Solutions for Exercise Round 2

Exercise 1. (Regression as State Estimation)

Consider the regression problem

$$\begin{aligned} y_k &= a_1\, s_k + a_2\, \sin(s_k) + b + \varepsilon_k, \qquad k = 1,\dots,T, \\ \varepsilon_k &\sim \mathrm{N}(0, r^2), \\ a_1 &\sim \mathrm{N}(0, \sigma_1^2), \\ a_2 &\sim \mathrm{N}(0, \sigma_2^2), \\ b &\sim \mathrm{N}(0, \sigma_b^2), \end{aligned} \tag{1}$$

where $s_k \in \mathbb{R}$ are the known regressor values, $r^2, \sigma_1^2, \sigma_2^2, \sigma_b^2$ are given positive constants, $y_k \in \mathbb{R}$ are the observed output variables, and ε_k are independent Gaussian measurement errors. The scalars a_1, a_2, and b are the unknown parameters to be estimated. Formulate the estimation problem as a linear Gaussian state space model.

Proof. Define the state variable $\mathbf{x}_k = \begin{bmatrix} a_1 & a_2 & b \end{bmatrix}^\top$ with the simple dynamics $\mathbf{x}_{k+1} = \mathbf{x}_k$ and initial distribution $\mathbf{x}_0 \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_b^2))$. The corresponding observation model is $\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \epsilon_k$ with $\mathbf{H}_k = \begin{bmatrix} s_k & \sin(s_k) & 1 \end{bmatrix}$. □

Exercise 2. (Linear Bayesian Estimation)

Assume that in the linear regression model from the previous exercise round (Ex. 1.2), we set independent Gaussian priors for the parameters θ_1 and θ_2 as follows:

$$\begin{aligned} \theta_1 &\sim \mathrm{N}(0, \sigma^2), \\ \theta_2 &\sim \mathrm{N}(0, \sigma^2), \end{aligned}$$

where the variance σ^2 is known. The measurements y_k are modeled as

$$y_k = \theta_1\, x_k + \theta_2 + \varepsilon_k, \qquad k = 1, 2, \dots, T,$$

where the ε_k are independent Gaussian error terms with mean 0 and variance 1, that is, $\varepsilon_k \sim \mathrm{N}(0, 1)$. The values x_k are fixed and known. The posterior distribution can be now written as

$$\begin{aligned} &p(\boldsymbol{\theta} \mid y_1, \dots, y_T) \\ &\propto \exp\left(-\frac{1}{2} \sum_{k=1}^{T} (y_k - \theta_1\, x_k - \theta_2)^2 \right) \exp\left(-\frac{1}{2\sigma^2} \theta_1^2 \right) \exp\left(-\frac{1}{2\sigma^2} \theta_2^2 \right). \end{aligned}$$

The posterior distribution can be seen to be Gaussian and your task is to derive its mean and covariance.

(a) Write the exponent of the posterior distribution in matrix form as in Exercise 1 on round 1 (in terms of $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\theta}$ and σ^2).

(b) Because a Gaussian distribution is always symmetric, its mean $\mathbf{m}$ is at the maximum of the distribution. Find the posterior mean by computing the gradient of the exponent and finding where it vanishes.

(c) Find the covariance of the distribution by computing the second derivative matrix (Hessian matrix) $\mathbf{H}$ of the exponent. The posterior covariance is then $\mathbf{P} = -\mathbf{H}^{-1}$ (why?).

(d) What is the resulting posterior distribution? What is the relationship with the least squares estimate in Exercise 2 on round 1?

Proof. Let $\mathbf{X} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_T & 1 \end{bmatrix}$ so that $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. We want to compute

$$
\begin{aligned}
p(\theta \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) && (2) \\
&= \exp(-\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2) \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|_2^2) && (3) \\
&\propto \exp\left[-\frac{1}{2}\boldsymbol{\theta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\theta} + \mathbf{y}^\mathsf{T}\mathbf{X}\boldsymbol{\theta} - \frac{1}{2\sigma^2}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta}\right] && (4)
\end{aligned}
$$

