CS 215 : Data Analysis and Interpretation

(Instructor : Suyash P. Awate)

Sample Questions

---

1. (a) For what kind of estimators, biased or unbiased, does the Cramer-Rao lower bound theorem apply ?

   (b) For what kind of estimators, biased or unbiased, does the Bayesian Cramer-Rao lower bound theorm apply ?

   (c) Is the maximum-likelihood estimator of the univariate Gaussian mean (when variance is known) an efficient estimator ? Prove or disprove.

   (d) Suppose you are given two N-sized data samples $\{x_n\}_{n=1}^{N}$ and $\{y_n\}_{n=1}^{N}$. It is known that each sample point x_n is drawn independently from a Gaussian with mean μ_X and variance σ^2. It is known that each sample point y_n is drawn independently from a Gaussian with mean μ_Y and variance σ^2. For a fixed finite sample size N, suppose you define a statistic that is the distance between the sample-mean estimates, i.e., $|\mu_X - \mu_Y|$.

   • Give a clear and precise way to estimate the distribution / histogram of $|\mu_X - \mu_Y|$ over different samples, for the fixed finite sample size N, when $\mu_X = \mu_Y$.

   • How can you use the aforementioned histogram to test if two N-sized samples have been drawn from Gaussians with the same mean or different means ? Note: you are given that the variances of the two Gaussian are equal to σ^2.

   • Can you use the aforementioned histogram to test if two M-sized, where $M \neq N$, samples have been drawn from Gaussians with the same mean or different means ? Note: you are given that the variances of the two Gaussians are equal to σ^2. Argue why or why not.

   • Can you use the aforementioned histogram to test if two N-sized, samples have been drawn from Gaussians with the same mean or different means, when you are told that variances of the two Gaussians are equal to α^2, where $\alpha \neq \sigma$? Argue why or why not.

   (e) Consider a Gaussian mixture probability density function (PDF) $P(X) := \sum_{n=1}^{N} w_n G(X; \mu_n, \sigma_n^2)$.

   (f) Consider Bayesian inference using the posterior probability density function $P(\theta)$ on $\theta \in \mathbb{R}$, using which you want to infer a value $\widehat{\theta} \in \mathbb{R}$ using a specific loss function deined as follows:

   $L(\widehat{\theta}|\theta) := w(\theta - \widehat{\theta})$, when $\theta \geq \widehat{\theta}$
   $L(\widehat{\theta}|\theta) := (1 - w)(\theta - \widehat{\theta})$, when $\theta < \widehat{\theta}$

   Here, w is a fixed specified real-valued scalar and $w \in [0, 1]$.

   For this loss function,

   • Define the risk function.

   • Derive the value $\widehat{\theta}$ that minimizes the risk function ? Specify $\widehat{\theta}$ in the simplest possible terms, without including any expectations.

(g) Suppose you want to build a classifier to seperate (i) pictures of human faces (say, passport-size photographs) from (ii) pictures (same size images as that of human faces) of faces of other living beings in the entire universe. The class of other living beings includes animals and birds on the earth and living beings (unseen) in other parts of the universe. You know that the distribution of human-face images is multivariate Gaussian. You cannot assume anything about faces that you havenąŕt seen, i.e., have no data. Describe clearly and precisely an implementable algorithm for:

- Training / learning a probabilistic model that will act as the classifier.
- Applying the learned model to classify an image as human or non-human.

(h) Suppose you want to classify a set $\mathcal{S}$ of high-quality images of handwritten digits in some script (0 to 9). Each image has N pixels and, thus, the space of representation of these images is $\mathbb{R}^N$. You are told that all images for any particular digit lie in some hyperplanar subspace $\mathbb{R}^D$ of $\mathbb{R}^N$, where $D \ll N$. Different digits correspond to different non-intersecting subspaces of possibly different dimensions D. You arenąŕt told anything about the distribution of the images of the digit within its subspace and you are forbidden to model the within-subspace distribution.

You are given a sufficiently large training set $\mathcal{T}$ of high-quality images of digits, with labels indicating the digit corresponding to the image. The set $\mathcal{S}$ isnąŕt associated with such labels. You must use all the aforementioned information (without any additional assumptions on the distributions of images) to solve the following questions.

- Design an algorithm to classify each image in the set $\mathcal{S}$ as even or odd.
- Now, assume that the images in another set $\mathcal{U}$ also need to be classified, but this image data is corrupted with measurement errors such that an independent random standard-normal perturbation gets incorporated into the measurement at each pixel. Design two PDFs $P(image|even)$ and $P(image|odd)$ that indicate probability densities of any image coming from the even set and the odd set, respectively. Use the PDFs designed in the previous part of this question to assign a probability to an image in $\mathcal{U}$ of being even (or odd).

(i) Consider a univariate real-valued random variable X with an associated probability density function (PDF) $P(X)$ having mean μ, median m, and standard deviation σ. In each of the following questions, you will be given two expressions. For each pair of expressions, prove or disprove (i) if the expressions are equal, (ii) if the first expression takes values always less than (or $\leq$) the other, (iii) the first expression takes values greater than (or $\geq$) the other, or (iv) there isnąŕt any such relationship between the expressions. You can use theorems derived in class, but state all theorems and arguments in the proof clearly.

