Informal Notes for Mathematics for Machine Learning
Imperial College London

Lecture 9: Bayesian Inference, Continued

Lecturer: Matthew Wicker

1 Learning Objectives

In our last lecture we introduced Bayes theorem and how we can use it to express uncertainty particularly in density estimation. In this lecture, we are going to specifically derive how Bayes theorem prescribes inference in a conditional density setup (i.e., supervised learning) using our linear regression model. We will end these notes by thinking about how having a posterior distribution over our parameters yields predictive uncertainty and in lecture we discussed how uncertainty allows us to perform model selection.

2 Background

It has been a few lectures since we last reviewed our linear regression model. So here, we will first write out some brief review notes, and then will recall Bayes theorem in the context of conditional density estimation.

2.1 Linear Regression Model

Recall that our basis expanded linear regression model takes the form:

$$\hat{\boldsymbol{y}} = \phi(\boldsymbol{X})\theta$$

Where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is our feature matrix and $\phi : \mathbb{R}^n \mapsto \mathbb{R}^k$ is a basis expansion function and finally $\theta \in \mathbb{R}^{k \times m}$ is the parameter that we would like to fit. We will throughout make the simplifying assumption that $m = 1$ and that the vector $\boldsymbol{y}$ is a column vector containing all of our response variables. Though the basis expansion function allows us to capture non-linear relationships in data we call this a linear model because the relationship between the basis-expanded features and the predictions is linear w.r.t. the parameter vector.It is helpful to recall that we have computed the maximum likelihood estimate of this model as:

$$\theta^{\text{MAP}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

2.2 Bayes Theorem

In our last lecture, we introduced Bayes theorem as critical component to a probabilistic learning paradigm:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

We used this formula to discuss conjugate priors and derive the posterior for Bernoulli random variables. In this lecture, we will deal with the slightly more complex setting of conditional density estimation where we would like to ultimately reason about $p(y|\boldsymbol{x}^*, \theta)$ where $\boldsymbol{x}^*$ is some unseen feature vector.

3 Bayesian Linear Regression

In any Bayesian machine learning model, there are two critical ingredients that we must carefully consider: the likelihood and the prior. To start, we will assume that we have a single dimensional response/label.

Likelihood The likelihood in supervised learning has the critical task of modelling the observational noise of a process. Throughout our exposition of linear models, we have focused on Gaussian noise (and will continue to do so). Our likelihood with respect to a given observation (x, y) can be expressed as:

$$p(y|x, \theta) = \mathcal{N}(\phi(x)\theta, \sigma^2)$$

We can write out the full equation in matrix form for all of our observations (and plugging into the Gaussian density equation):

$$p(\boldsymbol{y}|\boldsymbol{X}, \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left(\frac{(\boldsymbol{y} - \phi(\boldsymbol{X})\theta)^\top(\boldsymbol{y} - \phi(\boldsymbol{X})\theta)}{2\sigma^2}\right)$$

We will make the simplifying assumption that σ is known.

Prior As we have seen in our previous exercises the conjugate prior for the mean of a Gaussian with known variance is a Gaussian, so we know that in this case it will be analytically convenient to select a Gaussian prior:

$$p(\theta) = \mathcal{N}(0, \tau^2\mathbf{I})$$

where $\mathbf{I}$ is the $k \times k$ identity matrix, our prior is an isotropic Gaussian with variance τ^2. We have actually seen in previous lectures that we can express this distribution as:

$$p(\theta) = \frac{1}{(2\pi\tau^2)^{k/2}} \cdot \exp\left(\frac{\theta^\top\theta}{2\tau^2}\right)$$

We will now explore two ways of deriving the Bayesian posterior for our model.

