.) (1½)

First, we need to define our parameters as follows:

$$\mu_q = \mu_p = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \; ; \; \Sigma_q = \begin{pmatrix} \sigma_{q_1}^2 & 0 \\ 0 & \sigma_{q_2}^2 \end{pmatrix} \; ; \; \Sigma_p = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Then, our distributions $q(x)$ and $p(x)$ are:

$$q(x) = \mathcal{N}(x|\mu_q, \Sigma_q) \quad \text{and} \quad p(x) = \mathcal{N}(x|\mu_p, \Sigma_p)$$

The reverse KL divergence between two bivariate Gaussian distributions is given by:

$$KL(q\|p) = \int_{-\infty}^{\infty} q(x) \cdot \log\left(\frac{q(x)}{p(x)}\right) dx = \int q(x) \cdot \log\left[\frac{\frac{1}{2\pi \cdot |\Sigma_q|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\left(x^T \Sigma_q^{-1} x\right)\right\}}{\frac{1}{2\pi \cdot |\Sigma_p|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\left(x^T \Sigma_p^{-1} x\right)\right\}}\right] dx$$

$$= \int q(x) \cdot \log\left(\frac{|\Sigma_p|}{|\Sigma_q|}\right)^{1/2} dx + \int q(x) \cdot \left[-\frac{1}{2}\left(x^T \Sigma_q^{-1} x\right) + \frac{1}{2}\left(x^T \Sigma_p^{-1} x\right)\right] dx$$

$$= \frac{1}{2} \cdot \log\left(\frac{|\Sigma_p|}{|\Sigma_q|}\right) + \frac{1}{2} \int q(x) \cdot \left[\left(x^T \Sigma_p^{-1} x\right) - \left(x^T \Sigma_q^{-1} x\right)\right] dx \quad (1)$$

However, the scalar inner product can be reorder as follows:

$$x^T \Sigma^{-1} x = tr\left(x^T \Sigma^{-1} x\right) = tr\left(\Sigma^{-1} x^T x\right) = tr\left(\Sigma^{-1} x x^T\right) \quad (2)$$

tr = Trace operation

Then, replacing (2) in (1):

$$KL(q\|p) = \frac{1}{2} \cdot \log\left(\frac{|\Sigma_p|}{|\Sigma_q|}\right) + \frac{1}{2} \int q(x) \left[tr\left(\Sigma_p^{-1} x x^T\right) - tr\left(\Sigma_q^{-1} x x^T\right)\right] dx$$

As $x x^T = \Sigma_q$, we have:

$$KL(q\|p) = \frac{1}{2} \log\left(\frac{|\Sigma_p|}{|\Sigma_q|}\right) + \frac{1}{2} \int q(x) \cdot tr\left(\Sigma_p^{-1} \Sigma_q\right) dx - \frac{1}{2} \int q(x)\, tr\left(\underbrace{\Sigma_q^{-1} \Sigma_q}_{I_2 = \text{Identity matrix } 2\times 2}\right) dx$$

$$= \frac{1}{2} \log\left(\frac{|\Sigma_p|}{|\Sigma_q|}\right) + \frac{1}{2} tr\left(\Sigma_p^{-1} \Sigma_q\right) - \frac{1}{2} tr(I_2)$$

Therefore, the reverse KL divergence between two bivariate Gaussians distributions is:

$$KL(q\|p) = \frac{1}{2} \log\left(\frac{|\Sigma_p|}{|\Sigma_q|}\right) + \frac{1}{2} tr\left(\Sigma_p^{-1} \Sigma_q\right) - 1 \quad (3)$$

Now, optimizing the variance σ_q^2 means to minimize the KL divergence:

$$\frac{\partial KL(q\|p)}{\partial \Sigma_q} = -\frac{1}{2} \Sigma_q^{-1} + \frac{1}{2} \Sigma_p = 0 \quad (4)$$

To find the optimal σ_q^2, we have to solve (4). But, let us define the form of the inverse of a covariance matrix (i.e. precision matrix):

$$\Sigma^{-1} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}^{-1} = \frac{1}{\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2} \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix} \quad (5)$$

If we have a diagonal covariance matrix, we have:

$$\Sigma^{-1} = \begin{pmatrix} \sigma_{q_1}^2 & 0 \\ 0 & \sigma_{q_2}^2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\sigma_{q_1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q_2}^2} \end{pmatrix} \quad (6)$$

Replacing (5) as Σ_p^{-1} and (6) as Σ_q^{-1} in (4), we have:

$$\begin{pmatrix} \frac{1}{\sigma_{q_1}^2} - \frac{1}{\sigma_x^2 - \rho^2\sigma_x^2} & \frac{\rho}{\sigma_x\sigma_y - \rho^2\sigma_x\sigma_y} \\ \frac{\rho}{\sigma_x\sigma_y - \rho^2\sigma_x\sigma_y} & \frac{1}{\sigma_{q_2}^2} - \frac{1}{\sigma_y^2 - \rho^2\sigma_y^2} \end{pmatrix} = 0$$