- Expressions $|E[X-m]|$ and $E[|X-m|]$, where $|\cdot|$ is the absolute-value function.
- Expressions $E[|X-m|]$ and $\sqrt{E}[(X-\mu)^2]$.
- Expressions $E[|X-m|]$ and $E[|X-\mu|]$.
- Expressions $\mu - m$ and σ.

(j) Suppose you are given a data sample $\{x_n\}_{n=1}^N$, where each x_n is drawn from a probability density function (PDF) that is multivariate Gaussian. You want to build a sampler for generating more observations x from that same PDF. Describe, and clearly justify, an algorithm for the sampler.

(k) Consider two random variables U_1 and U_2 that are independent, with each having a uniform distribution over $(0,1)$. Consider the two transformed random variables $Z_1 :=$

$\sqrt{-2\log U_1}\cos(2\pi U_2)$ and $Z_2 := \sqrt{-2\log U_1}\sin(2\pi U_2)$. Using the theory of transformation of random variables, (i) derive the joint probability density function (PDF) for the bivariate random variable $Z := (Z_1; Z_2)$ and (ii) derive the marginal PDFs $P(Z_1)$ and $P(Z_2)$.

(l) Maximum-likelihood estimators are always unbiased ? Prove or disprove.

(m) Suppose you are performing principal component analysis (PCA) on data from several individuals, where the data from individual i involves D observations $x_i \in \mathbb{R}^D$, e.g., the individual's height, weight, blood cell counts, etc.

(i) Can changing the units of one of the measurements, say, height from meters to centimeters (i.e., the height changing from 1 to 100) modify the first principal component, if the other measurements' units remained unchanged ?

(ii) What is a (standard) technique to make units of different measurements commensurate in PCA ? Clearly state the algorithm and justify it briefly.

(n) Suppose you are performing principal component analysis (PCA) on data $\{x_i\}_{i=1}^{I}$ whose mean is non-zero. Suppose you compute the covariance without subtracting the mean from (i.e., centering) the data.

(i) Will this affect the computed principal components /directions ? If so, explain clearly or illustrate clearly with a picture. If not, prove it.

(ii) Will this affect the computed variances along the principal components ? If so, explain clearly or illustrate with a picture. If not, prove it.

(o) Suppose you are performing principal component analysis (PCA) on data from several individuals, where the data from individual i involves D observations $x_i \in \mathbb{R}^D$.

(i) Suppose the number of individuals is $N < D$. In this case, what can you say about the eigenvalues of the covariance matrix ?

(ii) Suppose, during PCA, you find that all eigenvalues of the covariance matrix are identical and non-zero ? What does this tell us about the data in the context of its (i) scatter plot, (ii) modes of variation, and (iii) relationships between the variables?

(p) Suppose you want to evaluate the Kullback-Leibler divergence between distributions $P(X)$ and $Q(X)$ in a real-world application, where X takes values in a high-dimensional space. Assume $P(X)$ and $Q(X)$ have support over the entire domain. However, (1) the underlying integral is analytically intractable and (2) numerical approximation of the integral via a Riemann sum (i.e., finite sum approximation to the area under the "curve") is also computationally infeasible because of the large dimensionality involved.

(i) In such a case, describe a computationally-efficient statistical method to approximate the value of the integral. The method must allow estimation of the integral value up to any arbitrary level of accuracy. Make reasonable assumptions on the real-world distributions.

(ii) Describe under what circumstances will the statistical method be significantly more efficient over a Riemann-sum approximation.

(q) Consider a continuous random variable X with the distribution $P(x) := [\pi(1+x^2)]^{-1}, \forall x \in (-\infty, \infty)$. Are the mean and variance of this distribution finite ? If so, find them. If not, prove so.

(r) To uniquely specify a **general** Poisson distribution (i.e., to be able to specify the probability density function of the distribution), what is the least number of parameter values (scalars) that you need to specify ? What are they ?

(s) To uniquely specify a **general** univariate Gaussian distribution, what is the least number of parameter values (scalars) that you need to specify ? What are they ?

(t) To specify a **general** bivariate Gaussian distribution, what is the least number of scalars that you need to specify to uniquely determine the Gaussian parameters ? What are they ?

(u) To specify a **general** multivariate Gaussian distribution in D dimensions, what is the least number of scalars that you need to specify to uniquely determine the Gaussian's parameters ? Justify your answer.

(v) In which of the following scenarios can Bayesian parameter estimation be significantly advantageous over maximum-likelihood parameter estimation: (i) when we have very small set of observed data, (ii) when we have a very large set of observed data. Justify briefly.

(w) Can Bayesian estimation ever perform worse than maximum-likelihood estimation ? If not, explain why. If so, give an example.

(x) Consider a random variable X having a Normal distribution with mean μ and variance σ^2. Consider a transformation $Y := |X|$. What is the distribution of the random variable Y; give a mathematical expression and justify it (you needn't give a mathematical derivation) ?