To get the mean and covariance, we can first define

$$
E(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\theta} - \mathbf{y}^\mathsf{T}\mathbf{X}\boldsymbol{\theta} + \frac{1}{2\sigma^2}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta} \tag{5}
$$

which is minus the exponent of the distribution above. Thus the minimum of this is going to give the maximum of the posterior distribution. We can set the gradient to zero:

$$
\nabla E(\boldsymbol{\theta}) = \mathbf{X}^\mathsf{T}\mathbf{X}\,\boldsymbol{\theta} - \mathbf{X}^\mathsf{T}\mathbf{y} + \frac{1}{\sigma^2}\,\theta = -\mathbf{X}^\mathsf{T}\mathbf{y} + \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right)\theta = 0 \tag{6}
$$

which is zero at the posterior mean $\mathbf{m} = \boldsymbol{\theta}_{\min} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ and its Hessian (the inverse covariance) is

$$
\nabla\nabla^\mathsf{T} E(\boldsymbol{\theta}) = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right) \tag{7}
$$

So that the covariance is given by

$$\mathbf{P} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right)^{-1} \tag{8}$$

The rationale of the above mean and covariance computation is that we first see with our "engineering eye" (or applied mathematician / physicist) that the posterior distribution has to be Gaussian, because it is e powered to a quadratic form. Hence if we knew the mean $\mathbf{m}$ and covariance $\mathbf{P}$, then we could write it in form

$$p(\theta \mid \mathbf{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^\mathsf{T}\mathbf{P}^{-1}(\boldsymbol{\theta} - \mathbf{m})\right] \tag{9}$$

The maximum of this is $\mathbf{m}$ and the Hessian of the negative exponent is $\mathbf{P}^{-1}$. Hence we can use these properties to extract the mean and covariance from the original expression (4). This can be see as a way to do "completion of the squares". We could equivalently just ad-hoc rewrite the expression in this kind of form.

If one takes $\sigma^2 \to \infty$ (very uninformative prior) we recover the least square solution. Note that we can also deduce that the posterior variance of the least squares estimate is $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$. □

Exercise 3. (Gaussian Identities)

Recall that the Gaussian probability density is defined as

$$\mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P}) = \frac{1}{(2\,\pi)^{n/2}\,|\mathbf{P}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\mathsf{T}\,\mathbf{P}^{-1}\,(\mathbf{x} - \mathbf{m})\right).$$

Derive the following Gaussian identities.

(a) Let $\mathbf{x}$ and $\mathbf{y}$ have the Gaussian densities

$$p(\mathbf{x}) = \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P}), \qquad p(\mathbf{y} \mid \mathbf{x}) = \mathrm{N}(\mathbf{y} \mid \mathbf{H}\,\mathbf{x}, \mathbf{R}),$$

then the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ is

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{m}\\ \mathbf{H}\,\mathbf{m}\end{pmatrix}, \begin{pmatrix}\mathbf{P} & \mathbf{P}\,\mathbf{H}^\mathsf{T}\\ \mathbf{H}\,\mathbf{P} & \mathbf{H}\,\mathbf{P}\,\mathbf{H}^\mathsf{T} + \mathbf{R}\end{pmatrix}\right)$$

and the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathrm{N}(\mathbf{H}\,\mathbf{m}, \mathbf{H}\,\mathbf{P}\,\mathbf{H}^\mathsf{T} + \mathbf{R}).$$

Hint: Use the properties of expectation $\mathrm{E}[\mathbf{H}\,\mathbf{x} + \mathbf{r}] = \mathbf{H}\,\mathrm{E}[\mathbf{x}] + \mathrm{E}[\mathbf{r}]$ and $\mathrm{Cov}[\mathbf{H}\,\mathbf{x} + \mathbf{r}] = \mathbf{H}\,\mathrm{Cov}[\mathbf{x}]\,\mathbf{H}^\mathsf{T} + \mathrm{Cov}[\mathbf{r}]$ (if $\mathbf{x}$ and $\mathbf{r}$ are independent).