3.1 Deriving the Posterior via Densities

A concrete way to compute the posterior distribution use the method of equating coefficients. The first step is to simply plug in the densities and use algebra to simplify until we arrive at a form that looks approximately like the posterior form we expect (from conjugacy). Here, let us consider computing the posterior by taking the log of both sides of our unnormalized posterior:

$$\log(p(\theta|\mathcal{D})) \propto \log(p(\theta|\mathcal{D})) + \log(p(\theta)) \tag{1}$$

$$= \frac{1}{2\sigma^2}((\boldsymbol{y} - \phi(\boldsymbol{X})\theta)^\top(\boldsymbol{y} - \phi(\boldsymbol{X})\theta)) + \frac{1}{2\tau^2}(\theta^\top\theta) \tag{2}$$

$$= \frac{1}{2\sigma^2}\left(\theta^\top\phi(\boldsymbol{X})^\top\phi(\boldsymbol{X})\theta - 2\boldsymbol{y}^\top\phi(\boldsymbol{X})\theta + \boldsymbol{y}^\top\boldsymbol{y}\right) + \frac{1}{2\tau^2}\theta^\top\theta \tag{3}$$

$$= \frac{1}{2\sigma^2}\theta^\top\phi(\boldsymbol{X})^\top\phi(\boldsymbol{X})\theta + \frac{2}{2\sigma^2}\boldsymbol{y}^\top\phi(\boldsymbol{X})\theta - \frac{1}{2\sigma^2}\boldsymbol{y}^\top\boldsymbol{y} + \frac{1}{2\tau^2}\theta^\top\theta \tag{4}$$

$$\propto \tau^2\theta^\top\phi(\boldsymbol{X})^\top\phi(\boldsymbol{X})\theta + 2\tau^2\boldsymbol{y}^\top\phi(\boldsymbol{X})\theta - \tau^2\boldsymbol{y}^\top\boldsymbol{y} + \sigma^2\theta^\top\theta \tag{5}$$

While perhaps unintuitive at first, we can see that the above equation is quadratic in θ which is the variable our posterior density is over. Thus, we know we ought to be able to proceed with equating coefficients with a Gaussian. To start, we group terms that look like $\theta^\top\mathbf{M}\theta$, and pull out the θs:

$$= \tau^2\theta^\top\phi(\boldsymbol{X})^\top\phi(\boldsymbol{X})\theta + \sigma^2\theta^\top\theta + 2\tau^2\boldsymbol{y}^\top\phi(\boldsymbol{X})\theta + \tau^2\boldsymbol{y}^\top\boldsymbol{y} \tag{6}$$

$$= \theta^\top(\tau^2\phi(\boldsymbol{X})^\top\phi(\boldsymbol{X}) + \sigma^2\mathbf{I})\theta + 2\tau^2\boldsymbol{y}^\top\phi(\boldsymbol{X})\theta - \tau^2\boldsymbol{y}^\top\boldsymbol{y} \tag{7}$$

Before doing anymore algebra, recall the reason we chose a Gaussian prior is because we knew it was the conjugate prior to a Gaussian posterior from our previous exercises. And we know that we can write any Gaussian distribution as:

$$\log\big(\mathcal{N}(\theta;\mu,\Sigma)\big) \propto (\theta-\mu)\Sigma^{-1}(\theta-\mu) \tag{8}$$

$$= \theta^\top\Sigma^{-1}\theta - 2\theta^\top\Sigma^{-1}\mu + \mu^\top\Sigma^{-1}\mu \tag{9}$$

$$= \theta^\top\Sigma^{-1}\theta - 2\theta^\top\Sigma^{-1}\mu + \text{const.} \tag{10}$$

So, now we have two equations, the one above (Equation (10)) is that of the log of a Gaussian, and our previous Equation (7) which is the log of our posterior which we also know is Gaussian. So what we would like to do is somehow get the two equal to each other. First, we can drop the constant in equation (10) because we only have proportionality. Then, we can see that the first term in (7) looks a lot like the first term in (10). So it is natural to first set:

$$\Sigma^{-1} = (\tau^2\phi(\boldsymbol{X})^\top\phi(\boldsymbol{X}) + \sigma^2\mathbf{I})$$