(y) In Bayesian statistics, what is the **posterior predictive** distribution ? Give its expression in terms of an integral. How does a conjugate prior help in evaluating this integral ?

(z) What is the motivation behind estimating unknown parameter values by maximizing the likelihood function ?

2. (a) When and why can Bayesian estimation be more useful: (i) when data is limited or (ii) when data is plenty ? Will parameter estimates will always be improved using Bayesian estimation, as compared to maximum likelihood estimation ?

(b) What is the idea behind conjugate priors ? When is a prior said to be a conjugate prior ? Why are conjugate priors useful ?

(c) What are the maximum-likelihood estimates for the mean and covariance matrix of the multivariate Gaussian ? Give mathematical expressions for both.

(d) Suppose you have a random variable X, with unknown distribution, to which you add an independent random variable Y_1 with a Gaussian distribution with zero mean and variance σ_1^2 and another independent random variable Y_2 with a Gaussian distribution with zero mean and variance σ_2^2. Thus, $Z := X + Y_1 + Y_2$.

Suppose you generate another random variable $W := X + Y_3$ where Y_3, independent of X, has a Gaussian distribution with zero mean and variance $\sigma_3^2 := \sigma_1^2 + \sigma_2^2$.

Will Z and W have the same distribution ? Why or why not ?

(e) Consider independent random variables X_1 and X_2. X_1 is Gaussian distributed with mean 1 and variance 1. X_2 is Gaussian distributed with mean 2 and variance 2. Define $Y_1 := 2X_1 + X_2$ and $Y_2 := X_1 - X_2$. Give a mathematical form for the joint distribution $P(Y_1, Y_2)$.

(f) Consider the product of two multivariate Gaussians $G(x; \mu_1, C)$ and $G(x; \mu_2, C)$ with means μ_1 and μ_2 and covariance C. Take the resulting function $f(x) := G(x; \mu_1, C) G(x; \mu_2, C)$ and normalize it (so that it integrates to 1) to give the probability density function $g(x)$. Give a mathematical expression for $g(x)$.

3. Consider data $x \in \mathbb{R}^3$ that is represented in a 3D space, i.e., using 3 coordinates (x^1, x^2, x^3). Consider that we observe a large dataset $x_1, x_2, \cdots, x_n$, where $n = 100$. Suppose each datum x_i is known to be such that its coordinates exhibit a functional relationshop such that

$ax_i^1 + bx_i^2 + cx_i^3 + d = 0$, for the same $a, b, c, d \in \mathbb{R}$. Thus, although the data lies in a 3D space, the underlying degrees of freedom in the variability of the data is lower than 3. Your goal is to represent each datum using just as many coordinates as the number of degrees of freedom, without any loss of information, i.e., the mapping from the 3D space to the lower-dimensional space should be linear and distance perserving (Euclidean distance between any 2 points in the original 3D space should equal Euclidean distance between the mapped points in the lower-dimensional space).

- How many degrees of freedom are present in the variability in the dataset ?

- Only using the concepts covered in our class, give a detailed algorithm to find the lower-dimensional representation of the 3D data.

4. Consider a continuous random variable U having a uniform distribution $P(U)$ over $[0, 1]$.

   - Define a random variable $Z := \tan(\pi(U - 0.5))$. Derive the distribution for Z. This distribution is important for several applications in physics. Let us call it the standard-$\mathcal{C}$ distribution.

   - Define a random variable $Y := 1/Z$. Derive the distribution for Y.

   - Define a random variable $X := a + bZ$, where a and b are real-valued constants with a as the location parameter and b as the scale parameter. Derive the distribution $P(X)$ for X. Let us call $P(X)$ an instance from the general-$\mathcal{C}$ family of distributions.

   - Know that the sum of two independent random variables Z_1 and Z_2, each with distribution identical to $P(Z)$, is another random variable $Z_3 := (Z_1 + Z_2)/2$ with distribution identical to $P(Z)$. Then, if $X_1, \cdots, X_K$ is a sequence of random independent variables with distributions identical to $P(X; a, b)$, what is the distribution of $\overline{X} := (1/K)(X_1 + \cdots + X_K)$.

   - Using the results used so far in this question, what will the convolution of any two general-$\mathcal{C}$ distributions be ? Give a mathematical expression for the type of the resulting distribution.

   - State the Central limit theorem.

   - Will the Central limit theorem be satisfied for random variables belonging to the general-$\mathcal{C}$ family of distributions ? Why or why not ? Justify your answer using a logical sequence of precise arguments (don't miss any step).

5. Consider that an individual's height Y is a linear function of the weight X such that $Y := MX + C$. Consider an experiment involving N individuals where the weight and height measurements are erroneous such that, for individual i, the measured height is $H_i := Y_i + \alpha_i$ and the measured weight is $W_i := X_i + \beta_i$, where the errors α_i and β_i are independent random numbers drawn from Gaussian distributions with mean zero and variances σ_α^2 and σ_β^2, respectively. You want to estimate the true values of M and C, using the N erroneous observations.