(b) Write down the explicit expression for the joint and marginal probability densities above:

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) = ?$$
$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} = ?$$

(c) If the random variables $\mathbf{x}$ and $\mathbf{y}$ have the joint Gaussian probability density

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}, \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{B} \end{pmatrix} \right),$$

then the conditional density of $\mathbf{x}$ given $\mathbf{y}$ is

$$\mathbf{x} \mid \mathbf{y} \sim \mathrm{N}(\mathbf{a} + \mathbf{C}\,\mathbf{B}^{-1}\,(\mathbf{y}-\mathbf{b}), \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\mathbf{C}^\mathsf{T}).$$

Hints:

- Denote inverse covariance as $\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{12}^\mathsf{T} & \mathbf{D}_{22} \end{pmatrix}$, and expand the quadratic form in the Gaussian exponent.
- Compute the derivative with respect to $\mathbf{x}$ and set it to zero. Conclude that due to symmetry, the point where the derivative vanishes is the mean.
- From the block matrix inverse formulas given in Theorem A.4 in the course book, we get that the inverse of $\mathbf{D}_{11}$ is

  $$\mathbf{D}_{11}^{-1} = \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\,\mathbf{C}^\mathsf{T}$$

  and that $\mathbf{D}_{12}$ can be then written as

  $$\mathbf{D}_{12} = -\mathbf{D}_{11}\,\mathbf{C}\,\mathbf{B}^{-1}.$$

- Find the simplified expression for the mean by applying the identities above.
- Find the second derivative of the negative Gaussian exponent with respect to $\mathbf{x}$. Conclude that it must be the inverse conditional covariance of $\mathbf{x}$.
- Use the Schur complement expression above for computing the conditional covariance.

Proof. **a)**
In other words, the aim is to determine the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ when

$$\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \mathbf{P}), \tag{10}$$
$$\mathbf{r} \sim \mathrm{N}(\mathbf{0}, \mathbf{R}), \tag{11}$$
$$\mathbf{y} = \mathbf{H}\,\mathbf{x} + \mathbf{r}, \tag{12}$$

where $\mathbf{r}$ is independent of $\mathbf{x}$ and $\mathbf{y}$.

Any linear combination of $\mathbf{x}$ and $\mathbf{y}$ is a Gaussian, so that they are jointly Gaussian. Furthermore, the joint Gaussian distribution of them has the form

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathrm{E}[\mathbf{x}] \\ \mathrm{E}[\mathbf{y}] \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}[\mathbf{x},\mathbf{x}] & \mathrm{Cov}[\mathbf{x},\mathbf{y}] \\ \mathrm{Cov}[\mathbf{x},\mathbf{y}]^\mathsf{T} & \mathrm{Cov}[\mathbf{y},\mathbf{y}] \end{pmatrix} \right),$$

where we need to determine each of the vector and matrix elements. The means are

$$\begin{aligned} \mathrm{E}[\mathbf{x}] &= \mathbf{m}, && (13) \\ \mathrm{E}[\mathbf{y}] &= \mathrm{E}[\mathbf{H}\,\mathbf{x} + \mathbf{r}] = \mathbf{H}\,\mathbf{m} && (14) \end{aligned}$$

and the covariances The mean of the

$$\begin{aligned}
\mathrm{Cov}[\mathbf{x},\mathbf{x}] &= \mathbf{P}, && (15) \\
\mathrm{Cov}[\mathbf{x},\mathbf{y}] &= \mathrm{E}\left[(\mathbf{x}-\mathbf{m})\,(\mathbf{y}-\mathbf{H}\,\mathbf{m})^\mathsf{T}\right] && (16) \\
&= \mathrm{E}\left[(\mathbf{x}-\mathbf{m})\,(\mathbf{H}\,\mathbf{x}+\mathbf{r}-\mathbf{H}\,\mathbf{m})^\mathsf{T}\right] && (17) \\
&= \mathrm{E}\left[(\mathbf{x}-\mathbf{m})\,(\mathbf{x}-\mathbf{m})^\mathsf{T}\mathbf{H}^\mathsf{T}\right] + \underbrace{\mathrm{E}\left[(\mathbf{x}-\mathbf{m})\,\mathbf{r}^\mathsf{T}\right]}_{=0} && (18) \\
&= \mathbf{P}\,\mathbf{H}^\mathsf{T} && (19) \\
\mathrm{Cov}[\mathbf{y},\mathbf{y}] &= \mathrm{E}\left[(\mathbf{y}-\mathbf{H}\,\mathbf{m})\,(\mathbf{y}-\mathbf{H}\,\mathbf{m})^\mathsf{T}\right] && (20) \\
&= \mathrm{E}\left[(\mathbf{H}\,\mathbf{x}+\mathbf{r}-\mathbf{H}\,\mathbf{m})\,(\mathbf{H}\,\mathbf{x}+\mathbf{r}-\mathbf{H}\,\mathbf{m})^\mathsf{T}\right] && (21) \\
&= \mathrm{E}\left[\mathbf{H}\,(\mathbf{x}-\mathbf{m})\,(\mathbf{x}-\mathbf{m})^\mathsf{T}\mathbf{H}^\mathsf{T}\right] + \underbrace{\mathrm{E}\left[\mathbf{H}\,(\mathbf{x}-\mathbf{m})\,\mathbf{r}^\mathsf{T}\right]}_{=0} && (22) \\
&\quad + \underbrace{\mathrm{E}\left[\mathbf{r}\,(\mathbf{x}-\mathbf{m})^\mathsf{T}\mathbf{H}^\mathsf{T}\right]}_{=0} + \mathrm{E}\left[\mathbf{r}\,\mathbf{r}^\mathsf{T}\right] && (23) \\
&= \mathbf{H}\,\mathbf{P}\,\mathbf{H}^\mathsf{T} + \mathbf{R} && (24)
\end{aligned}$$