. Finally, we want to take care of the next term where we have:

$$2\theta^\top\Sigma^{-1}\mu = 2\tau^2\boldsymbol{y}^\top\phi(\boldsymbol{X})\theta$$

and we know the final parameter that we must find equality for is the mean, μ. Notice that we do not need to worry about the $\boldsymbol{y}^\top \boldsymbol{y}$ term as this is a constant with respect to μ. Luckily, we already have Σ^{-1} in the terms of equation (7), so we can simply solve for μ to get that:

$$\begin{aligned} 2\theta^\top \Sigma^{-1} \mu &= 2\tau^2 \boldsymbol{y}^\top \phi(\boldsymbol{X})\theta \\ \mu &= \tau^2 \Sigma \phi(\boldsymbol{X})^\top \boldsymbol{y} \end{aligned}$$

So now that we have solved for the mean and covariance, we can say that the posterior of our Bayesian linear regression model is:

$$\mathcal{N}\big(\theta; \tau^2 \Sigma \phi(\boldsymbol{X})^\top \boldsymbol{y}, (\tau^2 \phi(\boldsymbol{X})^\top \phi(\boldsymbol{X}) + \sigma^2 \mathbf{I})^{-1}\big)$$

3.1.1 Interpreting our Posterior Values

Let us take a closer look at the posterior that we have computed. In particular, let's reason about if the expressions we have come up with for the mean and covariance matrix of our Bayesian linear regression make intuitive sense. This will not only deepen our understanding of the quantities we have just derived, but will also let us understand if there are conditions under which our inference is not well defined. Beginning with the posterior covariance matrix (in no small part because the mean relies on the covariance):

$$\Sigma = (\frac{1}{\sigma^2} \phi(\boldsymbol{X})^\top \phi(\boldsymbol{X}) + \sigma^2 \mathbf{I})^{-1}$$

We know that not all matrices are invertible, yet our covariance matrix relies on the matrix inverse of a particular quantity. We can start with the fact that $\sigma^2 \mathbf{I}$ and $\tau^2 \phi(\boldsymbol{X})^\top \phi(\boldsymbol{X})$ are positive semi-definite matrices. However, not all PSD matrices are invertible. It is a useful exercise, and is covered in lecture, to prove that the covariance we have computed is indeed invertible. Next, let's reason about the mean:

$$\mu = \tau^2 \Sigma \phi(\boldsymbol{X})^\top \boldsymbol{y}$$

One thing we know about normal distributions is that their mode, the value maximizing the p.d.f. coincides with the mean of the distribution. In previous lectures, we introduced the concept of a *maximum a posteriori* (MAP) estimate, which is the parameter which has the highest probability when we assume a prior distribution over our weights. So here, we would like to check that the quantity of our mean is equivalent to the MAP estimate we computed in previous lectures. Expanding our equation for the mean we have:

$$\begin{aligned} \mu &= \tau^2 \Sigma \phi(\boldsymbol{X})^\top \boldsymbol{y} \\ &= \tau^2 \big(\tau^2 \phi(\boldsymbol{X})^\top \phi(\boldsymbol{X}) + \sigma^2 \mathbf{I})^{-1}\big) \phi(\boldsymbol{X})^\top \boldsymbol{y} \\ &= (\phi(\boldsymbol{X})^\top \phi(\boldsymbol{X}) + \frac{\sigma^2}{\tau^2} \mathbf{I})^{-1}) \phi(\boldsymbol{X})^\top \boldsymbol{y} \end{aligned}$$

Now, if we return to lecture 5, we will see this exactly the same as the quantity we derived before. With both of these, we have done a pretty good sanity check that our posterior makes sense.

3.2 Deriving the Posterior via Joint Gaussian

Another method for arriving at the posterior distribution is to start by modelling the joint distribution of θ and $\boldsymbol{y}$ as a Gaussian distribution. It may not be clear exactly how that helps us, but recall that the product of densities that we reasoned about above was easy to interpret as a joint distribution. Moreover, there are many probability rules that allow us to manipulate joint distributions to compute their marginal and conditional distributions. Gaussian distributions have particularly nice rules for marginalizing and conditioning. We state them below:

Theorem 3.1. ***Marginalization*** *Given a Gaussian random variable:*

$$\mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}; \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{a,a}, \Sigma_{a,b} \\ \Sigma_{b,a}, \Sigma_{b,b} \end{bmatrix}\right)$$