   - Formulate the problem of inferring M and C from the data as a maximum-likelihood estimation problem. What is the likelihood function ? What is the optimization problem you need to solve ?

   - Derive a system of equations that you'll need to solve to get the optimal values of M and C. No need to solve the equations; just derive.

   - Suppose you have priors on M and C, which are Gaussians with means zero and variances σ_M^2 and σ_C^2 respectively. Formulate the problem of inferring M and C from the data as a Bayesian estimation problem. What is the posterior function ? What is the optimization problem you need to solve ?

- Suppose your friend feels that the relationship between height and weight should be linear with $C = 0$, i.e., $Y := MX$. Your friend performs Bayesian estimation to fit this new model (where C is fixed to zero) to the data. Now, we have two models, say, S_{MC} and S_M. How will we choose which model fits the data better and should be used for subsequent analysis — describe a Bayesian approach to pick one model over the other ?

6. A whitening transformation is a linear transformation that transforms a vector X of random variables with a known (invertible) covariance matrix into a vector Y of new variables whose covariance matrix is the identity matrix. Given a large number of observations of a random vector X that has a multivariate Normal distribution, derive a whitening transformation and provide a detailed algorithm for computing it.

7. Consider two independent random variables X and Y that have an exponential distribution with $\lambda = 1$. What is the distribution of the random variable $Z := X + Y$? Give a detailed derivation.

8. Is the convolution of two probability density functions (PDFs) $f(\cdot)$ and $g(\cdot)$ always guaranteed to (i) be real valued, (ii) be non-negative, and (iii) integrate to 1 ? Prove or disprove.

9. Consider continuous random variables X and Y, with probability density functions (PDFs) $P(X)$ and $P(Y)$, respectively.

   - Define a random variable $Z := X + Y$. What is the PDF $P(Z)$, in terms of $P(X)$ and $P(Y)$?

   - Just for this sub-question, assume $P(X)$ and $P(Y)$ to be Gaussians with some arbitrary means and variances. Define a random variable $Z := X + Y$. What is the parametric form of the PDF $P(Z)$?

   - Just for this sub-question, assume $P(X)$ and $P(Y)$ to be uniformly distributed over $[0, 1]$. Define a random variable $Z := X + Y$. What is the functional form of the PDF $P(Z)$?

   - Just for this sub-question, assume $P(X)$ and $P(Y)$ to be Gaussians. Define a PDF $P(Z) := \eta P(X)P(Y)$, where η is a normalization constant. What is the parametric form of the PDF $P(Z)$?

   - Just for this sub-question, assume you have data $(x_i, y_i) \in \mathbb{R}^2$ with the property that $x_i^2 + y_i^2 = 1$, for all i. Consider the data spread *uniformly* on a circle of radius 1. Suppose you fit a bivariate Gaussian to this data, what would be the estimate of the mean vector ? What would be the values of the non-diagonal terms in the 2×2 covariance matrix ? Is the bivariate Gaussian model a good fit to this data — why or why not ?

10. Let $\{X_i\}_{i=1}^n$ model a random sample drawn from the exponential distribution $P(x; \theta)$ with parameter θ. Assume a Gamma prior on the θ of the form $P(\theta; \alpha, \beta)$ with parameters α, β.

    (i) Find the posterior distribution on θ.

    (ii) Is the prior a conjugate prior ? Why or why not ?

    (iii) Given n observations $x_1, \cdots, x_n$, find the predictive distribution of the next observation X_{n+1}; simplify / evaluate the integral as much as as possible. Recall that the predictive distribution of X_{n+1} given observations $\{x_i\}_{i=1}^n$ is the conditional distribution $P(X_{n+1}|\{x_i\}_{i=1}^n$.

11. (i) Prove or disprove: If two random variables X_1, X_2 are uncorrelated, then they are independent.

    (ii) Prove or disprove: If a bivariate Gaussian random variable $Y := (Y_1, Y_2)$ has a diagonal covariance matrix, then the random variables Y_1, Y_2 are correlated.

(iii) Prove or disprove: If a bivariate Gaussian random variable $Z := (Z_1, Z_2)$ is such that the random variables Z_1, Z_2 are uncorrelated, then the random variables Z_1, Z_2 are independent.

12. Assume that random variable X a probability density function $P(X) = 3(1-x)^2$ for $x \in (0, 1)$. Consider the transformed random variable $Y := (1 - X)^3$.

    (i) Find an analytical formula for the cumulative distribution function (CDF) of Y, i.e., $P(Y \leq y)$, starting with the definition of the CDF.

    (ii) Find an analytical formula for the probability density function of Y, i.e., $P(Y = y)$, as the derivative of the CDF.

13. Assume a Gaussian distribution $G(X)$ with unknown mean μ and known variance σ^2. You are given observed data $x_1, x_2, \cdots, x_n$ drawn from the Gaussian distribution $G(X)$. Assume a prior on the unknown mean μ to be another Gaussian with mean μ_p and variance σ_p^2.

    (i) Find the analytical form of the posterior distribution $P(\mu | x_1, \cdots, x_n)$.