However, we want to find the results for a spherical Gaussian distribution ($\sigma_{q_1}^2 = \sigma_{q_2}^2 = \sigma_q^2$). Then, we have to take the trace in order to find the optimal σ_q^2 of a spherical Gaussian:

$$\frac{1}{\sigma_q^2} - \frac{1}{\sigma_x^2 - \rho^2\sigma_x^2} + \frac{1}{\sigma_q^2} - \frac{1}{\sigma_y^2 - \rho^2\sigma_y^2} = 0$$

$$\frac{2}{\sigma_q^2} = \left(\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2} \right)$$

Hence, the optimal variance (σ_q^2) is:

$$\sigma_q^2 = 2\left(\frac{\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2}{\sigma_x^2 + \sigma_y^2} \right)$$

(b) If we instead optimize the forward KL (1½)

$$\mathrm{KL}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x; \sigma_q^2)} \, \mathrm{d}x, \tag{26}$$

what would be the optimal value for σ_q^2? (Show your work.)

From item (a), we found the expression for the reverse KL divergence (Eq. 3). Similarly, we can get the forward KL divergence by changing the indexes ($q \to p$ and $p \to q$). Then, the forward KL divergence is:

$$KL(p \| q) = \frac{1}{2} \log\left(\frac{|\Sigma_q|}{|\Sigma_p|}\right) + \frac{1}{2} tr\left(\Sigma_q^{-1} \Sigma_p\right) - 1 \quad (7)$$

To find the optimal σ_q^2, we have to minimize the forward KL divergence:

$$\frac{\partial KL(q \| p)}{\partial \Sigma_q} = \frac{1}{2} \Sigma_q^{-1} + \frac{1}{2} \frac{\partial}{\partial \Sigma_q} tr\left(\Sigma_q^{-1} \Sigma_p\right)$$

According to the matrix cookbook:

$$\frac{\partial X^{-1} A}{\partial X} = X^{-1} A X \; ; \; \frac{\partial tr(X)}{\partial X} = \frac{tr(\partial X)}{\partial X} \; ; \; tr(x^T \Sigma^{-1} x) = x^T \Sigma^{-1} x$$

$$\Sigma_q^{-T} = \Sigma_q^{-1}$$

$$= \frac{1}{2} \Sigma_q^{-1} + \frac{1}{2}\left(\Sigma_q^{-1} \Sigma_p \Sigma_q^{-1}\right) = 0 \quad (8)$$

Replacing the definition of Σ_p and (6) as Σ_q^{-1} (both from item (a)), we have

$$\begin{bmatrix} \frac{1}{\sigma_{q1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q2}^2} \end{bmatrix} - \begin{bmatrix} \frac{1}{\sigma_{q1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q2}^2} \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_{q1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q2}^2} \end{bmatrix} = 0$$

$$\begin{bmatrix} \frac{1}{\sigma_{q1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q2}^2} \end{bmatrix} - \begin{bmatrix} \frac{\sigma_x^2}{\sigma_{q1}^2} & \frac{\rho\sigma_x\sigma_y}{\sigma_{q1}^2} \\ \frac{\rho\sigma_x\sigma_y}{\sigma_{q2}^2} & \frac{\sigma_y^2}{\sigma_{q2}^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_{q1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q1}^2} \end{bmatrix} = 0$$

$$\begin{bmatrix} \frac{1}{\sigma_{q1}^2} & 0 \\ 0 & \frac{1}{\sigma_{q2}^2} \end{bmatrix} - \begin{bmatrix} \frac{\sigma_x^2}{\sigma_{q1}^4} & \frac{\rho\sigma_x\sigma_y}{\sigma_{q1}^2\sigma_{q2}^2} \\ \frac{\rho\sigma_x\sigma_y}{\sigma_{q1}^2\sigma_{q2}^2} & \frac{\sigma_y^2}{\sigma_{q2}^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_{q1}^2} - \frac{\sigma_x^2}{\sigma_{q1}^4} & -\frac{\rho\sigma_x\sigma_y}{\sigma_{q1}^2\sigma_{q2}^2} \\ -\frac{\rho\sigma_x\sigma_y}{\sigma_{q1}^2\sigma_{q2}^2} & \frac{1}{\sigma_{q2}^2} - \frac{\sigma_y^2}{\sigma_{q2}^4} \end{bmatrix} = 0$$

Again, we want to find the optimal variance for a spherical Gaussian distribution ($\sigma_{q1}^2 = \sigma_{q2}^2 = \sigma_q^2$). Then, we have to take the trace in order to find the optimal σ_q^2 of a spherical Gaussian:

$$\frac{1}{\sigma_q^2} - \frac{\sigma_x^2}{\sigma_q^4} + \frac{1}{\sigma_q^2} - \frac{\sigma_y^2}{\sigma_q^4} = \frac{2}{\sigma_q^2} - \frac{(\sigma_x^2 + \sigma_y^2)}{\sigma_q^4} = 0$$

Hence, the optimal variance (σ_q^2) is:

$$\sigma_q^2 = \frac{(\sigma_x^2 + \sigma_y^2)}{2}$$