The marginal distribution of a is given by:

$$p(a) = \mathcal{N}(a; \mu_a, \Sigma_{a,a})$$

Theorem 3.2. ***Conditioning*** *Given a Gaussian random variable:*

$$\mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}; \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{a,a}, \Sigma_{a,b} \\ \Sigma_{b,a}, \Sigma_{b,b} \end{bmatrix}\right)$$

The conditional distribution of $p(a|b)$ is given by:

$$\begin{aligned} p(a|b) &= \mathcal{N}(a; \mu_{a|b}, \Sigma_{a|a}), \\ \mu_{a|b} &= \mu_a + \Sigma_{a,b}\Sigma_{b,b}^{-1}(b - \mu_b) \\ \Sigma_{a|b} &= \Sigma_{a,a} - \Sigma_{a,b}\Sigma_{b,b}^{-1}\Sigma_{b,a} \end{aligned}$$

Now that we have established these theorems, we will show how we can use them to compute the Bayesian linear regression posterior. We start by writing out the joint distribution:

$$p(\theta, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \theta \\ \boldsymbol{y} \end{bmatrix}; \begin{bmatrix} \mathbb{E}_{p(\theta)}[\theta] \\ \mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}] \end{bmatrix}, \begin{bmatrix} \mathbb{V}[\theta], \mathbb{C}[\theta, \boldsymbol{y}] \\ \mathbb{C}[\boldsymbol{y}, \theta], \mathbb{V}[\boldsymbol{y}] \end{bmatrix}\right)$$

Ultimately, as we show in our lecture slides, we can compute the values for each entry in the joint Gaussian above and plug into our conditioning theorem to arrive at the posterior as:

$$p(\theta|\mathbf{y}) = \mathcal{N}\Big(\theta; \Phi(X)^\intercal\big[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\big]^{-1}\mathbf{y}\,, \mathrm{I}_M - \Phi(X)^\intercal\big[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\big]^{-1}\Phi(X)\Big)$$

Notice, that this form of the posterior is different than the one we arrived at via equating coefficients. In your exercises, you will be asked to prove that these two are indeed identical.

4 Posterior Predictive Distribution

We know that when we make predictions about a new, unseen point x^* in linear regression we simply plug the value into our model, i.e., multiply by our parameters, and the result is a value that we call $\hat{y}$, our prediction. However, in our Bayesian linear regression model we do not have a single set of parameters, instead we have a continuous distribution over parameters, which makes the idea of "plugging" unseen values into our model a little bit less clear. However, the key thing to understand is that in a Bayesian framework (almost) everything is treated as a probability distribution including our prediction. So what is the exact form of the probability distribution that we seek? Well we know we want the probability distribution to be over the space of outputs y and we know that we have access to the unseen value x^* and our previous dataset, $\mathcal{D}$, so we can write out the distribution we are interested in as: $p(y|x^*, \mathcal{D})$. This is known as the posterior predictive distribution. To compute this, we of course can use our basic rules of probability to arrive at an equation for this distribution in terms of quantities we already know:

$$p(y|x^*, \mathcal{D}) = \int_{\theta \in \Theta} p(y|x^*, \theta) p(\theta|\mathcal{D}) d\theta$$

Notice that this uses slightly odd notation as x^* is not a random variable. However, we include it on the conditioning side to indicate that the distribution over outputs is taken with respect to the input x^*. The first is just our likelihood and the second is our posterior distribution, both of which we know. So we can plug in our densities and then think about how to attack this integral:

$$\begin{aligned}
p(y|x^*, \mathcal{D}) &= \int_{\theta \in \Theta} p(y|x^*, \theta) p(\theta|\mathcal{D}) d\theta \\
&= \int_{\theta \in \Theta} \mathcal{N}(\phi(x^*)\theta, \tau^2) \mathcal{N}(\theta; \mu, \Sigma) d\theta \\
&\propto \int_{\theta \in \Theta} \exp\Big(\frac{1}{\tau^2}(y - w^\top x)^2\Big) \exp\Big(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\Big) d\theta \\
&= \int_{\theta \in \Theta} \exp\Big(\frac{1}{\tau^2}\big(y^2 - 2(\theta^\top \phi(x^*))y + (\theta^\top \phi(x^*))^2\big)\Big) \exp\Big(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\Big) d\theta \\
&= \int_{\theta \in \Theta} \exp\Big(\frac{1}{\tau^2}\big(y^2 - 2(\theta^\top \phi(x^*))y + (\theta^\top \phi(x^*))^2\big) - \frac{1}{2}\big(\theta^\top \Sigma \theta - 2\theta^\top \Sigma^{-1}\mu + \mu^\top \Sigma^{-1}\mu\big)\Big) d\theta \\
&\propto \int_{\theta \in \Theta} \exp\Big(\frac{1}{\tau^2}\big(y^2 - 2(\theta^\top \phi(x^*))y + (\theta^\top \phi(x^*))^2\big) - \frac{1}{2}\big(\theta^\top \Sigma \theta - 2\theta^\top \Sigma^{-1}\mu\big)\Big) d\theta
\end{aligned}$$

We will pause the derivation here and make a key observation: the above expression is quadratic in y and so, as in our derivation of the posterior, we can complete the square to arrive at the closed form normal distribution for the posterior predictive. We do not complete the derivation here and instead leave it as an exercise problem.

A Lecture 9: Bayesian Inference, Continued

Question 1 (Vector Ordering in Gaussians)**.** Consider a joint Gaussian density on a vector $\mathbf{z}$ that can be split up as $\mathbf{z} = [\mathbf{x}^\intercal, \mathbf{y}^\intercal]^\intercal$ (this notation denotes stacking):

$$p(\mathbf{z}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\mathrm{A} & \mathrm{B}\\ \mathrm{B}^\intercal & \mathrm{C}\end{bmatrix}\right). \tag{11}$$

Now consider the permuted vector $\mathbf{z}' = [\mathbf{y}^\intercal, \mathbf{x}^\intercal]^\intercal$. Show that the Gaussian distribution of $p(\mathbf{z}')$ has a mean with rows, and a covariance matrix with swapped rows and columns.

Hint: Notice that $\mathbf{z}'$ is a permuted version of $\mathbf{z}$. We can write this mathematically as $\mathbf{z}' = \mathrm{P}\mathbf{z}$, where P is a permutation matrix:

$$\mathrm{P} = \begin{bmatrix}0 & \mathrm{I}\\ \mathrm{I} & 0\end{bmatrix}. \tag{12}$$

Note: This is an important skill! Notice that the Gaussian conditioning formula in the formula sheet is only written in one way.

Question 2 (Woodbury Identity)**.** We saw that the two different ways of deriving the Gaussian posterior, gave two different results:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\Big(\boldsymbol{\theta}; \Phi(X)^\intercal\left[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\right]^{-1}\mathbf{y}\,, \qquad \mathrm{I}_M - \Phi(X)^\intercal\left[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\right]^{-1}\Phi(X)\Big)\,,$$
$$p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\Big(\boldsymbol{\theta}; \left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}\frac{1}{\sigma^2}\Phi(X)^\intercal\mathbf{y}\,, \quad \left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}\Big). \tag{13}$$

Apply the Woodbury identity to $\left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}$ to show that the two solutions are equal.

Note: This is an excellent exercise for your matrix algebra skills. The main difficulty that most uninitiated have, is that multiplication no longer commutes. You need to be careful that you consistently pre/post multiply matrices.

Question 3 (BLR Predictive)**.** For a Bayesian Linear Regression model:

1. Find $p(\mathbf{y}^*, \mathbf{y})$. $y^* = \text{unseen}$
2. Find $p(\mathbf{y}^*|\mathbf{y})$.

Question 1 (Vector Ordering in Gaussians)**.** Consider a joint Gaussian density on a vector $\mathbf{z}$ that can be split up as $\mathbf{z} = [\mathbf{x}^\mathsf{T}, \mathbf{y}^\mathsf{T}]^\mathsf{T}$ (this notation denotes stacking):