    (ii) Ignoring the prior, describe a strategy to come up with an analytical form for the predictive distribution $P(x | x_1, x_2, \cdots, x_n)$ for any arbitrary unobserved x drawn from $G(X)$. Give an analytical expression for this predictive distribution $P(y | x_1, x_2, \cdots, x_n)$.

14. Consider a random variable X with a Gaussian probability density function with unknown mean μ and known variance σ^2.

    - Consider prior information on μ that indicates that μ can have exactly three possible values (say, μ_1, μ_2, μ_3) with equal probability. How will you use this prior information to compute $P(x)$ given that x is drawn / sampled from $P(X)$? Give a mathematical expression for $P(x)$.

    - Consider prior information on μ in the form of a distribution on μ, which is a uniform distribution over (a, b). How will you use this prior information to compute $P(x)$ given that x is drawn / sampled from $P(X)$? Give a mathematical expression for $P(x)$.

15. Consider U_1 and U_2 as independent random variables with a uniform distribution over $(0, 1)$. Define transformed random variables $V_1 := 2U_1 - 1$ and $V_2 := 2U_2 - 1$. Define another random variable $S := (V_1^2 + V_2^2)$ iff $V_1^2 + V_2^2 < 1$.

    - Given that $S < 1$, prove that S is uniformly distributed over $(0, 1)$.

    - Given that $S < 1$, prove that $\Theta := \arctan(V_2, V_1)$ is uniformly distributed over $(0, 2\pi)$.

16. Suppose you collect multivariate data from two groups, say, group A and group B, which are known to differ in some way. The data are generated independently. The probability density function for each datum in group A is known to be Gaussian with mean μ^A and covariance matrix C^A. The probability density function for each datum in group B is known to be Gaussian with mean μ^B and covariance matrix C^B. The data from group A is $\{x_1^A, x_2^A, \cdots, x_N^A\}$. The data from group B is $\{x_1^B, x_2^B, \cdots, x_M^B\}$.

    (a) For a new datum x whose group is unknown, how can you use the data to find the likelihood of y being drawn from each group, i.e., $P(x | x$ drawn from Group A$)$ and $P(x | x$ drawn from Group B$)$? Give mathematical expressions to evaluate these probabilities.

    (b) If you wanted to classify / label the new datum x to exactly one of those groups, how can you do so using the aforementioned likelihoods ? Describe an algorithm and justify it.

    (c) If the covariance matrices C^A, C^B are both identity, what is the locus of points x that have equal likelihoods of being drawn from group A and group B ? Derive an expression for this locus and describe its geometry in simple words.

(d) If the covariance matrices C^A, C^B are $(\sigma^A)^2 I, (\sigma^B)^2 I$, where I is the identity matrix, what is the locus of points x that have equal likelihoods of being drawn from group A and group B ? Derive an expression for this locus.

17. Prove that the Mahalanobis distance $d(x,y) := \sqrt{(x-y)^\top C^{-1}(x-y)}$, where C is any symmetric positive definite matrix, satisfies the triangular inequality, i.e., $\forall x, y, z : d(x,y) \leq d(y,z) + d(z,x)$.

18. Consider the data sample $\{x_i\}_{i=1}^N$ drawn from some (*not* necessarily Gaussian) distribution with zero mean. Consider that the covariance matrix has an eigen-decomposition, say, $Q\Lambda Q^{-1}$, with distinct positive eigenvalues.

• Derive an expression, in terms of Q and/or Λ, for the unit-norm direction vector v_1 that maximizes the variance of the projected data (onto subspace associated with v_1, i.e., $\{y : y = av_1, \forall a \in \mathbb{R}\}$).

• Derive an expression, in terms of Q and/or Λ, for the unit-norm direction vector v_2 that (i) is orthogonal to v_1 and (ii) maximizes the variance of the projected data (onto the subspace associated with v_2).

State each step in your derivation / proof very clearly. Make the expression as specific as possible.

19. Derive an algorithm for drawing points $x \in \mathbb{R}^D$ from a Gaussian with mean $\mu \in \mathbb{R}^D$ and covariance matrix $C \in \mathbb{R}^{D \times D}$. Justify your algorithm theoretically.

20. For a PDF $P(X;\theta)$ with parameter vector $\theta \in \mathbb{R}^D$, where $D > 1$, the Fisher information matrix with respect to the parameter vector θ is given by
$F := E_{P(X;\theta)}[(\nabla_\theta \log P(X;\theta))(\nabla_\theta \log P(X;\theta)^\top)]$,
where $\nabla_\theta \log P(X;\theta)$ is the gradient vector, represented as a $D \times 1$ column vector.

• Consider the task of estimating the mean μ and variance C of a univariate Gaussian (both parameters unknown). For this case, derive the Fisher information matrix F.

• What does F tell us about the (relative) difficulty of the tasks of estimating μ versus estimating C, for a given sample size ? i.e., what does the difficulty of each task depend on and which one is more difficult (less reliable) to estimate accurately ?

21. • Does the Cramer-Rao lower bound (CRLB) apply to biased and unbiased estimators both ? Does the Bayesian CRLB apply to estimators that are biased and unbiased ?