Furthermore, the marginal mean and covariance of $\mathbf{y}$ determine its Gaussian distribution and they indeed are $\mathrm{E}[\mathbf{y}]$ and $\mathrm{Cov}[\mathbf{y},\mathbf{y}]$ as given above.

b)

These are simply the explicit expressions for the multivariate Gaussian densities with the means and covariances derived above.

c)

We thus have

$$p(\mathbf{x},\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}-\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right]^\top \begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^\top & \mathbf{B}\end{pmatrix}^{-1}\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}-\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right]\right) \tag{25}$$

$$= \exp\left(-\frac{1}{2}\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}-\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right]^\top \begin{pmatrix}\mathbf{D}_{11} & \mathbf{D}_{12}\\ \mathbf{D}_{12}^\top & \mathbf{D}_{22}\end{pmatrix}\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}-\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right]\right) \tag{26}$$

$$= \exp\Bigg(-\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{D}_{11}(\mathbf{x}-\mathbf{a}) - \frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{D}_{12}(\mathbf{y}-\mathbf{b}) \tag{27}$$

$$-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{D}_{12}^\top(\mathbf{x}-\mathbf{a}) - \frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{D}_{22}(\mathbf{y}-\mathbf{b})\Bigg) \tag{28}$$

As we have $p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x},\mathbf{y})/p(\mathbf{y}) \propto p(\mathbf{x},\mathbf{y})$, we can conclude the the completion of squares for $p(\mathbf{x} \mid \mathbf{y})$ corresponds to completion of the squares in $p(\mathbf{x},\mathbf{y})$ above with respect to $\mathbf{x}$. It is now easiest to consider the negative exponent instead of the full distribution although the result is the same. We can set the derivative to zero:

$$\nabla_{\mathbf{x}}\Bigg(\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{D}_{11}(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{D}_{12}(\mathbf{y}-\mathbf{b}) \tag{29}$$

$$+\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{D}_{12}^\top(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{D}_{22}(\mathbf{y}-\mathbf{b})\Bigg) \tag{30}$$

$$= \mathbf{D}_{11}\,(\mathbf{x}-\mathbf{a}) + \mathbf{D}_{12}(\mathbf{y}-\mathbf{b}) = 0 \tag{31}$$

which gives, together with the given identifies

$$\mathbf{x} = \mathbf{a} - \mathbf{D}_{11}^{-1}\mathbf{D}_{12}(\mathbf{y}-\mathbf{b}) \tag{32}$$

$$= \mathbf{a} + \mathbf{D}_{11}^{-1}\mathbf{D}_{11}\,\mathbf{C}\,\mathbf{B}^{-1}(\mathbf{y}-\mathbf{b}) \tag{33}$$

$$= \mathbf{a} + \mathbf{C}\,\mathbf{B}^{-1}(\mathbf{y}-\mathbf{b}) \tag{34}$$

$$\tag{35}$$

The second derivative of the negative exponent w.r.t. $\mathbf{x}$ is $\mathbf{D}_{11}$ whose inverse gives the conditional covariance

$$\mathbf{D}_{11}^{-1} = \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\,\mathbf{C}^\top$$

which concludes the result. □