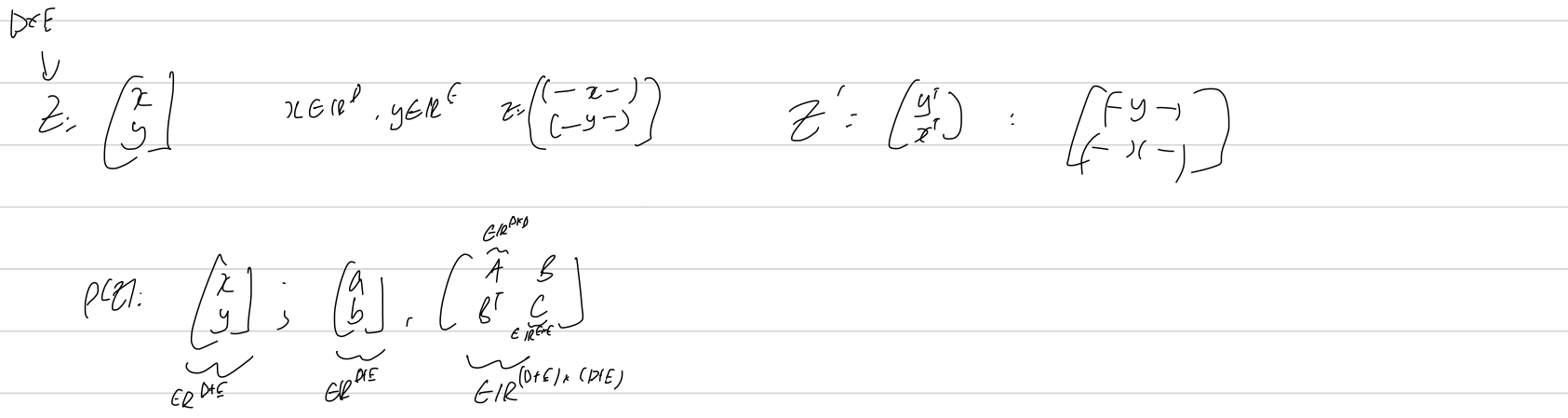
$$p(\mathbf{z}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\mathrm{A} & \mathrm{B}\\ \mathrm{B}^\mathsf{T} & \mathrm{C}\end{bmatrix}\right). \tag{11}$$

Now consider the permuted vector $\mathbf{z}' = [\mathbf{y}^\mathsf{T}, \mathbf{x}^\mathsf{T}]^\mathsf{T}$. Show that the Gaussian distribution of $p(\mathbf{z}')$ has a mean with rows, and a covariance matrix with swapped rows and columns.

Hint: Notice that $\mathbf{z}'$ is a permuted version of $\mathbf{z}$. We can write this mathematically as $\mathbf{z}' = \mathrm{P}\mathbf{z}$, where P is a permutation matrix:

$$\mathrm{P} = \begin{bmatrix}0 & \mathrm{I}\\ \mathrm{I} & 0\end{bmatrix}. \tag{12}$$

Note: This is an important skill! Notice that the Gaussian conditioning formula in the formula sheet is only written in one way.



Show $p(z') = \mathcal{N}\left(\begin{bmatrix}y\\ x\end{bmatrix}; \begin{bmatrix}b\\ a\end{bmatrix}, \begin{bmatrix}C & B^T\\ B & A\end{bmatrix}\right)$

$z' = Pz = \begin{bmatrix}0 & I\\ I & 0\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}y\\ x\end{bmatrix}$

If $x \sim \mathcal{N}(\mu, \Sigma) \rightarrow Ax \sim \mathcal{N}(A\mu, A\Sigma A^T)$ From Q1 sheet 6

Here $A = P$ $\therefore$ $\mu \rightarrow P\mu = \begin{bmatrix}0 & I\\ I & 0\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix} = \begin{bmatrix}b\\ a\end{bmatrix}$