• Suppose you have a maximum-likelihood (ML) estimator that is unbiased and efficient (as per CRLB). Suppose, for the same problem, we have a Bayes estimator that is biased. Can the Bayes estimator improve over the ML estimator ? If so, how ? If not, why not ?

• Conjugate priors are more useful for what kind of Bayes estimates: (i) posterior mode or (ii) posterior mean ? Justify your answer clearly.

22. Consider the estimation of the inverse-variance $(1/\sigma^2)$ (typically called the "precision" τ) parameter of a univariate Gaussian. That is, the Gaussian is represented as $G(x;\mu,\tau)$ in terms of τ (not σ^2). This re-parameterization (from variance to precision) is well known and indeed beneficial for certain modeling and estimation scenarios. Now, suppose you want to construct a conjugate prior on τ.

• Propose a conjugate prior.

• Derive the expression of the posterior for a data sample of size N, along with normalizing constants. Prove that the prior is indeed a conjugate prior.

23. Consider a bivariate Gaussian random variable $X := [X_1, X_2]^\top$ with mean $\mu := [\mu_1, \mu_2]^\top$ and covariance matrix C with the element in row i and column j represented as C_{ij}.

    (i) What are the marginal probability density functions of univariate random variables X_1 and X_2 ? Derive the expressions in terms of the components of μ and C.

    (ii) What is the conditional probability density function $P(X_1|X_2 = a)$? Derive the expression in terms of the components of μ and C.

24. Consider a multivariate Gaussian in D dimensions modeled by random vector $X := [X_1, X_2, \cdots, X_D]^\top$ with mean μ and covariance matrix $C := AA^\top$.

    (i) Derive the expression of the marginal probability density function of the set of $K < D$ random variables $Y := [X_1, X_2, \cdots, X_K]^\top$, in terms of the components of mean μ and covariance C.

    (ii) What is the mean and the covariance matrix of random vector Y ?

    (iii) What is the conditional probability density function $P(X_1|X_2 = a)$ when covariance matrix C is diagonal ? Derive the expression in terms of the components of μ and C.

25. Suppose we model a n-sized data sample as $\mathbf{X_n} := \{X_1, X_2, \cdots, X_n\}$ where the i-th observation X_i is an n-vector $X_i := [X_{i1}, \cdots, X_{in}]^\top$. We know that X_{ij} and X_{ik} are independent (for any j, k) and all $X_{i\cdot}$ have the normal probability density function $G(\mu_i, \sigma^2)$. We want to estimate parameters $\{\mu_i\}_{i=1}^n$ and σ^2 given data $\{x_1, \cdots, x_n\}$. This is a model where the number of parameters (i.e., $n + 1$) increase with sample size (i.e., n). Such situations indeed arise in real-world applications and have been well studied in the literature.

    (i) Derive the maximum-likelihood estimators for the parameters.

    (ii) As the sample size n tends to ∞, do the ML estimates converge to the true values ?

26. Suppose we model a n-sized data sample as $\mathbf{X_n} := \{X_1, X_2, \cdots, X_n\}$ where the i-th observation X_i is a 2-vector $X_i := [X_{i1}, X_{i2}]^\top$. We know that X_{i1} and X_{i2} are independent and both have the normal probability density function $G(\mu_i, \sigma^2)$. We want to estimate parameters $\{\mu_i\}_{i=1}^n$ and σ^2 given data $\{x_1, \cdots, x_n\}$. This is a model where the number of parameters $(n + 1)$ increase with sample size (n).

    (i) Derive the maximum-likelihood estimators for the parameters.

    (ii) As the sample size n tends to ∞, do the ML estimates converge to the true values ?

    (iii) Now consider that real problem is to estimate the parameter σ^2 and interpret μ_i as a nuisance variable that models an individual-specific bias in the observations (x_i, y_i) for individual i. You decide to use a Bayesian approach with some prior on the variance $P(\sigma^2)$; don't explicitly assume any particular form for $P(\cdot)$. The Bayesian approach will estimate σ^2 by maximizing the posterior $P(\sigma^2|\mathbf{x_n})$ that is obtained by integrating out the nuisance variables $\{\mu_i\}_{i=1}^n$ from $P(\sigma^2, \{\mu_i\}_{i=1}^n|\mathbf{x_n})$.

    • Derive an expression for the Bayes estimate of the variance maximizing posterior $P(\sigma^2|\mathbf{x_n})$.

    • As the sample size n tends to ∞, does the Bayes estimate converge to the true σ^2 ?

27. (i) Find the Kullback-Leibler divergence KL$(G_1\|G_2)$ between two Gaussian probability density function given by $G_1(x; \mu_1, \sigma_1^2)$ and $G_2(x; \mu_2, \sigma_2^2)$.

    (ii) In Bayesian statistics, for a prior $P(\theta)$ and posterior $Q(\theta)$, the Kullback-Leibler divergence between the prior distribution and the posterior, i.e., KL$(P\|Q)$, is termed as the "surprise".

- Find an expression for the surprise in the context of Bayesian estimation of a Gaussian's mean parameter μ, when the Gaussian's variance σ^2 is known, the observed sample size is N, the sample mean of the observed data is $\overline{m}$, and the prior on the mean is $G_1(x; \mu_1, \sigma_1^2)$.

28. Consider an experiment about tossing a coin N times, with the probability of obtaining a head being μ where $0 \leq \mu \leq 1$. Consider a Gaussian prior on μ with the probability density function $G(\mu; 0.5, \sigma^2)$. Suppose in $N = 3$ flips, we observe $N = 3$ heads. Derive an expression for the maximum-a-posteriori estimate for the parameter μ, as function of σ. Evaluate that expression (approximately; upto second place of decimal) when $\sigma = 0.1$.

29. Consider the prior density $P(\theta)$ that is known to be conjugate to the likelihood $L(\theta)$. Suppose you need to estimate θ. The prior density $P(\theta)$ is unimodal but the prior information you have indicates a distribution that is multimodal. Design a prior density $Q(\theta)$, using functions $P(\cdot)$, which allows you to model multimodal densities without losing the ease of parameter estimation with conjugate priors. Ensure that $Q(\theta)$ is also a conjugate prior.

30. Derive the Jeffreys prior, to its most simplified form (using simple polynomial or exponential functions only), for the following cases:

    - Mean μ for the univariate Gaussian probability density function (PDF), when variance is known.

    - Standard deviation σ for the univariate Gaussian probability density function (PDF), when mean is known.

    - If you reparametrize the univariate Gaussian PDF by substituting $\theta := \log \sigma^2$, then find the Jeffreys prior when the mean is known.

31. The entropy $H(X)$ of a random variable X is a measure of the spread of the distribution of the random variable, defined as $H(X) := E_{P(X)}[\log P(X)]$.

    - Derive the entropy of the Bernoulli random variable as a function of the associated parameter $\theta \in [0, 1]$. Find the parameter value θ for which the entropy is maximized.

    - Derive the entropy of the univariate Gaussian random variable as a function with parameters μ and σ^2. Find the parameter values for which the entropy is maximum.

    - Derive the entropy of the multivariate Gaussian random variable as a function with parameters μ and C. Find the parameter values for which the entropy is maximum.

32. Given a dataset $\{x_i\}_{i=1}^N$, where each x_i is known to be drawn independently from a D-variate Gaussian PDF with mean μ and covariance C, give a step-by-step algorithm to:

    - Compute the estimate of the mean.

    - Compute the estimate of the covariance.

    - Compute the principal modes / directions of variation and the variances along those modes / directions.

33. - Prove that

$$E_{P(X|\theta_{\text{true}})} \left[\left(\frac{\partial}{\partial \theta} \log P(X|\theta) \Big|_{\theta_{\text{true}}} \right)^2 \right] = -E_{P(X|\theta_{\text{true}})} \left[\frac{\partial^2}{\partial \theta^2} \log P(X|\theta) \Big|_{\theta_{\text{true}}} \right].$$

    - For a Bernoulli probability mass function (PMF) on $x \in \{0, 1\}$, given by $P(x|\theta) := \theta^x (1 - \theta)^{1-x}$, derive an expression of the Fisher information in the simplest possible form without involving any expectations or integrals.

• A canonical form of the Bernoulli PMF on $x \in \{0, 1\}$ is $P(x|\theta) := \exp(\theta x - \log(1 + e^{\theta}))$. For this form, derive an expression of the Fisher information in the simplest possible form without involving any expectations or integrals.

34. • Consider a dataset $\{(x_1^i, x_2^i) \in \mathbb{R}^2\}_{i=1}^N$ of size N, where random variable $X_1 := \theta \cos(\pi\theta)$, random variable $X_2 := \theta \sin(\pi\theta)$, and $\theta \sim U(0, 1)$.

Consider a continous invertible transformation $f(\cdot) : \mathbb{R}^2 \to \mathbb{R}^2$ of the data such that $(Y_1, Y_2) := f((X_1, X_2))$.

Does there exist an $f(\cdot)$ such that the mapped dataset $\{(y_1, y_2)\}_{i=1}^N$ has a covariance matrix that is singular ? If so, define $f(\cdot)$ and prove that the resulting covariance matrix will be singular. If not, give a thereotical argument for why there cannot exist such an $f(\cdot)$.

• Consider a dataset $\{(x_1^i, x_2^i) \in \mathbb{R}^2\}_{i=1}^N$ of size N, where random variable $X_1 := a \cos(2\pi\theta)$, random variable $X_2 := a \sin(2\pi\theta)$, $a \sim U(0, 1)$, and $\theta \sim U(0, 1)$.

Consider a continous invertible transformation $f(\cdot) : \mathbb{R}^2 \to \mathbb{R}^2$ of the data such that $(Y_1, Y_2) := f((X_1, X_2))$.

Does there exist an $f(\cdot)$ such that the mapped dataset $\{(y_1, y_2)\}_{i=1}^N$ has a covariance matrix that is singular ? If so, define $f(\cdot)$ and prove that the resulting covariance matrix will be singular. If not, give a thereotical argument for why there cannot exist such an $f(\cdot)$.