$\Sigma \rightarrow P\Sigma P^T = \begin{bmatrix}0 & I\\ I & 0\end{bmatrix}\begin{bmatrix}A & B\\ B^T & C\end{bmatrix}\begin{bmatrix}0 & I\\ I & 0\end{bmatrix} = \begin{bmatrix}B^T & C\\ A & B\end{bmatrix}\begin{bmatrix}0 & I\\ I & 0\end{bmatrix} = \begin{bmatrix}C & B^T\\ B & A\end{bmatrix}$

Question 2 (Woodbury Identity)**.** We saw that the two different ways of deriving the Gaussian posterior, gave two different results:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\Big(\boldsymbol{\theta}; \Phi(X)^\intercal\big[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\big]^{-1}\mathbf{y}\,, \qquad \mathrm{I}_M - \Phi(X)^\intercal\big[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\big]^{-1}\Phi(X)\Big)\,,$$

$$p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\Big(\boldsymbol{\theta}; \left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}\frac{1}{\sigma^2}\Phi(X)^\intercal\mathbf{y}\,, \quad \left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}\Big)\,. \tag{13}$$

Apply the Woodbury identity to $\left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}$ to show that the two solutions are equal.

Note: This is an excellent exercise for your matrix algebra skills. The main difficulty that most uninitiated have, is that multiplication no longer commutes. You need to be careful that you consistently pre/post multiply matrices.

Woodbury Identity

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$$

$\left(I_m + \frac{1}{\sigma^2}\phi(x)^T\phi(x)\right)^{-1}$

$A \quad + \quad U \quad C \quad V$

$\left(I_m + \phi(x)^T\frac{1}{\sigma^2}I_N\,\phi(x)\right)^{-1}$ $\qquad A = I_m \qquad C = \frac{1}{\sigma^2}I_N \Rightarrow C^{-1} = \sigma^2 I_N$

$A \quad U \quad C \quad V$

$$= I_m - I_m\phi(x)^T\left(\phi(x)\,I_m\,\phi(x)^T + \sigma^2 I_N\right)\phi(x)\,I_m$$

$$= I_m - \phi(x)^T\left[\phi(x)\phi(x)^T + \sigma^2 I_N\right)^{-1}\phi(x)$$

$$\left(I_m + C^{-1}\frac{1}{\sigma^2}\phi(x)^T y\right.$$

$$\left(I_m - \phi(x)^T\left[\phi(x)\phi(x)^T + \sigma^2 I_N\right)^{-1}\phi(x)\right)\frac{1}{\sigma^2}\phi(x)^T y$$

$$= \frac{1}{\sigma^2}\phi(x)^T y - \frac{1}{\sigma^2}\phi(x)^T Z^{-1}\phi(x)\,\phi(x)^T y$$

$$= \frac{1}{\sigma^2}\phi(x)^T\left(y - Z^{-1}\phi(x)\phi(x)^T y\right)$$

$$= \frac{1}{\sigma^2}\phi(x)^T\left(I - Z^{-1}\phi(x)\phi(x)^T\right)y$$

$$= \frac{1}{\sigma^2}\phi(x)^T Z^{-1}\left(Z - \phi(x)\phi(x)^T\right)y$$

$$= \frac{1}{\sigma^2}\phi(x)^T Z^{-1}\left[\phi(x)\phi(x)^T + \sigma^2 I_N - \phi(x)\phi(x)^T\right]y$$

$$= \frac{1}{\sigma^2}\phi(x)^T Z^{-1}\sigma^2 I_N\, y$$

$$= \phi(x)^T Z^{-1} y$$

$$= \phi(x)^T\left[\phi(x)\phi(x)^T + \sigma^2 I_N\right)^{-1} y$$

$\square$

Question 3 (BLR Predictive)**.** For a Bayesian Linear Regression model:

1. Find $p(\mathbf{y}^*, \mathbf{y})$.
2. Find $p(\mathbf{y}^*|\mathbf{y})$.

y^* = unseen

$$p(y|x^*, D) = \int p(y|x^*, \theta)\, p(\theta|D)\, d\theta$$

1. $p(y^*, y) = p(y|y^*)\, p(y^*)$

$$y = \phi\theta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I)$$

$$y \sim N(x\theta, \sigma^2)$$

$$y^* = \phi(x^*)\theta + \varepsilon$$