35. • For the Poisson probability mass function (PMF), define (i) a conjugate prior on its "parameter" and (ii) the associated posterior (including the normalizing constant) ? Justify that the posterior has the same form as the prior.

• Derive the posterior mean. Express the posterior mean as a weighted combination of the prior mean and the likelihood-driven mean.

• Derive the posterior mode, specifying any conditions on the prior parameters under which the expression is valid. Express the posterior mode as a weighted combination of the prior mode and the likelihood-driven mode.

36. Consider two univariate Gaussians $G_1(x; \mu_1, \sigma_1^2)$ and $G_2(x; \mu_2, \sigma_2^2)$, parametrized by means μ_1 and μ_2, and variances σ_1^2 and σ_2^2.

Consider a probability density function defined as $P(x) \propto G_1(x; \mu_1, \sigma_1^2) G_2(x; \mu_2, \sigma_2^2)$, upto the normalizing constant.

Consider a probability density function defined as $Q(x) \propto G_1(x; \mu_1, \sigma_1^2) + G_2(x; \mu_2, \sigma_2^2)$, upto the normalizing constant.

• Derive the range of possible values for the mean (analytical; not empirical) of the PDF $P(\cdot)$, in terms of the parameters of the Gaussians $G_1(\cdot)$ and $G_2(\cdot)$. State the range to be as compact as possible, without losing the generality of the problem setting.

• Derive the range of possible values for the variance (analytical; not empirical) of the PDF $P(\cdot)$, in terms of the parameters of the Gaussians $G_1(\cdot)$ and $G_2(\cdot)$. State the range to be as compact as possible, without losing the generality of the problem setting.

• Derive the range of possible values for the mean (analytical; not empirical) of the PDF $Q(\cdot)$, in terms of the parameters of the Gaussians $G_1(\cdot)$ and $G_2(\cdot)$. State the range to be as compact as possible, without losing the generality of the problem setting.

37. Consider random variables X_1 and X_2 with probability density function (PDF) as the standard Normal. Consider random variable E, taking values of zero or one, with a Bernoulli distribution with parameter 0.5. Random variables X_1, X_2, and E are all independent.

Define $X_3 := aX_1 + b$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}$.

Define $X_4 := aX_1 + b + X_2$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}$.

Define $X_5 := E(aX_1 + b) + (1 - E)(cX_1 + d)$, where $c \in \mathbb{R}$ and $d \in \mathbb{R}$.

- Derive the functional form of $P(X_1, X_3)$. Find its mean and its covariance matrix. Find the eigenvectors and eigenvalues of the covariance matrix.
- Derive the functional form of $P(X_1, X_4)$. Find its mean and covariance matrix.
- Derive the functional form of $P(X_1, X_5)$. Find its mean.

38. Consider an exponential probability density function $P(x; \lambda)$, with parameter $\lambda > 0$.
    - Derive an expression for the Fisher information for the parameter λ.
    - Derive an expression for the Jeffrey's prior for the parameter λ.
    - Propose a conjugate prior $P(\lambda; \theta)$, with parameter(s) θ. Justify that the prior is a conjugate prior.
    - Is the Jeffrey's prior for the parameter λ a special case of the conjugate prior for the parameter λ ? Justify.

39. Consider independent univariate random variables V and W, both having the probability density functions (PDFs) as the uniform PDF $U(0, 1)$.
    - Derive the PDF for $X := -\log(V)$, where $|\cdot|$ is the absolute-value function.
    - State the PDF for $X_1 := -\text{sgn}(W - 0.5)\log(|V|)$, where sgn$(\cdot)$ is the sign function. Justify your answer.
    - Is the PDF for $X_2 := -\text{sgn}(V - 0.5)\log(|V|)$ the same as the PDF for X_1. Justify your answer.
    - Derive the PDF for $X_3 := BX_1$, where $B > 0$.
    - State the PDF for $X_4 := A + X_3$, where $A \in \mathbb{R}$. Justify your answer.
    - Propose a conjugate prior on B, when A is known. Justify your answer.

40. Suppose X is a discrete random variable. Suppose $f(\cdot)$ is a convex function. Suppose that $g(\cdot)$ is a function.
    - Proves the Jensen's inequality: $f(E_{P(X)}[X]) \leq E_{P(X)}[f(X)]$.
    - Show that $f(E_{P(X)}[g(X)]) \leq E_{P(X)}[f(g(X))]$.

41. Consider a real non-singular matrix A of size $M \times M$.
    - For any A, are $A^\top A$ and $AA^\top$ both symmetric positive definite matrices ? Give mathematical derivations to support your answer.
    - For any A, are the eigenvalues and eigenvectors of $A^\top A$ and $AA^\top$ the same ? Give mathematical derivations to support your answer.

42. Consider a real symmetric matrix B of size $M \times M$.
    - For any B, does B have real eigenvalues ? Give mathematical derivations to support your answer.
    - For any B, are the eigenvectors of B corresponding to distinct eigenvalues orthogonal ? Give mathematical derivations to support your answer.